МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ХИМИКО-ФАРМАЦЕВТИЧЕСКИЙ УНИВЕРСИТЕТ

Кафедра высшей математики

Подольский В.А., Розовский Л.В., Травина Н.И., Ивановская Т.Ю.

Учебно-методическое пособие

Математические методы в статистике

Санкт-Петербург
"Издательство СПХФУ"
2021 г

Оглавление

Введение

Математическая статистика *– раздел математики, изучающий математические методы сбора, систематизации, обработки и интерпретации наблюдений с целью выявления статистических закономерностей.*

Любое понятие теории вероятностей есть отражение определённого физического процесса. Оно появляется в связи с требованием практики и представляет собой следствие некоторых экспериментов. Задачей математической статистики является обработка экспериментальных данных с целью изучения свойств случайных явлений и получения некоторой математической модели изучаемого объекта.

Математическая модель объекта – это система математических соотношений, которые описывают, с определённым уровнем приближения, свойства величин, характеризующих объект, а также их возможные связи. Математическая модель не полностью описывает реальный объект, однако по мере изучения самой модели (вместо объекта) она может уточняться, всё полнее отражая свойства этого объекта. Если в объекте имеют место случайные явления, то его математическая модель обязательно будет включать в себя некоторые понятия теории вероятностей. Изучить определённые закономерности случайных явлений можно только с помощью многократных повторений опыта с объектом и последующей обработкой результатов.

Если теория вероятностей изучает закономерности случайных явлений на основе абстрактного описания действительности (теоретической вероятностной модели), то математическая статистика оперирует непосредственно результатами наблюдений над случайным явлением, представляющими выборку из некоторой конечной или гипотетически бесконечной генеральной совокупности. Используя результаты, полученные теорией вероятностей, математическая статистика позволяет не только оценить значения искомых характеристик, но и выявить степень точности выводов, получаемых при обработке данных.

При большом числе наблюдений случайные воздействия в значительной мере нивелируются, и полученный результат оказывается практически неслучайным, предсказуемым. Этот принцип и является базой для практического использования вероятностных и математико-статистических методов исследования.

Кроме основной задачи – обработки экспериментальных данных – математическая статистика рассматривает и смежные задачи, такие, как методы регистрации, способы описания экспериментальных данных, а также само планирование эксперимента.

Задачу по изучению свойств некоторой случайной величины X с помощью экспериментальных данных можно поставить в разных вариантах. Во-первых, можно ограничиться лишь

оценкой ее числовых характеристик. В этом случае задача относится к группе задач оценки неизвестных параметров. В эту группу входят задачи выбора структуры оценок и задачи анализа качества этих оценок. Оценивать можно не только параметры, относящиеся к одной величине (математическое ожидание, дисперсию и т.п.), но и параметры, связанные с двумя или более величинами (коэффициент корреляции, корреляционное отношение и т.п.).

Более общая задача – установление закона распределения вероятностей величины *X*.

Пусть $x_1, x_2, \ldots, x_n$ - n измерений этой величины. При отсутствии иной информации остаётся предполагать, что эти значения являются единственно возможными значениями данной случайной величины, а вероятность каждого из них равна *1/n*. Исходя из этого, можно построить эмпирическую функцию распределения вероятностей величины *X* как дискретной случайной величины: $F^*(x) = k/n$, где k – число измерений, лежащих на числовой оси слева от *x*. Возникает вопрос, нельзя ли при конечном значении *n* провести сглаживание функции $F^*(x)$ так, чтобы получить функцию, близкую к истинной функции распределения? Этот вопрос рассматривается в группе задач, связанных с установлением закона распределения вероятностей по экспериментальным данным.

Следующая группа задач – это задачи проверки статистических гипотез. В качестве гипотез могут выступать гипотезы о виде закона распределения вероятностей, гипотезы о предположительном значении тех или иных числовых характеристик, гипотезы о существовании или отсутствии статистической связи между величинами и т.д.

Если рассматриваются несколько случайных величин в совокупности, то одной из задач математической статистики является изучение и аппроксимация статистических связей между ними. Рассмотрим, например, две случайные величины *X* и *Y*. Предположим теперь, что *X* и *Y* являются статистически зависимыми величинами. Тогда можно поставить следующую задачу: найти такую функцию $Y = \varphi(X)$, которая бы максимально отражала статистическую зависимость между величинами. С помощью такой функции можно было бы наиболее точно по заданным значениям *X* прогнозировать значения, которые будет принимать величина *Y*. Эта задача относится к группе задач регрессионного анализа.

Лабораторная работа 1

Статистическое распределение выборки.
Гистограмма. Точечные оценки.

1. Теоретические сведения.

Всякое каким-то образом выделенное множество объектов, которые могут отличаться друг от друга значением некоторой определенной характеристики (признака) X, называется ***генеральной совокупностью***. В математической статистике понятие *генеральной совокупности* трактуется как ***совокупность всех мыслимых наблюдений, которые могут быть произведены при данном реальном комплексе условий***, и в этом смысле его не надо смешивать с реальными совокупностями, подлежащими статистическому изучению. Понятие генеральной совокупности аналогично понятию ***случайной величины X***.

Число элементов генеральной совокупности называется ее ***объемом***.

Часть генеральной совокупности $\{X_1, X_2,..., X_n\}$, случайным образом отобранная для наблюдений, называется ***случайной выборкой*** или, для краткости, ***выборкой***.

Выборку можно рассматривать как некоторый эмпирический аналог генеральной совокупности. Элементы выборки можно считать независимыми одинаково распределёнными случайными величинами X_i, поскольку они являются результатом проведения последовательности независимых испытаний с одной и той же случайной величиной X.

Для того чтобы результаты исследований генеральной совокупности по выборке были объективными в достаточной степени, выборка должна быть случайной и репрезентативной (представительной). Случайность – любой из элементов генеральной совокупности обладает равной возможностью быть отобранным в выборку. Элементы выборки можно считать независимыми случайными величинами. На практике исследователь работает с конкретной реализацией выборки $\{x_1, x_2, ..., x_n\}$, где x_i являются значениями случайных величин X_i, распределение которых совпадает с распределением признака X. Число элементов выборки n называется ее ***объемом***, а конкретные значения реализации выборки x_i – ***вариантами***. Расположив варианты в порядке возрастания, получим ***вариационный ряд***.

По результатам наблюдений над выборкой можно вычислить точечные оценки неизвестных параметров распределения признака X.

Для неизвестного математического ожидания $M(X)$ (***генеральное среднее***) вычисляется точечная оценка – ***выборочное среднее***.

Для неизвестной дисперсии D(*X*) (***генеральная дисперсия***) вычисляется точечная оценка – ***выборочная или исправленная выборочная дисперсия***.

В соответствии с требованиями математической статистики эти оценки должны удовлетворять ряду критериев, основными из которых являются требования ***состоятельности*** и ***несмещенности***.

Оценка θ_n называется *состоятельной оценкой* параметра θ, если $\theta_n \to \theta$ по вероятности при $n \to \infty$, т.е. $P(|\theta_n - \theta| > \varepsilon) \to 0$ при любом $\varepsilon > 0$.

Другими словами, вероятность отклонения оценки от истинного значения параметра можно сделать сколь угодно малой, увеличивая объем выборки.

Оценка θ_n называется *несмещенной оценкой* параметра θ, если $M(\theta_n) = \theta$ при любом n, т.е. отклонение θ_n от θ не содержит систематической ошибки.

Доказывается, что $\overline{X} = \frac{1}{n}\sum_{1\le i\le n} x_i$ – *выборочное среднее* и $\overline{S}^2 = \frac{1}{n-1}\sum_{1\le i\le n}(x_i - \overline{X})^2$ – *исправленная выборочная дисперсия* являются *состоятельными и несмещенными* оценками M(*X*) и D(*X*), соответственно.

Если объем выборки недостаточно велик или интересующий нас признак генеральной совокупности *X* имеет непрерывное распределение, то для сокращения объема расчетов варианты вариационного ряда группируются в интервалы. Таким образом строится дискретная модель распределения изучаемого признака *X*, т.н. ***статистический интервальный ряд распределения***.

Типичная процедура группировки выглядит так.

Отрезок [a, b], содержащий все варианты вариационного ряда, делится на *k* интервалов (h_{i-1}, h_i), $i = 1, ..., k$,

$$a = h_o < h_1 < \ldots < h_k = b. \qquad (1.1)$$

Затем находится ***абсолютная частота*** *(*или просто частота) на каждом интервале, т.е. количество наблюдений n_i, попавших в *i*-й интервал. Для определенности будем полагать n_i равным числу вариант, принадлежащих полуинтервалу $[h_{i-1}, h_i)$, варианты, попавшие на правую границу, т.е. равные h_i, включаются в следующий промежуток при всех $i<k$.

Полученную группировку удобно представить в форме частотной таблицы, которая и является дискретной моделью - ***статистическим интервальным рядом распределения***:

Интервалы группировки	$[h_o, h_1)$	$[h_1, h_2)$	. . .	$[h_{k-2}, h_{k-1})$	$[h_{k-1}, h_k]$
Частоты	n_1	n_2	. . .	n_{k-1}	n_k
Отн.частоты	w_1	w_2	. . .	w_{k-1}	w_k

Табл. 1.1

Здесь *относительные частоты* $w_i = \frac{n_i}{n}$ (обращаем внимание на то, что правая граничная точка группировки $h_k = b$ включается в последний интервал, и напоминаем о том, что $n_1 + n_2 + ... + n_k = n$ – объему выборки, а $w_1 + w_2 + ... + w_k = 1$).

Наиболее информативной графической формой интервального ряда распределения является ***гистограмма относительных частот*** (или просто гистограмма), состоящая из прямоугольников с основаниями (h_{i-1}, h_i), высота которых равна *плотности относительных частот* $\omega_i = \frac{w_i}{h_i - h_{i-1}}$. Таким образом, площадь каждого прямоугольника равна w_i, а общая сумма этих площадей равна единице.

Заметим, что площадь той части гистограммы относительных частот, что лежит между h_i и h_m ($i<m$), равна относительному числу вариант, попавших в интервал $[h_i, h_m)$, и в соответствии с статистическим определением вероятности может быть интерпретирована как оценка вероятности $P(h_i \le X < h_m)$, где X - признак генеральной совокупности. Следовательно, с определенными оговорками, обусловленными дискретностью модели, можно утверждать, что гистограмма относительных частот является *выборочным аналогом графика плотности вероятностей распределения исследуемого случайного признака X*.

Обычно, $h_i - h_{i-1} = h$ при всех i, т.е. группировка осуществляется с шагом, равным h.

В этой ситуации будем руководствоваться следующими эмпирическими рекомендациями по выбору числа интервалов группировки k:

k полагаем равным наибольшему целому значению числа $(1 + 3.32 \lg n)$.

Например, если объем выборки $n = 50$, то $k = 7$.

Далее:

находим наименьший m и наибольший M элементы выборки; вычисляем размах выборки $R=M-m$, устанавливаем левую границу интервала группировки $a = m$,

шаг группировки $h = (M-a)/k$ (округляя при необходимости в большую сторону),

$h_i = a + i \cdot h$, $i = 0, 1, ..., k$ (таким образом, правая граница интервала группировки b становится равной $a + kh$).

Середины интервалов группировки $\bar{x}_i$, $i = 0, 1, ..., k-1$, находятся по формуле

$\bar{x}_i = h_i + h/2$.

Кроме статистического интервального ряда распределения, признак X можно описать *статистической функцией распределения* $F_n(x)$, которая, в рамках дискретной модели, является аналогом $F(x)$ - неизвестной функции распределения вероятностей признака X.

Статистической функцией распределения, построенной по случайной выборке X_1, X_2,..., X_n называется относительная частота того, что признак X примет значение меньшее заданного x:

$$F(x) = \sum_{1 \le i \le k} w_i, \qquad \text{если } \begin{cases} x_{k-1} < x \le x_k \\ k = 1,2 \cdots n \end{cases}$$

Выборочное среднее $\bar{X}$ и выборочная дисперсия $\bar{S}^2$ по данным табл.1.1 определяются формулами (для интервального вариационного ряда вместо вариант x_i следует взять середины частичных интервалов $\bar{x}_i$),

$$\bar{X} = \frac{1}{n}\sum_{1 \le i \le k} n_i \bar{x}_i = \sum_{1 \le i \le k} w_i \bar{x}_i, \quad S^2 = \frac{1}{n}\sum_{1 \le i \le k} n_i (\bar{x}_i - \bar{X})^2 = \sum_{1 \le i \le k} w_i (\bar{x}_i - \bar{X})^2, \tag{1.2}$$

или

$$S^2 = \frac{1}{n}\sum_{1 \le i \le k} n_i \bar{x}_i^2 - \bar{X}^2 = \sum_{1 \le i \le k} w_i \bar{x}_i^2 - \bar{X}^2, \quad \bar{S}^2 = \frac{n}{n-1} S^2 \tag{1.3}$$

Результаты обработки выборки следует свести в таблицу 1.2, приведенную ниже. Затем по данным 2-й и 7-й колонок построить гистограмму относительных частот.

№ интервала группировки	Границы интервала группировки	Середина интервала группировки $\bar{x}i$	Частота		$F_n(x)$	Плотность относит. частот ω_i
			абс. n_i	отн. w_i		
1	*2*	*3*	*4*	*5*	*6*	*7*
1	[*a*, *a*+*h*)	$\bar{x}_1$	n_1	w_1		ω_1
…	…	…	…	…		…
k	[*a*+(*k*-1)*h*, *b*]	$\bar{x}_k$	n_k	w_k		ω_k

Табл. 1.2

Выборочное среднее и выборочная дисперсия вычисляются по формулам (1.2), (1.3).

Пример. Найти статистический интервальный ряд распределения, построить гистограмму относительных частот с равным шагом, вычислить выборочное среднее и выборочную дисперсию по случайной выборке объема 50, вариационный ряд которой:

22 24 26 26 27 28 28 31 31 31
32 32 33 33 33 33 34 34 34 34
34 35 35 36 36 36 36 36 37 37
37 37 37 37 38 38 40 40 40 40
40 41 41 43 44 44 45 45 47 50

Решение. Объем выборки n=50, число интервалов группировки k = 7, наименьшее значение варианты m=22, наибольшее значение варианты M=50. Полагаем a=m=22, выбираем шаг группировки h=(50-22) / 7 = 4 и границы интервалов группировки $h_i = 22 + 4\,i$, i = 0, 1, ..., 7. Находим n_i , $w_i = n_i / 50$ и $\omega_i = w_i / 4$.

Таким образом, аналог таблицы 1.2 имеет вид

№ интервала группировки	Границы интервала группировки	Середина интервала группировки	Частота		$F_n(x)$	Плотность относит. частот ω_i
			абс. n_i	отн. w_i		
1	*2*	*3*	*4*	*5*	*6*	*7*
1	[22 , 26)	24	2	0.04	0.04	0.01
2	[26, 30)	28	5	0.1	0.14	0.025
3	[30, 34)	32	9	0.18	0.32	0.045
4	[34, 38)	36	18	0.36	0.68	0.09
5	[38, 42)	40	9	0.18	0.86	0.045
6	[42, 46)	44	5	0.1	0.96	0.025
7	[46, 50]	48	2	0.04	1.00	0.01

Табл. 1.3

Напоминаем, что при попадании на правую границу интервала варианта приписывается следующему интервал во всех случаях, кроме последнего.

Строим гистограмму относительных частот по данным третьего и пятого столбцов табл. 1.3 (Рис. 1.1):

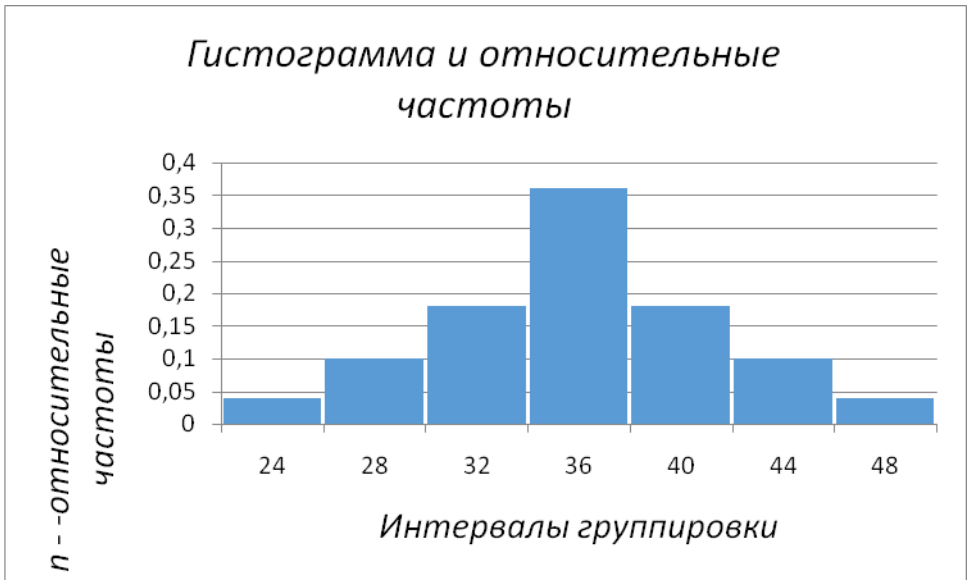


Рис. 1.1

Построим $F_n(x)$ по данным второго и шестого столбца табл. 1.3:

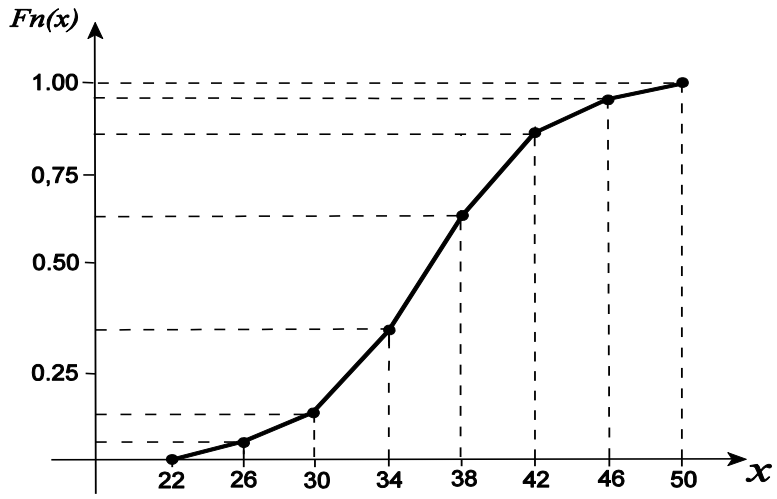


Рис. 1.2

Для интервального ряда ломаная начинается с точки, имеющей координаты (абсцисса = началу первого интервала, ордината = 0).

Точечные оценки для математического ожидания M(X) и дисперсии D(X) найдем по формулам (1.2) .

Имеем,

$$\bar{x} = \frac{1}{50}\cdot\left(2\cdot 24+5\cdot 28+9\cdot 32+18\cdot 36+9\cdot 40+5\cdot 44+2\cdot 48\right)=36, \qquad (1.4)$$

$$s^2 = \frac{1}{50}(2\cdot 24^2+5\cdot 28^2+9\cdot 32^2+18\cdot 36^2+9\cdot 40^2+5\cdot 44^2+2\cdot 48^2)-36^2 =$$

$$= 30.08 \qquad (1.5)$$

$$\bar{s}^2 = \frac{n}{n-1}S^2 = \frac{50}{49}\cdot 30.08 = 30.69$$

3. Ход выполнения работы.

Указанные пункты должны быть представлены в отчете по лабораторной работе.

Формулировка цели работы.

1. Анализ исходных данных. Подготовка данных для построения интервального ряда распределения:

1.1. Переписать исходные данные в тетрадь в виде таблицы 5×10 и построить вариационный ряд.

1.2. Найти наименьшее значение m и наибольшее M значение варианты.

1.3.Число интервалов группировки k вычислить по формуле k= 1 + 3.32 lg$n$, округляя до целого в большую сторону.

1.4. Выбрать левую границу интервала группировки a равной m, округляя при необходимости в пределах точности варианты в меньшую сторону.

1.5. Шаг группировки h найти по формуле $(M-m)/k$, округляя при необходимости в пределах точности варианты в большую сторону.

Таким образом, левая граница 1-го интервала группировки равна a, правая граница равна $a + h$, а его середина – $(a + h/2)$.

2 Построение интервального ряда распределения.

Заполнить таблицу

№ интервала группировки	Границы интервала группировки	Середина интервала группировки $\bar{x}i$	Частота		$F_n(x)$	Плотность относит. частот ω_i
			абс. n_i	отн. w_i		
1	2	3	*4*	*5*	*6*	*7*
1	$[a, a+h)$	$\bar{x}_1$	n_1	w_1		ω_1
…	…	…	…	…		…
k	$[a+(k-1)h, b]$	$\bar{x}_k$	n_k	w_k		ω_k

Рекомендации по заполнению таблицы.

Распределить варианты по интервалам группировки (варианты, попавшие на границу между интервалами группировки, включаются в правый интервал).

Для контроля следует проверить, что $n_1 + n_2 + ... + n_k = n$ – объему выборки.

Вычислить относительные частоты $w_i = n_i / n$ и их плотность $\omega_I = w_i / h$.

Заполнить 6-й и 7-й столбцы табл.1.2.

1. Построение гистограммы плотности относительной частоты.

По данным колонок 2 и 7 построить гистограмму.

При этом масштаб по осям x и ω следует выбирать таким образом, чтобы график занимал приблизительно две трети тетрадного листка.

2. Вычисление точечных оценок неизвестных параметров распределения изучаемого признака X.

По формулам (1.2) и (1.3) вычислить выборочные среднее и дисперсию.

3. Построение статистической функции распределения $F_n(x)$.
4. Вывод по проделанной работе.

4. Контрольные вопросы.

1. Что такое генеральная совокупность?
2. Что называется выборкой объема n?
3. Что такое варианта?
4. Как определяется вариационный ряд?

5. Какой график называется гистограммой?
6. Что называется генеральным средним?
7. Что называется генеральной дисперсией?
8. Что такое точечная оценка?
9. Какие оценки называются состоятельными и несмещенными?
10. Каким образом определяется выборочное среднее?
11. Каким образом определяется выборочная дисперсия?

5. Варианты выборок

№1

22.27	22.29	21.78	21.15	22.57	20.94	22.25	21.04	21.26	23.28
19.17	21.24	20.06	21.46	21.83	21.69	21.43	22.18	23.09	20.87
22.59	20.71	23.34	21.84	22.56	21.89	21.57	21.21	22.07	20.12
22.32	22.16	21.45	23.36	22.68	22.44	21.59	22.12	22.33	22.34
21.92	21.50	20.47	20.73	23.45	22.13	22.90	22.04	21.75	22.24

№2

20.01	21.52	19.95	19.50	20.71	20.45	17.27	17.87	18.95	20.21
20.66	19.16	20.18	20.88	20.36	20.24	19.41	19.10	20.15	20.54
23.21	21.09	20.37	20.04	20.71	21.62	19.94	21.30	18.60	20.02
19.34	20.44	20.14	19.19	18.60	19.13	20.14	18.59	19.32	20.62
20.00	18.24	21.67	19.70	21.65	20.73	20.37	20.75	19.60	20.19

№3

23.79	27.09	20.41	24.44	23.46	26.09	24.53	24.59	22.13	21.84
23.50	23.45	22.57	21.09	24.26	21.42	20.91	23.91	22.60	22.75
26.10	23.72	22.66	22.04	20.96	23.79	21.92	22.24	23.63	23.74
21.69	24.12	24.61	24.02	19.65	21.80	21.23	22.92	24.80	23.77
23.68	22.59	22.99	21.33	23.48	21.52	21.93	21.65	23.51	23.54

№4

19.13	16.31	18.12	16.72	19.54	17.88	18.66	18.47	17.84	18.84
16.80	18.37	17.83	17.61	18.09	17.63	16.84	17.18	16.68	19.02
18.80	20.10	18.25	17.87	17.18	18.49	17.07	17.09	18.01	18.03
17.94	18.14	17.90	18.96	16.26	16.03	20.01	19.08	16.97	16.34
19.07	17.76	16.57	16.24	17.35	17.92	18.13	20.42	18.72	18.06

№5

16.88	17.10	17.06	16.08	17.56	17.02	16.43	17.73	16.61	18.55
17.12	16.50	16.31	17.60	18.16	16.07	18.47	17.61	17.15	16.67
17.47	17.62	17.26	16.11	17.25	16.21	16.22	15.75	17.32	15.52
18.31	17.53	18.76	17.27	17.37	16.94	18.19	16.72	18.83	18.02
16.80	17.71	17.66	18.40	16.35	16.64	18.39	17.25	17.16	16.47

№6

17.58	18.86	17.34	16.39	17.65	18.54	19.08	18.75	17.45	15.63
17.17	15.87	18.53	16.40	16.77	15.06	16.67	18.78	15.99	15.26
17.48	13.61	16.24	16.11	16.91	17.17	17.11	18.02	14.24	17.88
16.27	15.61	15.98	16.56	16.05	15.94	17.66	14.80	19.51	17.70
16.86	17.49	16.13	16.40	14.47	19.88	16.93	12.80	17.34	16.62

№7

17.58	18.86	17.34	16.39	17.65	18.54	19.08	18.75	17.45	15.63
17.17	15.87	18.53	16.40	16.77	15.06	16.67	18.78	15.99	15.26
17.48	13.61	16.24	16.11	16.91	17.17	17.11	18.02	14.24	17.88
16.27	15.61	15.98	16.56	16.05	15.94	17.66	14.80	19.51	17.70
16.86	17.49	16.13	16.40	14.47	19.88	16.93	12.80	17.34	16.62

№8

15.41	14.82	14.59	15.21	14.45	15.40	15.77	16.45	14.24	13.92
14.58	15.09	16.91	12.18	14.93	15.67	15.17	14.49	15.28	16.31
16.26	14.76	14.69	14.70	15.01	15.45	16.75	15.07	15.79	12.58
15.62	14.22	14.56	15.73	14.03	15.53	13.81	14.05	13.27	15.91
16.62	15.24	14.18	14.87	14.61	15.59	15.18	14.45	13.99	14.93

№9

22.42	19.95	19.31	21.01	20.98	13.69	23.04	19.93	20.66	15.06
18.31	24.60	21.49	19.99	16.49	18.32	19.16	14.12	15.63	19.06
20.08	20.76	15.47	19.48	17.32	19.92	23.44	19.93	23.66	20.28
18.64	19.43	20.55	23.23	17.88	21.72	18.14	13.83	18.65	20.20
22.25	20.84	20.11	15.97	19.41	14.69	23.14	20.81	18.82	15.52

№10

24.28	21.20	25.38	24.13	21.78	29.29	27.67	22.04	24.55	25.94
21.95	22.43	28.21	26.67	27.19	28.19	24.05	24.71	31.02	22.87
23.08	26.97	25.78	24.74	19.96	21.05	26.49	23.73	26.05	23.23
22.86	26.47	27.41	21.49	21.27	23.27	32.78	23.47	20.82	27.15
25.57	29.54	25.26	25.79	26.91	26.72	19.06	23.57	26.58	27.74

№11

14.46	12.25	10.59	8.15	11.28	15.19	11.45	12.10	12.23	11.04
9.46	8.14	10.14	11.93	13.00	7.93	10.95	13.07	11.60	8.27
7.76	12.83	11.90	13.07	8.74	10.69	9.66	13.38	13.23	13.25
15.75	13.37	9.69	14.25	10.20	10.34	7.68	12.54	14.15	12.69
11.89	12.10	9.86	9.99	12.17	8.04	12.70	6.88	11.46	12.58

№12

8.73	13.50	11.27	12.75	13.97	10.36	9.37	11.43	11.70	15.99
15.43	15.51	9.69	15.42	10.80	11.59	14.81	6.83	13.93	13.40
11.09	11.51	12.39	13.80	15.66	13.87	10.89	10.77	12.78	15.47
11.23	13.41	13.84	10.97	12.39	11.96	11.36	12.16	10.43	15.78
10.97	11.97	12.55	11.34	9.44	14.44	11.08	12.25	10.64	10.73

№13

11.91	11.27	15.03	13.26	14.96	15.15	10.38	12.90	11.15	10.91
14.82	16.52	17.16	16.53	12.62	11.68	12.83	11.13	13.15	15.03
15.10	8.94	12.93	14.37	14.05	11.30	13.60	14.63	8.88	14.71
13.92	15.84	13.91	12.76	13.77	11.17	12.60	12.20	12.18	12.92
13.52	12.87	11.28	14.51	12.36	15.76	16.33	14.48	15.97	12.98

№14

15.36	10.60	14.96	13.20	12.61	15.06	15.60	13.50	9.91	15.80
14.71	15.67	12.88	15.76	11.49	14.17	14.31	14.96	14.28	16.62
18.07	16.59	16.52	17.74	15.18	16.93	14.79	15.40	13.78	15.40
14.78	14.58	19.13	10.55	15.33	17.37	13.61	10.05	14.38	15.02
16.74	12.69	15.00	15.88	15.16	14.38	15.40	17.43	14.62	12.18

№15

16.61	14.11	17.33	17.11	20.00	16.90	15.41	15.97	21.68	21.97
21.95	21.49	19.15	25.75	21.77	23.42	14.57	17.33	22.39	16.21
17.82	17.48	18.02	13.98	28.44	15.69	16.56	20.54	14.67	18.11
21.58	17.41	12.43	14.60	20.36	17.77	17.34	17.68	13.85	18.50
23.69	17.57	16.69	18.16	18.48	21.20	21.23	17.02	20.39	22.14

№16

13.80	6.10	7.42	8.49	6.73	9.34	13.17	12.37	15.72	10.76
12.66	5.83	12.30	10.82	9.76	9.66	8.67	5.65	13.00	16.31
9.73	11.05	11.63	7.97	10.05	11.54	11.59	7.46	11.91	11.30
5.47	3.32	16.26	13.21	14.58	12.80	10.75	6.85	10.90	12.78
11.71	16.44	4.12	8.86	13.15	10.96	11.69	5.34	10.79	9.01

№17

22,64	23,09	23,06	23,22	23,43	23,44	22,72	22,75	23,10	23,41
22,99	23,03	23,07	23,25	23,06	23,32	22,84	22,92	23,14	22,58
23,13	22,89	23,39	23,44	23,25	22,71	22,69	22,91	22,74	22,98
22,96	23,28	23,72	23,02	22,98	22,95	22,86	22,87	23,12	22,70
22,73	22,48	23,04	23,15	23,51	23,05	23,21	23,26	22,59	23,13

№18

42,11	41,74	41,04	42,06	42,05	41,90	41,53	41,76	41,89	42,41
41,87	42,14	42,17	42,79	41,67	42,71	42,21	41,29	42,51	41,64
41,65	41,96	41,83	41,97	42,07	41,34	42,32	42,29	41,76	42,03
42,21	41,59	41,76	41,84	42,44	42,55	41,96	41,76	41,96	41,99
42,60	42,89	41,87	41,70	43,08	42,22	41,70	42,32	41,62	41,94

№19

54,89	57,07	54,36	55,60	54,84	56,19	58,53	53,45	51,77	55,72
60,85	51,93	55,27	55,25	59,07	57,08	54,83	54,62	60,02	52,55
55,48	54,23	58,92	56,41	60,92	52,66	55,82	51,18	51,04	57,73
56,43	49,54	54,53	54,77	53,56	54,68	62,71	60,59	53,90	55,47
54,45	51,40	55,35	52,93	52,75	54,11	51,09	54,59	53,68	52,62

Лабораторная работа 2

Построение нормальной кривой по опытным данным

1. Постановка задачи.

