Федеральное агентство по образованию
Государственное образовательное учреждение
высшего профессионального образования
УФИМСКИЙ ГОСУДАРСТВЕННЫЙ АВИАЦИОННЫЙ
ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
Кафедра автоматизированных систем управления

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

к курсовому проектированию

по дисциплине «Статистика»

Уфа 2005

Составитель В.Ю. Арьков

УДК 311 (07)
ББК 60.6 (я7)

Методические указания к курсовому проектированию по дисциплине «Статистика» / Уфимск. гос. авиац. техн. ун-т; Сост. В.Ю.Арьков.– Уфа, 2005.– 37 с.

Приводится методика выполнения курсового проекта по дисциплине «Статистика». Даются указания по оформлению пояснительной записки. В приложениях приведены дополнительные справочные сведения.

Предназначены для студентов 2 курса специальности 080801 «Прикладная информатика (в экономике)».

Табл. 12 . Ил. 13. Библиогр.: 7 назв.

Рецензенты: канд. техн. наук Л.М.Бакусов
канд. техн. наук Р.В.Насыров

СОДЕРЖАНИЕ

ВВЕДЕНИЕ.....4
ИСХОДНЫЕ ДАННЫЕ 5
ЗАДАНИЕ И МЕТОДИКА РЕШЕНИЯ.....5
ОБЩИЕ РЕКОМЕНДАЦИИ.....31
ОФОРМЛЕНИЕ РЕЗУЛЬТАТОВ 31
ПРОЦЕДУРА ЗАЩИТЫ.....32
КОНТРОЛЬНЫЕ ВОПРОСЫ.....33
СПИСОК ЛИТЕРАТУРЫ 33
ПРИЛОЖЕНИЯ.....34

ВВЕДЕНИЕ

Общая теория статистики включает два крупных раздела: описательную статистику и аналитическую статистику. Методы описательной статистики охватывают сбор информации, группировку данных, построение статистических таблиц и графиков, а также вычисление основных статистических показателей. К аналитической статистике обычно относят такие темы, как выборка, вариация, корреляция, регрессия, динамика, структура и индексы.

Задачи в составе курсового проекта по статистике позволяют освоить базовые статистические методы. Для использования этих методов требуется понимание соответствующих разделов общей теории статистики. Кроме того, здесь используются результаты соответствующих разделов теории вероятностей и математической статистики.

Расчеты в процессе курсового проектирования выполняются с помощью калькулятора, графики строятся на миллиметровой бумаге. Выполнение всех операций «вручную» позволяет получить глубокое понимание статистических методов и освоить практические навыки статистической обработки данных. Такое понимание статистических «инструментов» позволяет в дальнейшем осмысленно и осознанно использовать любые компьютерные статистические программы и легко выявлять грубые ошибки.

ИСХОДНЫЕ ДАННЫЕ

В качестве номера варианта используется номер зачетной книжки студента. Текстовый файл с исходными данными содержит пять столбцов целых чисел (табл.1).

Т а б л и ц а 1

Описание набора исходных данных

№ столбца	Переменная	Описание
1	N	Номер элемента выборки
2	X	Значения признака x_i
3	Y	Значения признака y_i
4	Z	Значения признака z_i
5	G	Уровни ряда динамики g_t

Пример файла данных приводится в Приложении 1.

ЗАДАНИЕ И МЕТОДИКА РЕШЕНИЯ

Выполните статистическую обработку исходных данных, решая следующие задачи. При решении некоторых задач используются результаты решения предыдущих задач.

Задача 1

Вычислите показатели вариации по каждой из выборок *X*, *Y*, *Z*:

- среднее арифметическое;
- моду;
- медиану;
- размах вариации;
- дисперсию;
- стандартное отклонение;
- среднее линейное отклонение;
- коэффициенты осцилляции и вариации.

Методика решения

Показатели вариации вычисляются следующим образом.

Среднее значение – средняя арифметическая простая:

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

где n – объем выборки.

Мода – значение признака, встречающееся чаще всего. Для нахождения моды необходимо расположить все исходные данные в порядке возрастания. Повторяющиеся значения записывают столько раз, сколько они попадаются в исходном массиве. Затем нужно выбрать значение с максимальной частотой:

$$Mo = \arg\max_{x_i} n_i .$$

Медиана – центральное значение вариационного ряда. Используется построенный ранее ряд значений признака, отсортированных по величине. Если объем выборки нечетный, берем центральное значение; если объем выборки четный, берем среднее арифметическое двух центральных значений:

$$Me = \begin{cases} x_{\frac{n+1}{2}}, & \text{если } n - \text{нечетное} \\ \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{если } n - \text{нечетное} \end{cases}.$$

Размах вариации – разность максимального и минимального значений:

$$R = x_{\max} - x_{\min}.$$

Дисперсия – средний квадрат отклонения от среднего значения:

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

Стандартное отклонение – квадратный корень из дисперсии:

$$\sigma = \sqrt{D}.$$

Среднее линейное отклонение – средний модуль отклонения от среднего значения:

$$\bar{d} = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n}.$$

Относительные показатели вариации вычисляют как отношение абсолютного показателя к среднему значению, выраженное в процентах.

Коэффициент осцилляции:

$$V_R = \frac{R}{\bar{x}} \cdot 100\%.$$

Линейный коэффициент вариации:

$$V_d = \frac{\bar{d}}{\bar{x}} \cdot 100\%.$$

Коэффициент вариации:

$$V_\sigma = \frac{\sigma}{\bar{x}} \cdot 100\%.$$

Выборка считается *однородной*, если $V < 30$ %.

Для вычислений заполняется вспомогательная таблица (табл.2).

Т а б л и ц а 2

Расчет показателей вариации

№	x_i	$\lvert x_i - \bar{x} \rvert$	$(x_i - \bar{x})^2$
Σ			

Задача 2

По каждой из выборок *X*, *Y*, *Z*:

- проведите группировку данных по интервалам равной длины;
- составьте вариационный ряд;
- вычислите относительные частоты и накопленные частости;
- постройте полигон, гистограмму и кумуляту;
- нанесите на график кумуляты график накопленных частот без группировки.

Методика решения

Вариационный ряд – это значения признака (или интервалы значений) и их частоты. Вариационный ряд позволяет по фактическим данным оценить форму закона распределения.

При группировке данных вначале выбирают число интервалов группирования и границы интервалов. Интервалы должны полностью охватывать все значения признака в изучаемой выборке. Желательно

выбрать интервалы равной длины с «круглыми» границами. Например, если минимальное значение равно 22, а максимальное значение составляет 57, то можно выбрать следующие интервалы: (20 .. 30), (30 .. 40), (40 .. 50) и (50 .. 60).

Ориентировочное число интервалов можно определить по формуле Стерджесса:

$$k = 1 + 3{,}32 \cdot \lg n,$$

где k – число групп;

n – объем выборки (число единиц совокупности).

Поскольку группировка данных нужна для изучения формы распределения, используются следующие соображения. С одной стороны, число интервалов должно быть *достаточно* большим, чтобы изучить форму распределения. С другой стороны, число интервалов не должно быть *слишком* большим – тогда в каждый диапазон попадет несколько единиц совокупности. Желательно избегать получения малочисленных или «пустых» групп.

После формирования групп подсчитывают *абсолютные частоты* – число «попаданий» признака в каждый интервал, т.е. число объектов в каждой группе. Каждая единица совокупности x_i учитывается только один раз. Если значение оказывается на границе интервала, его относят к «левому» интервалу и не учитывают в «правом» интервале. Таким образом, интервалы выглядят следующим образом: [20 .. 30], (30 .. 40], (40 .. 50] и [50 .. 60] (табл.3). Здесь квадратные скобки означают включение границы в интервал; круглые скобки – игнорирование граничного значения.

Т а б л и ц а 3

Группировка данных

x_i	n_i	n_i, %	K_i, %
20 .. 30			
30 .. 40			
40 .. 50			
50 .. 60			100
Σ		100	–

Частости – относительные частоты, выраженные в процентах:

$$n_i(\%) = \frac{n_i}{n} \cdot 100\%.$$

Накопленные (кумулятивные) частости:

$$K_i = \sum_{j=1}^{i} n_j .$$

Каждая накопленная частость – сумма текущей частости и всех предыдущих.

