МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Курганский государственный университет»

Кафедра фундаментальной математики и методики преподавания математики

Математическая статистика
Часть 2

Материалы для практических занятий и самостоятельной работы
для студентов очной, очно-заочной и заочной форм обучения
37.03.01 «Психология», 37.05.02 «Психология служебной деятельности»

Курган 2018

Кафедра: «фундаментальной математики и методики преподавания математики»

Дисциплина: «Математическая статистика» 37.03.01 «Психология», 37.05.02 «Психология служебной деятельности»

Составил: ст. преподаватель Е .А. Лукерьянова

Утверждены на заседании кафедры «31» августа 2018 г.

Рекомендованы методическим советом университета «20» декабря 2017 г.

Содержание

Введение

Математические методы широко применяются в различных областях науки и техники, поэтому изучение дисциплины «Математическая статистика» важно для студентов, обучающихся по специальности «Психология».

Предметом данной дисциплины являются математические методы обработки количественной информации. Наиболее распространенным методом обработки экспериментальных данных является выборочный метод.

Цель курса «Математическая статистика» – познакомить студентов с методами обработки экспериментальных данных, полученных в результате наблюдений над случайными явлениями или в результате специально поставленных экспериментов; привить им практические навыки анализа данных с целью получения выводов, характеризующих изучаемые случайные явления.

Основной задачей курса должно стать овладение методами организации сбора, обработки данных статистического наблюдения.

Особая роль в подготовке студентов к профессиональной деятельности принадлежит самостоятельной работе, организуемой в процессе обучения. Настоящие материалы предназначены для организации самостоятельной работы по изучению курса «Математическая статистика».

Материалы для практических занятий и самостоятельной работы включают в себя планы занятий, вопросы для экзамена. Планы занятий содержат вопросы для повторения, задачи для решения в аудитории различного уровня сложности, задачи для самостоятельного решения.

Краткое содержание дисциплины

1 Выборки и их характеристики.

2 Статистическая оценка параметров распределения.

3 Проверка статистических гипотез.

4 Корреляционно регрессионный анализ.

Материалы для практических занятий и самостоятельной работы составлено в соответствии с учебным планом по дисциплине «Математическая статистика».

Раздел 3. Проверка статистических гипотез

Тема 5-6. Понятие статистической гипотезы. Виды статистических гипотез. Проверка гипотезы о нормальном распределении генеральной совокупности. Критерий согласия Пирсона

Теоретическая справка

Статистической называют гипотезу о виде неизвестного распределения или о параметрах известных распределений.

Нулевой (основной) называют выдвинутую гипотезу H_0.

Конкурирующей (альтернативной) называют гипотезу H_1, которая противоречит нулевой.

Различают гипотезы, которые содержат одно и более одного предположений.

Простой называют гипотезу, содержащую только одно предположение.

Сложной называют гипотезу, которая состоит из конечного или бесконечного числа простых гипотез.

В итоге проверки гипотезы могут быть допущены ошибки двух родов.

Ошибка первого рода состоит в том, что будет отвергнута правильная нулевая гипотеза. Вероятность ошибки первого рода называют *уровнем значимости* и обозначают через α.

Ошибка второго рода состоит в том, что будет принята неправильная нулевая гипотеза. Вероятность ошибки второго рода обозначают через β.

Статическим критерием (или просто *критерием*) называют случайную величину K, которая служит для проверки гипотезы.

Наблюдаемым (*эмпирическим*) значением $K_{набл}$ называют то значение критерия, которое вычислено по выборкам.

Критической областью называют совокупность всех значений критерия, при которых нулевую гипотезу отвергают.

Областью принятия гипотезы (*областью допустимых значений*) называют совокупность значений критерия, при которых нулевую гипотезу принимают.

Основной принцип проверки статистических гипотез: если наблюдаемое значение критерия принадлежит критической области, то нулевую гипотезу отвергают; если наблюдаемое значение критерия принадлежит области принятия гипотезы, то гипотезу принимают.

Критическими точками (*границами*) $K_{кр}$ называют точки, отделяющие критическую область от области принятия гипотезы.

Правосторонней называют критическую область, определяемую неравенством $K > K_{кр}$, где $K_{кр}$ – положительное число.

Левосторонней называют критическую область, определяемую неравенство $K < K_{кр}$, где $K_{кр}$ – отрицательное число.

Двусторонней называют критическую область, определяемую неравенством $K < k_1$, $K > k_2$, где $k_1 > k_2$. В частности, если критические точки симметричны относительно нуля, то двусторонняя критическая область определяется неравенствами (в предложении, что $k_{кр} > 0$)

$K < -k_{кр}$, $K > k_{кр}$, или равносильным неравенством $|K| > k_{кр}$.

Для отыскания критической области задаются уровнем значимости α и ищут критические точки, исходя из следующих соотношений:

а) для правосторонней критической области:

$$P(K > k_{кр}) = \alpha (k_{кр} > 0);$$

б) для левосторонней критической области:

$$P(K < k_{кр}) = \alpha (k_{кр} < 0);$$

в) для двусторонней симметричной области:

$$P(K > k_{кр}) = (\alpha/2)(k_{кр} > 0), P(K < -k_{кр}) = \alpha/2.$$

Мощностью критерия называют вероятность попадания критерия в критическую область при условии, что справедлива конкурирующая гипотеза. Другими словами, мощность критерия есть вероятность того, что нулевая гипотеза будет отвергнута, если верна конкурирующая гипотеза.