Будем считать, что изучаемый признак X рассматриваемой в первой работе генеральной совокупности, распределен по нормальному закону с некоторыми неизвестными параметрами a и σ ($X \sim N(a,\sigma)$). Построим графики плотности распределения вероятностей и функции распределения случайной величины, распределенной нормально,

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}}, \qquad F(x)=\frac{1}{2}+\Phi\left(\frac{x-a}{\sigma}\right)$$

заменив неизвестные числовые параметры распределения их точечными оценками.

Известно, что параметры a и σ^2 являются математическим ожиданием и дисперсией X соответственно. Точечные оценки этих неизвестных параметров рассчитывались в первой работе: $a=\bar{x}$ и $\sigma=\bar{S}$.

Нашей задачей является *визуальная* проверка близости изучаемого распределения признака генеральной совокупности к нормальному.

Для построения нормальной кривой воспользуемся тем, что в соответствии со статистическим определением вероятностей, относительная частота попадания вариант в любой интервал (α, β) с ростом объема выборки стремится к теоретической вероятности попадания в этот интервал, т.е. к вероятности

$$P(\alpha<X<\beta)=\Phi\left(\frac{\beta-\mu}{\sigma}\right)-\Phi\left(\frac{\alpha-\mu}{\sigma}\right), \qquad (2.1)$$

где $\Phi(z)$ - функция Лапласа (см. приложение 1) и μ=a.

2. Ход работы.

1. *Построение графика* p_i/h, *где* p_i - *вероятность попадания значения X в* i *-тый интервал группировки.*

1.1. Заполнить таблицу "*Построение* p_i/h"

i	α	β			$\Phi(z_i)$	$\Phi(z_{i-1})$	$p_i=\Phi(z_i)-\Phi(z_{i-1})$	p_i/h
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1								

…								
k								
							$\sum_i p_i = 1$	

Табл.2.1

Здесь i $(1 \le i \le k)$,$[h_{i-1}, h_i)$ i –й интервал группировки,

$\bar{x}$, $\bar{S}$ - точечные оценки параметров распределения, найденные в первой работе.

Вычислить значения вероятностей попадания случайной величины $X \sim N(\mu, \sigma)$ в интервалы группировки (-∞, h_1), $[h_1, h_2)$, …, $[h_{k-2}, h_{k-1})$, $[h_{k-1}, \infty)$по формуле (2.1)

при $\alpha = h_i$, $\beta = h_{i+1}$ для внутренних интервалов (с i=1,…,k-1),

при $\alpha = -\infty$, $\beta = h_1$ для крайнего левого интервала

и при $\alpha = h_{k-1}$, $\beta = \infty$ для крайнего правого интервала.

Учитывая, что функция Лапласа является нечетной и $\Phi(\infty) = 0.5$, получим

$$p_i = P(h_{i-1} \le X \le h_i) = \Phi\left(\frac{h_i - \bar{x}}{\bar{s}}\right) - \Phi\left(\frac{h_{i-1} - \bar{x}}{\bar{s}}\right), \qquad i = 2, \dots, k-1,$$

$$p_1 = P(X < h_1) = \frac{1}{2} + \Phi\left(\frac{h_1 - \bar{x}}{\bar{s}}\right), \qquad p_k = P(X \ge h_{k-1}) = \frac{1}{2} - \Phi\left(\frac{h_{k-1} - \bar{x}}{\bar{s}}\right) \qquad (2.4)$$

Для нахождения $\Phi(z_i)$ используем приложение 1

Замечание. Нормально распределенная случайная величина X может принимать любые сколь угодно большие по абсолютной величине значения, с очень малой вероятностью, так, в соответствии с правилом «3σ» $P(|X - \mu| > 3\sigma) = 0.0027$ *. Поэтому необходимо рассмотреть интервалы* (-∞, h_1) *вместо* $[h_0, h_1)$ и $[h_{k-1}, \infty)$ *вместо* $[h_{k-1}, h_k]$.

1.2. Построить гистограмму p_i / h

Гистограмму $w_i' = \frac{p_i}{h}$, $i = 1, \dots, k$, на (-∞, h_1), $[h_1, h_2)$, …, $[h_{k-2}, h_{k-1})$, $[h_{k-1}, \infty)$ по данным 2, 3-й и 9-й колонок таблицы 2.1 совместить с гистограммой плотности относительной частоты на интервалах группировки из первой работы.

2. *Построение графика плотности распределения вероятностей* $f(x)$,

$f(x_i) = \frac{1}{\sqrt{2\pi}S} e^{-\frac{(x_i - \bar{X})^2}{2S^2}}$, где параметры распределения заменены точечными оценками, $\bar{x}_i$ — середина соответствующего интервала группировки.

Построение графика функции распределения $F(x)$.

$F(x)=\frac{1}{2}+\Phi\left(\frac{\beta-\bar{x}}{\bar{S}}\right)$, где β – правая граница соответствующего интервала группировки.

2.1. Заполнить таблицу "*Построение* $f(x)$ *и* $F(x)$"

			$t=\left(\frac{\bar{x}_i-\bar{x}}{\bar{s}}\right)^2$	$f(\bar{x}_i)=q\cdot e^{-\frac{t}{2}}$	$\frac{\beta-\bar{x}}{\bar{S}}$	$\Phi\left(\frac{\beta-\bar{x}}{\bar{S}}\right)$	$F(x)$
i	(α,β)	$\bar{x}_i$	Построение $f(x)$, $q=\frac{1}{\bar{s}\cdot\sqrt{2\pi}}$		Построение $F(x)$		
...	...	...		...	...	...	...
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*

Табл.2.2

2.2. Построить графики $f(x)$ *и* $F(x)$, совместив с гистограммой плотности относительной частоты на интервалах, и с графиком статистической функции распределения (первая работа) соответственно.

3. Выводы по проделанной работе.

Пример выполнения работы.

1.1.Используя данные, полученные при решении примера из лаб. работы 1 (см. (1.3),(1.4) и табл. 1.3), находим $a=36$, $\bar{S}=\sqrt{30.69}\approx 5.54$.Заполняем таблицу.

i	α	β	$z_i=\frac{(\beta-\bar{x})}{\bar{S}}$	$z_{i-1}=\frac{(\alpha-\bar{x})}{\bar{S}}$	$\Phi(z_i)$	$\Phi(z_{i-1})$	$p_i=\Phi(z_i)-\Phi(z_{i-1})$	p_i/h
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1	$-\infty$	22	-2.53	$-\infty$	-0.4943	-0.5	0.0057	0.001
2	22	26	-1.81	-2.53	-0.4648	-0.4943	0.0295	0.007
3	26	30	-1.08	-1.81	-0.3599	-0.4648	0.1049	0.026
4	30	34	-0.36	-1.08	-0.1406	-0.3599	0.2193	0.055
5	34	38	0.36	-0.36	0.1406	-0.1406	0.2812	0.070
6	38	42	1.08	0.36	0.3599	0.1406	0.2193	0.055
7	42	46	1.81	1.08	0.4648	0.3599	0.1049	0.026
8	46	50	2.53	1.87	0.4943	0.4648	0.0295	0.007
9	50	$+\infty$	$+\infty$	2.53	0.5	0.4943	0.0057	0.001

$\sum_i p_i = 1$

Табл. 2.1

1.2. Построить гистограмму p_i/h

Гистограмма p_i/h строится по данным 2,3-го и 9-го столбцов табл. 2.2 (рис. 2.1).

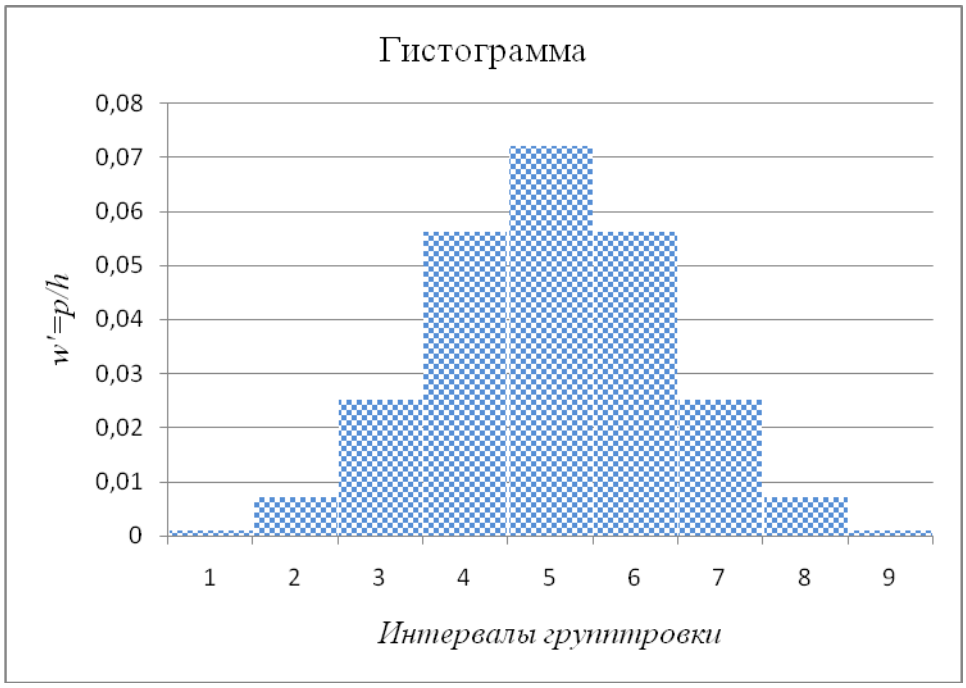


Рис 2.1

2.1. Заполнить таблицу "*Построение* $f(x)$ *и* $F(x)$"

i	(α, β)	$\bar{x}_i$	$f(\bar{x}_i)$	$\frac{\beta - \bar{x}}{\bar{S}}$	$\Phi\left(\frac{\beta - \bar{x}}{\bar{S}}\right)$	$F(x)$
...	...	...	...	...	...	...
1	*2*	*3*	*4*	*5*	*6*	*7*
1	(18, 22]	20	0.001	-2.53	-0.4943	0.0057
2	(22, 26]	24	0.007	-1.81	-0.4648	0.0352
3	(26, 30]	28	0.025	-1.08	-0.3599	0.1401
4	(30, 34]	32	0.056	-0.36	-0.1406	0.3594
5	(34, 38]	36	0.072	0.36	0.1406	0.6406
6	(38, 42]	40	0.056	1.08	0.3599	0.8599
7	(42, 46]	44	0.025	1.81	0.4648	0.9648
8	(46, 50]	48	0.007	2.53	0.4943	0.9943

9	(50,	52	0.001	$+\infty$	0.5	1.0000
…	…	…	…	…	…	…

2.2. Построить графики $f(x)$ *и* $F(x)$ (Рис. 2.2 и 2.3)

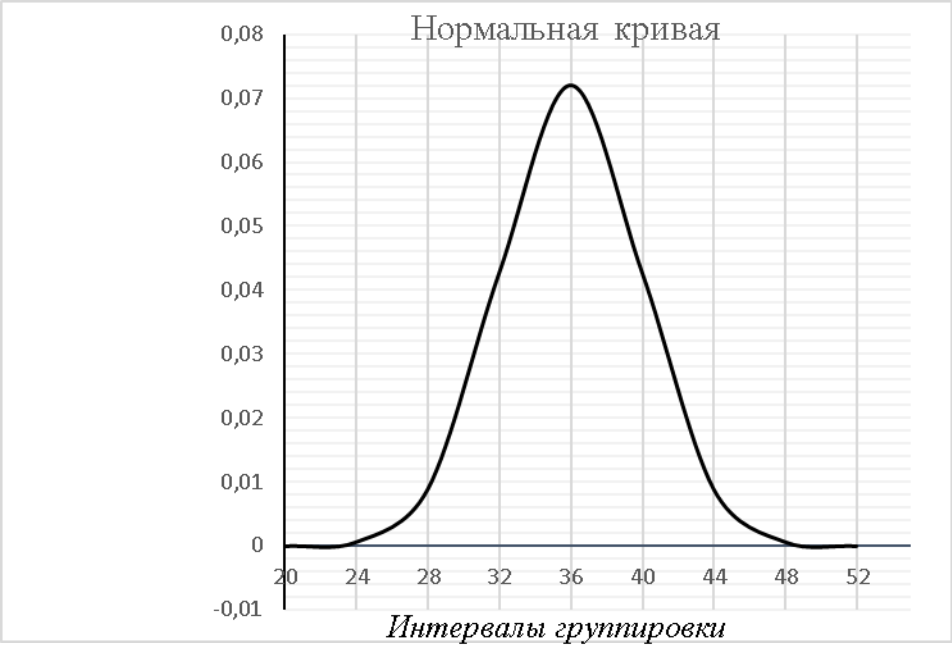
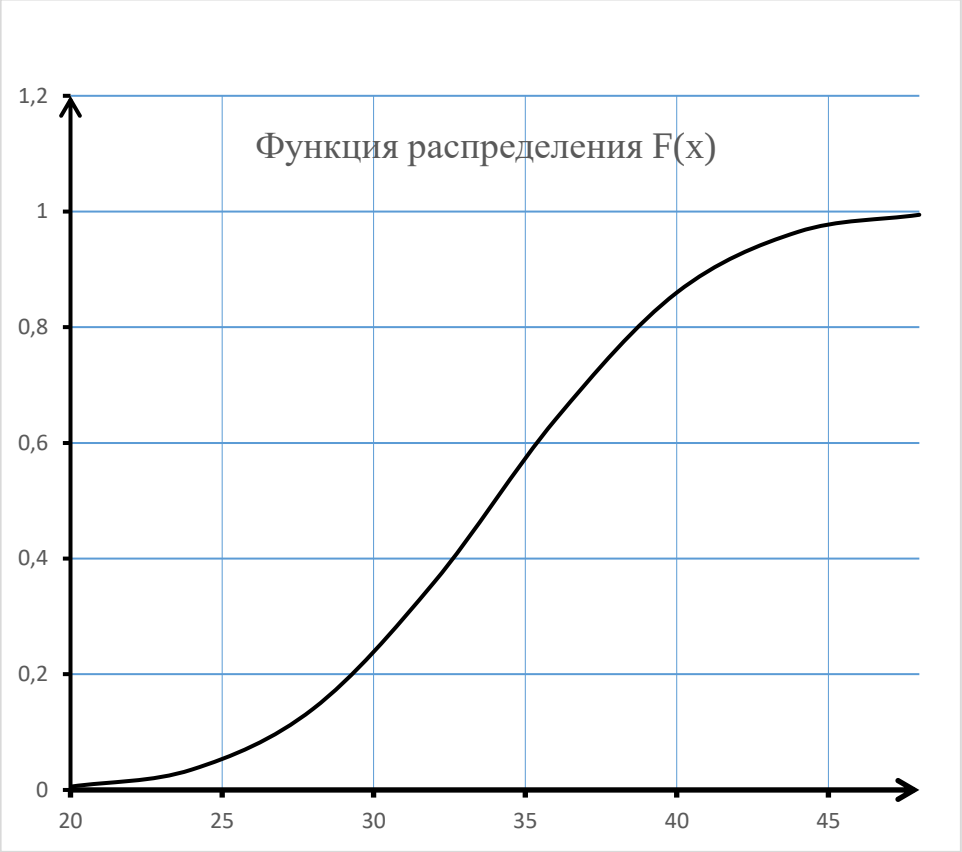


Рис. 2.2



Интервалы группировки

Рис 2.3

Визуальное сравнение этих графиков в случае их близости позволяет сделать вывод, что наше предположение о нормальности распределения признака X генеральной совокупности не противоречит полученным статистическим данным.

Замечание. Отметим, что математическая статистика располагает более формальными процедурами проверки гипотезы о нормальности распределения признака X, т.н. критериями согласия.

Лабораторная работа 3

Построение доверительного интервала

В лабораторной работе 3 рассматриваются задачи построения доверительного интервала для математического ожидания генеральной совокупности (генерального среднего) и генеральной дисперсии по выборке.

1. Основные понятия и формулы.

Поставим задачу оценки параметров в общем виде. Пусть распределение признака X– генеральной совокупности – задается функцией вероятностей $\varphi(x_i, \Theta) = P(X = x_i)$ (для дискретной случайной величины) или плотностью вероятностей $\varphi(x, \Theta)$ (для непрерывной случайной величины), которая содержит неизвестный параметр θ. Например, параметры a и σ^2 для нормального закона распределения.

По ряду причин не можем для вычисления параметра θ исследовать все элементы генеральной совокупности. Поэтому будем его оценивать по выборке, состоящей из значений $\{x_1, x_2, \ldots, x_n\}$, которые можно рассматривать как частные реализации n независимых случайных величин $\{X_1, X_2, \ldots, X_n\}$, каждая из которых имеет тот же закон распределения, что и случайная величина X.

В качестве такой оценки выберем функцию от элементов выборки $\theta_n^* = \theta_n^*(x_1, x_2, \ldots, x_n)$. Для конкретных значений элементов выборки эта оценка представляет собой одно число. Такие оценки называются *точечными* оценками параметров, так как на числовой оси они изображаются одной точкой. Задача состоит в том, чтобы найти такую оценку θ_n^*, которая была бы в определённом смысле наиболее близкой к оцениваемому параметру θ.

Как функция элементов выборки, оценка θ_n^* является случайной величиной. Определим её математическое ожидание. Оно, очевидно, будет зависеть от истинных числовых характеристик изучаемой величины X и от объёма выборки n.

Рассмотрим математическое ожидание этой оценки $M\{\vartheta_n^*\} = \theta + x(\theta, n)$, где $x(\theta, n)$ – некоторая функция истинного значения параметра θ. Желательно, чтобы функция $x(\theta, n)$ равнялась нулю. Тогда математическое ожидание оценки параметра будет равно истинному значению этого параметра. Оценка θ_n^*, обладающая таким свойством, называется *несмещённой* оценкой параметра θ.

Если при $n \to \infty$ оценка параметра сходится по вероятности к истинному значению параметра $(P(|\theta_n - \theta| > \varepsilon) \to 0)$, то θ_n^* называется *состоятельной* оценкой параметра θ. Если дисперсия оценки $D(\theta_n^*) = D(\theta, n)$ стремится к нулю при $n \to \infty$, то оценка будет *состоятельной.*

Различные оценки одного и того же параметра могут иметь разные дисперсии. Та из них, которая имеет наименьшую дисперсию, называется *эффективной* оценкой данного параметра.

Пусть параметр θ является числовой характеристикой случайной величины X (например, математическим ожиданием), а θ_n^* - некоторая оценка этого параметра,

полученная по выборке объема n. Поскольку θ_n^* является случайной величиной, желательно уметь оценивать "качество" равенства $\theta_n^* \approx \theta$ с тем, чтобы иметь представление к каким ошибкам может привести замена параметра θ его точечной оценкой и с какой степенью уверенности можно ожидать, что эти ошибки не выйдут за известные пределы.

Для ответа на эти вопросы в математической статистике используют интервальную оценку параметра.

Интервальной оценкой параметра θ называется числовой интервал $(\theta_n^{*(1)}, \theta_n^{*(2)})$, который с заданной вероятностью g (g - *доверительная вероятность*, уровень доверия или надежность оценки) накрывает неизвестное значение параметра θ. Границы интервала $(\theta_n^{*(1)}, \theta_n^{*(2)})$ и его величина находятся по выборочным данным и поэтому являются случайными величинами в отличие от оцениваемого не случайного параметра.

Доверительный интервал – это интервал со случайными границами, который накрывает оцениваемый параметр θ с заранее заданной вероятностью g. При этом границы доверительного интервала будут зависеть от вариант, объема выборки и *доверительной вероятности* g (заметим, что число g обычно выбирается достаточно близким к единице; в фармации, как правило, полагают γ = 0.95 или 0.99).

Выборочные распределения некоторых оценок θ_n^* (например, выборочной средней $\overline{X}$) симметричны относительно параметра θ, поэтому в этом случае целесообразно рассматривать симметричный относительно параметра θ доверительный интервал.

Наибольшее отклонение оценки θ_n^* от оцениваемого параметра θ, называется предельной ошибкой выборки $\Delta = |\theta_n^* - \theta|$. Она возникает потому, что исследуется не вся совокупность, а лишь ее часть, отобранная случайно.

Для построения доверительного интервала следует решить уравнение

$$P(|\theta_n^* - \theta| < \varepsilon_\gamma) = \gamma \qquad (3.1)$$

Не важно, каким образом были получены границы интервала $\theta_n^{*(1)}$ и $\theta_n^{*(2)}$, важен сам факт выполнения соотношения (3.1). Доверительный интервал даёт определённую информацию о точности оценки данного параметра.

С практической точки зрения можно утверждать, что если мы извлечем из генеральной совокупности, скажем, сто случайных выборок одинакового объема и построим по ним соответствующие доверительные интервалы (с одной и той же доверительной вероятностью g), то следует ожидать, что приблизительно в 100 g случаях (если g = 0.95, то в 95-ти), эти интервалы будут содержать параметр θ.

2. Доверительный интервал для математического ожидания генеральной совокупности при больших выборках.

2.1. Дисперсия D(*X*)= σ^2 известна.

В качестве оценки M(*X*)=*a* возьмем выборочную среднюю $\overline{X} = \frac{1}{n}\sum_{1\le i\le n} x_i$.

Эта оценка является несмещенной и состоятельной. Действительно, учитывая, что все X_i распределены с теми же параметрами, что и случайная величина X ($M[X_i] = a$ и $D[X_i] = \sigma^2$), получим:

$$M[\overline{X}] = M\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} M[X_i] = \frac{1}{n}a \cdot n = a \qquad (3.2)$$

$$D[\overline{X}] = D\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n} D[X_i] = \frac{1}{n^2}\sigma^2 \cdot n = \frac{\sigma^2}{n}, \qquad \lim_{n\to\infty} D[\overline{X}] = 0 \qquad (3.3)$$

На основании центральной предельной теоремы (цпт) при $n \to \infty$ распределение $\overline{X}$ неограниченно приближается к нормальному (практически, при n>40 распределение $\overline{X}$ можно считать приближенно нормальным), т.е. $\overline{X} \sim N\left(a, \frac{\sigma^2}{n}\right)$.

Для нахождения предельной ошибки выборки рассмотрим случайную величину

$$T = \frac{a - \overline{X}}{\frac{\sigma}{\sqrt{n}}} \qquad (3.3)$$

Она распределена по стандартному нормальному закону $T \sim N(0,1)$ (Рис. 3.1)

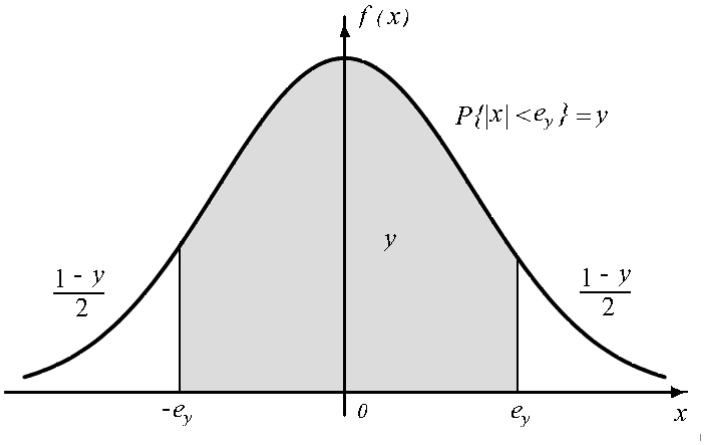


Рис. 3.1

$$M[T] = M\left[\frac{a-\bar{X}}{\frac{\sigma}{\sqrt{n}}}\right] = 0 \text{ и } D[T] = D\left[\frac{a-\bar{X}}{\frac{\sigma}{\sqrt{n}}}\right] = \frac{n}{\sigma^2} D[\bar{X}] = \frac{n}{\sigma^2} \cdot \frac{\sigma^2}{n} = 1$$

Для заданной доверительно вероятности γ решаем уравнение (3.1) относительно ε_γ:

$$P(|T| < \varepsilon_\gamma) = \gamma \sim P(-\varepsilon_\gamma < T < \varepsilon_\gamma) = \gamma \qquad (3.4)$$

$$P(-\varepsilon_\gamma < T < \varepsilon_\gamma) = 2\Phi(\varepsilon_\gamma)$$

Для определения ε_γ нужно, пользуясь таблицей функции Лапласа (приложением 1), решить уравнение

$$\Phi(\varepsilon_\gamma) = \frac{\gamma}{2} \qquad (3.5)$$

Подставим в (3.4) выражение (3.3) для T и найденное ε_γ

$$P\left(-\varepsilon_\gamma < \frac{a-\bar{X}}{\frac{\sigma}{\sqrt{n}}} < \varepsilon_\gamma\right)$$

Разрешая внутреннее неравенство относительно a, получим выражение:

$$P\left(\bar{X} - \frac{\varepsilon_\gamma \sigma}{\sqrt{n}} < a < \bar{X} + \frac{\varepsilon_\gamma \sigma}{\sqrt{n}}\right)$$

Итак, если $\bar{X}$ эмпирическое математическое ожидание нормально распределённой случайной величины с истинным математическим ожиданием a и известной дисперсией σ^2, то

$$a \in \left(\bar{X} - \frac{\varepsilon_\gamma \sigma}{\sqrt{n}};\ \bar{X} + \frac{\varepsilon_\gamma \sigma}{\sqrt{n}}\right) \qquad (3.6)$$

с вероятностью γ $\left(\text{здесь } \Delta = \frac{\varepsilon_\gamma \sigma}{\sqrt{n}}\right)$.

При этом величина ε_γ находится по таблицам нормального распределения, как решение уравнения (3.5).

2.2. Дисперсия D(*X*) = σ^2 неизвестна.

При большом объеме выборки практически достоверно, что $\sigma^2 \approx S^2$, где S^2 выборочная дисперсия (или исправленная выборочная дисперсия) . Учитывая, что при этом распределение $\bar{X}$ близко к нормальному, применим выше изложенную методику для нахождения доверительного ожидания, заменив в формуле (3.6) σ на S.

$$a \in \left(\bar{X} - \frac{\varepsilon_\gamma S}{\sqrt{n}};\ \bar{X} + \frac{\varepsilon_\gamma S}{\sqrt{n}}\right) \qquad (3.7)$$

3. Доверительный интервал для математического ожидания генеральной совокупности при малых выборках.

На практике часто приходится иметь дело с выборками небольшого объема $n<30$. В этом случае приведенный выше метод построения доверительного интервала для математического ожидания генеральной совокупности X является некорректным по двум причинам:

1) утверждение о нормальном распределении выборочной средней $\overline{X}$ является необоснованным, так как оно основано на центральной предельной теореме при больших n;
2) необоснованной становится замена неизвестной генеральной дисперсии σ^2 ее точечной оценкой S^2, которая возможна лишь при больших n.

3.1. Генеральная совокупность имеет *нормальное* распределение с известной σ^2.

Если признак X имеет нормально распределение с параметрами a и σ^2, т.е. $X \sim N(a, \sigma^2)$, то выборочная средняя $\overline{X}$ при любом n имеет нормальный закон распределения с параметрами a и σ^2/n.

