МИНИСТЕРСТВО ПУТЕЙ СООБЩЕНИЯ РФ
ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ПУТЕЙ СООБЩЕНИЯ



МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Методические указания по дисциплине "Математика" к выполнению контрольной работы для студентов заочного отделения всех специальностей

ИРКУТСК 2003

МИНИСТЕРСТВО ПУТЕЙ СООБЩЕНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

ИРКУТСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ПУТЕЙ СООБЩЕНИЯ

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

**Методические указания по дисциплине «Математика»
к выполнению контрольной работы
для студентов заочного отделения
всех специальностей**

Иркутск 2003

Содержат программу раздела «Математическая статистика», основные теоретические положения с примерами, образец выполнения контрольной работы и варианты индивидуальных заданий. Предполагается, что статистические расчёты могут выполняться студентами как вручную (на калькуляторе), так и на персональном компьютере. В связи с этим в пособии приводится руководство по решению задач математической статистики с помощью офисной программы EXCEL.

Приведены необходимые статистические таблицы. Пособие может быть использовано и на дневном отделении.

Ил. 7. Табл. 21. Библиогр: 5 назв.

Составитель канд. физ.-мат. наук, доц. Г.Д. Гефан.

Кафедра высшей математики

Рецензенты: д-р техн. наук, проф. Л.А. Астраханцев
(ИрГУПС),
канд. техн. наук, ст. преподаватель Н.А. **Рябченок**
(ИрГУПС)

1. Программа раздела «Математическая статистика»

1. Статистический метод исследований, его задачи и особенности. Генеральная и выборочная совокупности. Сплошной и выборочный методы наблюдения. Репрезентативность выборки. Случайные и систематические ошибки репрезентативности.
2. Статистическое распределение количественного признака. Варианты и частоты. Полигон и гистограмма.
3. Накопленные частоты. Эмпирическая функция распределения.
4. Выборочная и генеральная средние. Понятие оценки. Свойства оценок: несмещённость и состоятельность. Выборочная средняя как несмещённая оценка генеральной средней.
5. Характеристики вариации количественного признака: дисперсия и среднеквадратическое отклонение. Выборочная дисперсия — смещённая оценка генеральной дисперсии. Исправленная дисперсия, исправленное среднеквадратическое отклонение.
6. Выравнивание статистического ряда, эмпирические и теоретические частоты. Построение предполагаемого нормального распределения по данным наблюдений.
7. Интервальные оценки неизвестных параметров распределения. Понятия точности и надёжности оценки, доверительный интервал.
8. Интервальная оценка генеральной средней (математического ожидания) нормального распределения при известном генеральном среднеквадратическом отклонении.
9. Минимальный объём выборки, обеспечивающий заданную точность и надёжность интервальной оценки генеральной средней.
10. Интервальная оценка генеральной средней нормального распределения при неизвестном генеральном среднеквадратическом отклонении (малая выборка).
11. Понятие статистической гипотезы. Два рода ошибок, возникающих при проверке гипотез. Принципы проверки гипотез. Статистический критерий. Уровень значимости гипотезы.
12. Гипотеза о равенстве двух генеральных средних.
13. Гипотеза о виде распределения. Сравнение эмпирических и теоретических частот. Критерий согласия Пирсона. Проверка гипотезы о нормальном распределении признака в генеральной совокупности.
14. Понятие корреляции. Диаграмма рассеивания. Выборочный коэффициент линейной корреляции.
15. Проверка гипотезы о значимости коэффициента корреляции.
16. Понятие регрессии и регрессионного анализа. Метод наименьших квадратов. Выборочное уравнение линейной регрессии и его связь с коэффициентом корреляции.

2. Основные теоретические положения

(нумерация соответствует пунктам программы)

1. Применение *статистического метода* не предполагает глубокого индивидуального изучения объектов. Его сутью является систематизация, обработка, анализ и использование так называемых *статистических данных.* Последние представляют собой информацию о том, сколько объектов изучаемой совокупности обладают определёнными (качественными или количественными) признаками. Сами объекты при этом «обезличиваются». Целью статистических исследований являются научные и практические выводы об изучаемых явлениях и процессах.

Изучение совокупности однородных объектов относительно некоторого качественного или количественного признака начинается со сбора статистических данных. В зависимости от степени охвата объектов обследование совокупности (наблюдение) может быть *сплошным* или *выборочным*. При сплошном наблюдении производится регистрация значения признака для каждого объекта *генеральной* (т.е. полной) *совокупности*. При выборочном обследовании из генеральной совокупности выделяется *выборочная совокупность (выборка). Объемом выборки* называется число её объектов. Выборочный метод применяется тогда, когда проведение сплошного обследования нецелесообразно или вообще невозможно. Свойство выборки представлять характеристики генеральной совокупности называется *репрезентативностью* (по-русски — представительностью). Для того чтобы выборка была репрезентативной, желательно обеспечить случайность отбора объектов. Это означает, что включение каждого объекта генеральной совокупности в выборку должно быть равновероятным. Другим требованием является достаточно большой объём выборки — настолько большой, чтобы обеспечить нужную точность.

Статистическими ошибками обычно называют отклонения вычисленных показателей от истинных величин из-за недостаточной репрезентативности выборки. Они могут быть систематическими и случайными. *Систематические ошибки* связаны с неслучайным характером отбора. *Случайные ошибки* — следствие ограниченного объёма выборки.

2. Наблюдаемые значения количественного признака X, которые в дальнейшем обозначаются символом x_i, называются *вариантами*, а число появлений вариантов в выборке — выборочными *частотами* (n_i). *Статистическим распределением* выборки называется перечень вариантов и их частот (или относительных частот n_i / n). Если признак дискретный, то в качестве графика распределения строится *полигон частот* — ломаная линия, соединяющая точки (x_i, n_i).

Пример 1. Статистическое распределение числа баллов, набранных на трёх вступительных экзаменах, представлено ниже таблицей и графиком.

x_i (сумма баллов)	9	10	11	12	13	14	15	Объём совокупности $n = \sum n_i$
n_i (число студентов)	11	32	55	50	29	12	11	200

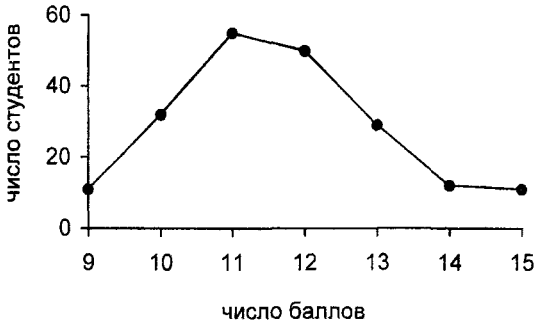


Рис. 1. Полигон частот

Полигон является статистическим аналогом многоугольника распределения дискретной случайной величины.

При большом числе вариантов или непрерывном характере признака вместо отдельных значений используются интервалы. Данные группируются, и определяется частота попадания значений количественного признака в каждый интервал. Графическим изображением статистического распределения в этом случае является *гистограмма частот* — ступенчатая фигура, каждый прямоугольник которой имеет в качестве основания частичный интервал, а в качестве высоты — соответствующую *плотность частоты* n_i / h, где h — длина интервала. Площадь отдельных столбиков численно равна соответствующим частотам, а площадь всей гистограммы равна объёму совокупности

$$S = \sum \frac{n_i}{h} h = n.$$

Иногда строится гистограмма относительных частот. В этом случае по оси ординат откладываются значения w_i / h. Площадь такой гистограммы будет равна единице.

Пример 2. Среднедушевой доход в 100 семьях составляет от 20 до 260 долларов. Если ввести 8 интервалов равной длины, то длина интервала составит $h = (260 - 20)/8 = 30$. Статистическое распределение выборки может быть представлено в виде таблицы или в виде графика:

Интервал, i	1	2	3	4	5	6	7	8
Начало интервала, $x_{i-\frac{1}{2}}$	20	50	80	110	140	170	200	230
Конец интервала, $x_{i+\frac{1}{2}}$	50	80	110	140	170	200	230	260
Середина интервала, x_i	35	65	95	125	155	185	215	245
Частота, n_i	6	7	23	20	18	14	5	7
Относительная частота, $w_i = n_i / n$	0.06	0 07	0 23	0.20	0.18	0 14	0 05	0.07

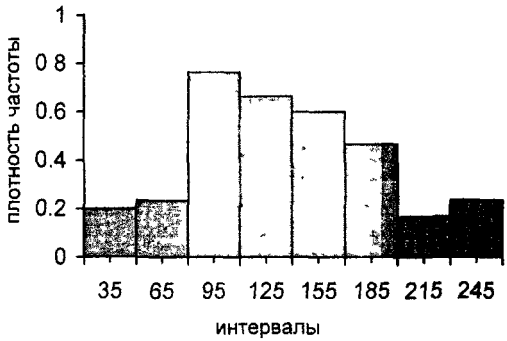


Рис. 2

Гистограмма является статистическим аналогом графика плотности вероятности непрерывной случайной величины. По виду полигона и гистограммы можно делать предположения о законе распределения количественного признака в генеральной совокупности.

3. Число наблюдений n_x, при которых значение количественного признака X оказалось меньше некоторого заданного значения x, называется *накопленной частотой*. Функция

$$F^*(x) = \frac{n_x}{n},$$

задаваемая для каждого значения x относительной накопленной частотой, называется *эмпирической функцией распределения*. В соответствии с теоремой Бернулли, $F^*(x)$ при $n \to \infty$ близка (сходится по вероятности) к теоретической функции распределения $F(x) = P(X < x)$.

Пример 3. (Продолжение примера 1). Выпишем ряд накопленных частот:

x (сумма баллов)	9	10	11	12	13	14	15	16
n_x (накопленная частота)	0	11	43	98	148	177	189	200

Соответствующая эмпирическая функция распределения имеет вид

$$F^*(x)=\begin{cases}0, & \text{если } x\le 9,\\ 0.055, & \text{если } 9<x\le 10,\\ 0.215, & \text{если } 10<x\le 11,\\ 0.49, & \text{если } 11<x\le 12,\\ 0.74, & \text{если } 12<x\le 13,\\ 0.865, & \text{если } 13<x\le 14,\\ 0.945, & \text{если } 14<x\le 15,\\ 1, & \text{если } x>15.\end{cases}$$

Таким образом, эмпирическая функция распределения есть ступенчатая функция, значения которой заключены в интервале от 0 до 1.

4. Если данные не сгруппированы, то *выборочная средняя* вычисляется по формуле

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

В том случае, когда определены частоты отдельных значений количественного признака, они выступают в качестве *весов*:

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

где k — число различных вариантов в выборке.

Пример 4. (Продолжение примера 1). Расчёт по последней формуле даёт следующий средний балл: $\bar{x}=11.67$.

Если данные сгруппированы в виде интервальных рядов, то в качестве вариантов обычно берут середины интервалов.