Для упрощения можно складывать текущую частость и предыдущую кумулятивную частоту:

$$K_i = n_i + K_{i-1},$$

При этом считаем, что кумулята начинается с нуля: $K_0 = 0$.

Алгоритм расчета кумуляты представлен на рис.1.

Если расчеты выполнены без ошибок, сумма частостей и последняя накопленная частота будут равны 100 %.

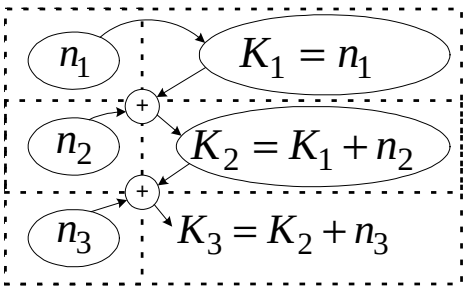


Рис. 1. Вычисление кумуляты

Графическое изображение вариационного ряда представляет собой эмпирическую оценку формы теоретического распределения. Гистограмма и полигон соответствуют плотности вероятности, а кумулята – функции распределения.

Гистограмма – столбиковая диаграмма частот. Основание каждого прямоугольника соответствует интервалу группировки. Высота столбика – частость.

Полигон частот – изображение вариационного ряда с помощью ломаной линии. Для построения полигона достаточно соединить отрезками прямых линий верхние стороны прямоугольников (рис.2).

Кумулята – изображение накопленных частостей, обычно в виде ломаной линии. По существу, кумулята – это интеграл от гистограммы (рис.3).

График накопленных частот может быть построен и без группировки данных. Для этого выборку упорядочивают по

возрастанию; каждое новое значение признака прибавляет к накопленной частоте величину:

$$\Delta n = \frac{1}{n} \cdot 100\% .$$

В этом случае приращение графика кумуляты происходит скачком в каждой точке x_i. Если встречается несколько одинаковых значений x_i, то величина приращения Δn умножается на число одинаковых элементов. График начинается с нуля и растет до 100 %.

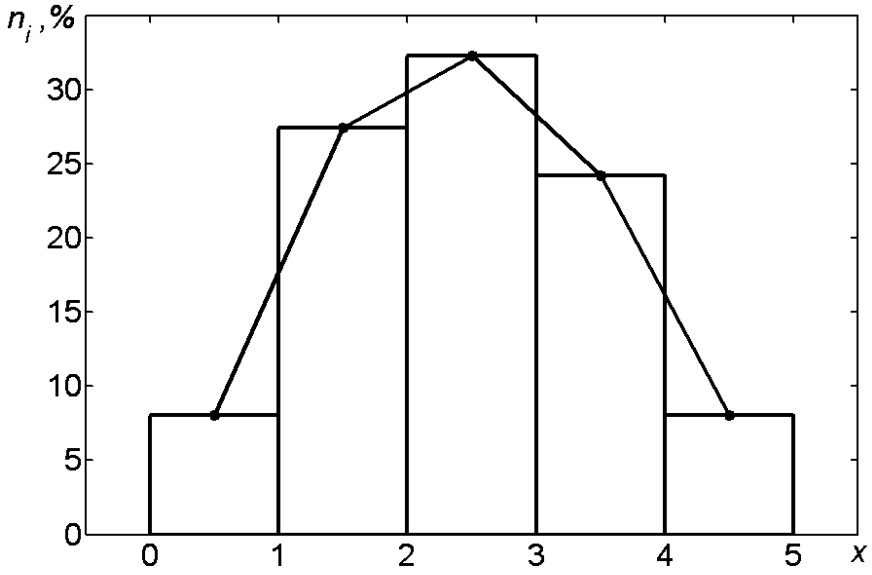


Рис. 2. Гистограмма и полигон

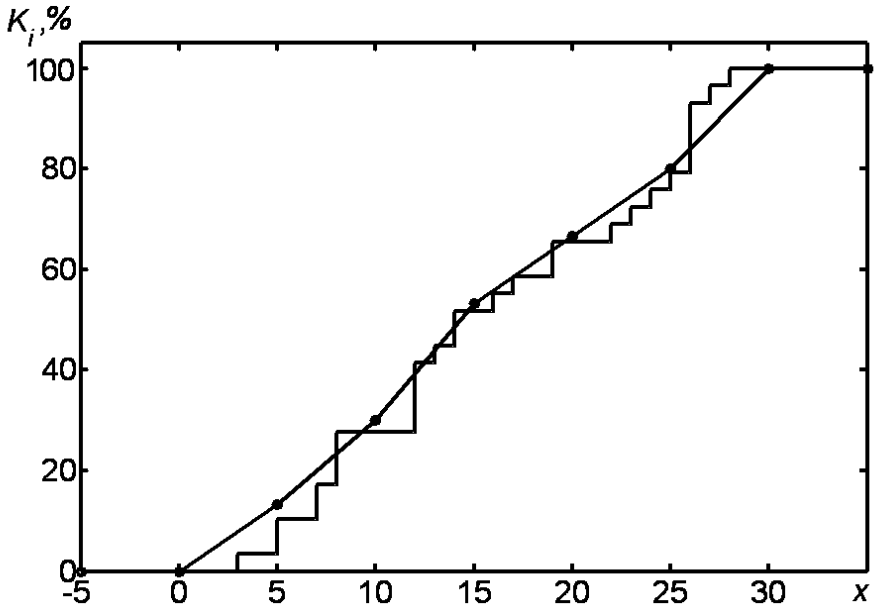


Рис. 3. Кумулята

Задача 3

По сгруппированным данным и графикам определите:

- среднее арифметическое;
- моду;
- медиану.

Сравните результаты с решением Задачи 1.

Методика решения

При расчетах по группированным данным учитывается относительная частота появления каждого варианта (табл.4).

Среднее значение – средняя арифметическая взвешенная:

$$\bar{x} = \frac{\sum_{i=1}^{k} (\bar{x}_i \cdot n_i)}{\sum_{i=1}^{k} n_i}.$$

Таблица 4

Расчет среднего значения

x_i	$\bar{x}_i$	n_i	$\bar{x}_i \cdot n_i$
10 .. 20	15		
…			
50 .. 60	55		
Σ	–		

Здесь в качестве среднего значения по интервалам $\bar{x}_i$ можно приближенно принять центр интервала группирования.

Мода – это координата основания самого высокого столбика гистограммы, т.е. *модального интервала*. Распределение может иметь несколько мод. В качестве оценки моды используют не середину модального интервала, а скорректированное значение (рис.4).

Медиана определяется как 50 %-й квантиль (рис.5). Это аргумент функции распределения, при котором вероятность равна 0,5 = 50 %:

$$Me: \quad F(Me) = 0{,}5 = 50\%;$$

$$Me = F^{-1}(0{,}5).$$

Медиана делит упорядоченную совокупность пополам.