В случае больших выборок из *любых* генеральных совокупностей нормальность распределения $\overline{X}$ обусловлена суммированием большого числа одинаково распределенных случайных величин. В случае же малых выборок, полученных из *нормальной* генеральной совокупности, нормальность распределения $\overline{X}$ связана с тем, что сумма любого числа нормально распределенных случайных величин также имеет нормальное распределение , $\overline{X} \sim N\left(a, \frac{\sigma^2}{n}\right)$. Таким образом, если известна σ^2, то доверительный интервал для a можно построить и для малых n, используя изложенную выше методику.

3.2. Генеральная дисперсия неизвестна.

На практике почти всегда генеральная дисперсия неизвестна (как и a).

Заменим ее исправленной выборочной дисперсией $\overline{S}^2 = \frac{1}{n-1}\sum_{1\le i\le n}(\overline{x}_i - \overline{X})^2$.

Рассмотрим статистику

$$t = \frac{a - \bar{X}}{\frac{\bar{S}}{\sqrt{n}}} \qquad (3.8)$$

Эта случайная величина имеет t-распределение Стьюдента с $k=n-1$ степенями свободы. Оно не зависит от неизвестных параметров распределения случайной величины X, а зависит только от числа k, и напоминает нормальное распределение (при $k \to \infty$ как угодно близко приближается к нему).

$t(n) = \frac{z}{\sqrt{\frac{\chi^2(n)}{n}}}$, $z \sim N(0,1)$, $\chi^2(n) = \sum_{i=1}^{n} z_i^2$, $z_i \sim N(0,1)$ –

распределение Стьюдента и χ^2 распределение.

Для определения предельной ошибки выборки решаем уравнение для заданной доверительной вероятности γ:

$$P(|t| < t_{\gamma,k}) = \Theta(t,k) = \gamma \qquad (3.9)$$

Функция $\Theta(t,k)$ табулирована (при $k \to \infty$ она неограниченно приближается к функции Лапласа $\Phi(t)$).

Рассматривая равенство $\Theta(t,k) = \gamma$, по таблице распределения Стьюдента для заданных k и γ находим $t_{\gamma,k}$.

Записывая (3.9) с учетом (3.8), получим

$$P\left(\left|\frac{a - \bar{X}}{\frac{\bar{S}}{\sqrt{n}}}\right| < t_{\gamma,k}\right) \qquad (3.10)$$

или, разрешая внутреннее неравенство относительно a:

$$P\left(\bar{X} - t_{\gamma,k}\frac{S}{\sqrt{n}} < a < \bar{X} + t_{\gamma,k}\frac{S}{\sqrt{n}}\right).$$

Следовательно, доверительный интервал для a:

$$a \in \left(\bar{X} - t_{\gamma,k}\frac{S}{\sqrt{n}} < a < \bar{X} + t_{\gamma,k}\frac{S}{\sqrt{n}}\right) \qquad (3.11)$$

Замечание. Для проведения выборочного наблюдения очень важно правильно установить объем выборки n. Для определения n необходимо задать надежность γ и предельную ошибку выборки Δ. Например, для случая большой выборки и известной σ^2, имеем $\Delta = \frac{\varepsilon_\gamma \sigma}{\sqrt{n}}$, отсюда $n = \left(\frac{\varepsilon_\gamma \sigma}{\Delta}\right)^2$.

4. Доверительный интервал для генеральной дисперсии.

Пусть распределение признака (случайной величины) X в генеральной совокупности является *нормальным* $X \sim N(a, \sigma^2)$. Рассмотрим статистику

$$\chi^2 = \frac{(n-1)\bar{S}^2}{\sigma^2} \qquad (3.12)$$

Эта случайная величина имеет распределение χ^2 с $k=n-1$ степенью свободы. Значения $\chi^2 \in [0, +\infty)$. Распределение χ^2 не является симметричным (Рис. 3.2) в отличие от нормального распределения $f(x)$ или распределения Стьюдента, поэтому невозможно использовать уравнение вида (3.4) для определения границ доверительного интервала.

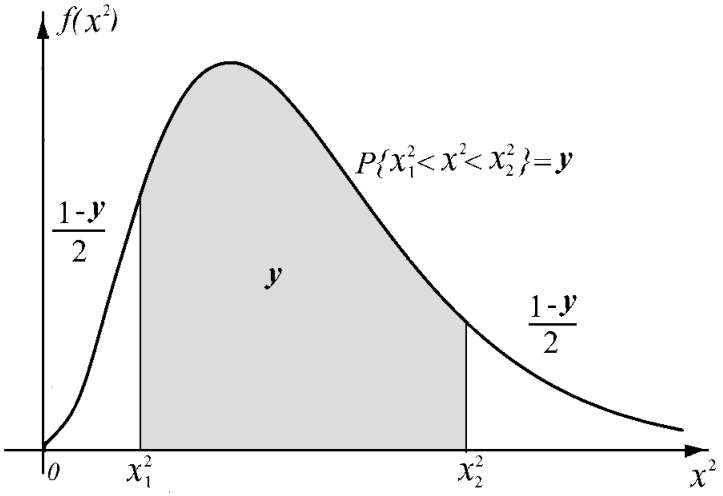


Рис. 3.2

Используем уравнение

$$P(\chi_1^2 \le \chi^2 < \chi_2^2) = \gamma \qquad (3.13)$$

Это уравнение содержит две неизвестные χ_1^2 и χ_2^2, поэтому имеет бесчисленное множество решений. Чтобы получить одно решение, нужно иметь одно дополнительное условие. Обычно χ_1^2 и χ_2^2 выбирают так, чтобы вероятности событий $\chi^2 < \chi_1^2$ и $\chi^2 > \chi_2^2$ были одинаковы. Вероятность попадания в интервал (χ_1^2 ; χ_2^2) равна γ. Выберем интервал так, чтобы оставшаяся вероятность $1 - \gamma$ была распределена поровну между интервалами $(0; \chi_1^2)$ и $[\chi_2^2 ; +\infty)$, т.е. положим

$P\{\chi^2 < \chi_1^2\} = (1-\gamma)/2$ и $P\{\chi^2 \ge \chi_2^2\} = (1-\gamma)/2$.

При этом $P\{\chi^2 \ge \chi_1^2\} = 1 - (1-\gamma)/2 = (1+\gamma)/2$. Границы χ_1^2 и χ_2^2 определяем из двух уравнений:

$$P(\chi^2 \ge \chi_1^2) = \frac{1+\gamma}{2} = \alpha_1 \quad \text{и} \quad P(\chi^2 \ge \chi_2^2) = \frac{1-\gamma}{2} = \alpha_2 \qquad (3.14)$$

По таблицам Приложения 4 решаются уравнения (3.14) относительно χ_1^2 и χ_2^2 при заданных значениях γ и $k=n-1$ используя таблицу значений χ_α^2, для которых $P\{\chi^2 \ge \chi_\alpha^2\} = \alpha$.

Учитывая (3.13), получим:

$$P\left(\chi_1^2 \le \frac{(n-1)S^2}{\sigma^2} < \chi_2^2\right) = \gamma \text{ или}$$

$$P\left(\frac{(n-1)S^2}{\chi_2^2} \le \sigma^2 < \frac{(n-1)S^2}{\chi_1^2}\right) = \gamma$$

Таким образом, доверительный интервал для дисперсии:

$$\sigma^2 \in \left(\frac{(n-1)\bar{S}^2}{\chi_2^2}; \frac{(n-1)\bar{S}^2}{\chi_1^2}\right) \text{ с вероятностью } \gamma \quad (3.15)$$

Доверительный интервал для среднего квадратичного отклонения:

$$\sigma \in \left(\frac{\sqrt{n-1}\cdot S}{\chi_2}; \frac{\sqrt{n-1}\cdot S}{\chi_1}\right) \text{ с вероятностью } \gamma \quad (3.16)$$

Пример 1. Найти доверительный интервал с доверительной вероятностью 0.95 для неизвестного математического ожидания по выборке объема 50 из примера, рассмотренного в лаб. работе 1 (большая выборка). Предположение о том, что теоретическая случайная величина имеет нормальное распределение, не делается.

Решение. Воспользуемся формулой (3.7). Учитывая, что в настоящем примере $\bar{s}^2 = 30.69$, $\bar{s} = 5.54$, $x = 36$, при g= 0.95 получим:

$$\Phi(\varepsilon_\gamma) = \frac{\gamma}{2} = 0.475, \quad \varepsilon_\gamma = 1.96, \quad \Delta = \frac{\varepsilon_\gamma \bar{S}}{\sqrt{n}} = 1.54 \text{ и } a \in (34.46; 37{,}54).$$

Пример 2. Найти доверительный интервал с доверительной вероятностью 0.95 для неизвестного математического ожидания по выборке объема 10 (малая выборка):

17.58 18.86 17.34 16.39 17.65 18.54 19.08 18.75 17.45 15.63

Предполагается, что $X \sim N(a, \sigma^2)$, генеральная дисперсия σ^2, как и a, неизвестны.

Решение.

$$x_i = \frac{1}{10}(17.58 + 18.86 + 17.34 + 16.39 + 17.65 + 18.54 + 19.08 + 18.75 + 17.45 + 15.63 = 17.7$$

$$s^2 = \frac{1}{n}\sum_{i=1}^{10} x_i^2 - \bar{x}^2 = = \frac{1}{10}(17.58^2 + 18.86^2 + 17.34^2 + 16.39^2 + 17.65^2 + 18.54^2 + 19.08^2 + 18.75 + 17.$$

$$\bar{s}^2 = \frac{n}{n-1}s^2 = \frac{10}{9}3.91 = 4.35, \quad \bar{s} = 2.09$$

Для g= 0.95 и k=n-1 =9 по таблице распределения Стьюдента находим $t_{\gamma,k} = 2.26$, тогда

$$\Delta = t_{\gamma,k}\frac{\bar{S}}{\sqrt{n}} = 2.26\frac{2.09}{3.16} = 1.49 \text{ и } a \in (16.24; 19.22).$$

Пример 3. Найти доверительный интервал с доверительной вероятностью 0.9 для неизвестной генеральной дисперсии по выборке объема 10 из примера 2.

Решение.

$P(\chi^2 \geq \chi_1^2) = 0.95$ и $P(\chi^2 \geq \chi_2^2) = 0.05$, $k = 9$. По таблице Приложения 4 находим:

$\chi_1^2 = \chi^2_{\frac{1+\gamma}{2},9} = \chi^2_{0.95,9} = 3.32$, $\chi_2^2 = \chi^2_{\frac{1-\gamma}{2},9} = \chi^2_{0.05,9} = 16.9$

Учитывая, что в настоящем примере

$\bar{s}^2 = 4.35$, $\bar{s} = 2.09$, $x = 17.73$, при g= 0.9 получим (3.15) и (3.16):

$$\frac{(n-1)\bar{S}^2}{\chi_1^2} = \frac{9 \cdot 4.35}{3.32} = 11.79 \text{ и } \frac{(n-1)\bar{S}^2}{\chi_2^2} = \frac{9 \cdot 4.35}{16.9} = 2.32,$$

$\sigma^2 \in (2.32; 11.79)$, $\sigma \in (1.52; 3.43)$ с вероятностью $\gamma = 0.9$.

5. Порядок выполнения работы

В настоящей работе требуется

1) построить доверительные интервалы для большой (n=50) выборки с доверительной вероятностью, приблизительно равной γ (γравно 0.9, 0.95 и 0.99). Так же, как в предыдущем задании, необходимо изобразить полученные интервалы на общей числовой оси. Исходные данные следует взять из лаб. работы 1. Границы доверительного интервала определяются по формуле (3.7) при $\sigma^2 \approx \bar{S}^2$ (см. пример 1).

2) построить доверительные интервалы с доверительными вероятностями равными 0.9, 0.95 и 0.99 по выборке объема 10 (малая выборка), отвечающей случайной величине $X \sim N(\mu, \sigma)$. Вычисления проводить с использованием формул (3.11) аналогично примеру 2, определяя значения $t_{\gamma,k}$ по приложению 2. Кроме того, следует изобразить полученные интервалы на общей числовой оси и сделать заключение о влиянии величины доверительной вероятности на ширину интервала.

3) найти доверительный интервал с доверительной вероятностью 0.9 для неизвестной генеральной дисперсии по выборке объема 10 (пример 3).

Номер варианта задания, содержащий 10 случайных чисел, следует получить у преподавателя.

6. Контрольные вопросы

12. Что называется доверительным интервалом?
13. Что такое доверительная вероятность?
14. Как вычисляется исправленная выборочная дисперсия?
15. Какое распределение называется распределением Стьюдента?

7.Варианты малых выборок

№1

19.17	21.24	20.06	21.46	21.83	21.69	21.43	22.18	23.09	20.87

№2

20.01	21.52	19.95	19.50	20.71	20.45	17.27	17.87	18.95	20.21

№3

23.79	27.09	20.41	24.44	23.46	26.09	24.53	24.59	22.13	21.84

№4

19.13	16.31	18.12	16.72	19.54	17.88	18.66	18.47	17.84	18.84

№5

17.47	17.62	17.26	16.11	17.25	16.21	16.22	15.75	17.32	15.52

№6

17.58	18.86	17.34	16.39	17.65	18.54	19.08	18.75	17.45	15.63

№7

17.58	18.86	17.34	16.39	17.65	18.54	19.08	18.75	17.45	15.63

№8

15.62	14.22	14.56	15.73	14.03	15.53	13.81	14.05	13.27	15.91

№9

18.31	24.60	21.49	19.99	16.49	18.32	19.16	14.12	15.63	19.06

№10

24.28	21.20	25.38	24.13	21.78	29.29	27.67	22.04	24.55	25.94

№11

14.46	12.25	10.59	8.15	11.28	15.19	11.45	12.10	12.23	11.04

№12

11.09	11.51	12.39	13.80	15.66	13.87	10.89	10.77	12.78	15.47

№13

11.23	13.41	13.84	10.97	12.39	11.96	11.36	12.16	10.43	15.78

№14

14.82	16.52	17.16	16.53	12.62	11.68	12.83	11.13	13.15	15.03

№15

18.07	16.59	16.52	17.74	15.18	16.93	14.79	15.40	13.78	15.40

№16

21.58	17.41	12.43	14.60	20.36	17.77	17.34	17.68	13.85	18.50

№17

13.80	6.10	7.42	8.49	6.73	9.34	13.17	12.37	15.72	10.76

№18

22,99	23,03	23,07	23,25	23,06	23,32	22,84	22,92	23,14	22,58

№19

42,60	42,89	41,87	41,70	43,08	42,22	41,70	42,32	41,62	41,94

№20

54,45	51,40	55,35	52,93	52,75	54,11	51,09	54,59	53,68	52,62

Лабораторная работа 4

Совместный закон распределения и числовые характеристики двух случайных величин

1. Основные понятия и формулы. Примеры

Часто результат испытания характеризуется не одной случайной величиной, а несколькими – *системой случайных величин* $X_1, X_2,..., X_n$, которую называют n – мерной случайной величиной или случайным вектором $X=(X_1, X_2,..., X_n)$. Любая случайная величина $i = \overline{1,n}$) является функцией элементарных событий ω, образующих пространство элементарных событий Ω $(\omega \in \Omega)$, поэтому и многомерная случайная величина есть функция элементарных событий ω: $(X_1, X_2, \dots, X_n) = f(\omega)$. Каждому элементарному событию ω ставится в соответствие n действительных чисел $x_1, x_2, \dots, x_n$, которые приняли случайные величины $X_1, X_2,..., X_n$ в результате испытания. Вектор $x=(x_1, x_2, \dots, x_n)$ есть *реализация* случайного вектора $X=(X_1, X_2,..., X_n)$. Случайные величины $X_1, X_2,..., X_n$ могут быть как *дискретными*, так и *непрерывными*. В этой работе будем рассматривать двумерные дискретные случайные величины (X,Y). Геометрически реализацию двумерную случайную величину можно изобразить случайной точкой $A(x,y)$ (или случайным вектором (x,y) плоскости xOy) при этом X и Y назовем компонентами двумерного вектора(X,Y).

Полным описанием двумерной случайной величины (X,Y) является *закон ее распределения*. Если множество возможных значений случайной величины (X,Y) конечно, такой закон может быть задан в форме таблицы, содержащей *все возможные сочетания значений* каждой из одномерных случайных компонент X и Y и *соответствующие им вероятности*.

Пусть дискретные случайные величины X и Y получают свои значения в результате одного и того же случайного эксперимента, и принимают значения $x_1, x_2, \dots, x_m$ и $y_1, y_2, \dots, y_n$, соответственно, где m и n некоторые целые положительные числа.

Для $1 \le i \le m$ и $1 \le j \le n$ определим вероятности P_{ij}, положив

$$P_{ij} = P(X = x_i , Y = y_j), \qquad (4.1)$$

где P_{ij} равно вероятности того, что события $\{X = x_i\}$ и $\{Y = y_j\}$ произойдут одновременно. Равенство (4.1) задает *совместный закон распределения* пары случайных величин X и Y, или закон распределения двумерного случайного вектора (X, Y).

Этот закон запишем в виде следующей таблицы:

X \ Y	y_1	y_2	…	y_j	…	y_n	P_x
x_1	P_{11}	P_{12}	…	P_{1j}	…	P_{1n}	$P_{1\cdot}$
x_2	P_{21}	P_{22}	…	P_{2j}	…	P_{2n}	$P_{2\cdot}$
…	…	…	…	…	…	…	…
x_i	P_{i1}	P_{i2}	…	P_{ij}	…	P_{in}	$P_{i\cdot}$
…	…	…	…	…	…	…	…
x_m	P_{m1}	P_{m2}	…	P_{mj}	…	P_{mn}	$P_{m\cdot}$
P_y	$P_{\cdot 1}$	$P_{\cdot 2}$		$P_{\cdot j}$		$P_{\cdot n}$	1

Табл. 4.1

в которой

$$P_{i\cdot} = \sum_{j=1}^{n} P_{ij} \quad , \quad 1 \le i \le m \,,$$

$$P_{\cdot j} = \sum_{i=1}^{m} P_{ij} \quad , \quad 1 \le j \le n \,,$$

- вероятности событий $(X = x_i)$ и $(Y = y_i)$ (4.2)

т.е. $P_x = P_{i\cdot}$ равна сумме вероятностей, расположенных в i-й строке, а $P_y = P_{\cdot j}$ равна сумме вероятностей из j-го столбца.

События $\{(X = x_i)(Y = y_j)\}(1 \le i \le m, 1 \le j \le n)$ несовместны и единственно возможны, т.е. образуют полную группу событий, поэтому

$$\sum_{i=1}^{m}\sum_{j=1}^{n} P_{ij} = \sum_{i=1}^{m} P_{i\cdot} = \sum_{j=1}^{n} P_{\cdot j} = 1 \quad \text{и} \; P_{i\cdot} = P(X = x_i), \quad P_{\cdot j} = P(Y = y_j). \qquad (4.3)$$

Индивидуальные законы распределения самих дискретных случайных величин X и Y можно записать как

X	x_1	x_2	…	x_i	…	x_m
P	$P_{1\cdot}$	$P_{2\cdot}$	…	$P_{i\cdot}$	…	$P_{m\cdot}$

Табл. 4.2

Y	y_1	y_2	…	y_j	…	y_n
P	$P_{\cdot 1}$	$P_{\cdot 2}$	…	$P_{\cdot j}$	…	$P_{\cdot n}$

Табл.4.3

Если зафиксировать значения одного из компонентов (X, Y), например, положить $X = x_i$, то полученное распределение случайной величины Y называется *условным* распределением Y при условии $X = x_i$. Вероятности $P_j(x_i) = P\left(\{Y = y_j\} \middle| \{X = x_i\}\right) \equiv p(j|i)$ этого распределения – *условные вероятности* события $\{Y = y_j\}$ при условии, что событие $\{X = x_i\}$ произошло.

$$P_j(x_i) = p(j|i) = \frac{P\{(X = x_i)(Y = y_j)\}}{P\{X = x_i\}} = \frac{P_{ij}}{P_{i\bullet}} \qquad (4.4)$$

Аналогично:

$$P_i(y_j) = p(i|j) = \frac{P\{(X = x_i)(Y = y_j)\}}{P\{Y = y_j\}} = \frac{P_{ij}}{P_{\bullet j}} \qquad (4.5)$$

Замечание. Закон распределения двумерного случайного вектора может быть задан и с помощью функции распределения вероятностей $F(x, y) = P(X < x, Y < y)$.

Для *непрерывного* случайного вектора - только с помощью $F(x, y)$ или совместной плотности распределения вероятностей

$\varphi_1(x) = \int_{-\infty}^{+\infty} \varphi(x, y)dy$ и $\varphi_2(y) = \int_{-\infty}^{+\infty} \varphi(x, y)dx$ – плотности распределения вероятности одномерных случайных величин *X* и *Y*, соответственно. Условные плотности распределения вероятностей можно в этом случае выразить через совместную плотность следующим образом:

$$\varphi_y(x) \equiv \varphi(x|y) = \frac{\varphi(x, y)}{\int_{-\infty}^{+\infty} \varphi(x, y)dx}, \qquad \varphi_x(y) \equiv \varphi(y|x) = \frac{\varphi(x, y)}{\int_{-\infty}^{+\infty} \varphi(x, y)dy} \quad (4.4', 4.5')$$

Условная плотность $\varphi(x|y)$ есть кривая, получаемая сечением поверхности $z = \varphi(x, y)$ плоскостью $Y = y$ параллельной плоскости xOz и отсекающей на оси *Oy* отрезок *y*.

Геометрически функция распределения $F(x, y)$ означает вероятность попадания случайной точки (*X*, *Y*) в заштрихованную область (Рис.6.1) – бесконечный квадрант, лежащий левее и ниже точки $A(x, y)$.

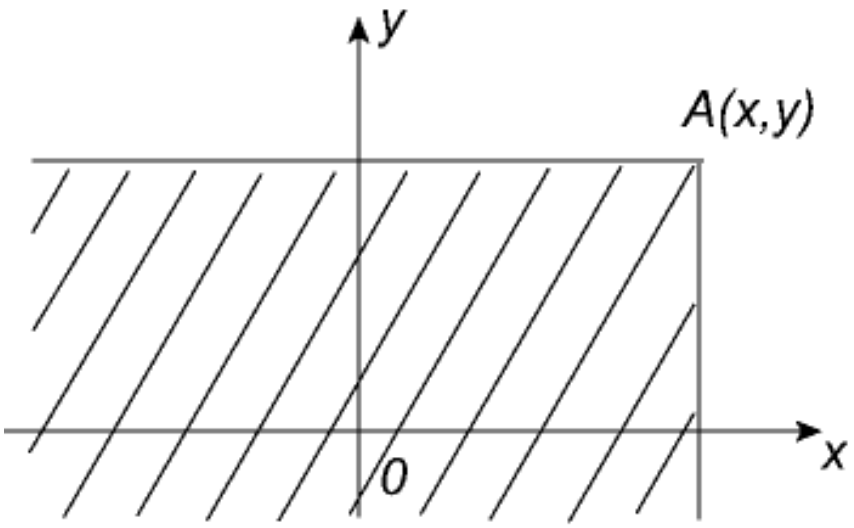


Рис.4.1

Для дискретной случайно величины (*X*, *Y*) функция распределения определяется по формуле:

$$F(x,y) = \sum_{x<x_i,\, y<y_j} P_{ij} \qquad (4.6)$$

Основные свойств функции распределения двумерной случайной величины аналогичны свойствам функции распределения одномерной случайной величины.

1. $F(x,y)$ – неотрицательная функция: $0 \le F(x,y) \le 1$
2. $F(x,y)$ – неубывающая функция по каждому из своих аргументов:
   $x_2 > x_1 \Rightarrow F(x_2,y) \ge F(x_1,y)$
   $y_2 > y_1 \Rightarrow F(x,y_2) \ge F(x,y_1)$
3. $F(x,-\infty) = F(-\infty,y) = F(-\infty,-\infty) = 0$
4. $F(x,+\infty) = F_x(x)$ и $F(+\infty,y) = F_y(y)$, где $F_x(x)$ и $F_y(y)$ – функции распределения одномерных случайных компонент X и Y т.е.
   $F_x(x) = P(X < x), \qquad F_y(y) = P(Y < y)$.
5. $F(+\infty,+\infty) = 1$.

Графиком двумерной случайной величины (X, Y) является ступенчатая поверхность, ступени которой соответствуют скачкам функции $F(x,y)$.

Вероятность попадания случайной точки (X, Y) внутрь прямоугольника с вершинами $A(x_1,y_2), B(x_2,y_2), C(x_2,y_1), D(x_1,y_1)$ (Рис.4.2)

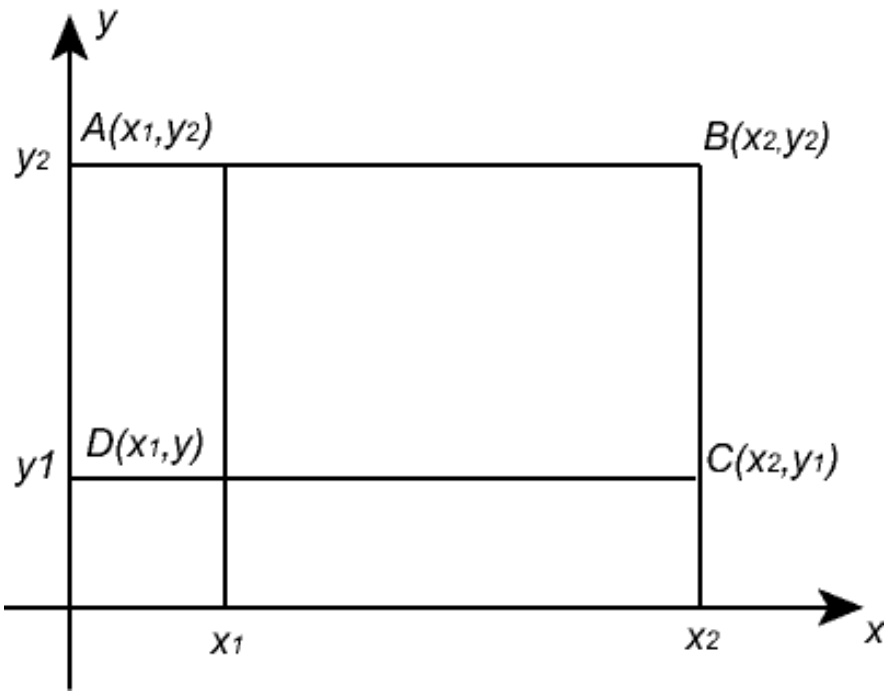


Рис.4.2

вычисляется по формуле:

$$P\{(x_1 \le X < x_2)(y_1 \le Y < y_2)\} = F(x_2,y_2) - F(x_1,y_2) - F(x_2,y_1) + F(x_1,y_1) \quad (4.7)$$

При изучении свойств двумерной случайной величины (*X, Y*) можно говорить о математических ожиданиях отдельных компонент M(*X*) и M(*Y*) , и об их дисперсиях D(*X*) и D(*Y*). Однако знание характеристик изолированных компонент не позволяет делать выводы о существовании статистической связи между этими компонентами и о характере этой связи. При изучении многомерных величин дополнительно привлекают такие характеристики, которые отражают статистическую связь. К таким характеристикам относятся смешанные моменты случайных величин.

Пусть (*X, Y*) двумерная дискретная случайная величина с возможными значениями (x_i, y_j), *i* = 1, 2, …, *m*, *j* = 1, 2, …, n, и с вероятностями P_{ij} , тогда:

Начальным моментом ν_{gh} порядка $k = g + h$ двумерной величины (X, Y) называется математическое ожидание произведения $X^g \cdot Y^h$:

$$\nu_{gh} = M[X^g \cdot Y^h] = \sum_{i,j}^{m,n} x_i^g y_j^h P_{ij} \qquad (4.7)$$

Здесь *g* и *h*– целые числа, $g \geq 0, h \geq 0, g + h = k, k = 0,1,2, \ldots$

Число моментов порядка *k* будет равно *k* + 1. При *k* = 0 имеется всего один момент ν_{00} , который всегда равен единице (условие нормировки вероятностей). При *k* = 1 имеется два момента ν_{10} и ν_{01} , которые совпадают с математическими ожиданиями величин *X* и *Y*:

$$\nu_{10} = M[X^1 \cdot Y^0] = M[X] = \sum_{i,j}^{m,n} x_i P_{ij} = \sum_{i=1}^{n} x_i P_{i\bullet} = m_x \qquad (4.8)$$

$$\nu_{01} = M[X^0 \cdot Y^1] = M[Y] = \sum_{i,j}^{m,n} y_j P_{ij} = \sum_{j=1}^{m} y_j P_{\bullet j} = m_y \qquad (4.9)$$

Для анализа статистической связи межу компонентами *X* и *Y* используются центральные моменты.

Центральный момент μ_{gh} порядка *k* двумерной величины (*X, Y*) - это математическое ожидание произведения $(X - m_x)^g (Y - m_y)^h$, такое, что *g* + *h* = *k*, т.е.

$$\mu_{gh} = M[(X - m_x)^g (Y - m_y)^h] = \sum_{(i,j)}^{(m,n)} [(x_i - m_x)^g (y_j - m_y)^h] P_{ij} \qquad (4.10)$$

В формуле (4.10) *g* и *h*– целые числа, $g \geq 0, h \geq 0, g + h = k, \quad k = 0,1,2, \ldots$)

Очевидно, что $\mu_{00} = \sum_{i,j}^{m,n} P_{ij} = 1, \quad \mu_{10} = \mu_{01} = 0$.