Пример 5. (Продолжение примера 2). Выборочная средняя для среднедушевого дохода: $\bar{x}=135.2$. Подлинный результат (по несгруппированным данным) был бы несколько иным.

Т.к. выборочная средняя принимает (в зависимости от выборки) разные значения, её можно рассматривать как случайную величину $\bar{X}$. Генеральная средняя есть математическое ожидание количественного признака (неслучайная величина, константа). Математическое ожидание выборочной средней рав-

но генеральной средней: $M(\overline{X}) = M(X)$. Практический смысл этого положения: мы имеем возможность *оценивать* генеральную среднюю по выборке значений признака, причём эта оценка не содержит систематической ошибки.

Величина Θ^* называется *несмещённой оценкой* величины Θ, если $M(\Theta^*) = \Theta$. Таким образом, выборочная средняя есть несмещённая оценка генеральной средней.

Согласно теореме Бернулли, выборочные относительные частоты отдельных значений признака (n_i / n) сколь угодно близки к генеральным относительным частотам при $n \to \infty$. Значит, и выборочная средняя с ростом объёма выборки сколь угодно близка (сходится по вероятности) к генеральной средней.

Оценка, сколь угодно близкая к точному значению оцениваемой величины для больших выборок, называется *состоятельной*. Выборочная средняя является состоятельной оценкой генеральной средней.

5. *Выборочной дисперсией* D и *выборочным среднеквадратическим отклонением* σ количественного признака X называются величины, определяемые формулами:

$$D = \sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{n} = \frac{\sum_{i=1}^{k} x_i^2 n_i}{n} - \overline{x}^2 = \overline{x^2} - \overline{x}^2.$$

Если данные не сгруппированы, то все $n_i \equiv 1$, а суммирование проводится от 1 до n. Можно показать, что выборочная дисперсия является *смещённой* оценкой *генеральной дисперсии* $D(X)$:

$$M(D) = \frac{n}{n-1} D(X).$$

Исправленной выборочной дисперсией называется величина

$$s^2 = \frac{n}{n-1} D,$$

являющаяся несмещённой оценкой генеральной дисперсии. Величина $s = \sqrt{s^2}$ называется исправленным среднеквадратическим отклонением. Считается, что при достаточно больших n (больше чем 30) выборочное и исправленное среднеквадратические отклонения близки друг к другу ($s \approx \sigma$) и различием между ними пренебрегают.

Пример 6. При выпуске или закупке швейных изделий необходимо учитывать распределение людей по размеру и по росту. Требуется найти оценки среднего значения, дисперсии и среднеквадратического отклонения роста женщин по выборочным данным. (Обследовано 50 человек; результаты сгруппиро-

ваны в интервалы длиной 4 см каждый, в таблице указаны середины интервалов.)

x_i	156	160	164	168	172	176	180
n_i	5	7	13	14	6	4	1

Расчёты дают: $\bar{x} = 166$; $D = 33.44$; $s^2 = \frac{50}{49} D = 34.12$; $\sigma = 5.783$; $s = 5.841$.

6. Изучая распределение выборки, можно сделать предположение об истинном законе распределения количественного признака. Подгонку выборочного распределения под предполагаемое генеральное распределение признака называют *выравниванием статистического ряда*.

Выборка, приведённая в примере 6, возможно, извлечена из *нормальной* генеральной совокупности. Действительно, распределение имеет почти симметричную форму с максимумом вблизи среднего значения признака.

Теоретическими частотами n_i' называются частоты, полученные в предположении о некотором законе распределения при заданном объёме выборки n:

$$n_i' = nP(x_{i-\frac{1}{2}} < X < x_{i+\frac{1}{2}}).$$

Наиболее часто в качестве предполагаемого закона распределения используется нормальный закон. Плотность нормального распределения описывается формулами:

$$f(x) = \frac{\varphi(z)}{\sigma}, \quad \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}, \quad z = \frac{x-a}{\sigma},$$

где $a = M(X)$, $\sigma = \sqrt{D(X)}$ — генеральные характеристики, оценками которых являются следующие величины: $a^* = \bar{x}$, $\sigma^* = s$. Если все частичные интервалы имеют одинаковую (и малую) длину h, то вероятность попадания в интервал i приблизительно равна $hf(x_i)$, что даёт следующую формулу для теоретических частот нормального распределения:

$$n_i' = nh\frac{\varphi(z_i)}{s}, \quad z_i = \frac{x_i - a^*}{s}.$$

Значения функции $\varphi(z)$ см. в приложении 1.

Пример 7. (Продолжение примера 6). Произведём выравнивание статистического ряда. Для $x_1 = 156$ получаем $z_1 = -1.71$, $\varphi(z) \approx 0.092$, $n_1' = 3.2$.

Аналогично вычисляются остальные теоретические частоты:

x_i	156	160	164	168	172	176	180
n_i'	3.2	8.1	12.9	12.9	8.1	3.2	0.7

7. Пусть Θ^* — найденная по результатам выборки точечная оценка неизвестного параметра Θ, с вероятностью γ удовлетворяющая условию:

$$|\Theta - \Theta^*| < \delta.$$

Тогда:

– полуширина δ симметричного относительно Θ^* интервала называется *точностью* оценки;

– вероятность γ называется *доверительной вероятностью* или *надёжностью* оценки;

– интервал $(\Theta^* - \delta, \Theta^* + \delta)$, который заключает в себе (покрывает) неизвестный параметр Θ с вероятностью γ, называют *доверительным интервалом*.

8. Пусть количественный признак X распределён в генеральной совокупности нормально с известным генеральным среднеквадратическим отклонением σ. Тогда доверительный интервал для оценки генеральной средней a по выборочной средней $\bar{x}$ с надёжностью γ при объёме выборки n ищется в виде

$$\bar{x} - \frac{t\sigma}{\sqrt{n}} < a < \bar{x} + \frac{t\sigma}{\sqrt{n}},$$

где величина t определяется по заданной надёжности из условия $\gamma/2 = \Phi(t)$;

$$\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-z^2/2} dz \ —$$

— функция Лапласа (См. приложение 2). В частности, для задаваемых обычно значений надёжности имеем (сравните с приложением 2):

γ	0.95	0.99	0.999
t	1.96	2.57	3.3

Пример 8. Сделана партия электроламп. Известно, что среднеквадратическое отклонение нормально распределённой продолжительности горения лампы равно 70 часов. По выборке определена средняя продолжительность горения — 548.53 часа. Требуется определить доверительный интервал для средней продолжительности горения.

Допустим, что объём выборки, по которой определена средняя продолжительность горения, равен 144. Зададим надёжность оценки $\gamma = 0.95$. Тогда $t = 1.96$, и полуширина доверительного интервала составит $\delta = \frac{1.96 \cdot 70}{12} = 11.43$. Таким образом, $537.1 < a < 559.96$.

Пусть объём выборки был равен 225. При той же надёжности получаем: $\delta = 9.15$, $539.38 < a < 557.68$.

Наконец, если при прежнем объёме выборки (144) задать $\gamma = 0.99$, то получим: $\delta = 14.99$, $533.54 < a < 563.52$.

Таким образом, при заданной надёжности с увеличением объёма выборки улучшается точность интервальной оценки. При постоянном объёме выборки с ростом требуемой надёжности растёт кратность ошибки и, следовательно, ухудшается точность интервальной оценки. Невозможно одновременно повышать надёжность и улучшать точность интервальной оценки, если не увеличивать объём выборки.

9. Пусть известно генеральное среднеквадратическое отклонение σ нормально распределённого признака, а также заданы требуемые точность δ и надёжность γ интервальной оценки генеральной средней a. Из формулы $\delta = t\sigma/\sqrt{n}$ следует, что *минимальный требуемый объем выборки* в этом случае равен

$$n_{\min} = \frac{t^2\sigma^2}{\delta^2}.$$

Пример 9. Известно, что для нормально распределённого контролируемого размера детали $\sigma = 0.2$. Какого объёма выборка требуется для того, чтобы с надёжностью $\gamma = 0.999$ определить средний размер деталей с точностью $\delta = 0.05$?

$$n \geq \frac{3.3^2 \cdot 0.2^2}{0.05^2} \approx 175.$$

Необоснованное наращивание объёма ведёт к дополнительным затратам времени и труда, а иногда и к материальным потерям. Поэтому в каждом конкретном случае приходится делать выбор между качеством оценки и затратами на неё. Например, если бы в примере 9 мы потребовали точности не 0.05, а 0.01, то необходимый объём выборки возрос бы в 25 раз!

10. Пусть генеральное среднеквадратическое отклонение σ неизвестно. При больших объёмах выборки ($n > 30$) для интервальной оценки генеральной средней пользуются значением выборочного среднеквадратического отклонения и функцией Лапласа (см. пункт 8). При *малых выборках* следует определять доверительный интервал по формуле

$$\bar{x} - \frac{t(\gamma, n)s}{\sqrt{n}} < a < \bar{x} + \frac{t(\gamma, n)s}{\sqrt{n}},$$

в которой генеральное среднеквадратическое отклонение заменено исправленным (s), а величина $t(\gamma, n)$, называемая коэффициентом Стьюдента, ищется в специальной таблице (приложение 3) по заданным n и γ.

Пример 10. Рыбак поймал 7 рыб. Эти рыбы весили (в килограммах): 0.65, 0.48, 0.96, 0.67, 1.24, 1.07, 0.87. С надёжностью 0.95 оценить средний вес рыб, попадающихся на крючок в этом водоёме. Предполагается, что имеет место нормальное распределение рыб по весу.

Обработка статистических данных даёт:

$$\bar{x} = 0.849;\ D = 0.060;\ s^2 = 0.070;\ s = 0.265.$$

При $t(0.95, 7) = 2.45$ точность составляет $\delta = 2.45 \times 0.265 / \sqrt{7} \approx 0.245$. Доверительный интервал для среднего веса рыбы $0.604 < a < 1.094$. Оценка имеет невысокую точность. Причин этому две: во-первых, слишком сильно рассеяны значения количественного признака, во-вторых, слишком мала выборка.

11. *Статистическими гипотезами* называются утверждения о виде или характеристиках распределений количественных признаков в генеральных совокупностях, выдвигаемые и проверяемые на основе обработки выборочных данных. Выдвинутая гипотеза H_0 называется нулевой или основной, а противоречащая ей H_1 — альтернативной или конкурирующей. При принятии или отклонении гипотезы возможны 4 различные ситуации. Проанализируем их, пользуясь понятием условной вероятности.