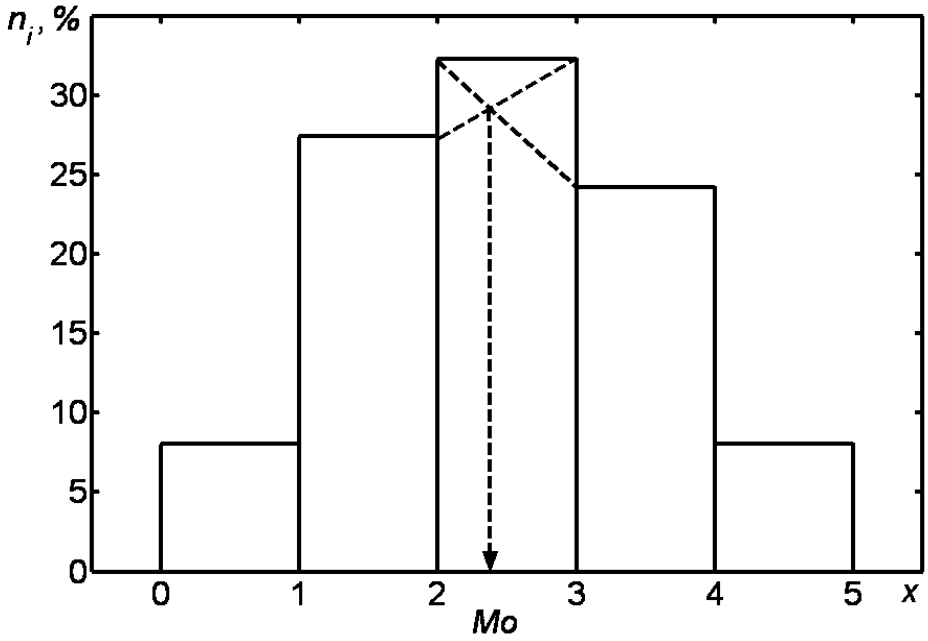


Рис. 4. Графическое определение моды

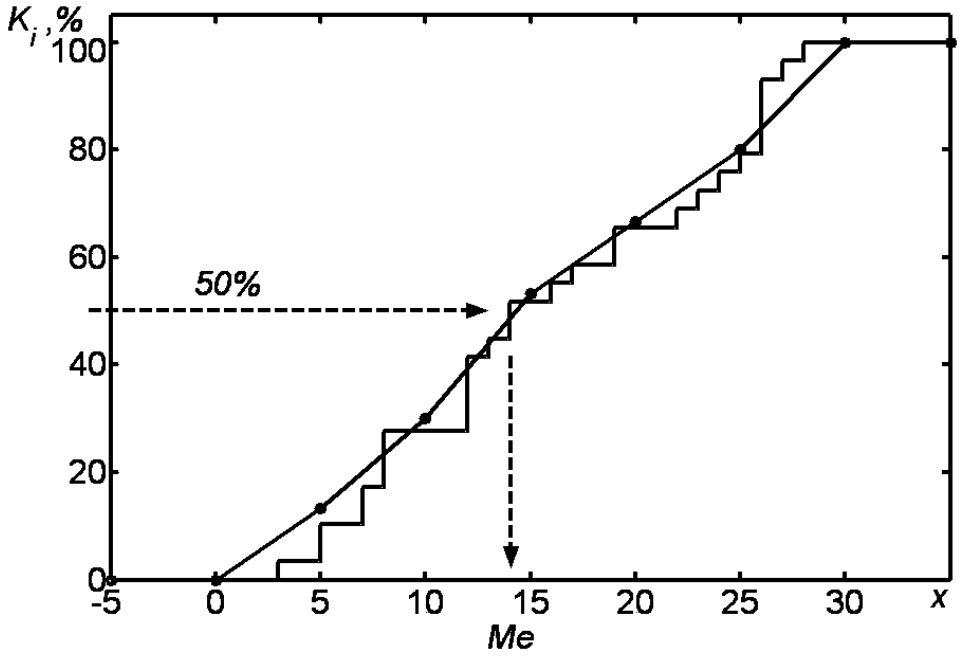


Рис. 5. Графическое определение медианы

Оценки моды и медианы, полученные по результатам группировки могут отличаться от оценок показателей, полученных без группировки. Группировка данных – это обобщение, укрупнение, при котором могут теряться отдельные мелкие подробности, но зато становится видна «картина в целом».

Задача 4

Постройте корреляционное поле. Проведите группировку Y и Z, используя X как группировочный признак. Вычислите условные средние $\overline{Y}_x$, $\overline{Z}_x$. Нанесите линию эмпирической регрессии на корреляционное поле.

Методика решения

Корреляционное поле (поле корреляции, диаграмма рассеяния) – это графическое изображение исходных данных. Каждый изучаемый объект (единица совокупности) характеризуется парой значений $\{x_i, y_i\}$ и изображается точкой. Поскольку объекты в составе выборочной совокупности считаются независимыми, точки на графике *не соединяют* линиями. График должен быть достаточно большим, чтобы его можно было легко анализировать. Масштаб по координатным осям выбирается так, чтобы эффективно использовать все доступное место на графике. Для этого определяют минимальное и максимальное значение для X и Y, затем округляют их в меньшую и большую сторону соответственно – до ближайшего «стандартного» круглого числа.

Группировка данных – это деление совокупности на группы единиц по какому-либо признаку. Этот признак называют *группировочным*. Группировка позволяет оценить характер зависимости между переменными путем вычисления условного среднего.

Условное среднее значение – это среднее значение одного признака при условии, что другой признак принимает заранее заданное фиксированное значение:

$$\overline{y}_X = \mu(y \mid x = X) \approx \frac{\sum_{i=1}^{k} y_i}{k} \quad \forall i: \quad x_i = X .$$

При вычислении условных средних значений подсчитывают средние $\overline{y}_X$ и $\overline{z}_X$ для каждой группы единиц в зависимости от значения группировочного признака x (табл.5).

Например, при изучении предприятий их можно сгруппировать по численности работников и найти средний доход для каждого вида предприятий. В этом случае численность работников X выступает в роли группировочного признака. При расчете среднего дохода по

каждой группе предприятий $\bar{y}_X$ используются значения признака *Y* только для тех единиц совокупности, которые относятся к выбранной группе по *X*.

Таблица 5

Условные средние

x_i	$\bar{x}_i$	n_i	$\sum y_j$	$\bar{y}_x$	$\sum z_j$	$\bar{z}_x$
10 .. 20						
…						
50 .. 60						

На поле корреляции наносят линию условного среднего (*эмпирической регрессии*). Для этого наносят точки с координатами $\{\bar{x}_i, \bar{y}_{x_i}\}$ и соединяют их отрезками прямых линий (рис.6).

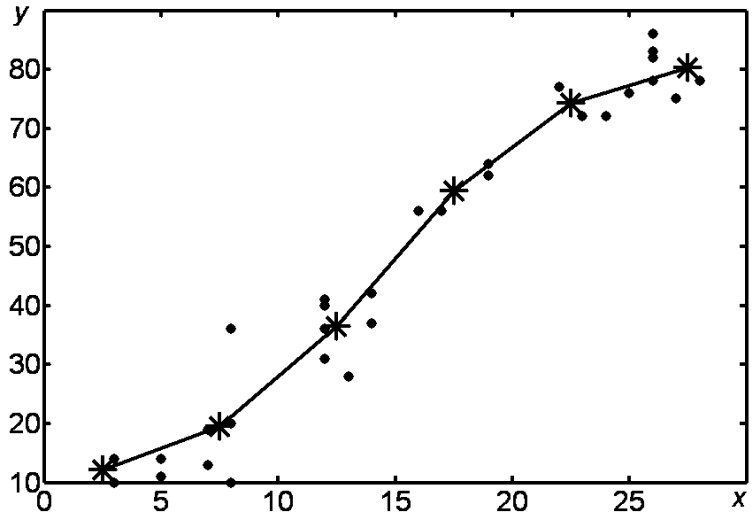


Рис. 6. Поле корреляции и эмпирическая регрессия

По внешнему виду графика можно выявить возможный характер взаимосвязи между признаками и оценить погрешность построения уравнения регрессии.

Задача 5

Найдите предельную ошибку выборки *X*, *Y*, *Z*; постройте доверительные интервалы для среднего, дисперсии и стандартного отклонения генеральной совокупности при доверительной вероятности *p* = 68 %; 95 %; 99,7 %.