Рассмотрим моменты 2-го порядка. Момент μ_{20} является дисперсией $D[X]$ случайной величины *X*, а момент μ_{02} – дисперсией $D[Y]$ величины *Y*:

$$\mu_{20} = M[(X - m_x)^2] = D[X] = \sum_{i=1}^{m} (x_i - m_x)^2 P_{i\bullet} = \sum_{i=1}^{n} x_i^2 P_{i\bullet} - m_x^2, \qquad \sigma_x = \sqrt{D[X]}$$

$$\mu_{02} = M\left[(Y - m_y)^2\right] = D[Y] = \sum_{j=1}^{n} (y_j - m_y)^2 P_{\bullet j} = \sum_{i=1}^{n} y_j^2 P_{\bullet j} - m_y^2, \qquad \sigma_y \sqrt{D[Y]} \quad (4.11)$$

Смешанный момент μ_{11} имеет первостепенное значение при изучении зависимости между случайными величинами. Этот момент принято называть *ковариационным моментом*, моментом связи или просто *ковариацией* и обозначать через K_{xy} .

$$K_{xy} = M\left[(X - m_x)(Y - m_y)\right] = \sum_{i,j}^{m,n} (x_i - m_x)(y_j - m_y) P_{ij} \qquad (4.12)$$

Пользуясь свойствами математического ожидания, (4.12) можно переписать в виде:

$$K_{xy} = M[XY] - m_x m_y \qquad (4.12')$$

Ковариационный момент характеризует линейную связь между рассматриваемыми величинами.

Случайные величины *X* и *Y* называются статистически независимыми, если их совместная функция распределения $F(x, y) = F_x(x) F_y(y)$ (при этом $P_{ij} = P_{i\bullet} P_{\bullet j}$).

Если *X* и *Y* являются статистически независимыми величинами, то ковариационный момент равен нулю. Действительно, если *X* и *Y* независимы, то $P_{ij} = P_i P_j$ и, следовательно,

$$\mu_{11} = \sum_{i=1}^{n} (x_i - m_x) P_{i\bullet} \sum_{j=1}^{m} (y_j - m_y) P_{\bullet j} = \mu_{10}\mu_{01} = 0$$

Таким образом, если величины *X* и *Y* являются статистически независимыми, то $K_{xy} = 0$. Обратное утверждение не является верным, т.е. если $K_{xy} = 0$, то это ещё не значит, что данные величины являются статистически независимыми. Равенство нулю ковариационного момента означает, что между величинами *X* и *Y* отсутствует *линейная* связь. Однако может существовать *нелинейная* связь. Форма связи при $K_{xy} = 0$ определяется моментами более высокого порядка.

Отсюда, в частности, следует, что если ковариация двух случайных величин не равна нулю, то они стохастически зависимы.

Приведем некоторые свойства корреляционного момента:

- $K_{xy} = K_{yx}$,
- $K_{xx} = M\left[(X - m_x)^2\right] = D(X)$,
- $|K_{xy}| \le \sqrt{K_{xx} K_{yy}} = \sqrt{D(X) D(Y)} = \sigma_x \sigma_y$

Коэффициент корреляции случайных величин *X* и *Y* определяется равенством

$$r = r(X,Y) = \frac{K_{xy}}{\sigma_x \sigma_y}. \qquad (4.\ 13)$$

Коэффициент корреляции является безразмерной характеристикой, которая используется в качестве меры *линейной зависимости* случайных величин. Очевидно, что $|r| \le 1$, при этом, чем его модуль ближе к единице, тем, вообще говоря, *теснее* линейная зависимость между величинами.

Понятно, что коэффициент корреляции и ковариация обращаются в нуль одновременно; следовательно, коэффициент корреляции независимых случайных величин равен нулю.

Напомним, что $|r| \le 1$; кроме того, $|r| = 1$ тогда и только тогда, когда $Y = a + bX$ с вероятностью единица, где a и b - некоторые постоянные.

Если коэффициент корреляции случайных величин X и Y равен нулю, то такие величины называются *некоррелированными*. На практике некоррелированность часто отождествляют с независимостью, что, вообще говоря, неправомерно.

Прямая, имеющая уравнение

$$y = m_y + r\frac{\sigma_y}{\sigma_x}(x - m_x) \quad , \qquad (4.14)$$

называется *прямой средней квадратичной регрессии* Y на X .

Эта прямая наилучшим (в смысле среднего квадратичного) образом приближает случайную величину Y линейной функцией $a + bX$.

Меняя в соотношении (6.8) X и Y ролями, мы получаем прямую средней квадратичной регрессии X на Y

$$x = m_x + r\frac{\sigma_x}{\sigma_y}(y - m_y)\,. \qquad (4.15)$$

Заметим, что линии регрессии (4.13) и (4.14) пересекаются в точке $M(m_x, m_y)$.

Функцией регрессии (или просто *регрессией*) Y на X называется *условное математическое ожидание* $M_x(Y) \equiv M(Y|X=x)$ случайной величины Y при условии, что случайная величина X приняла значение x.

График $M(Y|X=x)$, как функции x, называют *линией регрессии*.

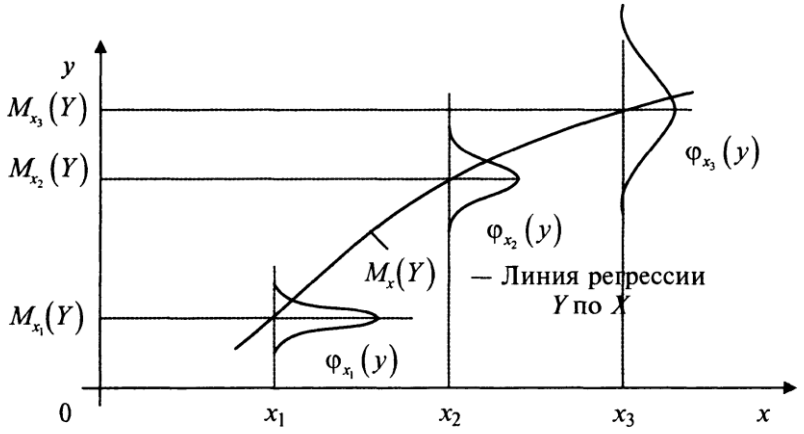


Рис. 4.3

На рис. 4.3 для непрерывного случайного вектора зависимость между X и Y проявляется в том, что с изменением x меняется как распределение Y (изменяется $\varphi_{X}(y)$), так и условное математическое ожидание $M_X(Y)$. На нем же показана зависимость $M_X(Y)$ от x, т.е. линия регрессии Y по X.

Регрессия Y на X обладает оптимальным свойством. Она, как функция X, наилучшим образом приближает случайную величину Y в смысле среднего квадратичного, т.е. дисперсия $D\left((Y-g(X))^2\right)$ принимает минимальное значение, когда $g(x) = M(Y|X=x)$.

В дискретном случае (6.1) регрессия Y на X определена лишь в точках x_i, $1 \le i \le m$, и принимает в них значения

$$M_{x_i}(Y) \equiv M(Y \mid X = x_i) = \sum_{j=1}^{n} y_j p(j \mid i) = \sum_{j=1}^{n} y_j \frac{P_{ij}}{P_{i\cdot}} = \frac{1}{P_{i\cdot}} \sum_{j=1}^{n} y_j P_{ij}, 1 \le i \le m, \qquad (4.16)$$

являясь, таким образом, средневзвешенной зависимостью Y от значений случайной величины X.

Пример 4.1. По закону распределения двумерного случайного вектора (X, Y)

Y \ X	3	5	P_x
2	0.3	0.1	0.4
3	0.4	0.2	0.6
P_y	0.7	0.3	1

Табл.4.4

найти:

1) математические ожидания и стандартные отклонения X и Y;
2) корреляционный момент и коэффициент корреляции X и Y;
3) уравнениепрямой средней квадратичной регрессии Y на X;
4) значения регрессии Y на X и X на Y;

5) сравнить между собой при каждом значении *X* приближения средних значений *Y*, полученные по функции регрессии и по уравнению прямой средней квадратичной регрессии.

Решение:

1) Формулы (4.8) и (4.9):

$M(X) = 2 \cdot 0.4 + 3 \cdot 0.6 = 2.6$

$M(Y) = 3 \cdot 0.7 + 5 \cdot 0.3 = 3.6$

$D(X) = 2^2 \cdot 0.4 + 3^2 \cdot 0.6 - 2.6^2 = 0.24$

$\sigma_x = \sqrt{0.24} = 0.4898$

$D(Y) = 3^2 \cdot 0.7 + 5^2 \cdot 0.3 - 3.6^2 = 0.84$

$\sigma_y = \sqrt{0.84} = 0.9165$

2) Формулы (4.12) или (4.12') и (4.13):

$K_{xy} = 2 \cdot (3 \cdot 0.3 + 5 \cdot 0.1) + 3 \cdot (3 \cdot 0.4 + 5 \cdot 0.2)) - 2.6 \cdot 3.6 = 0.04$,

$$r = \frac{0.04}{0.4898 \cdot 0.9165} = 0.0891$$

3) Формула (4.14) приводит к уравнению линейной регрессии

$$y_{лин.} = 3.6 + 0.0891 \frac{0.9165}{0.4898}(x - 2.6) = 3.157 + 0.1667x ;$$

$y_{лин.}\big|_{x=2} = 3.5$, $y_{лин.}\big|_{x=3} = 3.667$

4) Формула (4.16) дает:

$$M(Y \mid X = 2) = \frac{1}{0.4}(3 \cdot 0.3 + 5 \cdot 0.1) = 3.5$$

$$M(Y \mid X = 3) = \frac{1}{0.6}(3 \cdot 0.4 + 5 \cdot 0.2) = 3.667.$$

Таким образом,

X	$y_{лин.}$	*M(Y|X)*
2	3.5	3.5
3	3.667	3.667

Табл. 4.5

Заключение: величины, вычисленные путем подстановки возможных значений X в уравнение прямой средней квадратичной регрессии и в функцию регрессии, практически совпадают.

2. Порядок выполнения работы

Вариант задания – закон распределения пары дискретных случайных величин X и Y следует получить у преподавателя. Требуется найти:

математические ожидания и стандартные отклонения X и Y;

корреляционный момент и коэффициент корреляции X и Y;

уравнение прямой средней квадратичной регрессии Y на X;

значения регрессии Y на X и сравнить между собой при каждом значении X приближения средних значений Y, полученные по функции регрессии и по уравнению прямой средней квадратичной регрессии.

Порядок выполнения работы как в примере 4.1.

3. Контрольные вопросы

16. Какие случайные величины называется дискретными?
17. Что называется совместным законом распределения пары случайных величин X и Y?
18. Что такое ковариация?
19. Как вычисляется коэффициент корреляции?
20. Какие свойства коэффициента корреляции вы можете назвать?

4. Варианты самостоятельных работ

№1

X \ Y	1.1	1.8	2.3	2.5
2.5	0.08	0.01	0.31	0.10
3.1	0.29	0.17	0.03	0.01

№2

X \ Y	3.1	3.5	4.4	4.9
1.1	0.07	0.04	0.29	0.01
1.8	0.32	0.25	0.01	0.01

№3

X \ Y	1.3	2.7	4.4	5.2
2.2	0.03	0.22	0.18	0.11
4.7	0.13	0.21	0.08	0.04

№4

X \ Y	1.28	1.43	1.79	2.04
3.53	0.07	0.11	0.17	0.09
4.02	0.23	0.13	0.14	0.06

№5

Y / X	2.07	2.25	2.94	3.35
3.44	0.31	0.20	0.17	0.02
4.15	0.11	0.08	0.06	0.05

№6

Y / X	3.06	3.37	3.95	4.43
2.56	0.08	0.01	0.31	0.10
4.27	0.08	0.11	0.15	0.23

№7

Y / X	4.21	4.93	5.20	6.73
3.44	0.03	0.22	0.13	0.06
5.21	0.11	0.17	0.26	0.02

№8

Y / X	2.75	3.13	5.47	8.31
2.17	0.08	0.13	0.12	0.07
4.83	0.12	0.22	0.21	0.05

№9

Y / X	1.33	2.05	2.72	3.54
2.64	0.07	0.23	0.25	0.13
3.13	0.15	0.09	0.05	0.03

№10

Y / X	4.17	5.23	6.84	7.31
2.66	0.09	0.18	0.25	0.11
6.35	0.07	0.15	0.11	0.04

№11

Y / X	2.33	2.97	3.22	4.39
3.44	0.12	0.12	0.23	0.15
7.18	0.04	0.13	0.15	0.06

№12

Y / X	2.34	3.89	4.35	6.19
2.39	0.09	0.18	0.25	0.09
5.81	0.13	0.12	0.09	0.05

№13

Y / X	1.23	1.95	2.13	3.37
4.56	0.06	0.31	0.18	0.12
6.28	0.25	0.03	0.03	0.02

№14

Y / X	2.85	3.13	3.97	4.56
3.47	0.09	0.19	0.18	0.10
6.13	0.22	0.13	0.06	0.03

№15

Y / X	2.35	3.37	4.23	6.54
7.35	0.07	0.31	0.18	0.01
9.93	0.23	0.04	0.03	0.13

№16

Y / X	3.25	3.74	4.96	7.03
5.27	0.07	0.13	0.26	0.12
8.43	0.09	0.18	0.11	0.04

№17

Y / X	1.25	3.86	5.88	7.34
4.47	0.09	0.13	0.21	0.05
8.95	0.11	0.19	0.15	0.07

№18

Y / X	1.22	1.43	2.7	6.54
1.34	0.11	0.23	0.19	0.05
4.57	0.07	0.15	0.13	0.07

Лабораторная работа 5

Метод наименьших квадратов и сглаживание экспериментальных зависимостей

1. Основные понятия и формулы. Примеры

Пусть результаты некоторого эксперимента систематизированы в виде таблицы,

x	x_1	x_2	...	x_n
y	y_1	y_2	...	y_n

Табл. 5.1

в которой $y_i = y(x_i)$ является функцией, а x - аргументом, и требуется сгладить эту табличную зависимость, многочленом или некоторой другой функцией, известной нам с точностью до нескольких подлежащих определению параметров.

Задача сглаживания экспериментальных зависимостей достаточно типична для практики. Решая эту задачу, обычно рассчитывают освободить экспериментальные данные от случайных ошибок, допущенных в каждом отдельном опыте, и свести большое количество этих данных к нескольким параметрам (в частности, к коэффициентам многочлена), одновременно получив возможность обрабатывать полученную функциональную зависимость аналитически (например, дифференцировать или интегрировать).

Пусть, для определенности, функция, с помощью которой будет осуществляться сглаживание, является многочленом

$$Q_r(x) = \alpha_o + \alpha_1 x + ... + \alpha_r x^r \qquad (5.1)$$

известной степени $r \geq 1$ $(r < n)$ с числовыми коэффициентами $\alpha_o, \alpha_1, \ldots, \alpha_r$, подлежащими определению.

Составим разности (*невязки*)

$$\varepsilon_i = y_i - Q_r(x_i) \quad i = 1,...,n, \qquad (5.2)$$

характеризующие близость табличных данных и значений, полученных при помощи сглаживающей функции. Нам следует подобрать коэффициенты $\alpha_o, \alpha_1, \ldots, \alpha_r$ таким образом, чтобы невязки в совокупности были минимальными.

Эффективной процедурой для решения подобного сорта задач является *метод наименьших квадратов*, согласно которому наилучшими во многих отношениях оценками для $\alpha_o, \alpha_1, \ldots, \alpha_r$ являются оценки, минимизирующие сумму квадратов разностей $\sum_{1 \leq i \leq n} \varepsilon_i^2$. Иными словами, в качестве оценки для неизвестных параметров α_iследует взять такие значения α_i, при которых функция

$$S(\alpha_o, \alpha_1, ..., \alpha_r) = \sum_{1 \leq i \leq n} (y_i - Q_r(x_i))^2$$

достигает минимума. Поскольку $S = S(\alpha_o, \alpha_1, \ldots, \alpha_r)$ представляет дифференцируемую функцию r переменных, необходимым условием ее минимизации является равенство нулю частных производных $\frac{\partial S}{\partial \alpha_i}$, $0 \leq i \leq r$. Решение системы из $r+1$*нормального* уравнения

$$\frac{\partial S}{\partial \alpha_i} = 0, \quad 0 \leq i \leq r, \qquad (5.3)$$

в типичных случаях единственное, дает нам искомые коэффициенты $\alpha_0 = a_o, \alpha_1 = a_1, \ldots, \alpha_k = a_r$.

Заметим, что сумма $\sum_{1\le i\le n}\varepsilon_i$ оптимальных разностей равняется нулю, что может быть использовано для контроля правильности вычислений.

Рассмотрим более детально линейную аппроксимацию экспериментальных зависимостей между величинами, т.е. сглаживание с помощью функции $y(x)=\alpha_0+\alpha_1 x$. В этом случае $S(\alpha_0,\alpha_1)=\sum_{1\le i\le n}(y_i-\alpha_0-\alpha_1 x_i)^2$,

$$\frac{\partial S}{\partial \alpha_0}=\sum_{i=1}^{n}2\,(y_i-\alpha_0-\alpha_1 x_i)(-1), \qquad \frac{\partial S}{\partial \alpha_1}=\sum_{i=1}^{n}2\,(y_i-\alpha_0-\alpha_1 x_i)(-x_i)$$

вследствие чего система нормальных уравнений принимает вид

$$\begin{cases}\sum_{i=1}^{n}(y_i-\alpha_0-\alpha_1 x_i)=0,\\ \sum_{i=1}^{n}(y_i-\alpha_0-\alpha_1 x_i)\,x_i=0\end{cases}. \qquad (5.4)$$

Введя обозначения

$$\begin{aligned}&\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i, \quad \overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,\\ &\overline{xy}=\frac{1}{n}\sum_{i=1}^{n}x_i y_i, \quad \overline{x^2}=\frac{1}{n}\sum_{i=1}^{n}x_i^2,\end{aligned} \qquad (5.5)$$

перепишем уравнения (5.4) в виде системы двух линейных уравнений с двумя неизвестными

$$\begin{cases}\alpha_0+\overline{x}\alpha_1=\overline{y}\\ \overline{x}\alpha_0+\overline{x^2}\alpha_1=\overline{xy}\end{cases} \qquad (5.6)$$

Решив эту систему, найдем

$$\alpha_0=\frac{\overline{y}\cdot\overline{x^2}-\overline{x}\cdot\overline{xy}}{\overline{x^2}-\overline{x}^2}=\overline{y}-\frac{\overline{r}S_y}{S_x}\overline{x}, \quad \alpha_1=\frac{\overline{xy}-\overline{x}\cdot\overline{y}}{\overline{x^2}-\overline{x}^2}=\frac{l_{xy}}{S_x^2}=\frac{l_{xy}}{S_xS_y}\cdot\frac{S_y}{S_x}=\overline{r}\frac{S_y}{S_x}. \qquad (5.7)$$

Здесь $S_x^2=\overline{x^2}-\overline{x}^2$ и $S_y^2=\overline{y^2}-\overline{y}^2-$ выборочные дисперсии,

$l_{xy}=\overline{xy}-\overline{x}\overline{y}-$ выборочныйковариационный момент,

$\overline{r}=\frac{l_{xy}}{S_xS_y}-$ выборочный коэффициент корреляции.

Таким образом, наилучшая в смысле метода наименьших квадратов линейная сглаживающая функция выражается уравнением $y(x)=\alpha_0+\alpha_1 x$ с коэффициентами (5.7).

Замечание. Это уравнение можно записать в виде выборочного уравнения линейной регрессии *Y* на *X*:

$$y = \bar{y} + \alpha_1(x - \bar{x}) = \bar{y} + \frac{\bar{r}S_y}{S_x}(x - \bar{x})$$

Пример 1. Сгладить линейной зависимостью от *x* следующие табличные данные:

x	-6	-2	-1	1	3	5
y	0.2	1	1.5	2	3.4	3.9

Табл. 5.2

Вычислить разности и отобразить на графике табличные данные и сглаживающую прямую.

Решение. Используя формулы (5.5), найдем $\bar{x} = 0, \quad \bar{y} = 2, \quad \overline{xy} = 4.5, \quad \overline{x^2} = 12.5$ и, следовательно, $\alpha_0 = 2, \quad \alpha_1 = 0.36$. Отсюда, уравнение сглаживающей прямой имеет вид $y = 2 + 0.36x$. Теперь дополним табл. 5.2:

x	-6	-2	-1	1	3	5
y	0.2	1	1.5	2	3.4	3.9
2+0.36x	-0.16	1.28	1.64	2.36	3.08	3.8
Невязки	0.36	-0.28	-0.14	-0.36	0.32	0.1

Табл. 5.3

Заметим, что сумма невязок равна нулю.

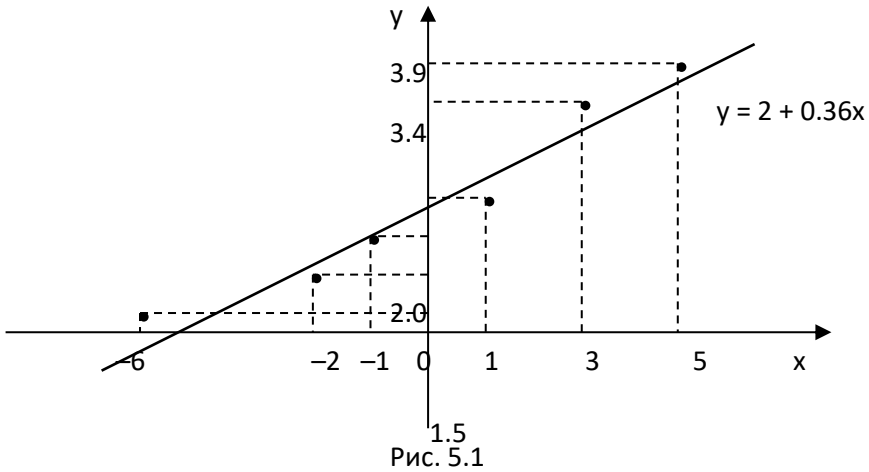


Рис. 5.1

Метод наименьших квадратов можно с некоторой потерей в точности использовать для сглаживания функциональных зависимостей, приводящихся к линейной с помощью замены переменных.

Так, зависимость

$$z = ae^{bt} \qquad (5.8)$$

после логарифмирования $\ln z = \ln a + bt$ может быть переписана в виде, $y = \alpha_0 + \alpha_1 x,$ где $y = \ln z, \quad x = t, \quad \alpha_0 = \ln a, \quad \alpha_1 = b.$ Применяя метод наименьших квадратов для нахождения коэффициентов линейного уравнения $y = \alpha_0 + \alpha_1 x$ и возвращаясь к

первоначальной зависимости, получим в качестве оценки коэффициенты $a = e^{\alpha_0}$, $b = \alpha_1$, т.е. $z = e^{\alpha_0 + \alpha_1 t}$.

Аналогично следует поступить при $z = at^b$. Если же

$$z = \frac{1}{at + b} \qquad (5.9)$$

или

$$z = \frac{t}{at + b}, \qquad (5.10)$$

то вводим $y = 1/z$ и, кроме того, в случае (5.10) вместо x берем $1/t$.

Пример 2. Пусть данные некоторых измерений представлены таблицей

t	-1	0	2	3	5
z	1.8	2	2.4	2.7	3.3

Табл. 5.4

Требуется сгладить их при помощи формулы $z = ae^{bt}$ и вычислить невязки с точностью до тысячных.

Решение. 1) Заменим числа z из табл. 5.4 числами $y = lnz$, $t = x$:

$x=t$	-1	0	2	3	5
$y = lnz$	0.5877	0.6931	0.8754	0.9932	1.193

Табл.5.5

2) По формулам (5.5) и (5.7) найдем коэффициенты $\alpha_0 = 0.6876$ и $\alpha_1 = 0.1006$ линейного уравнения $\alpha_o + \alpha_1 x$, наилучшим образом приближающего y.

3) Учитывая, что коэффициенты α_o и α_1 связаны с a и b соотношениями $a = e^{\alpha_0}$, $b = \alpha_1$ окончательно получим $a = 1.988$, $b = 0.1006$, и, следовательно, $z = 1.988\ e^{0.1006\ t}$.

Вычисленные значения y приведены в таблице 5.6 , графики на рис. 5.2

x	-1	0	2	3	5
y (теорет.)	1,798	1,988	2,431	2,688	3,288

Табл. 5.6

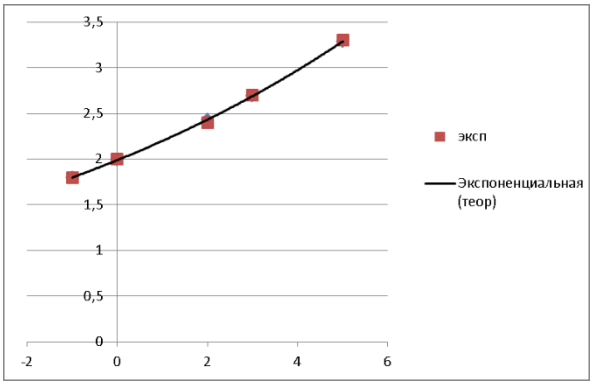


Рис. 5.2

Разности между табличными и сглаженными значениями сведены в таблицу:

t	-1	0	2	3	5
ε	0.002	0.012	-0.031	0.012	0.012

Табл. 5.6

Заметим, что сумма разностей здесь отлична от нуля.

2. Порядок выполнения работы

В настоящей работе, используя метод наименьших квадратов, требуется сгладить предложенную табличную зависимость их при помощи формул (5.8) – (5.10). Помимо этого, следует вычислить невязки с точностью до сотых и отобразить на графике табличные данные и сглаживающую кривую. Предварительно зависимость следует линеаризовать. Вариант задания следует получить у преподавателя. Порядок выполнения работы как в примере 2.

3. Варианты заданий

1. $z = \frac{1}{a + bt}$

t	0.4	0.6	0.7	0.9	1.0	1.2	1.4	1.5	2.0	3
z	2.36	1.9	1.75	1.5	1.39	1.22	1.09	1.04	0.82	0.59

2. $z = \frac{t}{a + bt}$

t	1.25	0.625	0.5	0.4	0.31	0.25	0.21	0.18
z	1.32	0.78	0.69	0.55	0.43	0.36	0.31	0.26

3. $z = ae^{bt}$

t	0	1	2	4	6	8	10
z	10.00	7.42	5.50	2.99	1.66	0.89	0.50

4. $z = \frac{1}{a + bt}$

t	0.5	1	1.5	2	2.5	3	3.5
z	0.65	0.7	0.83	0.98	1.12	1.48	1.96

5. $z = \frac{t}{a + bt}$

t	8	10	15	20	30	40	60	80
z	13	14	15.4	16.3	17.2	17.8	18.5	18.8

6. $z = ae^{bt}$

t	0.6	0.8	1.1	1.4	1.8	2
z	1.21	1.83	3.36	5.95	13.66	19.74

7. $z = \frac{1}{a + bt}$

t	0	0.2	0.5	1	1.5	2	2.5	3	3.5	4	4.5
z	1	0.833	0.667	0.5	0.4	0.33	0.286	0.25	0.22	0.2	0.18

8. $z = \frac{t}{a + bt}$

t	320	240	180	140	120	100	80	60
z	21.35	21.8	22.5	23.2	23.8	24.6	26.2	29

9. $z = ae^{bt}$

t	0.5	1	1.5	2	2.5	3	3.5
z	4.63	4.17	3.31	2.76	2.44	1.96	1.66

10. $z = \frac{1}{a + bt}$

t	0.5	1	1.5	2	2.4	2.8	3.4	3.8	3.5	4	6
z	1	0.63	0.51	0.43	0.37	0.33	0.3	0.27	0.24	0.22	0.17

11. $z = \frac{t}{a + bt}$

t	1.67	1.25	0.91	0.714	0.56	0.5
z	5.15	1.66	0.82	0.56	0.38	0.34

12. $z = ae^{bt}$

t	0	2.7	8.4	14.52	21.16	32.03
z	760	737	686	635	584	508

13. $z = \frac{\mathbf{1}}{a + bt}$

t	0.2	0.3	0.7	0.8	1.2	1.4	1.8
z	0.45	0.46	0.51	0.53	0.59	0.63	0.75

14. $z = \frac{t}{a + bt}$

t	2	1	0.67	0.5	0.4	0.33	0.29
z	0.65	0.7	0.84	0.98	1.12	1.48	1.97

15. $z = ae^{bt}$

t	1	1.5	1.8	2.2	2.7	3.2
z	1.28	1.91	2.3	3.14	4.84	7.16

16. $z = \frac{\mathbf{1}}{a + bt}$

t	0.8	1.6	2	2.5	3.2	4	4.7	5.6
z	1.32	0.78	0.69	0.55	0.43	0.36	0.31	0.26

17. $y = \frac{t}{a + bt}$

t	1	0.67	0.56	0.45	0.37	0.31
z	4.08	1.55	1.2	0.87	0.63	0.51

18. $z = ae^{bt}$

t	0.5	1	1.5	2	2.5	3	3.5

y	0.65	0.7	0.83	0.98	1.12	1.48	1.96

Лабораторная работа 6

Анализ временных рядов

1. Основные понятия и формулы. Примеры

Многие задачи науки и техники связаны с процессами $X(t)$, которые можно представить в виде совокупности измерений x_t на некотором интервале времени. Значения процесса x_t в каждый момент времени t является случайной величиной. Такие процессы называют *временными рядами*.

*Временным ряд*ом назовем (*динамическим* рядом) назовем последовательность наблюдений некоторого признака (случайной величины) X в последовательные (как правило, равноотстоящие) моменты времени t. Отдельные наблюдения называются уровнями ряда, обозначим их x_t ($t = 1,2, \dots , n$), где n – число уровней.