1. *Ошибка 1-го рода*: отклонена правильная гипотеза. Вероятность такого исхода равна

$$P(H_1 | H_0) = \alpha.$$

2. *Ошибка 2-го рода*: Принята неправильная гипотеза. Вероятность такого исхода

$$P(H_0 | H_1) = \beta.$$

3. Принята правильная гипотеза

$$P(H_0 | H_0) = 1 - \alpha.$$

4. Отклонена неправильная гипотеза

$$P(H_1 | H_1) = 1 - \beta.$$

Пример 11. Сделана большая партия деталей. Производитель предполагает, что разброс значений (дисперсия) контролируемого количественного признака соответствует паспортным данным того устройства, на котором изготовлены детали. Основная гипотеза имеет вид $H_0 \cdot D(X) = D_0$ при конкурирующей гипотезе H_1 $D(X) > D_0$. Выборочное обследование деталей либо подтверждает, либо опровергает основную гипотезу. Ошибка 1-го рода будет заключаться в том, что хорошая партия деталей окажется забракованной (''риск производителя''). Ошибка 2-го рода будет состоять в том, что плохая партия деталей будет признана годной (''риск потребителя'').

Если объём выборки жёстко ограничен, то вероятность ошибки одного рода можно снизить только ценой роста вероятности ошибки другого рода. В примере 11 проверяется выборка деталей, после чего вся партия принимается или бракуется. Если мы требуем очень строгого соответствия выборки известным требованиям, то снижаем вероятность ошибки 2-го рода β (т.е. маловероятно, что плохая партия будет признана годной). Но при этом повышается риск ошибки 1-го рода (α), т.к. возрастает вероятность забраковать хорошую в целом партию из-за неудачной выборки.

При проверке гипотезы прежде всего задаются вероятностью совершения ошибки 1-го рода α, которая называется *уровнем значимости* гипотезы. Обычно уровень значимости принимают равным 0.05, 0.01 или 0.001.

Статистическим критерием называется случайная величина K с известным законом распределения вероятностей, служащая для проверки нулевой гипотезы. Вся область значений K делится на критическую область, где H_0 отвергается, и область принятия гипотезы, где отвергать H_0 нет оснований. Названные области отделяются друг от друга критическими точками k_{cr}. Существует 3 вида критических областей: правосторонняя, левосторонняя и двусторонняя. Если допустить, что основная гипотеза верна, то вероятность попадания критерия в критическую область есть вероятность ошибки 1-го рода α. Из этого условия и исходят при отыскании критических точек. Вид критической области зависит от конкурирующей гипотезы. Сам критерий для заданного уровня значимости α выбирается так, чтобы вероятность ошибки 2-го рода β была минимальной. Вероятность отклонения неправильной основной гипотезы $(1 - \beta)$ называется мощностью критерия. Из всех возможных критериев с заданным уровнем значимости α выбирается наиболее мощный.

12. Пусть имеются два нормально распределённых в генеральных совокупностях количественных признака: X и Y. Генеральные дисперсии $D(X)$ и $D(Y)$ известны. Есть основания предположить, что генеральные средние $M(X)$ и $M(Y)$ равны друг другу. Нулевая гипотеза имеет вид $H_0 : M(X) = M(Y)$. В качестве критерия рассматривается величина

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{D(\overline{X}) + D(\overline{Y})}} = \frac{\overline{X} - \overline{Y}}{\sqrt{D(X)/n_X + D(Y)/n_Y}},$$

где n_X и n_Y — объёмы выборок значений количественных признаков X и Y соответственно. Большие (по модулю) значения случайной величины Z снижают ''шансы'' основной гипотезы. Очевидно, этому могут способствовать следующие факторы: (а) сильное различие выборочных средних $\bar{x}$ и $\bar{y}$; (б) слабое рассеивание признаков; (в) большие объёмы выборок. Последние два фактора как бы повышают наше доверие к выборочным данным и увеличивают значимость различия между x и y.

Для трёх случаев конкурирующей гипотезы имеем следующие правила проверки основной гипотезы $H_0: M(X) = M(Y)$ при известных генеральных дисперсиях на уровне значимости α..

1. Для конкурирующей гипотезы $H_1: M(X) \neq M(Y)$ находим правую границу двусторонней критической области из условия $\Phi(z_{cr}) = \frac{1-\alpha}{2}$. При $|Z| < z_{cr}$ нет оснований отклонить основную гипотезу; в противном случае H_0 отвергается.

2. Для конкурирующей гипотезы $H_1: M(X) > M(X)$ находим границу правосторонней критической области из условия $\Phi(z_{cr}) = \frac{1}{2} - \alpha$. При $Z < z_{cr}$ нет оснований отклонить основную гипотезу; в противном случае H_0 отвергается.

3. Для конкурирующей гипотезы $H_1: M(X) < M(Y)$ находим границу левосторонней критической области $(-z_{cr})$ из условия $\Phi(z_{cr}) = \frac{1}{2} - \alpha$. При $Z > -z_{cr}$ нет оснований отклонить основную гипотезу; в противном случае H_0 отвергается.

Можно обобщить описанную методику на случай произвольного (т.е. необязательно нормального) распределения количественных признаков X и Y. Но для этого необходимо располагать достаточно большими (n_X>30 и n_Y>30) независимыми выборками. Действительно, согласно центральной предельной теореме, выборочные средние в этом случае будут распределены почти нормально. Кроме того, даже если нам не известны генеральные дисперсии $D(X)$ и $D(Y)$, то их хорошими оценками могут служить выборочные дисперсии D_x и D_y.

Пример 12. В цехе, производящем посуду, поочерёдно работают два грузчика. Каждый их них отработал по 100 дней. В дни работы 1-го грузчика оказывалось повреждёнными в среднем за смену 36 изделий, в дни работы 2-го грузчика – 38 изделий. При этом выборочные дисперсии числа повреждённых изделий составили соответственно 12 изд^2 и 13 изд^2. Можно ли считать различие качества работы двух грузчиков незначимым?

Заметим, что количественные признаки здесь должны подчиняться закону Пуассона (закону редких событий). Тем не менее наличие двух больших независимых выборок позволяет нам воспользоваться критерием Z. Проверим основную гипотезу $H_0: M(X) = M(Y)$ на уровне значимости α = 0.02 при конкурирующей гипотезе $H_1: M(X) \neq M(Y)$. Имеем:

$$Z = \frac{36-38}{\sqrt{12/100+13/100}} = -4;$$

$\Phi(z_{cr}) = 0.49$; $z_{cr} = 2.33$; $|Z| > z_{cr}$ – нулевая гипотеза отвергается. 2-ой грузчик работает существенно хуже 1-го.

13. Пусть имеется частотное распределение выборки значений количественного признака X. Есть основания предположить, что в генеральной совокупности признак X распределён по нормальному (или равномерному, биномиальному, пуассоновскому и т.д.) закону. В качестве критерия проверки этой гипотезы логично выбрать величину, которая равнялась бы нулю при полном совпадении эмпирических и теоретических частот ($n_i = n_i'$ для всех i). При этом необходимо, чтобы противоположные по знаку отклонения не гасили друг друга. Этим требованиям удовлетворяет так называемый *критерий согласия*

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'}.$$

Распределение χ^2 зависит от числа степеней свободы $s = k - 1 - l$, где k – число групп выборки, т.е. число вариантов или частичных интервалов, l – число параметров гипотетического распределения (например, для нормального и равномерного распределений $l = 2$, для биномиального и пуассоновского распределений $l = 1$). Критическая точка $\chi^2_{cr}(\alpha, s)$ находится в специальной таблице (приложение 4), входом в которую являются уровень значимости α и число степеней свободы s. При $\chi^2 < \chi^2_{cr}(\alpha, s)$ нет оснований отклонить основную гипотезу о виде распределения. В противном случае H_0 отвергается.

Пример 13. (Продолжение примера 6). Проверим гипотезу о нормальном распределении, которая использовалась при получении теоретических частот. Получаем: $\chi^2 = 2.1$, $s = 7 - 1 - 2 = 4$, $\chi^2_{cr}(0.05, 4) = 9.5$. Нет оснований для отклонения гипотезы. В генеральной совокупности признак распределён по нормальному закону.

14. *Корреляционной связью* называется вероятностная или статистическая зависимость, не имеющая строгого функционального характера. Например, рыночная цена квартиры сильно зависит от полезной площади, но эта зависимость не выражается функционально. Существенное влияние оказывают и другие характеристики: район, этаж, планировка, состояние жилища, наличие удобств, срочность продажи и т.п.

Для визуального анализа корреляционной связи служит *диаграмма рассеивания*. Например, вид графика, изображённого на рис. 3, даёт основания предположить, что между двумя признаками существует *отрицательная линейная* корреляционная связь: величина *Y в среднем* приблизительно линейно убывает с ростом *X*. Степень близости точек на диаграмме рассеивания к некоторой функциональной зависимости (например, к прямой линии) принято называть *теснотой корреляции*. Измерителем тесноты линейной корреляции служит выборочный (эмпирический) *коэффициент корреляции*

$$r_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y},$$

где (в случае несгруппированных данных)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i; \quad \sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2, \quad \sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n} y_i - (\bar{y})^2.$$

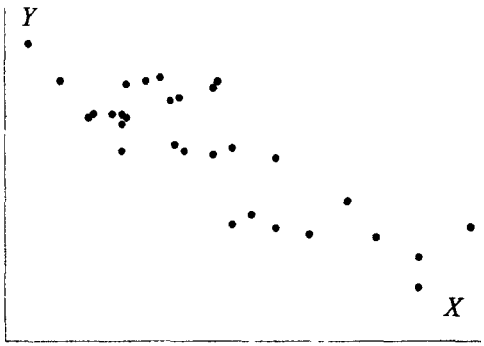


Рис. 3. Диаграмма рассеивания

Коэффициент корреляции может принимать значения в пределах от -1 до 1. Если количественные признаки тесно коррелированы (т.е. близки к линейной функциональной зависимости), то $r_{xy} \approx \pm 1$. В статистической практике принято считать корреляционную связь *заметной* при $|r_{xy}| > 0.5$ и *достаточно тесной* при $|r_{xy}| > 0.8$. Нет оснований ожидать, что для некоррелированных величин мы обязательно получим нулевое или близкое к нулю значение коэффициента r_{xy}.

Пример 14. В таблице приводятся выборочные данные о площади (X, кв. м) и цене (Y, тыс. долларов) 10-и квартир.

x_i	58	74	36	44	70	52	57	65	37	45
y_i	20	21	12	15	22	18	17	23	14	16

Найдём выборочный коэффициент корреляции r_{xy}. Расчёты дают:

$$\bar{x} = 53.8, \quad \bar{y} = 17.8, \quad \overline{xy} = 998.4, \quad \sigma_x = 12.6475, \quad \sigma_y = 3.4583, \quad r_{xy} = 0.9319.$$

Имеется тесная корреляционная связь между площадью квартиры и её ценой.