Методика решения

Выборочное среднее значение $\bar{x}$, вычисляемое по выборке ограниченного объема n, будет отличаться от идеального «точного» значения μ_x, которое можно было бы получить для бесконечно большой выборки. Разница между выборочным средним и математическим ожиданием (генеральным средним) называется ошибкой выборки:

$$\Delta = \left|\bar{x} - \mu_x\right|.$$

Ошибка выборочного наблюдения пропорциональна стандартному отклонению и обратно пропорциональна квадратному корню из объема выборки:

$$\Delta = t \cdot \sigma_{\bar{x}} = t \cdot \frac{\sigma}{\sqrt{n}}.$$

Стандартное отклонение выборочного среднего составляет:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Коэффициент доверия t находят по распределению Стьюдента (табл.6) с учетом объема выборки и заданного значения доверительной вероятности:

$$t = t\left(\frac{1+p}{2}, n\right).$$

Таблица 6

Процентные точки распределения Стьюдента $t(n, p)$ и нормального распределения $z = \Phi^{-1}(p)$

	p							
n	0,5	0,6	0,7	0,8	0,9	0,95	0,99	0,999
5	0	0,2672	0,5594	0,9195	1,4759	2,0150	3,3649	5,8934
10	0	0,2602	0,5415	0,8791	1,3722	1,8125	2,7638	4,1437
20	0	0,2567	0,5329	0,8600	1,3253	1,7247	2,5280	3,5518
30	0	0,2556	0,5300	0,8538	1,3104	1,6973	2,4573	3,3852
40	0	0,2550	0,5286	0,8507	1,3031	1,6839	2,4233	3,3069
50	0	0,2547	0,5278	0,8489	1,2987	1,6759	2,4033	3,2614
100	0	0,2540	0,5261	0,8452	1,2901	1,6602	2,3642	3,1737
1000	0	0,2534	0,5246	0,8420	1,2824	1,6464	2,3301	3,0984
z	0	0,2533	0,5244	0,8416	1,2816	1,6449	2,3263	3,0902

Для симметричного распределения достаточно определить одно значение квантиля, например, для верхней границы доверительного интервала. Противоположная граница симметрична относительно точки {0; 0,5}:

$$t\left(\frac{1-p}{2},n\right)=-t\left(\frac{1+p}{2},n\right).$$

Доверительный интервал для генерального среднего:

$$\bar{x}-t\cdot\frac{\sigma}{\sqrt{n}}\le\mu_x\le\bar{x}+t\cdot\frac{\sigma}{\sqrt{n}}.$$

Значение t определяет, «сколько сигм» нужно взять для построения доверительного интервала. Принцип построения доверительного интервала проиллюстрирован на рис.7.

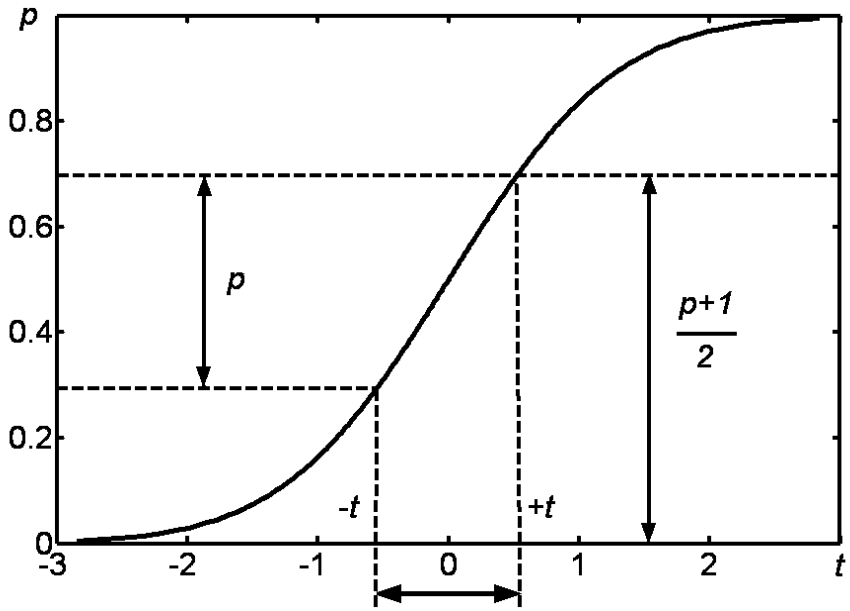


Рис. 7. Двусторонний доверительный интервал

При использовании табулированного распределения приходится проводить интерполяцию, т.е. находить приближенное значение функции между известными точками (рис.8).

Исходные данные для интерполяции:

$t_1(p_1)$ и $t_2(p_2)$.

Требуется найти значение t между точками p_1 и p_2:

$t(p)$: $p_1 < p < p_2$.

Искомое значение t для заданного p находим по формуле:

$$t = t_1 + \frac{t_2 - t_1}{p_2 - p_1} \cdot (p - p_1).$$

Интерполяция может проводиться дважды – по p, затем по n.

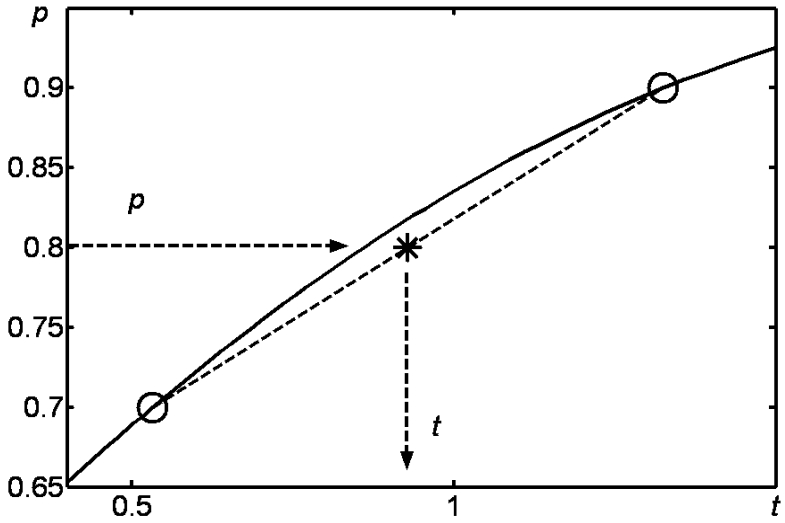


Рис. 8. Линейная интерполяция

Доверительный интервал для генеральной дисперсии:

$$\frac{(n-1)s^2}{\chi^2_{\frac{1+p}{2}}} \le \sigma^2 \le \frac{(n-1)s^2}{\chi^2_{\frac{1-p}{2}}},$$

где s^2 – выборочная дисперсия;

χ^2_p – квантиль распределения Пирсона (табл.7).

Доверительный интервал для генерального с.к.о.:

$$s\sqrt{\frac{n-1}{\chi^2_{\frac{1+p}{2}}}} \le \sigma \le s\sqrt{\frac{n-1}{\chi^2_{\frac{1-p}{2}}}}.$$

Задача 6

Постройте доверительные интервалы для генерального среднего μ_x, μ_y и μ_z при доверительной вероятности p = 68 %; 95 %; 99,7 % упрощенным способом: «одна/две/три сигмы».

Методика решения

Величину коэффициента t можно приближенно выбрать для стандартных значений вероятности (табл.8). Для этого используются

процентные точки стандартной функции нормального распределения с нулевым средним и единичной дисперсией:

$z \sim N(0;1)$.

Таблица 7

Процентные точки распределения χ^2_n

	n						
p	5	10	20	30	40	50	100
0,001	0,2102	1,478	5,921	11,58	17,91	24,67	61,92
0,010	0,5543	2,558	8,260	14,95	22,16	29,70	70,07
0,050	1,145	3,940	10,85	18,49	26,50	34,76	77,93
0,100	1,610	4,865	12,44	20,59	29,05	37,68	82,36
0,200	2,342	6,179	14,57	23,36	32,34	41,44	87,95
0,300	2,999	7,267	16,26	25,50	34,87	44,31	92,13
0,400	3,655	8,295	17,80	27,44	37,13	46,86	95,81
0,500	4,351	9,341	19,33	29,33	39,33	49,33	99,33
0,600	5,131	10,47	20,95	31,31	41,62	51,89	102,9
0,700	6,064	11,78	22,77	33,53	44,16	54,72	106,9
0,800	7,289	13,44	25,03	36,25	47,26	58,16	111,7
0,900	9,236	15,98	28,41	40,25	51,80	63,16	118,5
0,950	11,07	18,30	31,41	43,77	55,75	67,50	124,3
0,990	15,08	23,20	37,56	50,89	63,69	76,15	135,8
0,999	20,51	29,58	45,31	59,70	73,40	86,66	149,4

Форма распределения Стьюдента приближается к нормальному распределению при большом объеме выборки, начиная с нескольких десятков единиц. При этом погрешность от замены распределения Стьюдента нормальным не превышает единиц процентов.