Раньше вариационный ряд $x_1, x_2, \dots, x_n$ рассматривался как одна из реализаций случайной величины X, временной ряд $x_1, x_2, \dots, x_n$ можно рассматривать как одну из реализаций случайного процесса $X(t)$. Однако, имеется принципиальное различие между временным рядом x_t ($t = 1,2, \dots , n$) и последовательностью наблюдений $x_1, x_2, \dots, x_n$, образующих случайную выборку. В отличие от элементов выборки члены временного ряда, как правило, не являются статистически независимыми и одинаково распределенными.

Примером временного ряда может служить переменная x_t, полученная при наложении случайных флуктуаций на детерминированную (неслучайную) составляющую, которую называют *трендом* временного ряда (тренд изображен на рис.6.1 пунктирной линией).

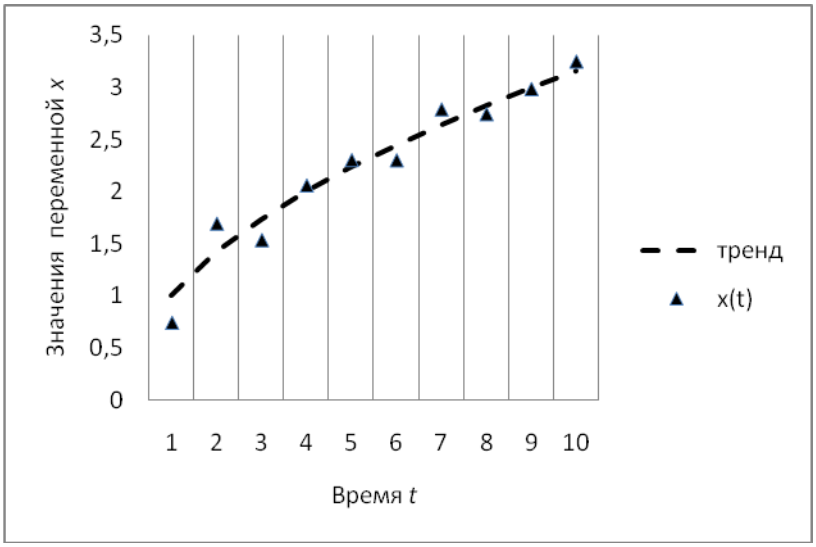


Рис. 6.1

В тех случаях, когда измерения могут регистрироваться (по крайней мере, теоретически) непрерывно, соответствующие временные ряды называют *непрерывными*. Если временной параметр t является дискретным, мы имеем *дискретные временные ряды*.

Для временных рядов основной интерес представляет моделирование их структуры с дальнейшим применением модели для экстраполяции или прогнозирования. При исследовании временных рядов необходим статистический подход из-за ошибок измерений и случайных флуктуаций, свойственных практически любой наблюдаемой системе, относится ли она к медицине, биологии, окружающей среде или технике. Важной составляющей статистических методов анализа временных рядов является оценка тренда, который, при наличии информации об его виде, можно моделировать при помощи компонент, являющихся детерминированными функциями времени. Отметим, что эксперименты, в которых осуществляются наблюдения, как правило, не являются независимыми, и последовательные ошибки модели должны, вообще говоря, рассматриваться как статистически связанные. Обычной практикой является предположение, что временной ряд наблюдается в равноотстоящие друг от друга моменты времени $t_1, t_2, \ldots, t_n$и что последовательные *ошибки* модели образуют *стационарный временной ряд*, вероятностные свойства которого не изменяются во времени.

Временной ряд x_t ($t = 1,2, \ldots, n$) называется *строго стационарным*, если совместное распределение вероятностей n наблюдений $x_1, x_2, \ldots, x_n$ такое же, как и n наблюдений $x_{1+\tau}, x_{2+\tau}, \ldots, x_{n+\tau}$ при любых n, t и τ. То есть, свойства строго стационарных временных рядов (закон распределения и числовые характеристики) не зависят от t. Следовательно, математическое ожидание $m_x(t) = m_x$ и дисперсия $D_x(t) = D_x$ могут быть оценены по наблюдениям x_t ($t = 1,2, \ldots, n$) по формулам:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + ... + x_n), \qquad (6.1)$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 . \qquad (6.2)$$

Состоятельной оценкой среднего квадратичного отклонения стационарного временного ряда является величина $S = \sqrt{S^2}$.

Моделью некоторых *нестационарных* временных рядов служат процессы вида

$$y_t = \mu(t) + x_t, \qquad (6.3)$$

где детерминированная функция $\mu(t)$ зависит лишь от t , а x_t- стационарный временной ряд с нулевым средним $m_x(t) = 0$. Детерминированная компонента $\mu(t)$ характеризует тенденцию изменения временного ряда y_t в среднем со временем, т.е. его тренд , а слагаемое x_t определяет случайные не зависящие от t ошибки модели и структуру их зависимости. Типичный вид графического изображения подобного нестационарного временного ряда был изображен на рис.6.1.

Как уже говорилось выше, одной из задач теории временных рядов является оценивание тренда временного ряда x_t - неслучайной составляющей $\mu(t)$ по результатам наблюдений $x_1, x_2, ..., x_n$ его отрезка длины n. Для решения этой задачи необходимо выбрать вид функции $\mu(t)$. Наиболее часто используются следующие функции:

линейная - $\mu(t) = a_0 + a_1 t$;

полиномиальная - $\mu(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_r t^r$;

экспоненциальная - $\mu(t) = e^{a_0 + a_1 t}$;

логистическая - $\mu(t) = \dfrac{a_0}{1 + a_1 e^{-a_2 t}}$.

При выборе функции $\mu(t)$ на практике используются визуальные наблюдения на основе графического изображения временного ряда.

Оценивание тренда временного ряда можно осуществить при помощи операции *сглаживания*, целью которой является уменьшение дисперсии временного ряда x_t, а, по существу, амплитуды случайных флуктуаций вокруг его детерминированной составляющей.

Можно провести сглаживание с помощью линейной функции $y = a_0 + a_1 t$.

Пусть явление рассматривалось в моменты времени $t_1, t_2, \dots, t_n$.

Используя методику построения прямых линий приближенной регрессии Y на X, получим:

$$y = \bar{y} + \frac{\bar{r} S_y}{S_t}(t - \bar{t}), \qquad (6.4)$$

где $\bar{y}, \bar{t}$ – выборочные средние,

$\bar{r}$ – выборочный коэффициент корреляции,

S_y, S_t – выборочные среднеквадратические отклонения.

Другим методом выравнивания временного ряда является *метод экспоненциального сглаживания* или его частный случаем - *метод скользящего среднего*. При любом использованном методе сглаживания исследователь рассчитывает, что полученный в результате сглаживания новый временной ряд y_t будет иметь более четко выраженный тренд, мало отличающийся от тренда первоначального ряда x_t и, следовательно, в первом приближении могущий его заменить.

Приведем формулу для метода скользящего среднего по трем точкам:

$$y_t = \frac{1}{3}(x_{t-1} + x_t + x_{t+1}), \qquad 2 \le t \le n-1. \qquad (6.5)$$

Этот метод сглаживания оставляет *линейный тренд* временного ряда x_t без изменения и, вообще говоря, уменьшают его дисперсию.

Отметим также, что при сглаживании, как правило, происходит некоторая потеря информации. Так, при использовании метода скользящего среднего по 3 точкам отрезок сглаженного ряда будет содержать n-2 элемента вместо n.

Пример 6.1. Данные о динамике роста объема производства x_t некоторого препарата (в тоннах) на фармацевтической фабрике за 10 последовательных лет представлены в табл. 5.1:

t	1	2	3	4	5	6	7	8	9	10
x_t	12	10	17	13	20	18	25	27	24	30

Табл.6.1

Провести сглаживание временного ряда x_t, используя формулу (6.5). На одном графике построить изображение ряда x_t и его сглаженного варианта.

Решение. Сглаженные значения y_t, подсчитанные при t=2,…,9 , внесем в табл. 6.2:

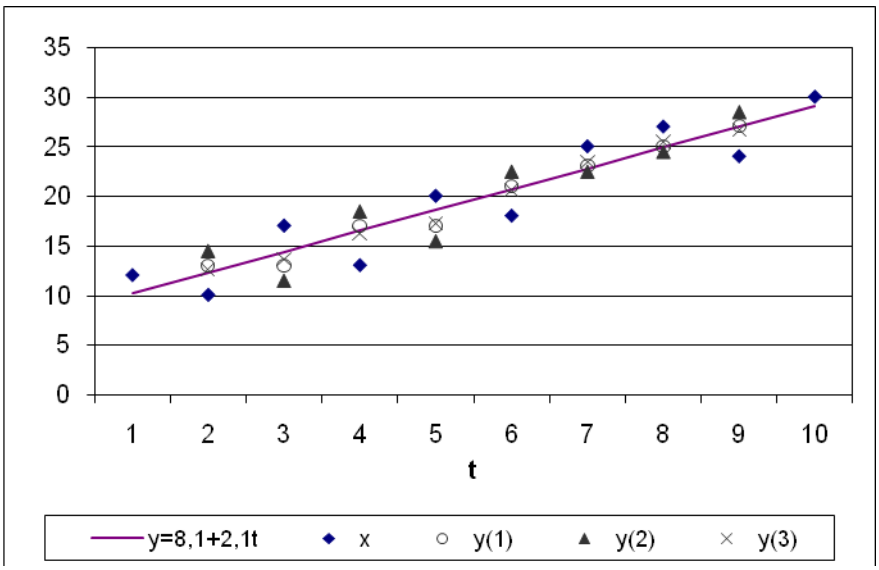
t	2	3	4	5	6	7	8	9
y_t	13	13	17	17	21	23	25	27

Табл.6.2

Например, y_2= (12+10+17)/3=13, y_9= (27+24+30)/3=27.

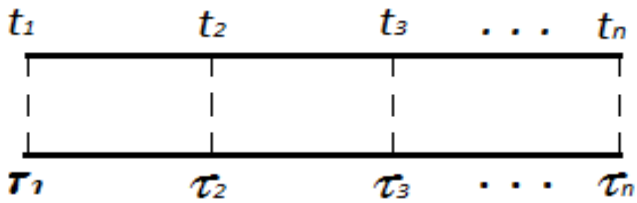
Графики, построенные по табл.6.1 и 6.2 , изображены на рис.6.2. Их визуальный анализ дает основания полагать, что наблюдаемый временной ряд имеет линейный тренд.

Рис. 6.2



Если с помощью метода скользящего среднего или из каких-либо других, например, теоретических, соображений сделан вывод, что тренд $\mu(t)$ временного ряда является функцией, известной с точностью до нескольких параметров, то оценки последних можно определить, используя метод наименьших квадратов (МНК).

В случае, когда тренд $\mu(t)$ является линейной или квадратичной функцией времени функцией времени, т.е. $\mu(t) = a_0 + a_1 t$ или $\mu(t) = a_0 + a_1 t + a_2 t^2$, то МНК-оценки параметров a_0 , a_1 и a_2 технически легче получить, рассматривая безразмерное время $\tau = (t - t_0)/h$ вместо t.



1) Если n (объем выборки) *нечетное* число, то полагаем $t_0 = t_{(n+1)/2}$ и $h = \Delta t$. Таким образом, t_0 совпадает со временем среднего члена выборки, h с шагом временного ряда, а безразмерное время принимает вид:

$$\tau_k = \frac{t_k - t_0}{h} = \frac{t_k - t_{\frac{n+1}{2}}}{h} = \frac{2k - (n+1)}{2}, \qquad (6.6)$$

т.е $\frac{1-n}{2}, \frac{3-n}{2}, \ldots$ -2, -1, 0, 1, 2, $\ldots \frac{n-3}{2}, \frac{n-1}{2}$

2) Если n *четное* число, то $t_0 = (t_{n/2} + t_{n/2+1})/2$ (т.е. равняется полусумме времен двух средних членов выборки), $h = \Delta t/2$ и безразмерное время:

$$\tau_k = \frac{t_k - t_0}{h} = \frac{t_k - \left(t_{n/2} + t_{n/2+1}\right)}{h} = 2k - (n+1), \qquad (6.7)$$

т.е. $1-n, 3-n, \ldots$ -3, -1, 1, 3, $\ldots n-3, n-1$

Если τ является безразмерным временем, то при линейном тренде $\mu_\tau = a_0 + a_1\tau$, то коэффициенты a_0, a_1 определяются МНК-формулами

$$a_o = \bar{x}, \qquad a_1 = \frac{\overline{\tau x}}{\overline{\tau^2}}, \qquad (6.8)$$

а в случае квадратичного тренда $\mu_\tau = a_0 + a_1\tau + a_2\tau^2$ коэффициенты a_0, a_1, a_2 вычисляются по МНК-формулам

$$a_o = \frac{\overline{\tau^4}\,\bar{x} - \overline{\tau^2}\,\overline{\tau^2 x}}{\overline{\tau^4} - \overline{\tau^2}^2}, \; a_1 = \frac{\overline{\tau x}}{\overline{\tau^2}}, \qquad a_2 = \frac{\overline{\tau^2 x} - \overline{\tau^2}\,\bar{x}}{\overline{\tau^4} - \overline{\tau^2}^2}, \qquad (6.9)$$

в которых

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{\tau x} = \frac{1}{n}\sum_{i=1}^{n} \tau_i x_i,$$

$$\overline{\tau^2} = \frac{1}{n}\sum_{i=1}^{n} \tau_i^2, \quad \overline{\tau^2 x} = \frac{1}{n}\sum_{i=1}^{n} \tau_i^2 x_i, \quad \overline{\tau^4} = \frac{1}{n}\sum_{i=1}^{n} \tau_i^4, \qquad (6.10)$$

а x_i является значением временного ряда в момент времени τ_i, i=1,2,...n, где n – объем выборки (напоминаем, что время τ является безразмерным).

Отметим, что величины $\overline{\tau^2}$, $\overline{\tau^4}$ и $\overline{\tau^4} - \overline{\tau^2}^2$ из формул (6.8), (6.9) зависят лишь от n и выражаются следующими формулами: при четном n

$$\overline{\tau^2} = (n^2-1)/3, \quad \overline{\tau^4} = (n^2-1)(3n^2-7)/15, \quad \overline{\tau^4} - \overline{\tau^2}^2 = 4(n^2-1)(n^2-4)/45; \qquad (6.11)$$

при нечетном n

$$\overline{\tau^2} = (n^2-1)/12, \quad \overline{\tau^4} = (n^2-1)(3n^2-7)/240, \quad \overline{\tau^4} - \overline{\tau^2}^2 = (n^2-1)(n^2-4)/180. \qquad (6.12)$$

Знание тренда $\mu(t)$ временного ряда позволяет с известной степенью надежности и точности *прогнозировать* его значения. Предположительные значения ряда для моментов времени τ_*, выходящих за границу проведенных наблюдений, считаются равными $\mu(t)$. При этом важно понимать, что оценки подобного сорта предполагают то, что основная тенденция изменения временного ряда в течение интервала времени между моментом наблюдения и моментом времени, для которого оценивается значение временного ряда, сохраняется. Кроме того, надо иметь в виду, что точность прогноза, как правило, снижается с ростом этого интервала.

Пример 6.2. Динамика производства серной кислоты:

Годы (t)	1960	1970	1980	1990	2000
Серная кислота в млн. т (x_t)	1,6	2,1	5,4	12,1	23,0

Табл. 6.4

Требуется дать прогноз объема производства в 2009 г., применяя сглаживание временного ряда с помощью как линейной, так и квадратичной функции.

Решение. В данном примере имеем нечетный объем выборки. Поэтому преобразуем годы t в безразмерное время по формуле $\tau = (t - t_0)/h$, положив $t_0 = 1980$ (средний член выборки) и h=10(шаг по времени). Тогда

Безразмерное время (τ)	– 2	– 1	0	1	2
Произв-во (млн. т)	1,6	2,1	5,4	12,1	23,0

Табл. 6.5

Коэффициенты в формуле $\mu_\tau = a_0 + a_1\tau$ определяем по формулам (6.8) (см. также табл.6.3): $a_o = 8{,}84, \quad a_1 = \frac{10{,}56}{2} = 5{,}28$, т.е. $\mu_\tau = 8.84 + 5.28\tau$.

Поскольку 2009г. соответствует $\tau_k = \frac{t_k - t_0}{h}$ = (2009-1980) / 10=2,9, линейный прогноз равен $\mu_{2.9} = 8{,}84 + 5{,}28 \times 2{,}9 = 24{,}152$

Коэффициенты a_0 и a_2 в формуле $\mu_\tau = a_0 + a_1\tau + a_2\tau^2$ определяем по (5.9), используя 3-й столбец табл.6.3 (коэффициент a_1 уже сосчитан и равен 5,28):

$$a_o = \frac{(34/5) \times 8{,}84 - 2 \times 22{,}52}{14/5} = 5{,}38, \quad a_2 = \frac{22{,}52 - 2 \times 8{,}84}{14/5} = 1{,}73,$$ следовательно,

$\mu_\tau = 5.38 + 5.28\tau + 1.73\tau^2$. Прогноз производства серной кислоты в млн. т. на 2009 год при квадратичной аппроксимации равен $\mu_{2.9} = 35{,}24$. Следует заметить, что к найденным прогнозам нельзя относиться как к абсолютной истине.

Безразмерное время (τ)	-2	-1	0	1	2
	-1.72	3.56	8.84	14.12	19.40
	1.74	1.83	5.38	12.39	22.86

Графики экспериментальных данных и линейного и квадратичного сглаживаний представлены на Рис 6.3.

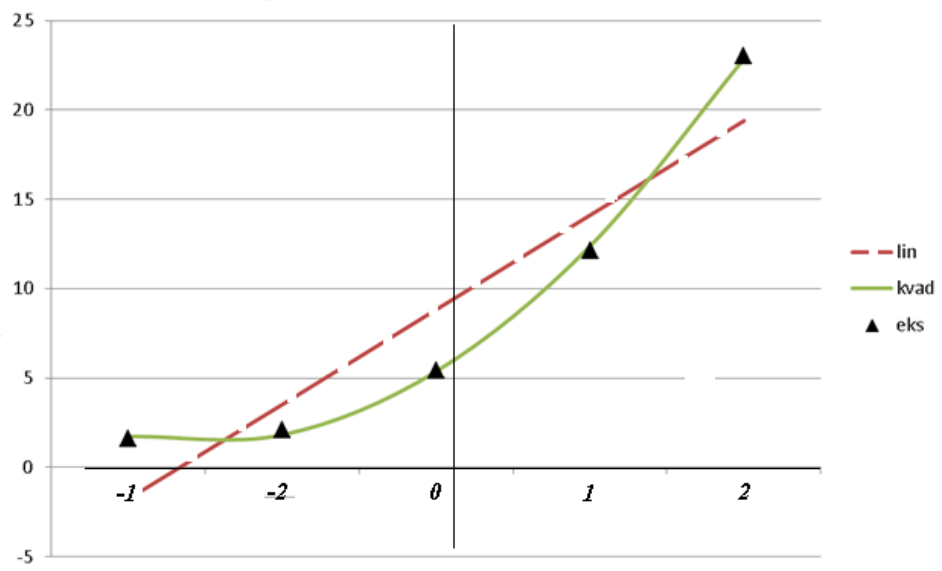


Рис 6.3

2. Порядок выполнения работы

В настоящей работе после получения варианта задания требуется:

1) сгладить предложенную табличную зависимость, используя метод скользящего среднего по трем точкам (см. пример 6.1).

2) дать соответствующий прогноз, применяя выявление тренда временного ряда методом наименьших квадратов с помощью линейной и квадратичной функции, используя переход к безразмерному времени как в примере 6.2.

3) На одном графике построить изображение ряда x_t, его сглаженного варианта, а также линейного и квадратичного трендов.

3. Контрольные вопросы

3. Какие процессы называют временными рядами?
4. Что называется трендом временного ряда?
5. Какой тренд называется линейным?
6. Какие временные ряды называются дискретными?
7. Какими свойствами должны обладать стационарные временные ряды?
8. Для чего осуществляется сглаживание временного ряда?
9. Каким образом прогнозируются значения временного ряда?

4. Варианты заданий

1.

Годы	1981	1986	1991	1996	2001
Розничный товарооборот в млн. руб.	73	119	173	235	305

Прогноз на 2003

2.

Годы	1981	1986	1991	1996
Кол-во стационарных рецептов в тыс. штук	11	15	30	44

Прогноз на 1998

3.

Годы	1980	1985	1990	1995	2000
Розничный товарооборот в тыс. руб.	30.6	63.4	100.4	141.6	187

Прогноз на 2004

4.

Годы	1976	1982	1988	1994
Оптовый товарооборот в тыс. руб.	20.8	36.7	56.0	78.7

Прогноз на 1997

5.

Годы	1985	1989	1993	1997	2001
Товарооборот в млн. руб.	26.8	110.4	170.8	241.6	323.8

Прогноз на 2003

6.

Годы	1985	1990	1995	2000
Кол-во рецептов в тыс. штук	27	111	242	417

Прогноз на 2004

7.

Годы	1976	1982	1988	1994	2000
Розничный товарооборот в тыс. руб.	40.4	74.2	115.6	164.6	221.2

Прогноз на 2003

8.

Годы	1985	1989	1993	1997
Кол-во рецептов в тыс. штук	82	162	254	358

Прогноз на 2000

9.

Годы	1988	1991	1994	1997	2000
Розничный товарооборот в тыс. руб.	28.8	42.8	72.0	132.8	169.8

Прогноз на 2003

10.

Годы	1989	1992	1995	1998
Оптовый товарооборот в млн. руб.	10.6	56.9	87.0	120.9

Прогноз на 2000

11.

Годы	1981	1986	1991	1996	2001
Товарооборот в тыс. руб.	29.8	52.2	79.8	112.6	150.4

Прогноз на 2005

12.

Годы	1988	1991	1994	1997	2000
Кол-во рецептов в тыс. штук	59	78	101	128	159

Прогноз на 2003

13.

Годы	1970	1980	1990	2000
Кол-во амбулаторных рецептов в тыс. штук	16	62	134	232

Прогноз на 2005

14.

Годы	1989	1992	1995	1998	2001
Товарооборот в млн. руб.	3.5	5.7	8.5	11.9	15.9

Прогноз на 2003

15.

Годы	1989	1993	1997	2001

Кол-во рецептов в тыс. штук	18	51	101	168

Прогноз на 2003

16.

Годы	1980	1985	1990	1995	2000
Розничный товарооборот в тыс. руб.	35.8	62.2	95.0	134.2	179.8

Прогноз на 2004

17.

Годы	1970	1980	1990	2000
Кол-во амбулаторных рецептов в тыс. штук	15.8	62.2	134.2	231.8

Прогноз на 2006

18.

Годы	1981	1986	1991	1996	2001
Оптовый товарооборот в тыс. руб.	32.5	53.6	80.3	112.3	149.7

Прогноз на 2005

Лабораторная работа 7

Построение выборочной линии регрессии

1. Основные понятия и формулы. Примеры

В естественных науках обычно речь идет о *функциональной зависимости*, когда каждому значению одной переменной отвечает *определенное* значение другой. Функциональная зависимость может существовать как между неслучайными переменными x и y, так и между случайными величинами X и Y. Однако очень часто между переменными величинами существуют зависимости, когда каждому значению одной переменной соответствует множество возможных значений другой переменной, т.е. каждому значению одной переменной соответствует *условное распределение* другой переменной. Такая зависимость называется *статистической* (стохастической, вероятностной) зависимостью. Статистическая связь обуславливается тем, что зависимая переменная испытывает влияние

некоторых неучтенных или неконтролируемых факторов, а также тем, что измерение любой переменной обязательно сопровождается случайными ошибками.

Вследствие неоднозначной (статистической) зависимости между *Y* и *X* рассматривается усредненная по *x* зависимость, т.е. изменение среднего значения – *условного математического ожидания* $M_x(Y) = M(Y|X = x)$ в зависимости от *x*.

Статистическая зависимость между двумя переменными, при которой каждому значению одной переменной *x* ставится в соответствие условное математическое ожидание $M(Y|X = x)$ другой, называется *корреляционной* зависимостью. Такая зависимость может быть представлена в виде:

$$M_x(Y) = M(Y|X = x) = \varphi(x) \qquad (7.1)$$

или

$$M_y(X) = M(X|Y = y) = \psi(y) \qquad (7.2)$$

Предполагается, что $\varphi(x) \neq const$ и $\psi(x) \neq const$, иначе корреляционная зависимость между величинами *X* и *Y* отсутствует.

Рассматривая различные виды зависимостей между *X* и *Y*, можно сказать, что при функциональной зависимости с изменением значений переменной *X* однозначно изменяется значение переменной *Y*, при ***корреляционной – условное математическое ожидание Y***, а при ***статистической – условное распределение Y*** (см. Лабораторную раб.4). Таким образом, наиболее общей является статистическая зависимость. Каждая корреляционная зависимость является статистической, но не каждая статистическая является корреляционной; так ковариационный момент может быть равен нулю – линейная зависимость между компонентами *X* и *У* отсутствует, но могут быть отличными от нуля центральные моменты μ_{gh} более высоких порядков, определяющие нелинейную зависимость между компонентами (см. Лабораторную раб.4) . Функциональная зависимость – частный случай корреляционной.

Уравнения (7.1) и (7.2) называются модельными уравнениями регрессии *Y* на *X* и *X* на *Y*, $\varphi(x)$ и $\psi(y)$ - модельными функциями регрессии, а их графики - модельными линиями регрессии.

Для отыскания модельных уравнений регрессии необходимо знать закон распределения двумерной случайной величины (X, Y). На практике, как правило, исследователь имеет лишь выборку пар значений (x_i, y_i) ограниченного объема *n*. В этом случае речь может идти об оценке по выборке функции регрессии. Наилучшей оценкой (в смысле метода наименьших квадратов) является *выборочная линия регрессии* *Y* на *X*:

$$y_x = \bar{\varphi}(x, \alpha_0, \alpha_1, \dots, \alpha_m), \qquad (7.3)$$

где y_x – условная средняя переменной Y при фиксированном значении переменной $X=x$, $\alpha_0, \alpha_1, \dots, \alpha_m$ – параметры кривой.

Аналогично определяется *выборочная линия регрессии* X на Y :

$$x_y = \bar{\psi}(y, \beta_0, \beta_1, \dots, \beta_m), \qquad (7.4)$$

Уравнения (7.3) и (7.4) – выборочные уравнения регрессии Y на X и X на Y, соответственно.

В настоящем параграфе исследуется случай, когда интересующие нас признаки объектов генеральной совокупности можно рассматривать как двумерный случайный вектор (X,Y) с частично или полностью неизвестным совместным законом распределения. Нашей целью вычисление статистических оценок основных характеристик этого распределения по наблюдениям выборки из генеральной совокупности, и, в частности, *выборочной ковариации* и *выборочного коэффициента корреляции*, а также *выборочных условных средних* (статистических аналогов функции регрессии), а также вычисление корреляционного отношения η_{yx}.

Определение. *Случайной выборкой* объема n, отвечающей паре случайных величин (X,Y), назовем набор n независимых пар случайных величин (X_1,Y_1), …, (X_n,Y_n), каждая из которых имеет такой же закон распределения, как и пара величин (X,Y).

Другими словами, случайной выборкой объема n можно считать величины (X_i,Y_i), $1 \le i \le n$, полученные в результате n независимых "одинаковых" случайных экспериментов. Оценками для математических ожиданий M(Y) и M(X), построенными по выборке (X_1,Y_1), …, (X_n,Y_n), являются выборочные средние $\bar{Y}$, $\bar{X}$ случайных величин Y и X , соответственно,

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{Y_1 + \dots + Y_n}{n}, \quad \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{X_1 + \dots + X_n}{n}, \qquad (7.5)$$

а оценками для дисперсий D(Y), D(X) – исправленные выборочные дисперсии Y и X:

$$S_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \qquad S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2; \qquad (7.6)$$

Несмещенной состоятельной оценкой ковариации K_{xy} случайных величин X и Y является исправленная выборочная ковариация

$$l_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}). \qquad (7.7)$$

В качестве оценки для коэффициента корреляции $r(X,Y)$ используется выборочный коэффициент корреляции

$$\bar{r} = \frac{l_{XY}}{\sqrt{S_X^2 S_Y^2}}, \qquad (7.8)$$

который является состоятельной оценкой коэффициента корреляции $r(X,Y)$.