15. Результаты рассмотрения примера 14 достаточно очевидны. Однако при значениях $|r_{xy}| \approx 0.5$ (и меньше) мы не могли бы с уверенностью утверждать, что признаки корреляционно связаны. Из $r_{xy} \neq 0$ ещё нельзя заключить, что не равен нулю и *генеральный* коэффициент корреляции $r(X,Y)$. Необходимо проверить гипотезу о том, что отклонение r_{xy} от нуля незначимо и случайно: $H_0 : r(X,Y) = 0$, $H_1 : r(X,Y) \neq 0$. В качестве критерия используется случайная величина

$$T_r = r_{xy}\sqrt{\frac{n-2}{1-r_{xy}^2}},$$

подчиняющаяся распределению Стьюдента. Строится двусторонняя критическая область. Правая критическая точка $t_{2\,cr}(\alpha,k)$, где $k = n - 2$, может быть найдена в таблице (приложение 5). На заданном уровне значимости α при $T_r < t_{2\,cr}(\alpha,k)$ нет оснований отклонить нулевую гипотезу; в противном случае H_0 отвергается.

Пример 15. По выборке объёма $n = 10$ парных значений двух признаков найден выборочный коэффициент корреляции $r_{xy} = 0.4$. Проверим гипотезу $H_0 : r(X,Y) = 0$ на уровне значимости α = 0.05. Найдём наблюдаемое значение критерия:

$$T_r = 0.4\sqrt{\frac{8}{1-0.16}} = 1.23.$$

По таблице критических точек распределения Стьюдента находим $t_{2.cr}(0.05,8) = 2.31$. Поскольку $|T_r| < t_{2.cr}$, нет оснований отклонить нулевую гипотезу. Выборочный коэффициент корреляции незначим. Между признаками нет линейной корреляции.

16. Цель *регрессионного метода* – отыскание параметров функциональной зависимости, наиболее точно описывающей поведение среднего значения количественного признака Y при изменении значения другого количественного признака X (или нескольких признаков).

Пусть изучается взаимозависимость двух количественных признаков (X,Y). В результате n опытов или наблюдений получены пары чисел: (x_i, y_i), где $i = 1, 2, ..., n$. Если эмпирическая картина рассеивания свидетельствует о линейной форме корреляции, то в качестве линии связи можно выбрать прямую

$$\bar{y}_x = ax + b,$$

где под $\bar{y}_x$ подразумевается среднее значение признака Y, соответствующее значению $X = x$, а оптимальные значения коэффициентов a и b должны быть подобраны. В *методе наименьших квадратов* критерием оптимизации служит условие

$$S(a,b) = \sum_{i=1}^{n}(y_i - ax_i - b)^2 \to \min.$$

Найденные из этого условия коэффициенты a и b обеспечивают минимальные отличия значений функции $\bar{y}_x(x_i) = ax_i + b$ от наблюдаемых ординат y_i:

$$a^* = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - (\bar{x})^2}, \quad b^* = \frac{\bar{y}\,\overline{x^2} - \bar{x}\,\overline{xy}}{\overline{x^2} - (\bar{x})^2}.$$

Уравнение вида $\bar{y}_x = ax + b$, в котором значения коэффициентов $a = a^*$ и $b = b^*$ вычислены на основании статистических данных методом наименьших квадратов, называется *выборочным уравнением линейной регрессии* Y по X. Угловой коэффициент a^* называется *выборочным коэффициентом регрессии.*

Сравнивая формулы для выборочных коэффициентов корреляции r_{xy} и регрессии a^*, нетрудно убедиться, что

$$a^* = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x^2} = \frac{\sigma_y}{\sigma_x} r_{xy}.$$

Можно показать, что уравнение регрессии представимо в виде

$$\bar{y}_x - \bar{y} = a(x - \bar{x}) \text{ или } \bar{y}_x - \bar{y} = \frac{\sigma_y}{\sigma_x} r_{xy}(x - \bar{x}).$$

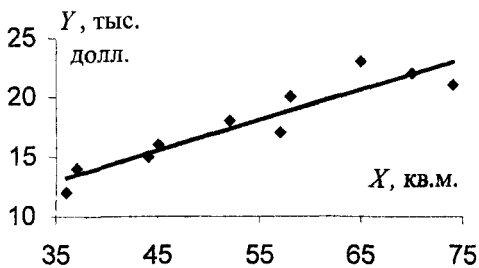


Рис. 4. Прямая линия регрессии

Таким образом, линия регрессии проходит через точку $(\bar{x}, \bar{y})$ с угловым коэффициентом (коэффициентом регрессии), прямо пропорциональным коэффициенту корреляции.

Пример 16. (Продолжение примера 14). Пусть требуется дополнительно: (1) записать уравнение линейной регрессии Y по X; (2) предсказать цену квартиры площадью 50 кв. м. Используя полученные ранее результаты, найдём:

$$\frac{\sigma_y}{\sigma_x} r_{xy} = 0.255; \quad \bar{y}_x - 17.8 = 0.255(x - 53.8); \quad \bar{y}_x = 0.255x + 4.09.$$

Точно такой же результат может быть получен непосредственно с использованием приведённых выше формул для a^* и b^*. Воспользовавшись найденным уравнением регрессии, получим $\bar{y}_x(50) = 16.84$. Итак, цена квартиры площадью 50 кв. м составит (в среднем) 16840 долларов.

3. Образцы заданий для контрольной работы

Задание 1. Для определения средней дальности грузоперевозок проведено наблюдение за 20 грузами. В таблице приведена масса каждого груза (в тоннах) и дальность перевозки (в км).

Масса	Дальн.
25	792
36	432
32	235
27	1030
44	1425
21	727
38	159
22	980
12	407
23	225
45	527
48	1299
57	290
10	64
13	1216
15	895
43	774
23	545
30	755
40	958
надёжн.	0.94

1. Найти минимальное и максимальное значения дальности перевозки в выборке. Построить гистограмму частот для дальности перевозок (без учёта масс перевезённых грузов), введя интервалы 0 – 200, 200 – 400, 400 – 600, 600 – 800, 800 – 1000, 1000 – 1200, 1200 – 1400, 1400 – 1600.
2. Найти точечную несмещённую оценку средней дальности перевозок:
   а) с учётом масс грузов;
   б) без учёта масс грузов.
3. Найти точечную несмещённую оценку дисперсии дальности перевозок в генеральной совокупности и исправленное среднеквадратическое отклонение (СКО) без учёта масс грузов.
4. Считая генеральное СКО известным (приняв его равным исправленному СКО), а распределение — нормальным, построить доверительный интервал для средней дальности перевозок с надёжностью, указанной в таблице.
5. Считая генеральное СКО неизвестным, построить доверительный интервал для средней дальности перевозок с надёжностью 0.99. Объяснить причины того, что доверительный интервал оказался шире, чем в пункте 4.

Решение.

1. $x_{\min} = 64$; $x_{\max} = 1425$. Составим статистический ряд распределения непрерывного количественного признака (см. пример 2 основных теоретических положений (ОТП)).

Интервал, i	1	2	3	4	5	6	7	8
Начало интервала, $x_{i-\frac{1}{2}}$	1	200	400	600	800	1000	1200	1400
Конец интервала, $x_{i+\frac{1}{2}}$	200	400	600	800	1000	1200	1400	1600
Середина интервала, x_i	100	300	500	700	900	1100	1300	1500
Частота, n_i	2	3	4	4	3	1	2	1

Для построения гистограммы частот необходимо для каждого интервала рассчитать значение величины $\frac{n_i}{h}$ ($h = 200$ — длина интервала). Результат изображён на рис. 5.

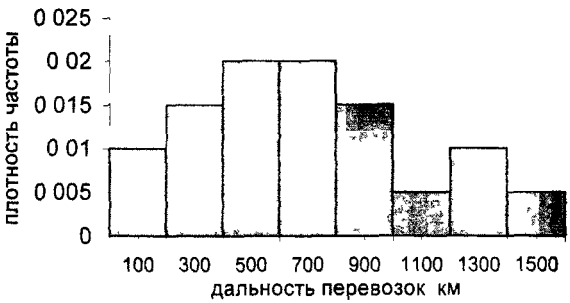


Рис. 5

2. Несмещённой точечной оценкой генеральной средней является выборочная средняя. (См. пункт 4 ОТП). Средняя дальность перевозок с учётом масс грузов (m_i) рассчитывается по формуле

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

где массы являются аналогом частот. Числитель представляет собой грузооборот (тонно-километры), знаменатель – объём перевозок (тонны). В нашем случае $\sum_{i=1}^{n} x_i m_i = 425567$, $\sum_{i=1}^{n} m_i = 604$, $\bar{x} = 704.58$.

Средняя дальность перевозок без учёта масс грузов может быть оценена двумя способами. Теоретически более точной оценкой является выборочная средняя, рассчитанная по несгруппированным данным: $\bar{x} = 686\ 75$. Рассчёт по сгруппированным данным с использованием в качестве вариантов середин интервалов даёт значение $\bar{x} = 690$.

3. Несмещённой оценкой генеральной дисперсии является исправленная дисперсия (п.5 ОТП). В нашем случае $s^2 = \frac{20}{19} D$, где D — выборочная дисперсия. Если последнюю величину рассчитать по несгруппированным данным, то получим следующие значения исправленной дисперсии и исправленного среднеквадратического отклонения: $s^2 = 156606$, $s = 395.734$.

4. При известном генеральном среднеквадратическом отклонении σ доверительный интервал для оценки генеральной средней (математического ожи-

дания) строится с помощью функции Лапласа (п.8 ОТП) по заданной надёжности γ. Полуширина этого интервала равна

$$\delta = \frac{t\sigma}{\sqrt{n}},$$

где $\Phi(t) = \gamma/2$. В нашем случае ($\gamma = 0.94$) по таблице (приложение 2) находим $t = 1.88$. Если, в соответствии с заданием, принять $\sigma = s$, то получим $\delta = 166.4$. Доверительный интервал для средней дальности перевозок

$$686.8 - 166.4 < a < 686.8 + 166.4 \quad \Rightarrow \quad 520.4 < a < 853.2.$$

5. На самом деле, использованное в предыдущем пункте приближение ($\sigma = s$) годится только для больших выборок ($n > 30$). В нашем случае правильнее считать генеральное среднеквадратическое отклонение неизвестным и строить доверительный интервал с помощью коэффициента Стьюдента (п.10 ОТП). Согласно приложению 3, $t(0.99,20) = 2.86$. Тогда $\delta = 253.1$. Доверительный интервал для средней дальности перевозок

$$686.8 - 253.1 < a < 686.8 + 253.1 \quad \Rightarrow \quad 433.7 < a < 939.9$$

оказался шире, чем в пункте 4, по двум причинам. Во-первых, он построен с более высокой надёжностью ($0.99 > 0.94$). Во-вторых, даже если бы в пункте 4 была задана надёжность 0.99, значение t составило бы только $2.6<2.86$, т.к. значение генерального среднеквадратического отклонения полагалось известным, а не оценивалось по выборке объёма 20.