Таблица 8

Стандартные квантили нормального распределения

Вероятность (округленно)	Вероятность	Коэффициент доверия	Ошибка выборки	Доверительный интервал
68 %	0,682689	1,000	одна сигма	$\mu = \bar{x} \pm \sigma$
95 %	0,954500	2,000	две сигмы	$\mu = \bar{x} \pm 2\sigma$
99,7 %	0,997300	3,000	три сигмы	$\mu = \bar{x} \pm 3\sigma$

Таким образом, получаем *приближенные* границы доверительных интервалов:

$$p = 68\%: \quad \mu = \bar{x} \pm \sigma_{\bar{x}}; \quad \bar{x} - \sigma_{\bar{x}} \le \mu \le \bar{x} + \sigma_{\bar{x}}$$
$$p = 95\%: \quad \mu = \bar{x} \pm 2\sigma_{\bar{x}}; \quad \bar{x} - 2\sigma_{\bar{x}} \le \mu \le \bar{x} + 2\sigma_{\bar{x}} \quad .$$
$$p = 99{,}7\%: \quad \mu = \bar{x} \pm 3\sigma_{\bar{x}}; \quad \bar{x} - 3\sigma_{\bar{x}} \le \mu \le \bar{x} + 3\sigma_{\bar{x}}$$

Задача 7

При уровне значимости α = 32 %; 5 %; 0,3 % проверьте гипотезы:

- $\sigma_x^2 = \sigma_y^2$;
- $\mu_x = \bar{x} + 5$;
- $\mu_x = \mu_y$.

Методика решения

Проверка статистических гипотез основана на использовании *стандартных распределений*. Изучаемый статистический показатель преобразуется к случайной величине с известным стандартным законом распределения. Затем задается вероятность, по которой находят квантиль.

Вероятность принятия ошибочного решения при проверке гипотез называют *уровнем значимости*:

$\alpha = 1 - p$.

Уровень значимости определяет, в каком проценте случаев возможна ошибка, если принять изучаемую гипотезу.

Можно считать, что *область принятия гипотезы* соответствует доверительному интервалу, а за его пределами находится *критическая область*. Для двустороннего интервала уровень значимости делится поровну между критическим областями.

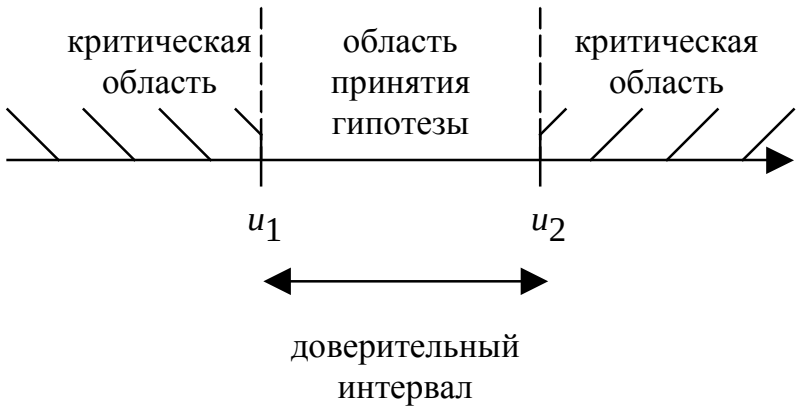


Рис. 9. Проверка статистических гипотез

Если фактическая статистика оказывается в критической области, например $|t_{\text{ф}}| > t_{\text{кр}}$, то гипотезу отвергают. Если фактическая статистика оказывается в области принятия гипотезы, то гипотезу принимают при заданном уровне значимости.

Сравнение дисперсий – проверка гипотезы о том, можно ли считать сравниваемые выборочные дисперсии s_x^2 и s_y^2 оценками одной и той же генеральной дисперсии. Используется распределение Фишера. При заданном уровне значимости α должно выполняться следующее неравенство:

$$F_{n_1,n_2}\left(\frac{\alpha}{2}\right) \le \frac{s_x^2}{s_y^2} \le F_{n_1,n_2}\left(1-\frac{\alpha}{2}\right)$$

Распределение Фишера обладает своеобразной «симметрией»:

$$F_{n_1,n_2}(1-p) = \frac{1}{F_{n_1,n_2}(p)}.$$

Поэтому в табл. 9 приводится только верхняя половина распределения.

Т а б л и ц а 9

Процентные точки распределения Фишера $F_{n_1,n_2}(p)$

для выборок равного объема: $n_1 = n_2$

	p							
n	0,5	0,6	0,7	0,8	0,9	0,95	0,99	0,999
5	1,0	1,2692	1,6410	2,2275	3,4530	5,0503	10,967	29,752
10	1,0	1,1787	1,4061	1,7316	2,3226	2,9782	4,8491	8,7539
20	1,0	1,1216	1,2684	1,4656	1,7938	2,1242	2,9377	4,2900
30	1,0	1,0978	1,2132	1,3641	1,6065	1,8409	2,3860	3,2171
40	1,0	1,0840	1,1817	1,3076	1,5056	1,6928	2,1142	2,7268
50	1,0	1,0747	1,1608	1,2706	1,4409	1,5995	1,9490	2,4413

Сравнение средних – проверка гипотезы о равенстве генеральных средних:

H_0: $\mu_1 = \mu_2$ или $\mu_1 - \mu_2 = 0$.

Гипотеза о равенстве средних отвергается, если фактическая статистика больше табличной:

$|t_{\text{факт}}| > t_{\text{табл}}$.

При изучении выборочных оценок используется распределение Стьюдента с числом степеней свободы $df = n_1 + n_2 - 2$:

$$|\bar{x} - \bar{y}| > t_{n_1+n_2-2}\left(1-\frac{\alpha}{2}\right)\cdot\sqrt{\frac{s_x^2}{n_1}+\frac{s_y^2}{n_2}}.$$

Проверка *гипотезы о среднем* значении:

$H_0: \quad \mu_1 = A.$

Рассматриваем случайную величину:

$$(\bar{x} - A) \sim t_{n-1}\frac{s}{\sqrt{n}}.$$

Для проверки гипотезы вычисляется фактическая *t*-статистика:

$$t_{\text{ф}} = t_{n-1}\left(1-\frac{\alpha}{2}\right) = \frac{|\bar{x} - A|}{s/\sqrt{n}}.$$

Гипотезу отвергаем, если фактическая статистика больше, чем табличное (критическое, теоретическое) значение:

$t_{\text{ф}} > t_{\text{кр}}.$

Задача 8

Вычислите линейные коэффициенты корреляции r_{yx} и r_{zx}. Сделайте вывод о тесноте линейной связи между признаками.

Методика решения

Линейный коэффициент корреляции вычисляется следующим образом:

$$r_{yx} = \frac{n\sum yx - \sum y\sum x}{\sqrt{(n\Sigma x^2 - (\Sigma x)^2)(n\Sigma y^2 - (\Sigma y)^2)}} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_y \cdot \sigma_x}.$$

Коэффициент корреляции принимает значение в диапазоне:

$-1 \le r \le +1.$

Знаки коэффициента корреляции и коэффициента регрессии совпадают. Знак говорит о направлении зависимости – положительной (прямой) или отрицательной (обратной).

Величина коэффициента корреляции говорит о тесноте связи:

- $|r| = 1$ – функциональная связь, все точки лежат на прямой линии;

- $0<|r|<1$ – линейная зависимость на фоне случайных отклонений;
- $|r|<0{,}3$ – слабая, несущественная линейная зависимость;
- $|r|>0{,}7$ – существенная линейная зависимость;
- $r=0$ – линейная взаимосвязь отсутствует, либо взаимосвязь не заметна на фоне случайных отклонений, либо связь есть, но она существенно нелинейна.

Задача 9

Вычислите коэффициенты корреляции рангов Спирмена и Кендалла $Y(X)$ и $Z(X)$. Сделайте вывод о тесноте связи.

Методика решения

Ранговый коэффициент корреляции характеризует общий характер нелинейной зависимости: возрастание или убывание результативного признака при возрастании факторного. Это показатель тесноты монотонной нелинейной связи.

Коэффициент корреляции рангов Спирмена:

$$\rho_{xy}=1-\frac{6\cdot\sum d^2}{n\cdot(n^2-1)},\quad -1\le\rho\le+1.$$

где $d_i=R(x_i)-R(y_i)$ – разность рангов X и Y;

n – число наблюдений (число пар, число разностей рангов);

6 – число шесть (не путать с сигмой!).