Существенность выборочного коэффициента корреляции ($|r| \simeq 1$) позволяет рассчитывать на то, что зависимость между величинами X и Y близка к линейной. Последнее возможно визуально оценить по виду *корреляционого поля* точек. Так, корреляционое поле точек, изображенных на рис.7.1, позволяет высказать гипотезу о линейной регрессии Y на X, т.е. дает основание рассчитывать на то, что теоретическое *уравнение регрессии* имеет вид

$$M_x(Y) = M(Y|X=x) = \alpha_0+\alpha_1 x. \qquad (7.9)$$

Следующей задачей является нахождение оценок коэффициентов регрессии α_0, α_1. Воспользуемся методом наименьших квадратов, и будем искать такие значения a_0, a_1величин α_0, α_1, которые минимизируют сумму

$$\sum_{i=1}^{n}(Y_i - \alpha_i - \alpha_i X_i)^2 . \qquad (7.10)$$

Имеем (см. (7.5) и [1, п. (8.4)])

$$a_0 = \overline{Y} - a_1\overline{X}, \qquad a_1 = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\overline{X^2} - \overline{X}^2}, \qquad (7.11)$$

где применены обозначения

$$\overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \quad , \quad \overline{Y^2} = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 \quad , \quad \overline{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i \ . \qquad (7.12)$$

Имея в виду формулы

$$\frac{S_y}{S_x} = \sqrt{\frac{\overline{Y^2} - \overline{Y}^2}{\overline{X^2} - \overline{X}^2}}, \qquad \bar{r} = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sqrt{(\overline{X^2} - \overline{X}^2)(\overline{Y^2} - \overline{Y}^2)}}, \qquad (7.13)$$

видим, что выборочное уравнение прямой средней квадратичной регрессии Y на X имеет вид

$$y = \overline{Y} + \bar{r}\,\frac{S_Y}{S_X}(x - \overline{X}) \qquad (7.14)$$

и отличается от своего теоретического аналога лишь заменой параметров m_y, m_x, r, σ_y, σ_x их оценками. То же можно сказать о выборочном уравнении прямой среднеквадратической регрессии X на Y

$$x = \overline{X} + \bar{r}\frac{S_X}{S_Y}(y - \overline{Y}) \ . \qquad (7.15)$$

Отметим, что прямые, имеющие уравнения (7.14) и (7.15) проходят через точку $(\overline{X},\overline{Y})$, а их угловые коэффициенты совпадают по знаку с $\bar{r}$ и что практическом плане коэффициенты этих уравнений удобно вычислять при помощи формул (7.13).

Пример 7.1. Изучалась зависимость между систолическим давлением *Y* мужчин в начальной стадии шока и возрастом *X*. Результаты наблюдений приведены в таблице в виде двумерной выборки объема 11:

X	68	37	50	53	75	66	52	65	74	65	54
Y	114	149	146	141	114	112	124	105	141	120	124

Табл. 7.1

Требуется вычислить выборочный коэффициент корреляции и найти выборочное уравнение линейной регрессии *Y* на *X*.

Решение. Применяя формулы (7.5) и (7.6), найдем, округляя до сотых:

$\overline{X^2} = 3711.73,\quad \overline{Y^2} = 1681.33,\quad \overline{XY} = 7471.2,\quad \overline{X} = 59.91,\quad \overline{Y} = 126.36.$

Отсюда, по (7.11) - (7.13)

$$\bar{r} = \frac{7471.2 - 59.91\cdot 126.36}{\sqrt{(3711.73 - 59.91^2)(1681.33 - 126.36^2)}} = \frac{-99.05}{\sqrt{122.53\cdot 214.33}} = \frac{-99.05}{162.06} = -0.61\ ,$$

$$a_1 = \frac{-99.05}{122.53} = -0.81,\qquad a_0 = 126.36 - (-0.61)\cdot 59.91 = 174.89\ .$$

Таким образом, выборочное уравнение прямой средней квадратичной регрессии *Y* на *X* имеет вид $y = 174.88 - 0.81x$.

2. Корреляционная таблица

Если объем *n* выборки достаточно велик, то перед нахождением тех или иных статистических оценок по наблюдениям выборки, обычно используют *корреляционную таблицу*.

При большом числе наблюдений одно и то же значение *x* может встретиться n_x раз, одно и то же значение *y* – n_y раз, одна и та же пара чисел (*x, y*) может наблюдаться n_{xy} раз. Поэтому данные наблюдений группируют, т. е. подсчитывают частоты n_x, n_y, n_{xy}. Все сгруппированные данные сводят в *корреляционную таблицу*:

X / Y	x_1	…	x_i	…	x_m	n_y
y_1	n_{11}	…	n_{i1}	…	n_{m1}	n_{*1}
…	…	…	…	…	…	…
y_j	n_{1j}	…	n_{ij}	…	n_{mj}	n_{*j}
…	…	…	…	…	…	…

y_k	n_{1k}	…	n_{ik}	…	n_{mk}	$n_{\cdot k}$
n_x	$n_{1\cdot}$	…	$n_{i\cdot}$	…	$n_{m\cdot}$	n

Табл. 7.2

Здесь

$n_{i*} = \sum_{j=1}^{k} n_{ij}$, $1 \le i \le m$, - сумма элементовi – го столбца,

$n_{*j} = \sum_{i=1}^{m} n_{ij}$, $1 \le j \le k$; - сумма элементов j –той строки, (7.16)

$$\sum_{i=1} n_{i*} = \sum_{j=1}^{k} n_{*j} = n;$$

$x_1, x_2, \ldots, x_m$ – середины интервалов группировки (см. п.1.4) по X; $y_1, y_2, \ldots, y_k$– середины интервалов группировки по Y, n_{ij} – число точек выборки, попавших в прямоугольник с центром (x_i, y_j).

Как правило, группировка осуществляется с равным шагом h_x по x и равным шагом h_y по y, т. е.

$$x_{i+1} - x_i = h_x, \quad 1 \le i \le m; \qquad y_{j+1} - y_j = h_y, \quad 1 \le j \le k. \qquad (7.17)$$

Аналоги формул по (7.5) и (7.12), полученные по данным корреляционной таблицы, выглядят так:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{m} x_i n_{i\cdot}, \ \bar{Y} = \frac{1}{n}\sum_{j=1}^{k} y_j n_{\cdot j}, \ \overline{XY} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{k} x_i y_j n_{ij}, \ \overline{Y^2} = \frac{1}{n}\sum_{j=1}^{k} y_j^2 n_{\cdot j}, \ \overline{X^2} = \frac{1}{n}\sum_{i=1}^{m} x_i^2 n_{i\cdot} \quad (7.18)$$

Оценки дисперсий и ковариации рассчитываются по формулам:

$$S_x^2 = \overline{X^2} - \bar{X}^2, \quad S_y^2 = \overline{Y^2} - \bar{Y}^2, \quad l_{xy} = \overline{XY} - \bar{X}\bar{Y}. \qquad (7.19)$$

Выборочный коэффициент корреляции $\bar{r}$, выборочные коэффициенты регрессии a_0, a_1 и выборочные уравнения прямой средней квадратичной регрессии находятся по тем же формулам (7.11), (7.13), (7.14) и (7.15), что и ранее, с учетом обозначений (7.18) и (7.19). Напоминаем, что при вычислениях удобно пользоваться равенствами (7.13).

Представление наблюдений выборки в виде корреляционной таблицы позволяет определить выборочный аналог функции регрессии

$$y_x = M(Y \mid X = x_i) = \bar{Y}(x_i) = \frac{1}{n_{i\cdot}}\sum_{j=1}^{k} y_j n_{ij}, \quad 1 \le j \le m. \qquad (7.20)$$

Графическое представление точек $(x_i, \overline{Y}(x_i))$ в случае, если стохастическая зависимость Y от X является регрессионной, позволяет проверить справедливость предположений о виде теоретической функции регрессии Y на X.

В самом деле, если выборочная линия регрессии имеет тенденцию к расположению точек вдоль прямой, то весьма вероятно, что теоретическая линия регрессии Y на X также является прямой (см. рис. 7.1), а если эти точки расположены вдоль, скажем, параболы, как на рис. 7.2, то теоретическая линия регрессии предположительно является параболой.

Для упрощения вычислений в табл. удобно от (x_i,y_j) перейти к новым переменным (u_i,v_j), положив

$$u_i = \frac{x_i - x_0}{h_1} \quad , \quad v_j = \frac{y_j - y_0}{h_2}, \qquad (7.16)$$

и выбирая числа x_0 , y_0 , h_1 , h_2 таким образом, чтобы по формулам (7.13), (7.14), в кото-

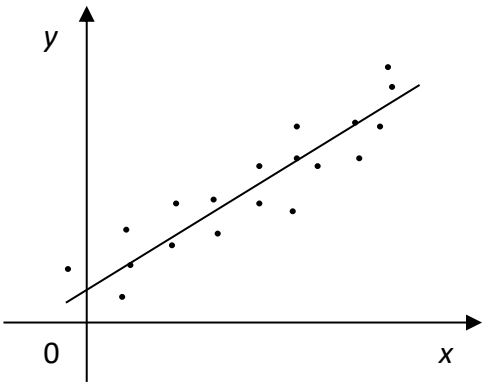


Рис. 7.1

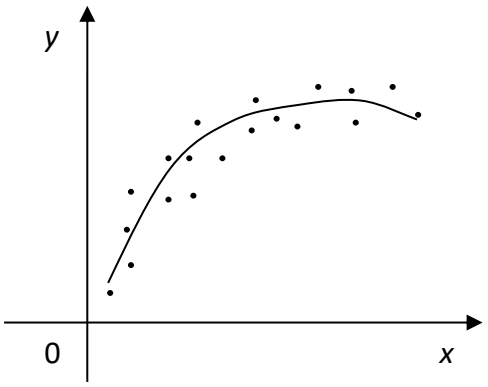


Рис. 7.2

рых x_i заменено на u_i, а y_j на v_j, было проще считать (при условии (7.12), например, удобно положить $h_1 = h_x$, $h_2 = h_y$, в качестве x_0 и y_0 обычно выбирают значения Y на X , отвечающие наибольшей n_{xy}). Обратный пересчет осуществляется по формулам

$$\overline{X} = h_1\overline{U} + x_0 \,, \quad \overline{Y} = h_2\overline{V} + y_0 \,, \; S_x^2 = h_1^2 S_u^2, \qquad S_y^2 = h_2^2 S_v^2, \qquad l_{xy} = h_1 h_2 l_{uv} \,. \qquad (7.17)$$

Заметим, что при линейном преобразовании (7.16) значение выборочного коэффициента корреляции $\bar{r}$ не изменяется.

Аналогично, M($Y|X = x_i$) (см. (7.15)), перейдя к новым координатам, можно вычислить по формуле

$$y_x = M(Y \mid X = x_i) = y_0 + h_2 \frac{1}{n_{i\cdot}} \sum_{j=1}^{k} v_j n_{ij} \quad , \quad 1 \le i \le m \,. \qquad (7.18)$$

Пример 7.2. Изучалась зависимость между количеством гемоглобина в крови Y(%) и массой животных X (кг). Результаты наблюдений приведены в виде корреляционной таблицы (пропуски означают нули):

X / Y	18	22	26	30	n_y
70	5				5
75	7	46	1		54
80		29	72		101
85			29	8	37
90				3	3
n_x	12	75	102	11	200

Табл. 7.3

Требуется определить выборочные аналоги функции регрессии и уравнения прямой средней квадратичной регрессии Y на X.

Решение. Для упрощения вычислений перейдем к новым переменным U и V, воспользовавшись формулами (7.16) при h_1=4, h_2=5, x_0=26, y_0=80 (в качестве x_0и y_0выбираем значения Y на X, отвечающие наибольшей n_{xy}). Для удобства перепишем табл. 7.3 в новых обозначениях, добавив справа столбец с суммой частот по строкам, а снизу строку с суммой частот по столбцам:

U / V	-2	-1	0	1	n_v
-2	5				5
-1	7	46	1		54
0		29	72		101
1			29	8	37
2				3	3
n_u	12	75	102	11	n=200

Табл.7.4

Имеем (см. (7.13), (7.14) при $x_i = u_i$ и $y_j = v_j$)

$$\overline{U} = \frac{1}{n}\sum_{i=1}^{m} u_i n_{ui} = \frac{1}{200}(-2\cdot 12 + (-1)\cdot 75 + 0\cdot 102 + 1\cdot 11) = -0.44,$$

$$\overline{V} = \frac{1}{n}\sum_{j=1}^{k} v_j n_{vj} = \frac{1}{200}(-2\cdot 5 + (-1)\cdot 54 + 0\cdot 101 + 1\cdot 37 + 2\cdot 3) = -0.105,$$

$$\overline{U^2} = \frac{1}{n}\sum_{i=1}^{m} u^2{}_i n_{ui} = \frac{1}{200}((-2)^2\cdot 12 + (-1)^2\cdot 75 + 0^2\cdot 102 + 1^2\cdot 11) = 0.67,$$

$$\overline{V^2} = \frac{1}{n}\sum_{i=1}^{m} v^2{}_i n_{vi} = = \frac{1}{200}((-2)^2\cdot 5 + (-1)^2\cdot 54 + 0^2\cdot 101 + 1^2\cdot 37 + 2^2\cdot 3) = 0.615$$

$$\overline{UV} = \frac{1}{n}\sum_{i=1}^{m}\sum_{j=1}^{k} u_i v_j n_{ij} = \frac{1}{200}((-2)\,((-2)\cdot 5 + (-1)\cdot 7) + (-1)((-1)\cdot 46 + 0\cdot 29) + 0\cdot((-1)\cdot 1 + 1\cdot 29) +$$
$$1\cdot(1\cdot 8 + 2\cdot 3)) = 0.47,$$

$$S_U^2 = 0.67 - (-0.44)^2 = 0.4764, \qquad S_U = 0.6902\ ,$$

$$S_V^2 = 0.615 - (-0.105)^2 = 0.604, \qquad S_V = 0.7772\ \textbf{,}$$

$$l_{UV} = \overline{UV} - \overline{U}\cdot\overline{V} = 0.47 - (-0.44)(-0.105) = 0.426,$$

$$\bar{r}(U,V) = \frac{l_{UV}}{S_U \cdot S_V} = \frac{0.426}{0.6902\cdot 0.7772} = 0.7941 \qquad .$$

Таким образом (см. (7.17)),

$\overline{X} = h_1\overline{U} + x_0 = 4\cdot(-0.44) + 26 = 24.24, \quad \overline{Y} = h_2\overline{V} + y_0 = 5\cdot(-0.105) + 80 = 79.475\ ,$

$S_X = h_1 S_U = 4\cdot 0.6902, \quad S_Y = h_2 S_V = 5\cdot 0.7772 = 3.886, \qquad \bar{r} = \bar{r}(U,V) = 0.7941;$

выборочное уравнение прямой средней квадратичной регрессии Y на X выражается формулой (7.9):

$$y_{рег.} = \overline{Y} + \bar{r}\frac{S_Y}{S_X}(x - \overline{X}) = 79.475 + 0.7941\frac{3.886}{2.7608}(x - 24.24) = 52.381 + 1.118x. \qquad (7.19)$$

Формула (7.18) дает:

$$M(Y \mid X = 18) = 80 + 5\cdot\frac{1}{12}((-2)\cdot 5 + (-1)\cdot 7) = 72.92,$$

$$M(Y \mid X = 22) = 80 + 5\cdot\frac{1}{75}((-1)\cdot 46 + 0\cdot 29) = 77.91,$$

$$M(Y \mid X = 26) = 80 + 5\cdot\frac{1}{102}((-1)\cdot 1 + 0\cdot 72 + 1\cdot 29) = 81.37,$$

$$M(Y \mid X = 30) = 80 + 5\cdot\frac{1}{11}(1\cdot 8 + 2\cdot 3) = 86.36.$$

Отсюда

X	18	22	26	30
$y_{лин.}$	72.5	76.98	81.45	85.92
$Y_{рег.}$	72.92	77.91	81.37	86.36

Табл. 7.5

Сопоставляя полученные результаты, приходим к выводу, что значения, вычисленные по уравнению выборочной регрессии и по линейной зависимости (7.19) хорошо согласуются.

3. Корреляционноеотношение

Введенный выше коэффициент корреляции является полноценным показателем тесноты связи Y и X лишь в случае *линейной* зависимости между переменными. Часто возникает необходимость в показателе интенсивности связи при *любой* форме зависимости. Таким показателем может служить *эмпирическое корреляционное отношение* Y по X:

$$\eta_{xy} = {S_{\text{межгр}}}\Big/{S_{\text{общ}}} \qquad (7.20)$$

Здесь $S^2_{\text{общ}}$ - общая дисперсия переменной Y

$$S^2_{\text{общ}} = \frac{1}{n}\sum n_y(y-\bar{y})^2 \qquad (7.21)$$

$S^2_{\text{межгр}}$ - межгрупповая дисперсия

$$S^2_{\text{межгр}} = \frac{1}{n}\sum n_x(\bar{y}_x-\bar{y})^2 \qquad (7.22)$$

Пусть данные наблюдений над количественными признаками X и Y сведены в корреляционную таблицу. Можно считать, что тем самым наблюдаемые значения Y разбиты на группы; каждая группа содержит те значения Y, которые соответствуют *определенному* значению X.

Например, дана корреляционная табл. 7.3.

X / Y	18	22	26	30	n_y
70	5				5
75	7	46	1		54
80		29	72		101
85			29	8	37
90				3	3
n_x	12	75	102	11	200

К первой группе относятся те 12 значений Y (5 раз наблюдалось $y_1 = 70$ и 7раз $y_2 = 75$), которые соответствуют $x_1 = 18$, ко второй − 75 значений Y, которые соответствуют $x_2 = 22$, и т.д.

Условные средние теперь можно назвать групповыми средними: групповая средняя первой группы:

Условные средние теперь можно назвать групповыми средними: групповая средняя первой группы:

$$\bar{y}_{x_1} = \frac{1}{n_1}\sum_{i=1}^{2} n_i y_i = \frac{1}{75}(5\cdot 70 + 7\cdot 75) = 72.92;$$

групповая средняя второй группы:

$$\bar{y}_{x_2} = \frac{1}{n_2}\sum_{i=1}^{2} n_i y_i = \frac{1}{75}(46\cdot 75 + 29\cdot 80) = 76.93;$$

групповая средняя третьей группы:

$$\bar{y}_{x_3} = \frac{1}{n_3}\sum_{i=1}^{3} n_i y_i = \frac{1}{128}(1\cdot 75 + 72\cdot 80 + 29\cdot 85) = 81.37;$$

групповая средняя четвертой группы группы:

$$\bar{y}_{x_4} = \frac{1}{n_4}\sum_{i=1}^{2} n_i y_i = \frac{1}{11}(8\cdot 85 + 3\cdot 90) = 86.36;$$

$$\bar{y} = \frac{1}{n}\sum_{j=1}^{k} y_j n_{*j} = \frac{1}{n}\sum_{j=1}^{k} y n_y = = \frac{1}{200}(70 \cdot 5 + 75 \cdot 54 + 80 \cdot 101 + 85 \cdot 37 + 90 \cdot 3)$$
$$= 79.475$$
$$S^2_{\text{общ}} = \frac{1}{n}\sum n_y (y - \bar{y})^2 =$$
$$= \frac{1}{200}[5(70 - 79.475)^2 + 54(75 - 79.475)^2 + 101(80 - 79.475)^2 + 37(85 - 79.475)^2 +$$
$$+3(90 - 79.475)^2 = 15.099$$
$$S^2_{\text{межгр}} = \frac{1}{n}\sum n_x \left(\bar{y}_{x_i} - \bar{y}\right)^2 =$$
$$= \frac{1}{200}[12(72{,}92 - 79.475)^2 + 75(76.93 - 79.475)^2 + 102(81.37 - 79.475)^2$$
$$+11(86.36 - 79.475)^2 = 9.446$$
$$S_{\text{общ}} = \sqrt{S^2_{\text{общ}}} = 3.866 \quad S_{\text{межгр}} = \sqrt{S^2_{\text{межгр}}} = 3.073$$
$$\eta_{xy} = {S_{\text{межгр}}}\Big/{S_{\text{общ}}} = 0.791$$

Справедливы следующие утверждения:

1) если *У* связан с *Х* функциональной зависимостью, то $\frac{S^2_{\text{межгр}}}{S^2_{\text{общ}}} = 1$;

2) если *У* связан с *Х* корреляционной зависимостью, то $\frac{S^2_{\text{межгр}}}{S^2_{\text{общ}}} < 1$

.

Основные свойства корреляционного отношения (при достаточно большом объеме выборки).

1. Корреляционное отношение – неотрицательная величина:$0 \le \eta_{xy} \le 1$
2. Если:$\eta_{xy} = 0$; то корреляционная связь отсутствует.
3. Если:$\eta_{xy} = 1$; то между переменными существует функциональная зависимость.
4. Если $\eta_{xy} \neq \eta_{yx}$, т.е. в отличие от коэффициента корреляции (для которого $r_{xy} = r_{yx} = r$) при вычислении корреляционного отношения существенно какую переменную считать независимой, а какую – зависимой.

Значение $\eta_{xy} = 0.791$ близко к величине $r = 0.794$ (см. выше), поэтому оправдано сделанное выше предположение о линейной корреляционной зависимости между переменными.

2. Порядок выполнения работы

1.По данным таблицы

X	…	…	…	…	…	…
Y	…	…	…	…	…	…

вычислить

a) выборочный коэффициент корреляции, применяя формулы (7.5), (7.6) и (7.11) - (7.13),

b) найти выборочное уравнение линейной регрессии Y на X.

2. Используя данные корреляционной таблицы

X / Y	x_1	…	x_i	…	x_m	n_y
y_1	n_{11}	…	n_{i1}	…	n_{m1}	n_{*1}
…	…	…	…	…	…	…
y_j	n_{1j}	…	n_{ij}	…	n_{mj}	n_{*j}
…	…	…	…	…	…	…
y_k	n_{1k}	…	n_{ik}	…	n_{mk}	n_{*k}
n_x	n_{1*}	…	n_{i*}	…	n_{m*}	n

a) определить выборочные аналоги функции регрессии и уравнения прямой средней квадратичной регрессии Y на X(для упрощения вычислений перейти к новым переменным U и V , воспользовавшись формулами (7.16),(7.17), (7.18)), переписав табл. в новых обозначениях,

b) вычислить эмпирическое корреляционное отношение Y по X (формулы (7.20) – (7.22)).

3. Контрольные вопросы

1. Функциональная, корреляционная и стохастическая зависимость между величинами.

2. Модельные уравнения регрессии Y на X и X на Y.

3. Случайная выборка объема n, отвечающей паре случайных величин (X,Y).

4. Оценка ковариации K_{xy} и оценка коэффициента корреляции $r(X,Y)$ случайных величин X и Y.

5. Выборочное уравнение прямой средней квадратичной регрессии Y на X.

6. Корреляционная таблица.

7. Корреляционное отношение.

5. Варианты заданий

1. В 100 частях воды растворяется следующее число условных частей азотнокислого натрия $NaNO_3$ (признак Y) при соответствующих температурах (X):

X	0	4	10	15	21	29	36	51	68
Y	66.7	71.0	76.3	80.6	85.7	92.9	99.4	113.6	125.1

2. На количество растворившегося$NaNO_3$ влияют случайные факторы. Предполагается наличие стохастической линейной зависимости между температурой и количеством растворившегося $NaNO_3$ вида (8.32). Найти МНК-оценку коэффициентов линейной модели.
3. В нижеследующих задачах требуется вычислить выборочный коэффициент корреляции между переменными Y и X и найти выборочное уравнение прямой средней квадратической регрессии Y на X.
4. Изучалась зависимость между содержанием коллагена Y и эластина X в магистральных артериях головы (г/100 г сухого вещества)(возраст 51-75 лет). Результаты наблюдений приведены в таблице в виде двумерной выборки объема 5:

5.

X	13.50	13.09	6.45	7.26	8.80
Y	33.97	38.07	53.98	46.00	48.61

6. Изучалась зависимость между массой новорожденных павианов-гамадрилов X (кг) и массой их матерей Y (кг). Результаты наблюдений приведены в таблице в виде двумерной выборки объема 9:

X	0.7	0.73	0.75	0.7	0.65	0.7	0.61	0.70	0.63
Y	10	10.8	11.3	10	11.1	11.3	10.2	13.5	12

7. Изучалась зависимость между объемом грудной клетки мужчин Y (см) и ростом X (см). Результаты наблюдений приведены в таблице в виде двумерной выборки объема 7:

X	162	164	179	172	182	188	168
Y	88	94	98	100	102	108	112

8. Изучалась зависимость между минутным объемом сердца Y (л/мин) и средним давлением в левом предсердии X (мм рт.ст.). Результаты наблюдений приведены в таблице в виде двумерной выборки объема 5:

X	4.8	6.4	9.3	11.2	17.7
Y	0.4	0.69	1.29	1.64	2.4

По таблицам сгруппированных данных вычислить выборочный коэффициент корреляции X и Yи написать уравнение линейной регрессии Y на X, найти корреляционное отношение η_{yx}.

№ 1

	X						
Y	**2**	**10**	**18**	**26**	**34**	**42**	**50**
45	1	1					
50		1	1	5	4		
55			5	3	4		
60			10	3	3	2	
65			2		1		1
70	1	1		1			

№ 2

	X						
Y	**22**	**32**	**42**	**52**	**62**	**72**	**82**
15	1	1	2				
19	2		3				
23		4	2	10			
27		2		3	7	2	
31					5	4	
35					1		1

№ 3

	X						
Y	**4**	**8**	**12**	**16**	**20**	**24**	**28**

115			2		2		
120	1		2	2			1
125				13		4	
130	2	2	4	2	1		1
135		2	1		5	4	
140					1	2	

№ 4

	X						
Y	**2**	**5**	**8**	**11**	**14**	**17**	**20**
115	2		2				
130		2		2		2	
145			3	12	3		1
160		2		4	5		
175		1	4		3		
190	1						1

№ 5

	X						
Y	**14**	**24**	**34**	**44**	**54**	**64**	**74**
135		2	2				
145	1	1	3				
155		4	3	9			
165		2		3	4	6	
175			3		3		
185							2

№ 6

	X						
Y	**47**	**50**	**53**	**56**	**59**	**62**	**65**
112		2				1	
120	1			2		3	
128				5	11		2
136		3	6	1	2		1
144				1	2	4	
152	1				1		1

№ 7

	X						
Y	**42**	**50**	**58**	**66**	**74**	**82**	**90**

45						1	
50	2		2	3	2		1
55			2	10	4	2	
60		3	4		3	2	
65	1	1		3			
70					1		3

№ 8

	X						
Y	**18**	**27**	**36**	**45**	**54**	**63**	**81**
105		2				1	
110	1			2		3	
115				5	11		2
120		3	6	1	2		1
125				1	2	4	
130	1					1	1

№ 9

	X						
Y	**3**	**6**	**9**	**12**	**15**	**18**	**21**
95		2	2				
100	1		3				
105		4	3	9			
110		2	1	3	6	3	
115					5	4	
120						1	1

№ 10

	X				
Y	**3**	**5**	**7**	**9**	**11**
2	4	3			
4	5	5	6		
6		10	12	5	
8		1	10	16	
10		2	4	10	3
12				2	2

№ 11

	X						
Y	**10**	**20**	**30**	**40**	**5 0**	**60**	**70**

2	4	3	3				
4	1	10	8	10	3		
6	1	3	12	12	8	1	
8	1		4	5	5	1	
10						3	2

№ 12

	X						
Y	**13**	**14**	**15**	**16**	**17**	**18**	**19**
14	4	1	1	2			
15	3	3	5	4	3	1	
16		10	11	12	10	2	
17		3	9	9	4	2	
18							1

Лабораторная работа 8

Проверка статистических гипотез

В лабораторной работе 8 по наблюдениям выборки проверяется предположение о равенстве средних двух нормально распределенных генеральных совокупностей с неизвестными одинаковыми дисперсиями и приводится критерий равенства самих дисперсий. При некоторых дополнительных предположениях проводится также проверка значимости коэффициента корреляции. По данным выборки из Лабораторной работы 1 проверяется гипотезу о нормальности распределения генеральной совокупности признака X (используются результаты лабораторных работ 1 и 2, n=50).

1. Основные понятия

В прикладных задачах часто требуется по наблюдениям выборки высказать некоторое суждение (*гипотезу*) относительно интересующих экспериментатора характеристик генеральной совокупности, из которой эта выборка извлечена.

Статистическая гипотеза - это утверждение о виде неизвестного распределения или параметрах известного распределения. Статистические гипотезы проверяются по

результатам выборки статистическими методами в ходе эксперимента (эмпирическим путем) с помощью статистических критериев. В таких случаях говорят, что речь идет о *проверке статистических гипотез*.

В тех случаях, когда известен закон, но неизвестны значения его параметров (дисперсия или математическое ожидание) в конкретной ситуации, статистическую гипотезу называют *параметрической*.

Когда закон распределения генеральной совокупности не известен, но есть основания предположить, каков его конкретный вид, выдвигаемые гипотезы о виде его распределения называются *непараметрическими.*

По содержанию статистические гипотезы можно классифицировать:

1. Гипотезы о *типе вероятностного закона распределения случайной величины*, характеризующего явление или процесс.
2. Гипотезы об *однородности двух или более обрабатываемых выборок*. Изучаемое свойство исследуется с помощью двух или более генеральных совокупностей. Гипотеза в этом случае может заключаться в следующем: исследуемые выборочные характеристики различаются между собой статистически значимо или нет.
3. Гипотезы о *свойствах числовых значений параметров* исследуемой генеральной совокупности. Больше ли значения параметров некоторого заданного номинала или меньше.
4. Гипотезы о *вероятностной зависимости двух или более признаков*, характеризующих различные свойства рассматриваемого явления или процесса. При этом определяется характер этой зависимости.

Пример:

Увеличение числа заболевших некоторым заболеванием дает возможность выдвинуть гипотезу о наличии эпидемии. Для сравнения доли заболевших в обычных и экстремальных условиях используются статистические данные, на основании которых делается вывод о том, является ли данное массовое заболевание эпидемией. Предполагается, что существует некоторый критерий- уровень доли заболевших, критический для этого заболевания, который устанавливается по ранее имевшимся случаям.

Правила, позволяющие выяснить, соответствует или нет интересующая нас гипотеза опытным данным, называются *статистическими критериями* Λ или просто *критериями* – это случайная величина (статистика) определенная на выборке.