Задание 2. В таблице указаны протяжённость сети и размеры перевозок по железным дорогам 13 стран в 1969 году (БСЭ, т 9, с.139).

Страны	Эксплуатационная длина сети, тыс км	Густота сети на 100 кв км терр км	Грузовые перевозки			Пассажирские перевозки			
			объём перевозок млн т	грузооборот млрд т км	средн дальн перевоз км	объём перевозок млн чел	пассажирооборот, млрд пассажиро-км	средн дальн поездки, км	число поездок на 1 жителя
США	334 0	4 3	1440 0	1150 0	778	302	19 9	66	1 5
Великоб	19 5	8 5	209 0	27 0	113	805	29 6	37	14 5
Франция	36 5	6 6	243 0	69 0	277	607	39 1	64	12 1
ФРГ	33 8	13 6	380 0	69 0	183	1019	37 1	36	17 6
Италия	20 1	6 7	62 9	18 1	288	461	32 5	70	8 5
Япония	26 1	7 1	253 0	61 0	239	6541	181 0	20	65 0
СССР	134 6	0 6	2759 0	2367 0	858	283	261 3	92	11 8
Болгария	4 2	3 8	62 7	12 6	201	105	6 1	54	12 4
Чехосл	13 3	10 4	225 6	53 2	236	572	20 2	35	39 6
ГДР	14 9	13 7	252 0	39 4	156	636	17 6	28	36 8
Венгрия	9 3	10 0	112 4	18 4	164	549	16 4	30	53 2
Польша	26 6	8 5	373 7	95 0	254	1048	37 0	35	32 0
Румыния	11 0	4 6	155 4	39 8	256	306	16 7	55	15 0

1. Найти выборочный коэффициент корреляции между указанной парой показателей X, Y.
2. Проверить гипотезу о значимости коэффициента корреляции при уровне значимости гипотезы $\alpha = 0.05$.
3. Найти выборочное уравнение линейной регрессии Y по X и построить соответствующий график.

X — густота сети на 100 кв. километров территории;

Y — средняя дальность грузоперевозок.

Решение.

1. Для нахождения выборочного коэффициента корреляции требуется рассчитать ряд характеристик (п.14 ОТП). Найдём выборочные средние количественных признаков: $\bar{x} = 7.569$, $\bar{y} = 307.923$. Среднее значение произведения $\overline{xy} = 1776.631$. Выборочные дисперсии: $\sigma_x^2 = 13.3698$, $\sigma_y^2 = 49943.4556$. В результате получаем $r_{xy} = -0.678$. Между густотой сети и средней дальностью грузоперевозок обнаружена заметная отрицательная корреляция. (*Объяснение.* Густота сети обычно выше в странах с небольшой территорией. Средняя дальность грузоперевозок, напротив, с территорией страны растёт).

2. Основная и конкурирующая гипотезы имеют вид

$$H_0 : r(X,Y) = 0,\ H_1 : r(X,Y) \neq 0,$$

где $r(X,Y)$ — генеральный коэффициент корреляции (п.15 ОТП). При объёме выборки $n = 13$ наблюдаемое значение критерия $T_r = -3.06$. Критическую точку распределения Стьюдента находим в приложении 5: $t_{2\,cr}(0.05, 11) = 2.2$. Т.к. $|T_r| > 2.2$, основная гипотеза отвергается. Коэффициент корреляции значим.

3. Запишем уравнение регрессии в виде (п.16 ОТП)

$$y_x - \bar{y} = (\sigma_y / \sigma_x) r_{xy} (x - \bar{x}).$$

После арифметических преобразований получим $\bar{y}_x = -41.4x + 621.6$. К такому же результату можно придти непосредственно, найдя оценки коэффициентов уравнения регрессии a^*, b^* (п.16 ОТП). На рис. 6 показана найденная линия регрессии.

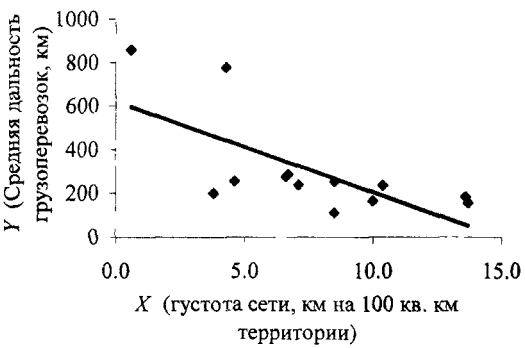


Рис. 6

Задание 3. С целью изучения прочности некоторого изделия исследованы образцы, для каждого из которых определён предел прочности на разрыв. Весь интервал значений (от $40 \cdot 10^7$ до $58 \cdot 10^7$ Н/м^2) разбит на 9 интервалов равной длины, и определены частоты попадания в каждый интервал. В таблице указаны середины интервалов (в 10^7 Н/м^2) и часто́ты.

Сер интерв	Часто-ты
41	0
43	5
45	8
47	12
49	10
51	17
53	13
55	6
57	4
Объем	75

1. Полагая, что в генеральной совокупности количественный признак (предел прочности на разрыв) распределён нормально, произвести выравнивание статистического ряда. На одном графике показать эмпирические и теоретические частоты.

2. Проверить гипотезу о нормальном распределении, задавшись уровнем значимости $\alpha = 0.05$.

Решение.

1. Нахождение теоретических частот нормального распределения описано в п.6 и примере 7 ОТП. Поскольку в данном случае объём выборки велик, не будем «исправлять» значения выборочной дисперсии и выборочного среднеквадратического отклонения. Необходимые для расчёта величины составляют: $n = 75$, $h = 2$, $x = 49.9$, $s \approx \sigma = 3.73$. В результате получаем следующее распределение теоретических частот:

x_i	41	43	45	47	49	51	53	55	57
n_i'	0.9	2.9	6.8	11.8	15.6	15.4	11.4	6.3	2.6

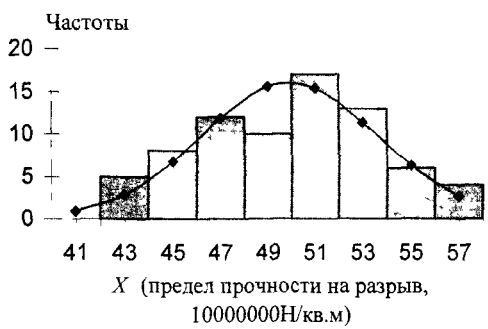


Рис. 7

Совместный график эмпирических и теоретических частот (рис. 7) показывает качественное согласие распределений.

2. Проверка гипотезы о виде распределения описана в п.13 ОТП. Наблюдаемое значение критерия χ^2 в нашем случае составляет 5.8. В таблице (приложение 4) находим критическую точку $\chi^2_{cr}(0.05, 9-1-2) = \chi^2_{cr}(0.05, 6) = 12.6$. Т.к. $\chi^2 < \chi^2_{cr}$, нет оснований для отклонения гипотезы. В генеральной совокупности признак распределён по нормальному закону.

4. Руководство по выполнению заданий на компьютере

Статистика относится к тем областям научно-практической деятельности, которые более всего нуждались в появлении компьютеров, поскольку ручная обработка и анализ статистической информации крайне неэффективны. В данном разделе рассматривается решение заданий контрольной работы на компьютере с помощью офисной программы Excel операционной системы Windows. Microsoft Excel привлекает тем, что это пакет общего назначения, применяемый во многих практических задачах и знакомый подавляющему большинству специалистов и студентов. Описание ориентировано на версию Excel 97, но без труда может быть адаптировано к более поздним версиям.

В качестве примеров рассматриваются образцы заданий с **теми же исходными данными**, что и в предыдущем разделе (тексты заданий не повторяются).

Выполнение задания 1.

После запуска программы Excel указатель ячейки, как правило, находится в ячейке A1. Данная ячейка является активной, т.е. пользователь может вводить в неё данные.

Введём заголовок: *Статистическое изучение дальности грузоперевозок.* Длина вводимого текста превышает ширину ячейки A1. Поэтому текст будет выступать за правый край ячейки. После нажатия клавиши [Enter] указатель переместится в ячейку A2.

Нажмём клавишу [↓] для перемещения указателя ячейки из A2 в A3 или активизируем ячейку A3 посредством щелчка. Запишем в эту ячейку заголовок *Масса*, а в ячейки A4 – A23 — массы грузов.

Переместимся в ячейку B3 (это можно сделать с помощью клавиш [↑] и [→], но проще — посредством щелчка). Поместим в этой ячейке заголовок *Дальн.*, а в ячейках B4 – B23 — значения дальности каждой перевозки. В ячейке A24 введём *Надёжн.*, а в ячейке B24 — заданное значение надёжности интервальной оценки (0.94). Столбцы цифр могут быть выровнены по левому краю, по центру, по правому краю. Соответствующие кнопки расположены на панели форматирования.

Если расположение полученной таблицы на экране вас не устраивает, вы можете улучшить её внешний вид, используя изменение масштаба и шрифта. Для того чтобы таблица полностью помещалась на экране, можно установить масштаб 75%.

Найдём минимальное и максимальное значения дальности в выборке. Конечно, когда данных немного, это можно сделать и «вручную». Однако, как правило, статистические ряды довольно велики, и имеет смысл воспользовать-

ся так называемыми **функциями рабочего листа**. Итак, введём в ячейку D4 текст *Минимальная дальность*, а в ячейку D5 — *Максимальная дальность*. Активизируем ячейку H4. Этим мы готовим её для помещения минимального значения из списка. Щёлкнем по кнопке **Вставка функций**, находящейся на панели инструментов, откроется соответствующее диалоговое окно. Теперь необходимо выбрать нужную нам функцию. В списке категорий выбираем **Статистические**, в списке функций находим МИН. После щелчка на кнопке **ОК** открывается диалоговое окно с описанием функции МИН, где должны быть заданы аргументы. Курсор ввода находится в поле ввода первого аргумента. Задать аргументы можно различными способами. Можно ввести адрес диапазона ячеек, содержащих значения, среди которых мы ищем минимальное (в нашем случае это B4:B23). Можно просто выделить этот диапазон в таблице (эта операция выполняется с помощью удерживания нажатой левой кнопки мыши). В результате выделенный диапазон будет окружён движущейся штриховой рамкой, а в поле ввода появится адрес B4:B23. Наконец, можно просто ввести сами значения в поле ввода. (Конечно, в данном случае это будет совершенно нерациональным способом, т.к. придётся заново набрать весь список. Однако, когда нужно вводить всего 1-2 числа, такой способ удобен.) Так или иначе, после ввода аргументов в нижней части окна появится результат (64). Щёлкнув на кнопке **ОК**, мы перенесём его в ячейку H4.

Имейте в виду, что теперь эта ячейка всегда будет показывать минимальное значение из диапазона B4:B23, «отслеживая» все происходящие в нём изменения. Если вы забудете о содержимом этой ячейки, вы всегда можете щелчком активизировать её и прочесть в **Строке формул** (над таблицей) формулу или функцию, по которой вычислено значение, хранящееся в ячейке.