Для вычисления ранговых коэффициентов исходные данные ранжируют, т.е. расставляют по порядку возрастания или убывания, а затем нумеруют (табл.10). Ранг – это порядковый номер. Если встречаются два одинаковых значения, им присваивают одинаковое значение ранга, равное среднему арифметическому рангов этих значений.

Таблица 10

Пример ранжирования данных

R	x	y	z
1	10	-240	25
2	11	-230	28
…	…	…	…
n	78	278	41

Затем рассматриваются пары значений $\{x_i, y_i\}$ исходных данных, каждое значение со своим рангом. В табл.11 столбцы *x* и *y* соответствуют исходным, неупорядоченным данным.

Т а б л и ц а 11

Коэффициент корреляции рангов Спирмена

x	y	R_x	R_y	d	d^2

Ранговый коэффициент корреляции Кендалла:

$$\tau = \frac{2S}{n(n-1)},$$

где n – число наблюдений;

$S = P - Q$ – разность *сумм* числа последовательностей и инверсий результативного признака.

В процессе вычислений факторный признак *X* упорядочивается по возрастанию с присвоением порядкового номера (*ранжируется*). Затем результативный признак *Y* упорядочивается по возрастанию факторного признака *X*.

Число последовательностей P для каждого ранга *Y* – это число следующих рангов, превышающих эту величину.

Число инверсий Q для каждого ранга *Y* – это число следующих рангов, меньших выбранного.

Задача 10

Постройте уравнения регрессии *Y*(*X*), *Z*(*X*) графическим способом.

Методика решения

При построении линии регрессии на корреляционном поле проводят линию регрессии с помощью линейки, «на глаз» – по местам «сгущения» точек. Отдельные точки, далеко отстоящие от «облака рассеяния» (аномальные данные), игнорируют (рис.10).

На линии регрессии выбирают две точки, ближе к краям диапазона значений. Составляем систему уравнений – два уравнения с двумя неизвестными:

$$\begin{cases} y_1 = a + b \cdot x_1 \\ y_2 = a + b \cdot x_2 \end{cases}.$$

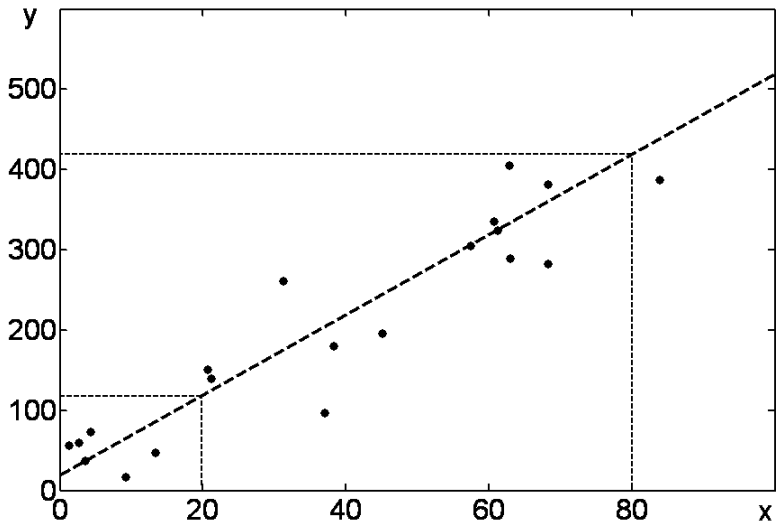


Рис. 10. Графическое построение уравнение регрессии

Решая систему, получаем оценки неизвестных коэффициентов $\hat{a}$ и $\hat{b}$. Затем записываем уравнение регрессии, подставляя найденные коэффициенты:

$$y = \hat{a} + \hat{b} \cdot x.$$

Задача 11

С помощью метода наименьших квадратов (МНК) постройте уравнения регрессии *Y*(*X*), *X*(*Y*), *Z*(*X*), *X*(*Z*). Нанесите линии регрессии на корреляционное поле.

Методика решения

Построение парной линейной регрессии по МНК сводится к решению системы нормальных уравнений. Например, для уравнения

$$y = a + b \cdot x$$

нужно решить следующую систему:

$$\begin{cases} an + b\Sigma x = \Sigma y \\ a\Sigma x + b\Sigma x^2 = \Sigma xy \end{cases}.$$

Решая систему, получаем оценки неизвестных коэффициентов $\hat{a}$ и $\hat{b}$. Затем записываем уравнение регрессии, подставляя найденные коэффициенты:

$$y = \hat{a} + \hat{b} \cdot x .$$

Задача 12

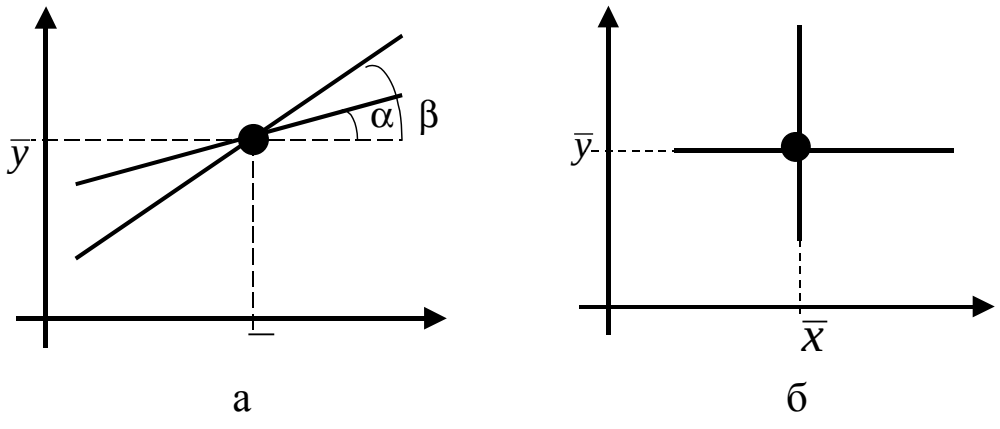
После определения коэффициентов корреляции и построения уравнения регрессии разными способами провести сравнение полученных оценок и построенных графиков.

Методика решения

Если вычисления сделаны правильно, то результаты, полученные разными способами, должны совпадать с некоторым приемлемым уровнем погрешности.

При отсутствии взаимосвязи линия регрессии проходит горизонтально: при изменении значения фактора результат в среднем остается постоянным (рис.11).

Геометрический смысл коэффициента корреляции: r показывает, насколько различается наклон двух линий регрессии: $y(x)$ и $x(y)$, насколько сильно различаются результаты минимизации отклонений по x и по y. Эти линии пересекаются в точке $\{\bar{x}, \bar{y}\}$.



а б

Рис. 11. Линии регрессии $y(x)$ и $x(y)$: (а) $r \to 1$; (б) $r \to 0$

Если линии совпадают, то коэффициент корреляции близок к 1. Чем больше угол между линиями, то тем больше r (рис.12).

Коэффициент линейной корреляции можно приблизительно оценить по виду диаграммы рассеяния. *Знак* коэффициента корреляции совпадает со знаком коэффициента регрессии и

определяет наклон линии регрессии, т.е. общую направленность зависимости (возрастание или убывание). *Абсолютная величина* коэффициента корреляции определяется степенью близости точек к линии регрессии.