Различают три вида критериев:

1. *Параметрические критерии* - критерии значимости, которые служат для проверки гипотез о *параметрах распределения* генеральной совокупности *при известном виде распределения*.

2. *Критерии согласия* - позволяют проверить гипотезы о соответствии распределений генеральной совокупности *известной теоретической модели*.

3. *Непараметрические критерии* - используются в гипотезах, когда не требуется знаний о *конкретном виде распределения*.

Проверка *параметрических* гипотез проводится на основе *критериев значимости*, а *непараметрических - критериев согласия*.

Задача проверки статистических гипотез сводится к исследованию генеральной совокупности по выборке. Множество возможных значений критерия может быть разделено на два непересекающихся подмножества - критическую область и область принятия гипотезы.

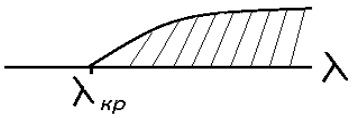
Областью принятия гипотезы или областью допустимых значений называют совокупность значений критерия, при которых эту гипотезу принимают.

Критической областью называют множество значений критерия, при котором гипотезу отвергают.

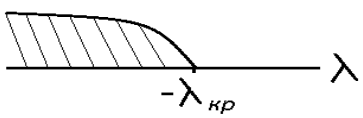
Наблюдаемые значения критерия (статистика) называют такое значение критерия, которое находится по данным выборки.

Границы критической области, отделяющие ее от области принятия гипотезы, называют критическими точками и обозначают. Критические точки разграничивают область значений Λ на критическую область и область принятия гипотезы. Выделяют *одностороннюю* и *двустороннюю* критические области.

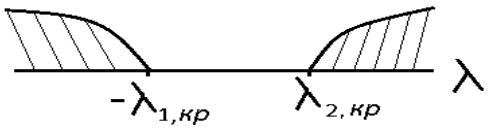
Правосторонняя критическая область определяется неравенством $\Lambda > \Lambda_{кр}$, где граница $\Lambda_{кр}$ определяется по таблице распределения данного критерия (статистике Λ).



Аналогично, *левосторонняя критическая область* определяется неравенством $\Lambda < -\Lambda_{кр}$,



а *двусторонняя критическая область* определяется неравенствами $\Lambda > \Lambda_{2,кр}$ и $\Lambda < -\Lambda_{1,кр}$.



Место для формулы.

Гипотеза, справедливость которой мы хотим проверить, носит название *нулевой гипотезы* и обозначается H_o.

Наряду с нулевой гипотезой необходимо также рассмотреть *конкурирующую гипотезу* H_1, которая является альтернативной по отношению к H_o, т.е. принимается в том случае, если H_o не верна.

Например, если $X \sim N(\mu,1)$, где параметр μ предполагается неизвестным, то в качестве нулевой гипотезы можно выбрать H_o: $\mu = 0$, а в качестве конкурирующей гипотезы рассмотреть H_1: $\mu \neq 0$.

Для определения критической области задается уровень значимости α - некая (малая) вероятность попадания критерия в критическую область .

Уровень значимости α - вероятность принятия конкурирующей гипотезы H_1, тогда как справедлива основная H_o.

С помощью уровня значимости определяются границы критической области.

Основной принцип проверки статистических гипотез состоит в следующем: если наблюдаемое значение статистики критерия попадает (не попадает) в критическую область , то гипотеза H_0 отвергается (принимается), а гипотеза H_1 принимается (отвергается) в качестве одного из возможных решений с формулировкой "гипотеза H_0 противоречит (не противоречит) выборочным данным на уровне значимости α".

Пусть нам необходимо проверить справедливость гипотезы H_o при альтернативе H_1. Поскольку выборочные наблюдения, на основе которых вопрос решается, являются случайными величинами, то абсолютно достоверно этого сделать нельзя. Всегда остается риск отвергнуть истинную гипотезу H_o, тем самым совершив так называемую *ошибку первого рода*, или же принять ложную гипотезу H_o, сделав *ошибку второго рода*.

H_o	Принимается	Отвергается
Верная	Правильное решение	Ошибка I рода
Неверная	Ошибка II рода	Правильное решение

Уровень значимости α – вероятность допустить ошибку I рода, т.е. $P(\Lambda_{набл} \in \Lambda_{крит} | H_0) = \alpha$. Вероятность допустить ошибку II рода обозначим β . Вероятность $1 - \beta$ не допустить ошибку II рода (принять H_o, когда она неверна, т.е. $P(\Lambda_{набл} \notin \Lambda_{крит} | H_1) = 1\text{-}\beta$) называется *мощностью критерия*. Желательно сделать α и β как можно малыми, но при фиксированном объеме выборки уменьшение одной из них приводит к увеличению другой. Критическую область выбирают так, чтобы при заданном уровне значимости α мощность критерия $1\text{-}\beta$ была максимальной.

Для отыскания *правосторонней* критической области необходимо найти $\lambda_{кр}$. Зададим уровень значимости α. Тогда при условии, что H_o является справедливой, $P(\Lambda_{набл} > \lambda_{крит}|H_0) = \alpha$. Для каждого статистического критерия Λ существуют таблицы, по которым определяют критическую точку, удовлетворяющую заданному уровню значимости.

Левосторонняя критическая область задается уравнением $P(\Lambda_{набл} < \lambda_{крит}|H_0) = \alpha$.

Двусторонняя критическая область задается уравнением

$$P(\Lambda_{набл} < \lambda_{крит}|H_0) + P(\Lambda_{набл} > \lambda_{крит}|H_0) = \alpha.$$

Для многих задач проверки статистических гипотез разработан не один статистический критерий, а целый ряд. Чтобы выбрать из них определенный критерий для использования в конкретной практической ситуации, проводят сравнение критериев по различным показателям качества. В качестве примера рассмотрим лишь два показателя качества критерия проверки статистической гипотезы – состоятельность и несмещенность.

Пусть объем выборки n растет, а $\Lambda(n)$ и $\lambda_{крит}(n)$ – статистики критерия и критические области, соответственно.

Критерий называется *состоятельным,* если

$$\lim_{n\to\infty} P(\Lambda(n) \in \lambda_{крит}(n)|H_0) = 1$$

т.е. вероятность отвергнуть нулевую гипотезу стремится к 1, если верна альтернативная гипотеза.

Статистический критерий называется *несмещенным*, если для любого Λ_0, удовлетворяющего H_0, и любого Λ_1 , удовлетворяющего H_1, справедливо неравенство

$$P(\Lambda_0 \in \Lambda_{крит}|H_0) < P(\Lambda_1 \in \Lambda_{крит}|H_1)$$

т.е. при справедливости H_0 вероятность отвергнуть H_0 меньше, чем при справедливости H_1.

При наличии нескольких статистических критериев в одной и той же задаче проверки статистических гипотез следует использовать состоятельные и несмещенные критерии.

Поступим следующим образом.

Зафиксируем некоторое малое положительное число α (например, $\alpha = 0.01$) - *уровень значимости*.

Выберем некоторую функцию Λ, зависящую от выборки, которую будем называть *статистикой критерия*.

Среди всех возможных значений, принимаемых Λ, определим множество K_α

(зависящее от вида статистики критерия Λ и от уровня значимости α), для которого вероятность события ($\Lambda \in K_\alpha$), если верна гипотеза H_o, равна α, т.е. $P(\Lambda \in K\alpha | H_0) = \alpha$

Статистический критерий состоит в следующем: если статистика критерия Λ, подсчитанная по выборке $\Lambda_{набл}$, попадает в множество K_α($\Lambda_{набл} \in K_\alpha$), то гипотеза H_o отвергается с вероятностью ошибки первого рода, равной α, в противном случае (т.е. когда $\Lambda_{набл} \notin K_\alpha$), она принимается. В силу этого, K_α назовем *критической областью размера* α.

Сконструированный таким способом критерий отвергает истинную гипотезу H_o лишь в 100α случаях из 100 (при уровне значимости $\alpha = 0.01$, например, лишь в одном случае из ста).

Согласно требованиям фармакопеи в биологических исследованиях принимается $\alpha = 0.05$, а при разработке биологических стандартов – $\alpha = 0.01$.

Замечание. Принцип проверки статистических гипотез не дает логического доказательства ее верности или неверности. Принятие гипотезы H_o по сравнению с альтернативной H_1 не означает, что мы уверены в абсолютной правильности H_o или что высказанное в H_o утверждение является наилучшим; просто H_o не противоречит имеющимся у нас выборочным данным. Таким же свойством наряду с H_o могут обладать и другие гипотезы (H_o может быть отвергнута в сравнении с другой альтернативой H_1).

2. Сравнение средних

Имеются независимые выборки $X_1, X_2, ..., X_m$ и $Y_1, Y_2, ..., Y_n$ объема m и n, извлеченные из *нормально* распределенных с неизвестными параметрами (μ_x, σ) и (μ_y, σ) генеральных совокупностей. Неизвестные генеральные дисперсии по предположению *равны*.

Проверяется нулевая гипотеза о равенстве генеральных средних μ_x и μ_y,

т. е. H_o: $\mu_x = \mu_y$.

1) Вычислим выборочные средние выборок $\overline{X}$ и $\overline{Y}$, а также их исправленные выборочные дисперсии S_x^2 и S_y^2 (см. [1, формулы (8.7), (8.8)]).

2) Образуем результирующую оценку общей дисперсии

$$S^2 = \frac{(m-1)S_x^2 + (n-1)S_y^2}{m+n-2} \qquad (8.1)$$

и статистику критерия

$$\Lambda = \frac{\overline{X} - \overline{Y}}{S\sqrt{\frac{1}{m} + \frac{1}{n}}}, \qquad (8.2)$$

которая при нулевой гипотезе имеет распределение Стьюдента с $m + n - 2$ степенями свободы.

Для проверки нулевой гипотезы H_o: $\mu_x = \mu_y$ следует по формуле (8.2) вычислить наблюдаемое значение статистики критерия $\Lambda_{набл}$.

1) При конкурирующей гипотезе H_1: $\mu_x \neq \mu_y$ следует по таблице критических точек распределения Стьюдента (уровень значимости α, число степеней свободы $k=m+n-2$, двусторонняя область) из приложения 2 определить величину $t_{кр}(\alpha, m+n-2)$. Если

$$|\Lambda_{набл}| > t_{кр}(\alpha, m+n-2), \qquad (8.3)$$

то нулевая гипотеза отклоняется на уровне α; в противном случае нулевая гипотеза принимается.

2) При конкурирующей гипотезе H_1: $\mu_x > \mu_y$ нулевая гипотеза отклоняется на уровне α, если

$$|\Lambda_{набл}| > t_{кр}(2\alpha, m+n-2); \qquad (8.4)$$

в противном случае нулевая гипотеза принимается.

Пример 8.1. Две группы детей, одинаковых по оценке умственных способностей, независимо обучались по двум различным методикам преподавания. Затем их подвергли выборочному тестированию, давшему следующие результаты: объем выборки из первой группы равен 20, $\overline{X} = 29.233$, $S_x^2 = 5.62$; объем выборки из второй группы равен 10, $\overline{Y} = 27.562$, $S_y^2 = 2.19$.

В предположении, что изучаемые показатели в каждой группе имеют *нормальное* распределение с неизвестными средними и неизвестными, но *одинаковыми дисперсиями*, проверить при уровне значимости 0,05 существенно ли отличаются средние показания групп?

Решение. Проверяем H_o: $\mu_x = \mu_y$ при двусторонней альтернативе

H_1: $\mu_x \neq \mu_y$. Наблюдаемые значения статистики критерия

$$\Lambda_{набл} = \frac{29.233 - 27.562}{\sqrt{\frac{(19 \cdot 5.62 + 9 \cdot 2.19)}{28} \cdot \left(\frac{1}{20} + \frac{1}{10}\right)}} = 2.03,$$

а критическая точка $t_{кр}(0.05, 28)$ распределения Стьюдента, соответствующая двусторонней области (см. приложение 2) равна 2.05.

Поскольку $|\Lambda_{набл}| = 2.03 < t_{кр}(0.05; 28) = 2.05$, у нас нет оснований отвергать нулевую гипотезу о равенстве средних в группах.

Пример 8.2. По данным примера 8.1 проверить при уровне значимости 0.05 нулевую гипотезу о равенстве среднихH_o: $\mu_x = \mu_y$ при конкурирующей гипотезе H_1: $\mu_x > \mu_y$.

Решение. Поскольку наблюдаемое значение статистики $\Lambda_{\text{набл}} = 2.03$, а соответствующее значение критической точки $t_{\text{кр}}(0.10, 28) = 1.7$, находим $\Lambda_{\text{набл}} > t_{\text{кр}}(0.10, 28)$. Таким образом (см. (8.4)), наблюдаемое значение статистики попадает в критическую область и, следовательно, нулевая гипотеза отвергается в пользу предположения $\mu_x > \mu_y$. Другими словами, выборочные данные значимо отклоняются от нулевой гипотезы, и мы в отличие от вывода примера 8.1 вынуждены заключить, что средние показания первой группы существенно превышают средние показания второй группы.

Отметим, что сравнение примеров 8.1 и 8.2 демонстрирует важность выбора конкурирующей гипотезы при проверке статистических гипотез.

3. Критерий равенства двух дисперсий

Пусть имеются две независимых выборки $X_1, X_2, ..., X_m$ и $Y_1, Y_2, ..., Y_n$ объема m и n, извлеченные из *нормально* распределенных генеральных совокупностей с неизвестными параметрами (μ_x, σ_1)и(μ_y, σ_2), соответственно.

Требуется проверить, согласуются ли выборочные данные с нулевой гипотезой

H_o: $\sigma_1^2 = \sigma_2^2$.

Поскольку $\sigma_1^2 = D(X)$, а $\sigma_2^2 = D(Y)$, то нулевая гипотеза предполагает равенство генеральных дисперсий $D(X)$ и $D(Y)$, т. е. H_o: $D(X) = D(Y)$.

По выборкам, отвечающим случайным величинам X и Y, вычислим исправленные выборочные дисперсии S_x^2 и S_y^2, и составим статистику критерия

$$\Lambda = \frac{S_x^2}{S_y^2}, \qquad (8.5)$$

считая без потери общности, что

$$S_x^2 \geq S_y^2 \qquad (8.6)$$

(в противном случае следует поменять обозначения X и Y местами).

При *нулевой* гипотезе статистика критерия Λ является случайной величиной, имеющей *распределение Фишера-Снедекора* (F – распределение) с $m - 1$ и $n - 1$ степенями свободы

(напоминаем, что m – объем выборки, соответствующей числителю S_x^2 дроби Λ, в то время как n – объем выборки, определяющей знаменатель S_y^2 этой дроби).

1) Для проверки на уровне значимости α нулевой гипотезы H_0: $D(X) = D(Y)$ при конкурирующей гипотезе H_1: $D(X) > D(Y)$ следует по выборочным данным с учетом договоренности (8.6) вычислить наблюдаемое значение статистики критерия $\Lambda_{\text{набл}}$ и сравнить его с величиной $f_{\text{кр}}(\alpha; m-1, n-1)$, которая является критической точкой F-распределения, отвечающей уровню значимости α и степеням свободы $k_1 = m - 1$ и $k_2 = n - 1$. Значения $f_{\text{кр}}(\alpha; k_1, k_2)$ при $\alpha = 0.01, 0.05$ и некоторых k_1, k_2 содержатся в Приложении 3.

Если

$$\Lambda_{\text{набл}} \le f_{\text{кр}}(\alpha; m-1, n-1), \qquad (8.7)$$

то нулевая гипотеза принимается;

если же $\Lambda_{\text{набл}} > f_{\text{кр}}(\alpha; m-1, n-1)$, то нулевая гипотеза отклоняется на уровне α (и, следовательно, принимается конкурирующая гипотеза H_1: $D(X) > D(Y)$).

2) В случае конкурирующей гипотезы H_1: $D(X) \neq D(Y)$ нулевая гипотеза принимается, если

$$\Lambda_{\text{набл}} \le f_{\text{кр}}(\frac{\alpha}{2}; m-1, n-1), \qquad (8.8)$$

и отвергается на уровне α в противоположном случае.

Пример 8.3. По данным примера 8.1 проверить предположение о равенстве дисперсий в тестируемых группах при $H_1 : \sigma_1 \neq \sigma_2$ и уровне значимости $\alpha = 0.05$.

Решение. Имеем, $\Lambda_{\text{набл}} = \frac{5.62}{2.19} = 2.57$. Значения $f_{\text{кр}}(0.025; 19, 9)$ в таблицах из Приложения 3 нет, однако $f_{\text{кр}}(0.025; 19, 9) > f_{\text{кр}}(0.05; 19, 9) > f_{\text{кр}}(0.05; 20, 9) = 2.93$. Поскольку $\Lambda_{\text{набл}} = 2.57 < 2.93$, условие (8.8) заведомо выполняется, откуда следует, что оснований отвергать предположение о равенстве дисперсий в тестируемых группах у нас нет.

4. Проверка значимости коэффициента корреляции

Предположим, что интересующие нас характеристики X и Y объектов генеральной совокупности имеют *двумерное нормальное* распределение с неизвестным коэффициентом корреляции ρ. Требуется по наблюдениям выборки объема n, извлеченной из этой

совокупности, проверить нулевую гипотезу H_o:ρ =0 при конкурирующей гипотезе H_1 :$\rho\neq0$, т.е. выяснить будут ли случайные величины X и Y некоррелированными или нет.

Обозначим через r выборочный коэффициент корреляции (см. [1, 856] или лаб.раб. 7, (7.8)) и положим

$$\Lambda = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \qquad (8.9)$$

Известно, что при нулевой гипотезе случайная величина Λ имеет распределение Стьюдента с n-2 степенями свободы.

Статистический критерий формулируется следующим образом.

По извлеченным из генеральной совокупности наблюдениям выборки объема n следует вычислить выборочный коэффициент корреляции $r_{набл}$ и соответствующее ему наблюдаемое значение статистики критерия $\Lambda_{набл} = \frac{r_{набл}\sqrt{n-2}}{\sqrt{1-r^2_{набл}}}$.

Если $|\Lambda_{набл}| > t_{кр}(\alpha,n-2)$, то нулевая гипотеза отклоняется на уровне α, в противном случае, при $|\Lambda_{набл}| \le t_{кр}(\alpha,n-2)$, она принимается.

Пример 8.4. По данным примера 7.1, предполагая, что пара случайных величин X и Y имеют двумерное нормальное распределение, при уровне значимости 0.01, проверить нулевую гипотезу о равенстве нулю коэффициента корреляции ρ между X и Y при конкурирующей гипотезе H_1 :$\rho\neq0$.

Решение. В примере 7.1 $n=11$, $r_{набл}$= *-0.61*. Отсюда, $\Lambda_{набл} = \frac{-0.61\sqrt{11-2}}{\sqrt{1-0.61^2}} = -2.31$.

В таблице критических точек распределения Стьюдента (приложение 2) по уровню значимости 0.01 и числу степеней свободы 11-2=9 находим критическую точку двусторонней критической области $t_{кр}(0.01,9)$ =3.25 . Поскольку $|\Lambda_{набл}|$=2.31 < 3.25, оснований отвергать нулевую гипотезу нет. Таким образом, выборочные данные не противоречат предположению о некоррелированности случайных величин X и Y.

5. Статистическая проверка непараметрических гипотез. Критерий согласия Пирсона.

Если закон распределения генеральной совокупности неизвестен, но есть основания предположить, что он имеет определенный вид (назовем его $F(x)$), то проверяют нулевую гипотезу

H_o: генеральная совокупность распределена по закону $F(x)$.

Проверка гипотезы о предполагаемом законе неизвестного распределения производится аналогично проверкам гипотез о параметрах распределения, при помощи т.н. критерия согласия.

Критерием согласия называют критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Имеется несколько критериев согласия : χ^2 («хи квадрат») К . Пирсона, Колмогорова, Смирнова и др. Ограничимся описанием применения критерия Пирсона к проверке гипотезы о нормальном распределении генеральной совокупности (критерий аналогично применяется и для других распределений). С этой целью будем сравнивать эмпирические (наблюдаемые) и теоретические (вычисленные для предполагаемого распределения) частоты. Обычно эмпирические и теоретические частоты различаются.

Расхождение может быть случайным (незначимым) и объясняется либо малым числом наблюдений, либо способом их группировки, либо другими причинами. Возможно, что расхождение частот неслучайно (значимо) и объясняется тем, что теоретические частоты вычислены исходя из неверной гипотезы о распределении генеральной совокупности. Критерий Пирсона отвечает на поставленный выше вопрос. При этом надо понимать, что как и любой критерий, он не доказывает справедливость гипотезы, а лишь устанавливает на принятом уровне значимости ее согласие или несогласие с данными наблюдений.

Эмпирические и выравнивающие (теоретические) частоты

Дискретное распределение. Рассмотрим дискретную случайную величину *X*, закон распределения которой неизвестен. Пусть произведено *n* испытаний, в которых величина *X* приняла n_1 раз значение x_1, n_2 раз значение x_2, … , n_m раз значение x_m, причем $\sum_{i=1}^{m} n_{i=n}$

Эмпирическими частотами называют фактически наблюдаемые частоты n_i (варианты) .

Пусть имеются основания предположить, что изучаемая величина *X* распределена по некоторому определенному закону *F*. Чтобы проверить, согласуется ли это предположение с данными наблюдений, вычисляют частоты наблюдаемых значений, т. е. находят теоретически n_i' каждого из наблюдаемых значений в предположении, что величина *X* распределена по предполагаемому закону.

Выравнивающими (теоретическими) в отличие от фактически наблюдаемых эмпирических частот называют частоты n_i', найденные теоретически (вычислением). Выравнивающие частоты находят с помощью равенства

$$n_i' = \mathrm{n}P_i,$$

где *n*- число испытаний; -P_i вероятность наблюдаемого значения $\mathrm{x_i}$, вычисленная при допущении, что *X* имеет предполагаемое распределение.

Итак, *выравнивающая частота наблюдаемого значения* x_i *дискретного распределения равна произведению числа испытаний на вероятность этого наблюдаемого значения*.

Непрерывное распределение. В случае непрерывного распределения, вероятности отдельных возможных значений равны нулю. Поэтому весь интервал возможных значений делят на *m* непересекающихся интервалов и вычисляют вероятности P_i попадания *X* в *i*-й частичный интервал, а затем, как и для дискретного распределения, умножают число испытаний на эти вероятности .

Итак, *выравнивающую частоту* непрерывного распределения находят по равенству

$$\mathrm{n_i^{'}} = \mathrm{n}P_\mathrm{i},$$

где *n* - число испытаний ; P_i - вероятность попадания *X* в *i*-й частичный интервал, вычисленная при допущении, что X имеет предполагаемое распределение.

Допустим, что для предполагаемого распределения генеральной совокупности вычислены теоретические частоты n_i'. При уровне значимости α требуется проверить нулевую гипотезу H_o: генеральная совокупность распределена по предполагаемому закону *F(x).*

В качестве критерия проверки нулевой гипотезы примем случайную величину

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - n_i')^2}{n_i'} = \sum_{i=1}^{m} \frac{(n_i - \mathrm{n}P_i)^2}{\mathrm{n}P_i} \qquad (*)$$

Эта величина случайная, так как в различных опытах она принимает различные, заранее неизвестные значения. Ясно, что чем меньше различаются эмпирические и теоретические частоты, тем меньше величина критерия, и, следовательно, он в известной степени характеризует близость эмпирического и теоретического распределений.

Доказано, что при $n \to \infty$ закон распределения случайной величины (*) независимо от того, какому закону распределения подчинена генеральная совокупность, стремится к закону распределения χ^2 с *k* степенями свободы. Поэтому случайная величина обозначена через χ^2, а сам критерий называют критерием согласия «хи квадрат».

Число степеней свободы находят по равенству *k* = *m*- 1 -*r*, где *m* - число групп (частичных интервалов) выборки ; *r* - число параметров предполагаемого распределения, которые оценены по данным выборки.

В частности, если предполагаемое распределение - нормальное, то оценивают два параметра (математическое ожидание и среднее квадратичное отклонение), поэтому *r* = 2 и число степеней свободы *k* = *m*- 1 -*r* = *m* - 1 - 2 = *m*- 3.

Поскольку односторонний критерий более «жестко» отвергает нулевую гипотезу, чем двусторонний, построим правостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α:

$$P\left(\chi^2 > \chi^2_{\text{кр}}(\alpha, k)\right) = \alpha$$

Таким образом, правосторонняя критическая область определяется неравенством $\chi^2 > \chi^2_{\text{кр}}(\alpha, k)$, а область принятия нулевой гипотезы - неравенством$\chi^2 < \chi^2_{\text{кр}}(\alpha, k)$. Обозначим значение критерия, вычисленное по данным наблюдений, через и сформулируем правило проверки нулевой гипотезы.

Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу H_o : генеральная совокупность распределена, например, нормально,

надо сначала вычислить теоретические частоты, а затем наблюдаемое значение критерия:

$$\chi^2_{\text{набл}} = \sum_{i=1}^{m} \frac{(n_i - n_i')^2}{n_i'} = \sum_{i=1}^{m} \frac{(n_i - \text{n}P_i)^2}{\text{n}P_i} = \sum_{i=1}^{m} \frac{{n_i}^2}{n_i'} - n$$

и по таблице критических точек распределения χ^2, по заданному уровню значимости α и числу степеней свободы $k = m$ -3 найти критическую точку $\chi^2_{\text{кр}}(\alpha, k)$.

Если $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$ - нет оснований отвергнуть нулевую гипотезу.

Если $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$- нулевую гипотезу отвергают.

Замечание. Объем выборки должен быть достаточно велик, не менее 50. Каждая группа должна содержать не менее 5-8 вариант; малочисленные группы следует объединять в одну, суммируя частоты.

Методика вычисления теоретических частот нормального распределения

1. Весь интервал наблюдаемых значений X (выборки объема n) делят на m частичных интервалов (x_i, x_{i+1}) одинаковой длины. Находят середины частичных интервалов $x_i^* = \frac{x_i + x_{i+1}}{2}$; в качестве частоты n_i варианты x_i^* принимают число вариант, которые попали в i-й интервал.

В итоге получают последовательность равноотстоящих вариант и соответствующих им частот:

x_1^*	x_2^*	…	x_m^*
n_1	n_2	…	n_m

При этом $\sum_{i=1}^{m} n_{i=n}$.

2. Вычисляют выборочную среднюю $\overline{x^*} = \frac{1}{n}\sum_i n_i x_i^*$ и выборочное среднее квадратичное отклонение $\overline{\sigma^*} = \sqrt{\frac{1}{n}\sum_i n_i x_i^2 - (\overline{x^*})^2}$.

3. Нормируют случайную величину *X*, т. е. переходят к величине $Z = \frac{X - \overline{x^*}}{\overline{\sigma^*}}$ и вычисляют концы интервалов (z_i, z_{i+1}):

$$z_i = \frac{x_i - \overline{x^*}}{\overline{\sigma^*}}, \qquad z_{i+1} = \frac{x_{i+1} - \overline{x^*}}{\overline{\sigma^*}},$$

причем наименьшее значение Z, т. е. z_1, полагают равным $-\infty$, а наибольшее, т. е. z_m, полагают равным ∞.

4. Вычисляют теоретические вероятности p_i попадания *X* в интервалы (x_i, x_{i+1})по равенству $p_i = \Phi(z_{i+1}) - \Phi(z_i)$

5. Находят искомые теоретические частоты $n_i^{'} = np_i$.

Критерий Пирсона позволяет производить проверку согласия эмпирической функции распределения с гипотетической функцией $F(x)$, принадлежащей к некоторому множеству функций определенного вида (нормальных, показательных, биномиальных и т.д.).

Пусть случайный признак *X* имеет функцию распределения $F(x)$, принадлежащую некоторому классу функций. Из генеральной совокупности извлечена выборка объема $n \geq 50$.

Разобьем весь диапазон полученных результатов на *k* частичных интервалов равной длины, и пусть в каждом частичном интервале оказалось n_i измерений, причем $\sum_{i=1}^{m} n_{i=n}$. Составим сгруппированный статистический ряд распределения частот (см. Лабораторная раб. 1).

Требуется на основе имеющейся информации с заданным уровнем значимости α проверить нулевую гипотезу о том, что генеральная совокупность распределена по гипотетическому закону $F(x)$.

Нулевой непараметрической гипотезой называется гипотеза относительно общего вида функции распределения.

При проверке нулевой гипотезы с помо$\sum_{i=1}^{m} n_{i=n}$щью критерия согласия придерживаются следующей последовательности действий:

1) На основании гипотетической функции $F(x)$ вычислим вероятности попадания с.в. X в частичные интервалы $[h_{i-1}, h_i)$, , для нормального закона (см. Лаб.раб. 3):

$$p_i = P(h_{i-1} \le X \le h_i) = \Phi\left(\frac{h_i - \bar{x}}{\bar{s}}\right) - \Phi\left(\frac{h_{i-1} - \bar{x}}{\bar{s}}\right), \qquad i = 2, \dots, k-1,$$

.

2) Умножая полученные вероятности p_i на объем выборки n, получим теоретические частоты np_i частичных интервалов $[h_{i-1}, h_i)$, т.е. частоты, которые следует ожидать, если нулевая гипотеза справедлива;

3) Расчеты оформим в виде таблицы

x_i	…	$[h_{i-1}, h_i)$	…	…
n_i	…	…	…	…
p_i	…	…	…	…
$n \cdot p_i$	…	…	…	…

4) Вычислим выборочную статистику (критерий) :

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} \qquad (8.10).$$

5) Выберем уровень значимости $\alpha=0.05$. Рассчитаем ν – число степеней свободы: $\nu = m-r-1$ (r - число параметров предполагаемого распределения, m – число интервалов).