Теперь поместите в ячейку H5 максимальное значение из того же списка. Единственное отличие в ваших действиях будет заключаться в том, что вместо функции МИН вы должны использовать функцию МАКС. Результатом будет число 1425.

Приступим к нахождению частотного распределения дальности перевозок. Введём в ячейках D7, E7 и F7 тексты *Нач., Сер., Кон.* соответственно, что будет означать начала, середины и концы интервалов. Теперь, согласно условию, в диапазоне D8:D15 должны быть введены числа 0, 200, 400, …, 1400, в диапазоне E8:E15 — числа 100, 300, 500, …, 1500, в диапазоне F8:F15 — числа 200, 400, 600, …, 1600.

Следующий столбец таблицы будет заполнен эмпирическими частотами, для нахождения которых мы воспользуемся одной из функций рабочего листа.

В ячейке G7 введём текст *Частота*. Выделим ячейки G8:G15. Щёлкнем на кнопке **Вставка функций**. В открывшемся диалоговом окне выберем: **Статистические**, ЧАСТОТА. Мы выбрали функцию, которая, как вы видите, «возвращает распределение частот в виде вертикального массива». Эта функция, в отличие от МИН и МАКС, обязательно должна иметь два аргумента: массив данных и двоичный массив. В первое окно ввода мы вводим все наблюдавшие-

ся значения дальности (диапазон B4:B23). Переместив с помощью мыши курсор во второе окно ввода, мы выделяем ячейки F8:F15, в которых находятся концы интервалов. (Обратите внимание, что все эти действия отражаются в строке формул). В правой части окна появился результат: *{2,3,4,4,3,1,2,1,0}*. Это и есть искомые частоты. Для того чтобы поместить их в таблицу, щёлкнем на кнопке **ОК**. В ячейке G8 появилось значение *2*, но остальные ячейки столбца остались незаполненными. Дело в том, что результатом обращения к функции ЧАСТОТА является не просто одно число, как это было с функциями МИН или МАКС, а массив. Это требует от нас следующих дополнительных действий. Нажмём клавишу [F2]. В строке состояния (она находится в нижней части экрана) появится надпись **Правка,** а в активной ячейке — информация о её содержимом. Теперь нажмём на комбинацию клавиш [Ctrl]+[Shift] и, не отпуская, на клавишу [Enter]. Все ячейки столбца должны заполниться числами.

Приступим к построению гистограммы частот. Для этого в каждом интервале необходимо определить плотность частоты, поделив соответствующую частоту на длину интервала (на 200). В ячейке H7 введём текст *Плотн.* Для определения плотности воспользуемся заданием **Формул** в соответствующих ячейках рабочего листа. Например, мы хотим поместить в ячейку H8 формулу для вычисления плотности частоты внутри 1-го интервала. Активизируем эту ячейку, введём с клавиатуры знак =. Соответствующая частота находится в ячейке G8. Мы можем либо ввести с клавиатуры G8/200, либо воспользоваться выделением нужной ячейки с помощью мыши (арифметические знаки всё равно придётся вводить с клавиатуры). Так или иначе, после этого в **Строке формул** и в самой ячейке H8 должна появиться надпись: = G8/200. Это и есть **Формула**, которая будет автоматически «отслеживать» все изменения, происходящие в ячейке G8. После нажатия на клавишу [Enter] вместо формулы в ячейке появится результат вычисления. Естественно, в дальнейшем вы всегда можете, активизировав ячейку, прочесть (и при желании, отредактировать) хранящуюся в данной ячейке формулу.

Казалось бы, теперь нам придётся вводить формулу в ячейки данного столбца, каждый раз изменяя адрес ссылки, что, конечно, заняло бы немало времени. На самом деле всё гораздо проще: можно воспользоваться процедурой **Автозаполнения**. Активизировав ячейку H8, поместите курсор мыши на маркер заполнения в правом нижнем углу ячейки. Курсор приобретает форму чёрного крестика. Нажав на левую кнопку мыши, переместите курсор на ячейку H15. Как только вы отпустите кнопку, выделенные ячейки автоматически заполнятся числами. Обратите внимание: программа «поняла», что мы хотим посчитать каждый элемент столбца по одной и той же формуле, но с разными частотами, расположенными в соответствующих строках.

В программе Excel существует широкий круг средств графического представления числовых данных. Гистограмма является частным случаем (типом) диаграммы, часто используемым в статистике. Прежде всего, необходимо выделить в рабочем листе диапазон ячеек, данные из которого должны быть

представлены в гистограмме. Удалим название *Сер.* из ячейки E7 (чтобы программа воспринимала этот столбец как ряд значений аргумента, а не функции). Выделим несмежные диапазоны ячеек: E7:E15 и H7:H15. Для этого придётся после выделения первого диапазона нажать и удерживать кнопку Ctrl.

Активизируем кнопку **Мастер диаграмм** на панели инструментов. На экране появится первое диалоговое окно Мастера диаграмм, в котором можно выбрать тип диаграммы. Мы выберем **Гистограмму** (обычную) и нажмём кнопку **Далее.** Во втором окне необходимо подтвердить, что ряды данных находятся в столбцах. В третьем диалоговом окне задаётся описание диаграммы. Можно добавить легенду, дать название диаграмме и дать названия по осям (назовём категории по оси *X – Дальность перевозок, км*, по оси *Y – Плотность частоты*). В последнем окне определим размещение диаграммы. Вставим её в имеющийся лист.

В рабочем листе гистограмма может находиться в закрытом состоянии (для этого нужно щёлкнуть мышью вне области диаграммы) или в активном, выделенном состоянии (щёлкнуть мышью в области диаграммы). Удерживая левую кнопку мыши нажатой, можно позиционировать диаграмму в любом месте листа. Обратите также внимание на маркеры, расположенные на границе выделенной диаграммы. Перемещая эти маркеры с помощью мыши, можно изменять размеры окна диаграммы. Поместим гистограмму в диапазоне ячеек J7:P21.

Для обработки гистограммы с помощью специальных средств следует активизировать её с помощью двойного щелчка мышью. Форматирование диаграммы может включать в себя вставку или изменение легенды, задание цвета и узора, помещение текста на осях, форматирование осей и т.п. Допустим, что нас не устраивают какие-то элементы графика, например, зазор между отдельными столбиками гистограммы. Установив курсор на любом из столбиков, с помощью щелчка на *правой* кнопке можно открыть контекстное меню и выбрать в нём **Формат данных**. В диалоговом окне активизируем кнопку **Параметры**, установим ширину зазора, равную нулю, и щелкнем на кнопке **ОК**. Желаемое изменение в график внесено.

Добейтесь, чтобы ваша гистограмма выглядела так же, как на рис. 5.

Перейдём к статистическим оценкам. Средняя дальность перевозок с учётом масс грузов представляет собой частное от деления грузооборота на объём перевозок. Введем в ячейке D17 текст *Грузооборот*, а в ячейке H17 — функцию СУММПРОИЗВ (она находится в списке **Математические**). В качестве аргументов этой функции зададим два массива: масс грузов и дальностей. Результат: 425567 (тонно-километров). В ячейке D18 введем текст *Объём перевозок*, а в ячейке H18 — функцию СУММ (она находится в том же списке). Результат: 604 (тонны). В ячейке D19 введем текст *Ср.дальность (с учётом масс)*, а в ячейке H19 — формулу, определяющую эту величину (с адресами числителя и знаменателя). Результат в этом случае окажется нецелым числом, причём количество знаков после точки (или запятой) будет зависеть от размера ячейки,

шрифта и масштаба. Допустим, мы хотим, чтобы были указаны только сотые. Активизируем ячейку H19, откроем меню **Формат**, в нём выберем **Ячейки...** В открывшемся диалоговом окне установим **числовой формат** с двумя десятичными знаками. Результат: 704.58 (км). Это вовсе не означает, что мы, пользуясь в дальнейшем данной ячейкой, потеряем точность: изменился лишь вид числа на экране.

В ячейке D20 введем текст *Ср.дальность (без учёта масс)*, а в ячейке H20 — функцию СРЗНАЧ (она находится в списке **Статистические**). Её аргументом будет массив дальностей. Результат: 686.75. Более грубый вариант оценки основан на интервальном ряде. В ячейке D21 введем текст *Ср.дальность (по сгрупп.данным)*, а в ячейке H20 — уже знакомую нам функцию СУММПРОИЗВ. В данном случае её аргументами будут следующие два массива: середины интервалов и часто́ты. Теперь скорректируем формулу в ячейке D21, добавив деление на объём выборки (на 20). Результат: 690 (км).

В ячейке D22 введем текст *Исправленная дисперсия,* а в ячейке H22 — функцию ДИСП с массивом дальностей в качестве аргумента. Результат: 156606 ($км^2$). *Исправленное среднекв.откл.* (этот текст введём в ячейке D23) вычисляется как квадратный корень из исправленной дисперсии, но в Excel для этой цели существует специальная функция СТАНДОТКЛОН (введите её в ячейке H23). Результат: 395.734 (км).

Построение интервальной оценки генеральной средней (математического ожидания) можно выполнить с помощью функции ДОВЕРИТ, которая даёт значение полуширины доверительного интервала по считающемуся известным генеральному среднеквадратическому отклонению. Функция ДОВЕРИТ имеет 3 аргумента: уровень значимости «альфа» ($\alpha = 1 - \gamma$, где γ — заданное значение надёжности), «стандартное отклонение» (исправленное среднеквадратическое отклонение) и «размер» (объём выборки). В нашем случае $\alpha = 0.06$, объём выборки составляет 20. В качестве второго аргумента («стандартное отклонение») удобнее задавать не конкретные числа, полученные с помощью функции СТАНДОТКЛОН, а адреса ячеек, куда эти значения помещены. Введите текст *Полуширина доверит.интерв.* в ячейке D24 и функцию ДОВЕРИТ с соответствующими аргументами в ячейке H24. Результат: 166.429 (км).

В ячейках D25 и D26 введите текст *Нижн.граница доверит.интерв.* и *Верхн.граница доверит.интерв.* соответственно. Для определения этих границ в ячейке H25 должна быть введена формула =H20-H24, а в ячейке H26 формула =H20+H24. Отметим, что эти формулы можно задать и с помощью функций. Например, мы могли бы использовать для нижней границы доверительного интервала формулу =СРЗНАЧ(...) - ДОВЕРИТ(...), где в скобках должны присутствовать аргументы функций.

Другой вариант построения доверительного интервала для генеральной средней применяется, когда генеральное среднеквадратическое отклонение неизвестно. В этом случае необходимо воспользоваться таблицей коэффициентов Стьюдента (приложение 3). Введите значение коэффициента в рабочий лист и

завершите выполнение задания самостоятельно. (Для вычисления квадратного корня из объёма выборки воспользуйтесь функцией КОРЕНЬ из списка **Математические**.)