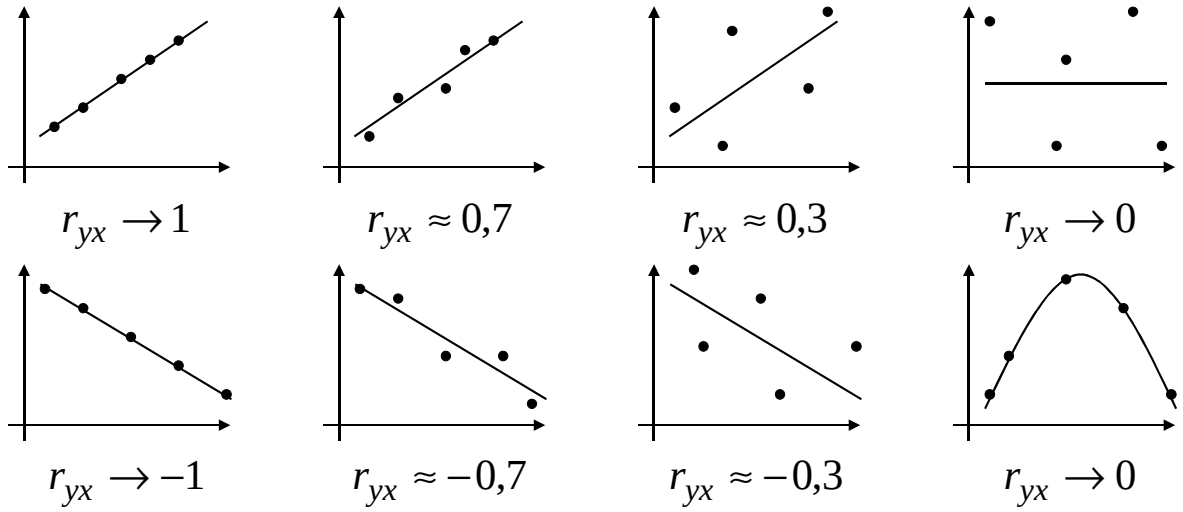


Рис. 12. Степень линейной корреляции на диаграмме рассеяния

Задача 13

Проведите сглаживание ряда динамики G_t с помощью простой и взвешенной скользящей средней, а также скользящей медианы по трем, пяти и двенадцати точкам. В качестве номера месяца t используется столбец N. Постройте графики исходного ряда динамики (ИРД) и сглаженных рядов следующим образом.

Для нечетных вариантов:

1) ИРД, ССП(3), ССП(5), ССП(12);
2) ИРД, ССВ(3), ССВ(5), ССВ(12);
3) ИРД, СМ(3), СМ(5), СМ(12).

Для четных вариантов:

1) ИРД, ССП(3), ССВ(3), СМ(3);
2) ИРД, ССП(5), ССВ(5), СМ(5);
3) ИРД, ССП(12), ССВ(12), СМ(12).

Сравните результаты сглаживания.

Методика решения

Ряд динамики – последовательность значений показателя во времени, в хронологическом порядке. Отдельное значение ряда называется *уровнем*.

Обычно выделяют три *компонента* ряда динамики: тренд, сезонные колебания и случайную составляющую. Для выделения основной тенденции используют скользящую среднюю или скользящую медиану.

Скользящая средняя – среднее значение за интервал фиксированной длины, причем начало интервала сдвигается («скользит») по времени. Определяют среднее значение по нескольким точкам и «привязывают» его к середине интервала. Затем начало интервала сдвигают на один шаг по времени вправо.

Если длина интервала сглаживания – четное число, то сглаженное значение относят к целому отсчету по времени, ближайшему к середине интервала. Таким образом, среднее по 12 уровням можно отнести к моменту времени 6 или 7, считая от начала интервала сглаживания.

Усреднение на небольшом интервале удаляет случайную составляющую. Усреднение на интервале 12 месяцев удаляет и случайную составляющую, и сезонные колебания.

Простая скользящая средняя – пример расчета по трем точкам:

$$\bar{y}_t = \frac{y_{t-1} + y_t + y_{t+1}}{3}.$$

Скользящая средняя взвешенная – пример расчета по трем точкам:

$$\bar{y}_t = \frac{y_{t-1} + 2y_t + y_{t+1}}{4} = \frac{\sum w \cdot y_t}{\sum w}.$$

В качестве весов w_k часто используют биномиальные коэффициенты (рис.13). В этой «пирамиде» каждое число является суммой двух верхних «соседей». Каждая строчка содержит набор весовых коэффициентов.

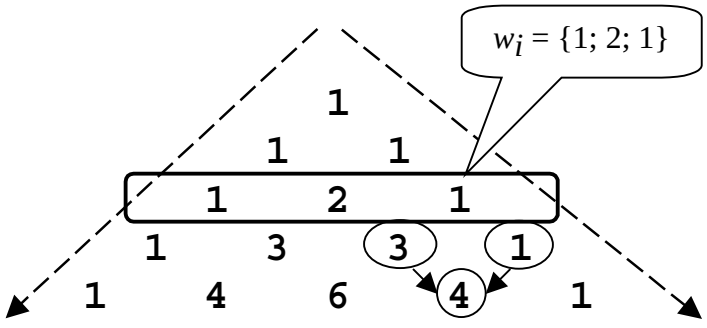


Рис. 13. Биномиальные весовые коэффициенты

Скользящая медиана вычисляется аналогично скользящей средней – по интервалу, который «скользит» по оси времени. Скользящая медиана менее чувствительна к аномальным наблюдениям, чем скользящая средняя.

При вычислении скользящей медианы и скользящей средней теряется несколько начальных и конечных уровней ряда. Поэтому сглаженный ряд оказывается короче, чем исходный

Т а б л и ц а 12

Сглаживание ряда динамики

t	G_t	СС(3)	СС(5)	СС(12)	…
1	10	–	–	–	
2	20	15	–	–	
3	15	21,7		–	
4	30			–	
…					

При сравнении результатов сглаживания следует обратить внимание на изменение свойств сглаженного ряда с изменением интервала сглаживания, а также на результаты работы разных методов сглаживания с одинаковым интервалом усреднения.

Задача 14

Вычислите показатели динамики для ряда *G*:

- средний уровень ряда динамики;
- абсолютный прирост;
- темп (коэффициент) роста;
- темп прироста;
- средний абсолютный прирост;
- средний темп (коэффициент) роста;
- средний темп прироста.

Методика решения

Показатели динамики характеризуют изменение уровня со временем. Сравнение производится в форме разности (*абсолютные* показатели динамики) или отношения (*относительные* показатели динамики). Сравнение текущего уровня с предыдущим дает *цепные*

показатели, сравнение с начальным (базисным) уровнем дает *базисные* показатели.

Средний уровень интервального ряда динамики вычисляется как средняя арифметическая простая:

$$\bar{y} = \frac{\sum_{t=1}^{n} y_t}{n}.$$

Абсолютный прирост:

$\Delta y_t = y_t - y_{t-1}$ или $\Delta y_t = y_t - y_1$.

Коэффициент роста (если отношение больше единицы):

$$K_{\text{р}} = \frac{y_t}{y_{t-1}} \text{ или } K_{\text{р}} = \frac{y_t}{y_1}.$$

Темп роста (если отношение меньше единицы):

$$T_{\text{р}} = \frac{y_t}{y_{t-1}} \cdot 100\% \text{ или } T_{\text{р}} = \frac{y_t}{y_1} \cdot 100\%$$

Темп прироста:

$$T_{\text{пр}} = \frac{y_i - y_{i-1}}{y_{i-1}} \cdot 100\%.$$

Средние показатели вычисляют, как правило, по цепным показателям динамики.

Средний абсолютный прирост:

$$\overline{\Delta y} = \frac{\sum \Delta y_t}{n-1}.$$

Средний коэффициент роста:

$$K_{\text{р}} = \sqrt[n-1]{\frac{y_n}{y_1}}$$

Средний темп роста:

$$\overline{T}_{\text{р}} = \sqrt[n-1]{\frac{y_n}{y_1}} \cdot 100\%$$

Средний темп прироста:

$\overline{T}_{\text{пр}} = \overline{T}_{\text{р}} - 100$.

Задача 15

Постройте уравнение тренда с помощью МНК двумя способами и нанесите линию тренда на график исходного ряда динамики. Определите величину остаточной дисперсии.

Методика решения

Уравнение тренда строят методами регрессионного анализа. Линейный тренд описывается с помощью линейного уравнения относительно времени:

$y(t) = a + b \cdot t$.

Первый способ. Составляется система нормальных уравнений по МНК:

$$\begin{cases} an + b\Sigma t = \Sigma y \\ a\Sigma t + b\Sigma t^2 = \Sigma yt \end{cases}.$$

Второй способ. Вычисления упрощаются, если сместить начало координат в середину ряда:

$t^* = t - \bar{t}$: $\{\ldots; -3; -2; -1; 0; 1; 2; 3; \ldots\}$.