6) По таблице χ^2 распределения находим критическое значение $\chi^2_{\nu,\alpha}$, удовлетворяющее условию $P(\chi^2 \ge \chi^2_{\nu,\alpha}) = \alpha$.

7) Сравним наблюдаемое значение выборочной статистики χ^2, вычисленное по формуле (8.10), с критическим значением , принимаем одно из двух решений:

а) если набл $\chi^2 \ge \chi^2_{\nu,\alpha}$, то нулевая гипотеза отвергается в пользу альтернативной, т.е. считается, что гипотетическая функция не согласуется с результатами эксперимента;

2) если набл $\chi^2 < \chi^2_{\nu,\alpha}$, то считается, что нет оснований для отклонения нулевой гипотезы, т.е. гипотетическая функция согласуется с результатами эксперимента.

Замечание. При применении критерия необходимо, чтобы в каждом частичном интервале было не менее 5 элементов. Если число элементов (частота) меньше 5, то рекомендуется объединять такие частичные интервалы со смежными.

x_i	[22 , 26)	[26, 30)	[30, 34)	[34, 38)	[38, 42)	[42, 46)	[46, 50)
n_i	2	5	9	18	9	5	2

Таблица 8.1

В соответствие с замечанием крайние интервалы следует объединить, заменив их интервалами (∞ ,30] и [42,∞) (Таблица 8.2):

Пример 8.5. По данным выборки Лабораторной работы 1 проверить гипотезу о нормальности распределения генеральной совокупности признака *X* (использовать результаты лабораторных работ 1 и 2 n=50) с уровнем значимости α = 0.05.

H_o: случайный признак *X* имеет нормальное распределение, $X \sim N(36,30.69)$.

H_1: распределение *X* существенно отличается от нормального.

Решение. Используем интервальный вариационный ряд лабораторной работы 1:

Таблица 8.1

В соответствие с замечанием крайние интервалы следует объединить, заменив их интервалами (∞ ,30] и [42,∞) (Таблица 8.2):

x_i	$(-\infty, 30)$	[30, 34)	[34, 38)	[38, 42)	[42, ∞)
n_i	7	9	18	9	7
p_i					
np_i					
$n_i - np_i$					
$(n_i - np_i)^2$					
$\frac{(n_i - np_i)^2}{np_i}$					

Таблица 8.2

Вычислим наблюдаемое значение χ^2=1.81, *m*= 5, *r*=2 ν = *m-r-1*=5-2-1=2.

По таблицам Приложения 4 находим $\chi^2_{2,0.05} = 6$.

$\chi^2 < \chi^2_{\nu,\alpha} \Rightarrow$ Нет оснований отвергать нулевую гипотезу, гипотеза не противоречит статистическим данным.

6. Контрольные вопросы

5. Что называется статистическим критерием?
6. Что такое нулевая гипотеза?
7. Что такое конкурирующая гипотеза?
8. Как выбирается уровень значимости?
9. Поясните смысл ошибки первого рода?

7.Порядок выполнения работы

По данным выборки лабораторной работы №1 проверить гипотезу о нормальном распределении генеральной совокупности, из которой была проведена выборка, с уровнем значимости α = 0.05, применяя критерий согласия Пирсона (использовать результаты лабораторных работ №1 и №2 *n*=50).

Решить задачу, предложенную преподавателем (см. Варианты самостоятельных работ).

8. Варианты самостоятельных работ

1. Даны результаты измерений пульса 11 студентов, проведенных сразу после окончания занятий по физкультуре (выборка $X1$), и 10 студентов – через 30 минут после окончания занятий по физкультуре (выборка $X2$). $X1_{ср}$ = 140 уд/мин, $X2_{ср}$ = 74 уд/мин. Расчетное значение t- критерия составило $t_{эксп}$ = - 0,6. При уровне значимости $\alpha \le 0{,}05$ определить значимость различия значения пульса.

2. Даны результаты измерений роста (в см) 19 детей (выборка $X1$) и массы (в кг) 15 детей из той же группы: оценки дисперсий соответственно равны $S_1^2 = 130{,}39$, $S_2^2 = 32{,}98$. При уровне значимости $\alpha \le 0{,}05$ проверить гипотезу о равенстве генеральных дисперсий по результатам проведенных измерений.

3. Даны результаты измерений пульса 13 студентов, проведенных перед началом занятий по физкультуре (выборка $X1$), и 12 студентов – после окончания занятий по физкультуре (выборка $X2$): $X1_{ср}$ = 86 уд/мин, $X2_{ср}$ = 152 уд/мин. Расчетное значение t- критерия составило $t_{эксп}$ = - 1,32. При уровне значимости $\alpha \le 0{,}05$ определить значимость различения средних значений пульса.

4. Даны результаты измерений систолического давления в начальной стадии шока (мм рт. ст.) у 21 больного, оставшихся в живых (выборка $X1$), и у 12 больных, умерших после шока (выборка $X2$): $X1_{ср}$ = 124, $X2_{ср}$ = 102. Расчетное значение t- критерия составило $t_{эксп}$ = 0,37. При уровне значимости $\alpha \le 0{,}05$ определить значимость различения средних значений.

5. Даны результаты измерений систолического давления в начальной стадии шока (мм рт. ст.) у 13 больных, оставшихся в живых после шока (выборка $X1$), и у 12 больных, умерших после шока (выборка $X2$): $X1_{ср}$ = -66, $X2_{ср}$ = 173. Расчетное значение t- критерия составило $t_{эксп}$ = 3,59. При уровне значимости $\alpha \le 0{,}05$ определить значимость различения средних значений.

6. Даны результаты измерений среднего артериального давления в начальной стадии шока (мм рт. ст.) у 15 больных в начальной стадии шока (выборка $X1$), и у 15 больных в конечной (выборка $X2$) стадии шока: $X1_{ср}$ = 99, $X2_{ср}$ = 63. Расчетное значение t- критерия составило $t_{эксп}$ = 2,47. При уровне значимости $\alpha \le 0{,}05$ определить значимость различия средних значений.

7. Даны результаты измерений среднего артериального давления (мм рт. ст.) у 23 больных в начальной стадии шока (выборка *X*1), и у тех же 23 больных в конечной (выборка *X*2) стадии шока: $X1_{ср} = 85$, $X2_{ср} = 87$. Расчетное значение *t*- критерия составило $t_{эксп} = -0,03$. При уровне значимости $\alpha \le 0,05$ определить значимость различия средних значений.

8. Даны результаты измерений систолического давления в начальной стадии шока (мм рт. ст.) у 14 больных, оставшихся в живых после шока (выборка *X*1), и у 11 больных, умерших после шока (выборка *X*2): $X1_{ср} = -137,3$, $X2_{ср} = 86,5$. Расчетное значение *t*- критерия составило $t_{эксп} = 3,8$. При уровне значимости $\alpha \le 0,05$ определить значимость различия средних значений.

9. Даны результаты измерений диастолического давления (мм.рт. ст.) у 18 мужчин и (выборка *X*1), и у 13 женщин (выборка *X*2): $X1_{ср} = 63,6$, $X2_{ср} = -31,5$. Расчетное значение *t*- критерия составило $t_{эксп} = 2,19$. При уровне значимости $\alpha \le 0,05$ определить значимость различия средних значений.

10. Даны результаты измерений частоты сердечных сокращений 11 студентов, проведенных сразу после окончания занятий по физкультуре (выборка X1), и 10 студентов – через 30 минут после окончания занятий по физкультуре (выборка X2). оценки дисперсий соответственно равны $S_1^2 = 139,9$, $S_2^2 = 74,2$. При уровне значимости $\alpha \le 0,05$ проверить гипотезу о равенстве генеральных дисперсий по результатам проведенных измерений.

11. Даны результаты измерений систолического давления в начальной стадии шока (мм рт. ст.) у 14 больных, оставшихся в живых (выборка X1), и у 11 больных, умерших после шока (выборка X2): оценки дисперсий соответственно равны $S_1^2 = 106,6$ $S_2^2 = 40,9$. При уровне значимости $\alpha \le 0,05$ проверить гипотезу о равенстве генеральных дисперсий по результатам проведенных измерений.

12. Даны результаты измерений систолического давления в начальной стадии шока (мм рт. ст.) у 21 больного, оставшихся в живых (выборка X1), и у 12 больных, умерших после шока (выборка X2): оценки дисперсий соответственно равны $S_1^2 = 172,8$, $S_2^2 = 161,4$. При уровне значимости $\alpha \le 0,05$ проверить гипотезу о равенстве генеральных дисперсий по результатам проведенных измерений.

13. Даны результаты измерений частоты сердечных сокращений у 13 студентов, перед началом занятий по физкультуре (выборка X1), и 12 студентов после окончания занятий по физкультуре (выборка X2). оценки дисперсий соответственно равны $S_1^2 = 79{,}6,\ S_2^2 = 125{,}2.$ При уровне значимости $\alpha \le 0{,}05$ проверить гипотезу о равенстве генеральных дисперсий по результатам проведенных измерений.

14. Даны результаты измерений систолического давления в начальной стадии шока (мм рт. ст.) у 13 больных, оставшихся в живых (выборка X1), и у 12 больных, умерших после шока (выборка X2): оценки дисперсий соответственно равны $S_1^2 = 73{,}8,\ S_2^2 = 116{,}9.$ При уровне значимости $\alpha \le 0{,}05$ проверить гипотезу о равенстве генеральных дисперсий по результатам проведенных измерений.

15. Даны результаты измерений среднего артериального давления (мм рт. ст.) у 15 больных в начальной стадии (выборка X1), и у тех же 15 больных в конечной (выборка X2) в конечной стадиях шока): оценки дисперсий соответственно равны $S_1^2 = 28{,}8,\ S_2^2 = 40{,}2.$ При уровне значимости $\alpha \le 0{,}05$ проверить гипотезу о равенстве генеральных дисперсий по результатам проведенных измерений.

16. Даны результаты измерений среднего артериального давления (мм рт. ст.) у 23 больных в начальной (выборка X1), и у тех же 23 больных в конечной (выборка X2) стадиях шока: оценки дисперсий соответственно равны $S_1^2 = 270{,}2,\ S_2^2 = 233{,}9.$ При уровне значимости $\alpha \le 0{,}05$ проверить гипотезу о равенстве генеральных дисперсий по результатам проведенных измерений.

17. Средняя масса таблеток, найденных по выборке объемом 35 из первой партии, составила 0,5 г; по выборке объемом 40 из второй партии – 0, 51 г. Расчетное значение значение t-критерия составило $t_{эксп} = -1{,}52$. При $\alpha \le 0{,}05$ выяснить, можно ли считать различие в средних значениях масс таблеток случайным.

18. Изучали влияние кобальта на массу тела кроликов. Опыт проводился на двух группах животных: опытной объемом 8 и контрольной объемом 9. Подопытные кролики, в отличие от контрольных, ежедневно получали добавку к рациону в виде хлористого кобальта по 0,06 г на 1 кг массы. За время опыта животные дали следующие прибавки живой массы тела: X_1

= 638 г при дисперсии $S_1^2 = 2596$ г^2 против X_2 = 626 г при дисперсии $S_2^2 = 3579$ г^2 у контрольной группы. Можно ли для оценки достоверности этой разности использовать критерий Стьюдента? Привести обоснование – расчеты с использованием критерия Фишера при $\alpha \le 0{,}05$.

19. Изучали влияние эндотоксина на выживаемость облученных животных. В опытной группе было 36 животных, выжило 23 (63,9%). В контрольной группе было 14 животных, выжило после облучения 3 (21,4%). Можно ли судить о положительном влиянии эндотоксина на выживаемость животных, если наблюдаемое значение t- критерия Стьюдента $t_{набл}$ = -2,71. Уровень доверительной вероятности принять $P \ge 0{,}95$.

20. Изучали влияние туберкулина на состав периферической крови низших обезьян: понижалось количество эозинофилов после введения туберкулина у большинства обезьян. Используя критерий знаков, получили следующие результаты: из 14 наблюдений 2 разности – нулевые, 10 – положительные, 2 – отрицательные. При $P \ge 0{,}95$ $n_{кр}$=12. Можно ли утверждать, что введение туберкулина вызывает снижение эозинофилов в периферической крови обезьян?

21. Изучали данные о годовых удоях коров и их потомства по второму и третьему отелам. При использовании критерия знаков получили следующие результаты: из 25 парных наблюдений положительная разница – в 12 случаях, отрицательная – в 13 случаях. При $P \ge 0{,}95$ $n_{кр}$=10. Сделайте вывод о достоверности различия между удоями коров материнского поколения и их потомства.

22. В таблице приведены эмпирические и вычисленные по нормальному закону частоты распределения длины тела у 267 мужчин. Из приведенных данных видно, что между этими частотами нет полного совпадения. При $\alpha \le 0{,}05$ нужно установить, случайны или закономерны эти различия, т.е. следует ли это распределение роста мужчин нормальному закону.

Расчет χ^2 - критерия дал значение 1,47.

Эмпирические частоты 12 31 71 82 46 19 6.

Теоретические частоты 12 34 68 78 51 20 5.

23. В таблице приведены эмпирические и вычисленные по нормальному закону частоты распределения урожайности фасоли для 200 семян. При $\alpha \le 0{,}05$ нужно установить,

случайны или закономерны эти различия, т.е. следует ли это распределение нормальному закону. Расчет χ^2 - критерия дал значение 20,09.

Эмпирические частоты 1 5 17 45 70 51 10 1 0.

Теоретические частоты 1 3 7 22 88 69 7 2 1.

Литература

Основная:

1. Основы высшей математики и математической статистики. М.:Геотар,2003.

1’. Математика. М.:Геотар,2013.

Дополнительная:

2. Кремер Н.Ш. Теория вероятностей и математическая статистика. М.: ЮНИТИ, 2010.
3. Лисьев В.П. Теория вероятностей и математическая статистика. М. 2006.
4. Гмурман В. Е.Теория вероятностей и математическая статистика. — М.: Высшее образование, 2008. — 480 с.
5. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. М.: Высшая школа, 1979.
6. Вентцель Е. С. Теория вероятностей. — М.: КноРус, 2010. — 480 с.

Приложение 1.

Функция Лапласа $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{t^2}{2}}dt$

x	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0,0	0,0000	0040	0080	0120	0159	0199	0239	02799	0319	0359
0,1	0398	0438	0477	0517	0557	0596	0636	0675	0714	0753
0,2	0793	0832	0870	0909	0948	0987	1026	1064	1103	1141
0,3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0,4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0,5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0.6	2257	2291	2324	2356	2389	2421	2454	2486	2517	2549
0,7	2580	2611	2642	2673	2703	2731	2764	2793	2823	2852
0,8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0,9	3159	3186	3212	3238	3264	3289	3315	3240	3365	3389
1,0	3413	3437	3461	3485	3508	3531	3554	3577	3599	3621
1,1	3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1,2	3849	3869	3888	3906	3925	3943	3962	3980	3997	4014
1,3	4032	4049	4066	4082	4099	4145	4131	4147	4162	4177
1,4	4192	4207	4222	4236	4251	4265	4277	4292	4306	4319
1,5	4332	4345	4357	4370	4382	4394	4406	4418	4429	4441
1,6	4452	4463	4474	4484	4595	4505	4515	4525	4535	4545

1,7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1,8	4641	4648	4656	4664	4671	4678	4686	4693	4699	4706
1,9	4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2,0	4772	4778	4783	4788	4793	4798	4803	4807	4812	4817
2,1	4861	4826	4830	4834	4838	4842	4846	4850	4854	4857
2,2	4861	4864	4868	4871	4874	4873	4881	4884	4887	4890
2,3	4893	4896	4898	4901	4904	4906	4909	4911	4913	4916
2,4	4918	4920	4922	4924	4927	4929	4930	4932	4934	4936
2,5	4938	4940	4941	4943	4945	4946	4948	4949	4957	4952
2,6	4953	4955	4956	4957	4958	4960	4961	4962	4963	4964
2,7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2,8	4974	4975	4976	4977	4977	4978	4979	4979	4980	4981
2,9	4281	4982	4982	4983	4984	4984	4985	4985	4986	4986
3,0	4986		3,3	4995		4,0	49997		5,0	499999
3,1	4990		3,4	4997		4,5	499997			
3,2	4993		3,5	4998						

Пример:$\Phi(1{,}32) = 0{,}4066$, $z_{0,9} = \left\{ x : \Phi(x) = \frac{0{,}9}{2} \right\} \approx 1{,}645$.

Приложение 2

Критические точки распределения Стьюдента при уровне значимости α в случае двусторонней критической области или α/2 в случае односторонней критической области (число степеней свободы *k*).

α / ***k***	**0,10**	**0,05**	**0,025**	**0,02**	**0,01**	**0,005**	**0,003**	**0,002**	**0,001**
1	6,314	12,706	25,452	31,821	63,657	127,3	212,2	318,3	636,6
2	2,920	4,303	6,205	6,965	9,925	14,089	18,216	22,327	31,600
3	2,353	3,182	4,177	4,541	5,841	7,453	8,891	10,214	12,922
4	2,132	2,776	3,495	3,747	4,604	5,597	6,435	7,173	8,610
5	2,015	2,571	3,163	3,365	4,032	4,773	5,376	5,893	6,869
6	1,943	2,447	2,969	3,143	3,707	4,317	4,800	5,208	5,959
7	1,895	2,365	2,841	2,998	3,499	4,029	4,442	4,785	5,408
8	1,860	2,306	2,752	2,896	3,355	3,833	4,199	4,501	5,041
9	1,833	2,262	2,685	2,821	3,250	3,690	4,024	4,297	4,781
10	1,812	2,228	2,634	2,764	3,169	3,581	3,892	4,144	4,587
11	1,796	2,20	2,6	2,72	3,11	3,5	3,8	4,03	4,44
12	1,782	2,179	2,560	2,681	3,055	3,428	3,706	3.930	4,318
13	1,771	2,16		2,65	3,01			3,85	4,22
14	1,761	2,145	2,510	2,624	2,977	3,326	3,583	3,787	4,140
15	1,753	2,13		2,60	2,95			3,73	4,07
16	1,746	2,120	2,473	2,583	2,921	3,252	3,494	3,686	4,015
17	1,740	2,11		2,57	2,90			3,65	3,96
18	1,734	2,101	2,445	2,552	2,878	3,193	3,428	3,610	3,922
19	1,729	2,09		2,54	2,86			3,58	3,88
20	1,725	2,086	2,423	2,528	2,845	3,153	3,376	3,552	3,849
21	1,721	2,08		2,52	2,83			3,53	3,82
22	1,717	2,074	2,405	2,508	2,819	3,119	3,335	3,505	3,792
23	1,714	2,07		2,50	2,81			3,49	3,77
24	1,711	2,064	2,391	2,492	2,797	3,092	3,302	3,467	3,745
25	1,708	2,06		2,49	2,79			3,45	3,72
26	1,706	2,056	2,379	2,479	2,779	3,067	3,274	3,435	3,704

27	1,703	2,05		2,47	2,77			3,42	3,69
28	1,701	2,048	2,369	2,467	2,763	3,047	3,250	3,408	3,674
29	1,699	2,05		2,46	2,76			3,40	3,66
30	1,697	2,042	2,360	2,457	2,750	3,030	3,230	3,386	3,646
40	1,684	2,02		2,42	2,70			3,31	3,55
60	1,671	2,00		2,39	2,66			3,23	3,46
120	1,658	1,98		2,36	2,62			3,17	3,37
+ ∞	1,645	1,960	2,241	2,326	2,576	2,807	2,968	3,090	3,291

Приложение 3

Критические точки распределения Фишера–Снедекора при уровне значимости α в случае правосторонней критической области или α/2 в случае двусторонней критической области (k_1 — число степеней свободы большей дисперсии, k_2 — число степеней свободы меньшей дисперсии).

α = 0,01

k_1 / k_2	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	4052	4999	5403	5625	5764	5889	5928	5981	6022	6056
2	98,49	99,01	90,17	99,25	99,33	99,30	99,34	99,36	99,36	99,40
3	34,12	30,81	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,54
5	16,26	13,27	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05
6	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82
9	10,56	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85
11	9,86	7,20	6,22	5,67	5,32	5,07	4,88	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39	4,30
13	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,28	6,01	5,09	4,58	4,25	4,01	3,85	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,71	3,56	3,45	3,37

k_1 / k_2	**11**	**12**	**14**	**16**	**20**	**24**	**30**	**40**	**50**
1	6082	6106	6142	6169	6208	6234	6258	6286	6302
2	99,41	99,42	99,43	99,44	99,45	99,46	99,47	99,48	99,48
3	27,13	27,05	26,92	26,83	26,69	26,60	26,50	26,41	26,35
4	14,45	14,37	14,24	14,15	14,02	13,93	13,83	13,74	13,69
5	9,96	9,89	9,77	9,68	9,55	9,47	9,38	9,29	9,24
6	7,79	7,72	7,60	7,52	7,39	7,31	7,23	7,14	7,09
7	6,54	6,47	6,35	6,27	6,15	6,07	5,98	5,90	5,85
8	5,74	5,67	5,56	5,48	5,36	5,28	5,20	5,11	5,06
9	5,18	5,11	5,00	4,92	4,80	4,73	4,64	4,56	4,51

10	4,78	4,71	4,60	4,52	4,41	4.33	4,25	4,17	4,12
11	4,46	4,40	4,29	4,21	4,10	4,02	3,94	3,86	3,80
12	4,22	4,16	4,05	3,98	3,86	3,78	3,70	3,61	3,56
13	4,02	3,96	3,85	3,78	3,67	3,59	3,51	3,42	3,37
14	3,86	3,80	3,70	3,62	3,51	3,43	3,34	3,26	3,21
15	3,73	3,67	3,56	3,48	3,36	3,29	3,20	3,12	3,12
16	3,61	3,55	3,45	3,37	3,25	3,18	3,10	3,01	2,96
17	3,52	3,45	3,35	3,27	3,16	3,08	3,00	2,92	2,86
18	3,44	3,37	3,27	3,19	3,07	3,00	2,91	2,83	2,78
19	3,36	3,30	3,19	3,12	3,00	2,92	2,84	2,76	2,70
20	3,30	3,23	3,13	3,05	2,94	2,86	2,77	2,69	2,63

α=0,05

k_1 / k_2	1	2	3	4	5	6	7	8	9	10
1	161	200	216	225	230	234	237	239	241	242
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4.21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,55	2,48	2,43	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,52	2,45	2,40	2,35

k_1 / k_2	11	12	14	16	20	24	30	40	50
1	243	244	245	246	248	249	250	251	252
2	19,40	19,41	19,42	19,43	19,44	19,45	19,46	19,47	19,47
3	8,76	8,74	8,71	8,69	8,66	8,64	8,62	8,60	8,58
4	5,93	5,91	5,87	5,84	5,80	5,77	5,74	5,71	5,70
5	4,70	4,68	4,64	4,60	4,56	4,53	4,50	4,46	4,44
6	4,03	4,00	3,96	3,92	3,87	3,84	3,81	3,77	3,75
7	3,60	3,57	3,52	3,49	3,44	3,41	3,38	3,34	3,32
8	3,31	3,28	3,23	3,20	3,15	3,12	3,08	3,05	3,03
9	3,10	3,07	3,02	2,98	2,93	2,90	2,86	2,82	2,80
10	2,94	2,91	2,86	2,82	2,77	2,74	2,70	2,67	2,64
11	2,82	2,79	2,74	2,70	2,65	2,61	2,57	2,53	2,50
12	2,72	2,69	2,64	2,60	2,54	2,50	2,46	2,42	2,40
13	2,63	2,60	2,55	2,51	2,46	2,42	2,38	2,34	2,32
14	2,56	2,53	2,48	2,44	2,39	2,35	2,31	2,27	2,24
15	2,51	2,48	2,43	2,39	2,33	2,29	2,25	2,21	2,18
16	2,45	2,42	2,37	2,33	2,28	2,24	2,20	2,16	2,13
17	2,41	2,38	2,33	2,29	2,23	2,19	2,15	2,11	2,08
18	2,37	2,34	2,29	2,25	2,19	2,15	2,11	2.07	2,04
19	2,34	2,31	2,26	2,21	2,151	2,11	2,07	2,02	2,00
20	2,31	2,28	2,23	2,18	2,12	2,08	2,04	1,99	1,96

Приложение 4

Правые критические точки распределения χ^2 (уровень значимости α, число степеней свободы k).

α / k_1	**0,99**	**0,98**	**0,95**	**0,90**	**0,80**	**0,70**	**0,50**	**0,30**	**0,20**
1	0,00016	0,0006	0,0039	0,016	0,064	0,148	0,455	1,07	1,64
2	0,020	0,040	0,103	0,211	0,446	0,713	1,386	2,41	3,22
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366	3,66	4,64
4	0,30	0,43	0,71	1,06	1,65	2,19	3,36	4,9	6,0
5	0,55	0,75	1,14	1,61	2,34	3,00	4,35	6,1	7,3
6	0,87	1,13	1,63	2,20	3,07	3,83	5,35	7,2	8,6
7	1,24	1,56	2,17	2,83	3,82	4,67	6,35	8,4	9,8
8	1,65	2,03	2,73	3,49	4,59	5,53	7,34	9,5	11,0
9	2,09	2,53	3,32	4,17	5,38	6,39	8,34	10,7	12,2
10	2,56	3,06	3,94	4,86	6,18	7,27	9,34	11,8	13,4
11	3,1	3,6	4,6	5,6	7,0	8,1	10,3	12,9	14,6
12	3,6	4,2	5,2	6,3	7,8	9,0	11,3	14,0	15,8
13	4,1	4,8	5,9	7,0	8,6	9,9	12,3	15,1	17,0
14	4,7	5,4	6,6	7,8	9,5	10,8	13,3	16,2	18,2
15	5,2	6,0	7,3	8,5	10,3	11,7	14,3	17,3	19,3
16	5,8	6,6	8,0	9,3	11,2	12,6	15,3	18,4	20,5
17	6.4	7,3	8,7	10,1	12,0	13,5	16,3	19,5	21,6
18	7,0	7,9	9,4	10,9	12,9	14,4	17,3	20,6	22,8
19	7,6	8,6	10,1	11,7	13,7	15,4	18,3	21,7	23,9
20	8,3	9,2	10,9	12,4	14,6	16,3	19,3	22,8	25,0
21	8,9	9,9	11,6	13,2	15,4	17,2	20,3	23,9	26,2
22	9,5	10,6	12,3	14,0	16,3	18,1	21,3	24,9	27,3
23	10,2	11,3	13,1	14,8	17,2	19,0	22,3	26,0	28,4
24	10,9	12,0	13,8	15,7	18,1	19,9	23,3	27,1	29,6
25	11,5	12,7	14,6	16,5	18,9	20,9	24,3	28,1	30,7
26	12,2	13,4	15,4	17,3	19,8	21,8	25,3	29,3	31,8
27	12,9	14,1	16,2	18,1	20,7	22,7	26,3	30,3	32,9
28	13,6	14,8	16,9	18,9	21,6	23,6	27,3	31,4	34,0
29	14,3	15,6	17,7	19,8	22,5	24,6	28,3	32,5	35,1
30	15,0	16,3	18,5	20,6	23,4	25,5	29,3	33,5	36,3

α / k_1	**0,10**	**0,05**	**0,02**	**0,01**	**0,005**	**0,002**	**0,001**
1	2,7	3,8	5,4	6,6	7,9	9,5	10,83
2	4,6	6,0	7,8	9,2	11,6	12,4	13,8
3	6,3	7,8	9,8	11,3	12,8	14,8	16,3
4	7,8	9,5	11,7	13,3	14,9	16,9	18,5
5	9,2	11,1	13,4	15,1	16,3	18,9	20,5
6	10,6	12,6	15,0	16,8	18,6	20,7	22,5
7	12,0	14,1	16,6	18,5	20,3	22,6	24,3
8	13,4	15,5	18,2	20,1	21,9	24,3	26,1
9	14,7	16,9	19,7	21,7	23,6	26,1	27,9
10	16,0	18,3	21,2	23,2	25,2	27,7	29,6
11	17,3	19,7	22,6	24,7	26,8	29,4	31,3
12	18,5	21,0	24,1	26,2	28,3	31	32,9
13	19,8	22,4	25,5	27,7	29,8	32,5	34,5
14	21,1	23,7	26,9	29,1	31	34	36,1
15	22,3	25,0	28,3	30,6	32,5	35,5	37,7
16	23,5	26,3	29,6	32,0	34	37	39,2

17	24,8	27,6	31,0	33,4	35,5	38,5	40,8
18	26,0	28,9	32,3	34,8	37	40	42,3
19	27,2	30,1	33,7	36,2	38,5	41,5	43,8
20	28,4	31,4	35,0	37,6	40	43	45,3
21	29,6	32,7	36,3	38,9	41,5	44,5	46,8
22	30,8	33,9	37,7	40,3	42,5	46	48
23	32,0	35,2	39,0	41,6	44,0	47,5	49,7
24	33,2	36,4	40,3	43,0	45,5	48,5	51,2
25	34,4	37,7	41,6	44,3	47	50	52,6
26	35,6	38,9	42,9	45,6	48	51,5	54,1
27	36,7	40,1	44,1	47,0	49,5	53	55,5
28	37,9	41,3	45,4	48,3	51	54,6	56,9
29	39,1	42,6	46,7	49,6	52,5	56	58,3
30	40,3	43,8	48,0	50,9	54	57,5	59,7