Окончательно ваш рабочий лист (кроме гистограммы) должен иметь следующий вид:

	A	B	C	D	E	F	G	H
1	Задание 1. Статистическое изучение дальности грузоперевозок							
2								
3	Масса	Дальн.						
4	25	792		Минимальная дальность				64
5	36	432		Максимальная дальность				1425
6	32	235						
7	27	1030		Нач.		Кон.	Частоты	Плотн.
8	44	1425		0	100	200	2	0.01
9	21	727		200	300	400	3	0.015
10	38	159		400	500	600	4	0.02
11	22	980		600	700	800	4	0.02
12	12	407		800	900	1000	3	0.015
13	23	225		1000	1100	1200	1	0.005
14	45	527		1200	1300	1400	2	0.01
	48	1299		1400	1500	1600	1	0.005
	57	290						
	10	64		Грузооборот				425567
	13	1216		Объём перевозок				604
	15	895		Ср.дальность (с учётом масс)				704.58
	43	774		Ср.дальность (без учёта масс)				686.75
	23	545		Ср.дальность по сгрупп.данным				690
	30	755		Исправленная дисперсия				156606
	40	958		Исправленное среднекв.откл				395.734
	Надёжн.	0.94		Полуширина доверит.интерв. (1)				166 429
				Нижн.граница доверит.интерв. (1)				520.321
				Верхн.граница доверит.интерв. (1)				853.179
				Коэффициент Стьюдента				2.86
				Полуширина доверит.интерв. (2)				253.078
				Нижн.граница доверит.интерв. (2)				433.672
				Верхн.граница доверит.интерв. (2)				939.828

Выполнение задания 2.

Откройте новый рабочий лист. Расположите ряды данных X и Y в два столбца. Для того чтобы получить значение выборочного коэффициента корреляции, достаточно обратиться к функции КОРРЕЛ из списка **Статистические**, задав в качестве аргументов два соответствующих друг другу массива данных. Сделав это, вы получите значение -0.6781

При объёме выборки $n=13$ наблюдаемое значение критерия $T_r=-3.06$. Критическую точку распределения Стьюдента находим в приложении 5: $t_{2\,cr}(0.05,11)=2.2$. Т.к. $|T_r|>2.2$, основная гипотеза отвергается. Коэффициент корреляции значим.

Снова выделите ряды данных X и Y В диалоговом окне **Мастера диаграмм** выберите тип диаграммы **Точечная**, вид первый. При необходимости последовательно отформатируйте оси абсцисс и ординат, щёлкнув на них дважды мышью и активизировав кнопку **Шкала** в диалоговом окне **Формат оси** (для этого необходимо отключить режим **Авто** при определении некоторых параметров шкалы).

В Excel имеется возможность интерпретировать ряды данных в виде **Линий тренда**, представляющих собой аппроксимации, полученные на основе регрессионного анализа. Активизируем диаграмму двойным щелчком, а затем установим курсор мыши на любую точку графика и сделаем щелчок на правой кнопке. В открывшемся контекстном меню выберем **Добавить линию тренда,** а в соответствующем диалоговом окне — тип аппроксимации **Линейная**. Теперь в этом же окне активизируем кнопку **Параметры** и установим опцию **Показывать уравнение на диаграмме**. После нажатия на кнопку [OK] на диаграмме появится линия тренда и соответствующее ей уравнение линейной регрессии:

$$y=-41.445x+621.63.$$

Добейтесь полного сходства полученной диаграммы с рис.6.

Конечно, для того чтобы получить уравнение регрессии в Excel, вовсе не обязательно пользоваться графическими средствами. Имеется функция НАКЛОН из списка **Статистические**, которая позволяет найти выборочный коэффициент регрессии a^* по массивам известных значений Y и X Обратитесь к ней, и вы получите значение $a^*=-41.445$. Функция ПРЕДСКАЗ из того же списка даёт значение линейного тренда (иначе говоря, значение Y в уравнении регрессии) для любого заданного значения X. В качестве первого аргумента задаётся значение $X=x$, а в качестве второго и третьего аргументов — массивы известных значений Y и X (именно в таком порядке!). Если в качестве первого аргумента этой функции задать 0, то мы получим оценку другого коэффициента уравнения регрессии $b^*=621.629$.

В качестве дополнительного упражнения: убедитесь в справедливости формулы

$$a^* = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x^2} = \frac{\sigma_y}{\sigma_x} r_{xy},$$

используя известные вам функции Excel.

Выполнение задания 3.

Подготовьте в новом рабочем листе следующую таблицу:

	A	B	C	D	E
1	Варианты	Эм-пир част	Теор част.	Слагаемые	
2	41	0			
3	43	5			
4	45	8			
5	47	12			
6	49	10			
7	51	17			
8	53	13			
9	55	6			
10	57	4			
11	Суммы				
12					
13	средняя				
14	среднекв отклонение				

В данном случае мы имеем дело со сгруппированными данными, и функции СРЗНАЧ, ДИСП, СТАНДОТКЛОН не годятся. Однако можно применить известную нам функцию СУММПРОИЗВ. Прежде всего, введите в ячейке B11 функцию СУММ, которая будет, суммируя эмпирические частоты, давать объём выборки. Результат: 75. В ячейке B13 должна быть введена формула, рассчитывающая среднюю по сгруппированным данным. Сделайте это с помощью функции СУММПРОИЗВ. Деля на объём выборки, набирайте не конкретное число, а адрес ячейки. Результат: 41.91.

Т.к. объём выборки достаточно велик, не будем «исправлять» значения выборочной дисперсии и выборочного среднеквадратического отклонения. Найдём последнюю величину как

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 n_i}{n} - \bar{x}^2},$$

для чего введём в ячейке D14 формулу

=КОРЕНЬ(СУММПРОИЗВ(A2:A10;A2:A10;B2:B10)/B11-D13^2).

Результат: 3.728.

В ячейках следующего столбца таблицы должны быть помещены формулы для расчёта теоретических частот. Для этого воспользуемся функцией НОРМРАСП из списка **Статистические**, которая вычисляет значение плотности вероятности нормального распределения. Как вы видите из описания синтаксиса этой функции, в качестве первых трёх её аргументов должны быть введены значения x, $\bar{x}$, σ, а на месте четвертого аргумента необходимо набрать *Ложь* (мы ищем не интегральную функцию распределения, а функцию плотности).

Используем формулу массива. Выделим диапазон C2:C10 и введём формулу массива:

=НОРМРАСП(A2:A10;D13;D14;ЛОЖЬ)*2*B11.

(Здесь мы учли длину интервала $h = 2$ и адреса ячеек, в которых введены объём выборки, средняя и среднеквадратическое отклонение.) Далее поступаем точно так, как мы делали ранее (Задание 1) при обращении к функции ЧАСТОТА, когда результатом являлся массив (через клавишу [F2] и комбинацию клавиш [Ctrl]+[Shift]+[Enter]). В результате столбец C2:C10 заполнится теоретическими частотами нормального распределения.

Постройте диаграмму, показывающую распределение эмпирических и теоретических частот. Среди **нестандартных** диаграмм выберите **График/гистограмма** и отформатируйте. Результат должен быть примерно таким, как на рис.7.

Для вычисления слагаемых наблюдаемого значения χ^2 в ячейку D2 введём формулу =(B2-C2)^2/C2 и с помощью **Автозаполнения** заполним остальные ячейки столбца (до D10 включительно). Теперь применим **Автозаполнение** к ячейкам C11, D11 от «источника» B11, где введена функция СУММ. В результате получим наблюдаемое значение критерия $\chi^2 = 5.845$. В таблице (приложение 4) находим критическую точку $\chi^2_{cr}(0.05, 9-1-2) = \chi^2_{cr}(0.05, 6) = 12.6$. Т.к. $\chi^2 < \chi^2_{cr}$, нет оснований для отклонения гипотезы. В генеральной совокупности признак распределён по нормальному закону.

5. Варианты заданий контрольной работы

К заданию 1:

Вариант	1	Вариант	2	Вариант	3	Вариант	4	Вариант	5
Масса	Дальн	Масса	Дальн	Масса	Дальн	Масса	Дальн	Масса	Дальн
30	135	47	723	11	567	17	1238	35	692
12	1063	10	1281	49	501	43	720	16	422
42	385	49	1030	15	836	10	916	22	130
13	878	10	1204	11	1579	10	310	47	830
10	711	10	1179	12	1010	19	305	14	195
49	1024	48	1166	11	915	15	1237	11	787
45	391	14	459	10	819	10	1334	38	659
25	1357	11	789	15	227	35	1190	10	969
45	1346	10	993	10	817	17	407	10	507
37	1056	11	1125	10	667	24	693	13	225
11	1086	16	523	36	525	31	818	10	557
34	883	42	806	10	424	13	1428	48	1099
43	686	15	810	10	704	14	580	17	390
38	410	11	918	22	721	24	1281	10	32
32	1099	15	965	44	458	47	674	10	1116
10	657	10	1092	39	1082	13	964	10	495
10	1136	10	214	14	226	42	642	11	674
15	387	15	1492	49	626	20	718	14	645
18	1273	27	211	34	507	27	1252	30	955
15	486	16	1160	12	1378	10	987	24	1001
надёжн	0 9	надёжн	0 91	надёжн	0 92	надежн	0 93	надёжн	0 94

Вариант	6	Вариант	7	Вариант	8	Вариант	9	Вариант	10
Масса	Дальн	Масса	Дальн	Масса	Дальн	Масса	Дальн	Масса	Дальн
46	677	15	663	10	1094	10	1009	16	1001
34	813	37	1295	28	889	40	431	12	859
22	458	11	1487	13	850	13	769	10	1050
10	969	17	959	19	359	36	653	28	76
34	1068	11	1121	36	184	10	1068	12	220
46	607	30	1159	10	970	11	630	27	484
41	200	30	1458	48	1427	25	879	10	1410
11	279	10	772	43	777	41	1045	11	933
24	423	41	1253	22	794	46	255	49	782
48	1193	13	1381	10	663	18	515	10	119
10	1568	19	652	24	1124	47	1021	12	800
47	497	11	264	23	968	15	592	49	760
25	711	17	679	11	588	38	503	19	936
11	734	44	842	25	670	17	418	17	417
49	994	27	714	48	1091	10	790	19	948
10	1033	21	554	10	1060	29	1095	24	719
10	531	11	466	11	151	15	576	23	406
14	1266	13	71	43	264	12	742	33	1116
42	692	10	1430	15	330	49	524	44	332
48	1268	29	969	34	1265	21	1066	13	552
надёжн	0 95	надёжн	0 96	надёжн	0 97	надёжн	0 98	надёжн	0 99

К заданию 2:

Вариант	Показатель (количественный признак) X	Показатель (количественный признак) Y
1	Длина сети	Объём грузоперевозок
2	Длина сети	Средняя дальность грузоперевозок
3	Густота сети	Средняя дальность поездки
4	Объём грузоперевозок	Средняя дальность грузоперевозок
5	Объём грузоперевозок	Пассажирооборот
6	Грузооборот	Пассажирооборот
7	Грузооборот	Средняя дальность грузоперевозок
8	Средняя дальность грузоперевозок	Средняя дальность поездки
9	Объём пассажирских перевозок	Число поездок на 1 жителя
10	Средняя дальность поездки	Число поездок на 1 жителя

К заданию 3:

Вариант	1	Вариант	2	Вариант	3	Вариант	4	Вариант	5
Сер интерв	Часто-ты	Сер интерв	Часто-ты	Сер интерв	Часто-ты	Сер интерв	Часто-ты	Сер интерв	Часто-ты
41	5	41	5	41	8	41	5	41	3
43	6	43	4	43	2	43	9	43	3
45	8	45	9	45	5	45	11	45	7
47	11	47	15	47	12	47	15	47	14
49	22	49	16	49	15	49	20	49	15
51	15	51	17	51	17	51	13	51	19
53	10	53	7	53	12	53	5	53	5
55	1	55	5	55	3	55	10	55	6
57	4	57	1	57	4	57	4	57	1
Объем	82	Объём	79	Объём	78	Объём	92	Объём	73

Вариант	6	Вариант	7	Вариант	8	Вариант	9	Вариант	10
Сер интерв	Часто-ты	Сер интерв	Часто-ты	Сер интерв	Часто-ты	Сер интерв	Часто-ты	Сер интерв	Часто-ты
41	0	41	9	41	7	41	4	41	7
43	8	43	10	43	4	43	8	43	9
45	*8*	*45*	*9*	*45*	*14*	*45*	*14*	*45*	*8*
47	13	47	11	47	16	47	11	47	14
49	15	49	13	49	13	49	22	49	21
51	15	51	18	51	17	51	10	51	16
53	9	53	13	53	5	53	8	53	9
55	4	55	8	55	8	55	5	55	2
57	2	57	9	57	1	57	9	57	1
Объём	74	Объём	100	Объём	85	Объём	91	Объём	87

Библиографический список

1. Браунли К.А. Статистическая теория и методология в науке и технике. – М.: Наука, 1977. – 407 с.
2. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 1977. – 479 с.
3. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 1977. – 400 с.
4. Колемаев В. А., Калинина В.Н. Теория вероятностей и математическая статистика. – М.: ИНФРА-М, 1997. – 301 с.
5. Статистические методы повышения качества /Под ред. Х.Куме. – М.: Финансы и статистика, 1990. – 304 с.
6. Ниворожкина Л.И., Морозова З.А. Основы статистики с элементами теории вероятностей для экономистов. – Ростов-на-Дону: Феникс, 1999. – 320 с.
7. Колесников А. Excel 97 для Windows 95. - Киев: BHV, 1998. – 479 с.

Приложение 1

Таблица значений функции

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

z	$\varphi(z)$	z	$\varphi(z)$	z	$\varphi(z)$
0.00	0.399	1.00	0.242	2.00	0.054
0.05	0.398	1.05	0.230	2.05	0.049
0.10	0.397	1.10	0.218	2.10	0.044
0.15	0.394	1.15	0.206	2.15	0.040
0.20	0.391	1.20	0.194	2.20	0.035
0.25	0.387	1.25	0.183	2.25	0.032
0.30	0.381	1.30	0.171	2.30	0.028
0.35	0.375	1.35	0.160	2.35	0.025
0.40	0.368	1.40	0.150	2.40	0.022
0.45	0.361	1.45	0.139	2.45	0.020
0.50	0.352	1.50	0.130	2.50	0.018
0.55	0.343	1.55	0.120	2.55	0.015
0.60	0.333	1.60	0.111	2.60	0.014
0.65	0.323	1.65	0.102	2.65	0.012
0.70	0.312	1.70	0.094	2.70	0.010
0.75	0.301	1.75	0.086	2.75	0.009
0.80	0.290	1.80	0.079	2.80	0.008
0.85	0.278	1.85	0.072	2.85	0.007
0.90	0.266	1.90	0.066	2.90	0.006
0.95	0.254	1.95	0.060	2.95	0.005

Таблица значений функции Лапласа

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{z^2}{2}} dz$$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.05	0.020	1.05	0.355	2.05	0.480
0.10	0.040	1.10	0.364	2.10	0.482
0.15	0.060	1.15	0.375	2.15	0.484
0.20	0.079	1.20	0.385	2.20	0.486
0.25	0.098	1.25	0.394	2.25	0.488
0.30	0.118	1.30	0.403	2.30	0.489
0.35	0.137	1.35	0.412	2.35	0.490
0.40	0.155	1.40	0.419	2.40	0.492
0.45	0.174	1.45	0.427	2.45	0.493
0.50	0.192	1.50	0.433	2.50	0.494
0.55	0.209	1.55	0.439	2.60	0.495
0.60	0.226	1.60	0.445	2.70	0.497
0.65	0.242	1.65	0.451	2.80	0.4974
0.70	0.258	1.70	0.455	2.90	0.4981
0.75	0.273	1.75	0.460	3.00	0.4987
0.80	0.288	1.80	0.464	3.25	0.4994
0.85	0.302	1.85	0.468	3.50	0.4997
0.90	0.316	1.90	0.471	4.00	0.49997
0.95	0.329	1.95	0.474	4.50	0.49999
1.00	0.341	2.00	0.477	5.00	0.49999

Коэффициенты Стьюдента $t(\gamma, n)$

Объём выборки	Надёжность оценки γ		
n	0.95	0.99	0.999
5	2.78	4.60	8.61
6	2.57	4.03	6.86
7	2.45	3.71	5.96
8	2.37	3.50	5.41
9	2.31	3.36	5.04
10	2.26	3.25	4.78
11	2.23	3.17	4.59
12	2.20	3.11	4.44
13	2.18	3.06	4.32
14	2.16	3.01	4.22
15	2.15	2.98	4.14
16	2.13	2.95	4.07
17	2.12	2.92	4.02
18	2.11	2.90	3.97
19	2.10	2.88	3.92
20	2.09	2.86	3.88
25	2.06	2.80	3.75
30	2.05	2.76	3.66
35	2.03	2.72	3.60
40	2.02	2.71	3.56
45	2.02	2.69	3.53
50	2.01	2.68	3.50
60	2.00	2.66	3.46
70	2.00	2.65	3.44
80	1.99	2.64	3.42
90	1.99	2.63	3.40
100	1.98	2.63	3.39
120	1.98	2.62	3.37
∞	1.96	2.58	3.29

Критические точки распределения χ^2

Число степеней свободы s	Уровень значимости α		
	0.05	0.01	0.001
1	3.84	6.63	10.83
2	5.99	9.21	13.82
3	7.81	11.34	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46
7	14.07	18.48	24.32
8	15.51	20.09	26.13
9	16.92	21.67	27.88
10	18.31	23.21	29.59
11	19.68	24.73	31.26
12	21.03	26.22	32.91
13	22.36	27.69	34.53
14	23.68	29.14	36.12
15	25.00	30.58	37.70
16	26.30	32.00	39.25
17	27.59	33.41	40.79
18	28.87	34.81	42.31
19	30.14	36.19	43.82
20	31.41	37.57	45.32
21	32.67	38.93	46.80
22	33.92	40.29	48.27
23	35.17	41.64	49.73
24	36.42	42.98	51.18
25	37.65	44.31	52.62
26	38.89	45.64	54.05
27	40.11	46.96	55.48
28	41.34	48.28	56.89
29	42.56	49.59	58.30
30	43.77	50.89	59.70

Приложение 5

Критические точки распределения Стьюдента

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0.10	0.05	0.02	0.01	0.002	0.001
1	6.31	12.7	31.82	63.7	318.3	631.0
2	2.92	4.30	6.97	9.92	22.33	31.6
3	2.35	3.18	4.54	5.84	10.22	12.9
4	2.13	2.78	3.75	4.60	7.17	8.61
5	2.01	2.57	3.37	4.03	5.89	6.86
6	1.94	2.45	3.14	3.71	5.21	5.96
7	1.89	2.36	3.00	3.50	4.79	5.40
8	1.86	2.31	2.90	3.36	4.50	5.04
9	1.83	2.26	2.82	3.25	4.30	4.78
10	1.81	2.23	2.76	3.17	4.14	4.59
11	1.80	2.20	2.72	3.11	4.03	4.44
12	1.78	2.18	2.68	3.05	3.93	4.32
13	1.77	2.16	2.65	3.01	3.85	4.22
14	1.76	2.14	2.62	2.98	3.79	4.17
15	1.75	2.13	2.60	2.95	3.73	4.07
16	1.75	2.12	2.58	2.92	3.69	4.01
17	1.74	2.11	2.57	2.90	3.65	3.96
18	1.73	2.10	2.55	2.88	3.61	3.92
19	1.73	2.09	2.54	2.86	3.58	3.88
20	1.73	2.09	2.53	2.85	3.55	3.85
21	1.72	2.08	2.52	2.83	3.53	3.82
22	1.72	2.07	2.51	2.82	3.51	3.79
23	1.71	2.07	2.50	2.81	3.49	3.77
24	1.71	2.06	2.49	2.80	3.47	3.74
25	1.71	2.06	2.49	2.79	3.45	3.72
26	1.71	2.06	2.48	2.78	3.44	3.71
27	1.71	2.05	2.47	2.77	3.42	3.69
28	1.70	2.05	2.46	2.76	3.40	3.66
29	1.70	2.05	2.46	2.76	3.40	3.66
30	1.70	2.04	2.46	2.75	3.39	3.65
40	1.68	2.02	2.42	2.70	3.31	3.55
60	1.67	2.00	2.39	2.66	3.23	3.46
120	1.66	1.98	2.36	2.62	3.17	3.37
∞	1.64	1.96	2.33	2.58	3.09	3.29
	0.05	0.025	0.01	0.005	0.001	0.0005
	Уровень значимости α (односторонняя критическая область)					

ОГЛАВЛЕНИЕ

1. Программа раздела «Математическая статистика»..... 3
2. Основные теоретические положения..... 4
3. Образцы заданий для контрольной работы..... 19
4. Руководство по выполнению заданий на компьютере..... 24
5. Варианты заданий контрольной работы..... 33

Библиографический список..... 35

Приложения 36

Редактор Л.И.Рубанова

Лицензия № 021231 от 23.07.97
Подписано в печать 06.02.03
Формат 60x84/16.Печать офсетная
Объем в печ. л. - 1,5
План 2003г. Позиция плана 3
Заказ 1163 / 387

Глазковская типография, г.Иркутск, ул.Гоголя, 53