В результате сумма значений нового аргумента t^* равна нулю, а система нормальных уравнений принимает вид:

$$\begin{cases} a^* n = \Sigma y \\ b\Sigma t^{*2} = \Sigma yt^* \end{cases}.$$

При смещении начала координат меняется только значение свободного члена. Поэтому после проведения оценивания следует вычислить исходное значение коэффициента a.

Остаточная дисперсия – мера точности аппроксимации, т.е. приближения линии регрессии к исходным данным:

$$D_{\text{ост}} = \frac{\Sigma\left(y(t) - y_t\right)^2}{n - p - 1}$$

где $y(t)$ – прогноз по уравнению тренда

y_t – исходный ряд динамики;

n – количество точек;

p – число коэффициентов уравнения регрессии.

ОБЩИЕ РЕКОМЕНДАЦИИ

Число значащих разрядов при записи результатов вычислений должно быть в разумных пределах – не слишком мало, не слишком много. При грубом округлении теряется полезная информация, при выписывании двадцати разрядов теряется наглядность. Кроме того, если исходные данные имеют два значащих разряда, то результат вычислений вряд ли будет содержать десять достоверных значащих цифр.

«*Прикиньте*» *ожидаемый результат*, сделайте примерную, приблизительную оценку перед выполнением вычислений. Например, определяя двусторонний квантиль распределения Стьюдента, можно заранее оценить его как одну, две или три «сигмы» нормального распределения. Если задана доверительная вероятность $p = 95\ \%$, то квантиль стандартного нормального распределения составит $u = 2$, т.е. две сигмы. В этом случае квантиль t-распределения не может быть 0,47. Это грубая ошибка, которую нужно исправить.

Составление пояснительной записки – это упражнение по оформлению грамотной инженерной документации. Текст должен быть понятен не только составителю, но и читателю. Документ должен содержать описание процесса вычислений со всеми промежуточными результатами, рассуждениями, обоснованиями, заключениями. Результаты вычислений и выводы сопровождаются пояснениями.

ОФОРМЛЕНИЕ РЕЗУЛЬТАТОВ

Пояснительная записка оформляется аккуратно, без помарок и зачеркиваний. Исправления вносят путем закрашивания корректирующей жидкостью. Текст пишется разборчиво, высота букв – достаточного размера (не менее 3 мм).

Бумага – стандартного размера 210*297 мм (формат А4). Страницы нумеруются. Листы скрепляют по левому краю. Отступы – по ГОСТ или СТП УГАТУ.

Титульный лист отчета должен содержать всю информацию, необходимую для однозначной идентификации работы и ее автора, как показано в Приложении 2. Для этого на титульном листе указывают вид документа, название дисциплины, вариант задания, номер группы, фамилии и инициалы студента, должность, фамилию и

инициалы преподавателя и т.п. (в соответствии со стандартом УГАТУ на оформление текстовых документов).

Графики строят на миллиметровой бумаге, во весь лист. Графический образ занимает все поле рисунка. Графики кумуляты по исходным и сгрупированным данным строят на одном графике. Линии регрессии наносят на корреляционное поле. При этом линии регрессии строят только в пределах имеющихся исходных данных. В этом случае по виду графика можно будет судить о качестве построенного уравнения регрессии, т.е. насколько «близко» к точкам проходит линия регрессии.

Таблицы не должны иметь пустых клеток. Если клетка не подлежит заполнению, в этой ячейке ставится прочерк. Каждая таблица сопровождается заголовком.

Заголовки должны быть составлены для всех рисунков и таблиц. Если таблица продолжается на следующей странице, то для второй части таблицы делается соответствующий заголовок, например «Продолжение табл. 7».

Если на графике имеется несколько линий, нужно обозначить каждый элемент графика.

ПРОЦЕДУРА ЗАЩИТЫ

В результате выполнения курсового проекта приобретается понимание теории статистики и способность практического применения статистических методов.

При защите курсовой работы проверяется и оценивается следующее.

- корректность математических выкладок и правильность вычислений;
- аккуратность оформления текста, таблиц и рисунков;
- соответствие оформления требованиям ГОСТ и СТП;
- грамотное написание технических терминов, орфография и грамматика русского языка;
- понимание использованных методов и терминов;
- способность быстро оценить ожидаемый результат вычислений;
- способность интерпретировать полученные результаты.

КОНТРОЛЬНЫЕ ВОПРОСЫ

- Что означает … (любой термин из работы)?
- Какие статистические показатели обозначаются как μ и σ?
- Как читаются греческие буквы, использованные в работе?
- Что такое медиана и как она вычисляется?
- Что такое квантиль; как найти квантиль по графику?
- Как связаны уровень значимости и доверительная вероятность?
- Как вычислялось значение … (любой показатель из работы)?
- Как вычислялись значения в этой таблице?
- Что изображено на этом рисунке?
- О чем говорит разный наклон линий регрессии $x(y)$ и $y(x)$?
- Что такое метод наименьших квадратов?
- Как выводится система нормальных уравнений?
- Как оценить коэффициент корреляции по диаграмме рассеяния?
- Что такое тренд и сезонные колебания?
- Что такое сглаживание?
- Что такое скользящая средняя?
- Что такое уравнение тренда?
- Можно ли считать данную выборку однородной?

СПИСОК ЛИТЕРАТУРЫ

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика в задачах и упражнениях.– М.: Юнити-Дана, 2001.– 270 с.
2. Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика.– М.: Юнити-Дана, 2001.– 656 с.
3. Гмурман В.Е. Теория вероятностей и математическая статистика.– М.: Высш. школа, 2004.–479 с.
4. ГОСТ 2.304-81. Шрифты чертежные.
5. Кремер Н.Ш. Теория вероятностей и математическая статистика.– М.: ЮНИТИ-ДАНА, 2004.–573 с.
6. Шмойлова Р.А. Практикум по теории статистики.– М.: Финансы и статистика, 2004.– 416 с.
7. Шмойлова Р.А. Теория статистики.– М.: Финансы и статистика, 2004.– 656 с.

Приложение 1

Пример исходных данных

*******d1020301.txt********

N	X	Y	Z	G
1	12	-41	-26	55
2	14	-37	14	54
3	8	-10	11	51
4	24	-72	-62	51
5	3	-10	-20	47
6	7	-19	-25	43
7	8	-20	38	43
8	19	-62	-51	42
9	19	-64	10	50
10	26	-78	-27	62
11	22	-77	-16	73
12	26	-83	-30	70
13	14	-42	31	75
14	26	-86	-90	70
15	5	-14	-23	69
16	13	-28	-46	61
17	8	-36	-88	57
18	27	-75	-86	55
19	12	-36	-40	49
20	12	-31	-64	63
21	25	-76	-48	62
22	17	-56	-10	73
23	12	-40	25	76
24	28	-78	-76	88
25	7	-13	-14	88
26	3	-14	-55	87
27	5	-11	-14	82
28	16	-56	-55	77
29	23	-72	-34	75
30	26	-82	-31	71

Приложение 2

Пример титульного листа

Федеральное агентство по образованию
УФИМСКИЙ ГОСУДАРСТВЕННЫЙ АВИАЦИОННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Кафедра АСУ

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
к курсовому проекту
по дисциплине «Статистика»
Вариант 250145

Выполнил:
студ. гр. ПИЭ–243
Суриков А.В.

Проверил:
проф. Арьков В.Ю.

Уфа 2005

Приложение 3

Греческий алфавит

Так называемая «инженерная грамотность» (результат высшего образования) включает умение правильно писать и называть греческие буквы в математических формулах. К защите необходимо знать написание всех букв и их названия по-русски – только для использования в математических формулах.

Т а б л и ц а П.3.1

Греческие буквы, используемые в математических формулах

Прописные	Строчные	Название
A	α	альфа
B	β	бета
Γ	γ	гамма
Δ	δ	дельта
E	ε	эпсилон
Z	ζ	дзета
H	η	эта
Θ	θ, ϑ	тета
I	ι	йота
K	κ , $\varkappa$	каппа
Λ	λ	лямбда
M	μ	мю
N	ν	ню
Ξ	ξ	кси
O	o	омикрон
Π	π	пи
P	ρ	ро
Σ	σ, ς	сигма
T	τ	тау
Υ	υ	ипсилон
Φ	ϕ, φ	фи
X	χ	хи
Ψ	ψ	пси
Ω	ω	омега