Сравнение дисперсий двух генеральных совокупностей

По независимым выборкам, объемы которых n_1, n_2, извлеченным из нормальных генеральных совокупностей, найдены исправленные выборочные дисперсии s_x^2 и s_y^2. Требуется сравнить эти дисперсии.

Правило 1. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0: D(x) = D(Y)$ о равенстве генеральных дисперсий нормальных совокупностей при конкурирующей гипотезе H_1: $D(X) > D(Y)$, надо вычислить наблюдаемое значение критерия (отношение большей исправленной дисперсии к меньшей):

$$F_{набл} = s_Б^2 / s_М^2.$$

И по таблице критических точек распределения Фишера – Снедекора, по заданному уровню значимости α и числам степеней свободы

$k_1 = n_1 - 1, k_2 = n_2 - 1$ (k_1 – число степеней большей исправленной дисперсии) найти критическую точку $F_{кр}(\alpha; k_1; k_2)$. Если $F_{набл} < F_{кр}$ – нет оснований отвергнуть нулевую гипотезу. Если $F_{набл} > F_{кр}$ – нулевую гипотезу отвергают.

Правило 2. При конкурирующей гипотезе $H_1: D(x) \neq D(Y)$ критическую точку $F_{кр}(\alpha/2; k_1; k_2)$ ищут по уровню значимости $\alpha/2$ (вдвое меньше заданного) и числам степеней свободы
k_1 и k_2 (k_1 – число степеней сводобы большей дисперсии).

Если $F_{набл} < F_{кр}$ – нет оснований отвергнуть нулевую гипотезу.

Если $F_{набл} > F_{кр}$ – нулевую гипотезу отвергают.

Пример. По двум независимым выборкам, объемы которых

$n_1 = 11$ и $n_2 = 14$, извлеченным из нормальных генеральных совокупностей X и У, найдены исправленные выборочные дисперсии $s_X^2 = 0{,}76$ и $s_Y^2$0,38. При уровне значимости $\alpha = 0{,}05$, проверить нулевую гипотезу $H_0: D(x) = D(Y)$ о равенстве генеральных дисперсий при конкурирующей гипотезе

$$H_1: D(X) > D(Y).$$

Решение. Найдем отношение большей исправленной дисперсии к меньшей:

$$F_{набл} = 0{,}76/0{,}38 = 2.$$

По условию конкурирующая гипотеза имеет вид $D(X) > D(Y)$, поэтому критическая область — правосторонняя.
По таблице критических точек, по уровню значимости $\alpha = 0{,}05$ и числам степеней свободы $k_1 = n_1 - 1 = 11 - 1 = 10$ и $k_2 = n_2 - 1 = 14 - 1 = 13$
находим критическую точку: $F_{кр} = (0{,}05; 10; 13) = 2{,}67$.

Так как $F_{набл} < F_{кр}$ — нет оснований отвергнуть гипотезу о равенстве генеральных дисперсий. Другими словами, выборочные исправленные дисперсии различаются незначимо.

Проверка гипотезы о нормальном распределении генеральной совокупности

Пусть эмпирическое распределение задано в виде последовательности равноотстоящих вариант и соответствующих им частот:

x_i	x_1	x_2	...	x_N
n_i	n_1	n_2	...	n_N

Требуется, используя критерий Пирсона, проверить гипотезу о том, что генеральная совокупность *X* распределена нормально.

Правило 1. Для того чтобы при заданном уровне значимости α проверить гипотезу о нормальном распределении генеральной совокупности, надо:

1) вычислить непосредственно (при малом числе наблюдений) или упрощенным методом (при большом числе наблюдений), например методом произ-

ведений или сумм, выборочную среднюю $\overline{x_B}$ и выборочное среднее квадратическое отклонение σ_B;

2) вычислить теоретические частоты

$$n_i = \frac{nh}{\sigma_B} \cdot \varphi(u_i),$$

где n – объем выборки (сумма всех частот), h — шаг (разность между двумя соседними вариантами),

$$u_i = \frac{x_i - \overline{x_B}}{\sigma_B}, \quad \varphi(u) = \frac{1}{\sqrt{2\pi}} \cdot e^{-u^2/2};$$

3) сравнить эмпирические и теоретические частоты с помощью критерия Пирсона. Для этого:

а) составляют расчетную таблицу, по которой находят наблюдаемое значение критерия

$$\chi^2_{\text{набл}} = \sum \frac{(n_i - n_i')^2}{n_i'};$$

б) по таблице критических точек распределения χ^2, по заданному уровню значимости α и числу степеней свободы k = s—3 (s — число групп выборки) находят критическую точку $\chi^2_{\text{кр}}(\alpha; k)$ правосторонней критической области.

Если $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}$ — нет оснований отвергнуть гипотезу о нормальном распределении генеральной совокупности. Другими словами, эмпирические и теоретические частоты различаются незначимо (случайно).

Если $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$ — гипотезу отвергают. Следовательно, эмпирические и теоретические частоты различаются значимо.

Пусть эмпирическое распределение задано в виде последовательности интервалов (x_i, x_{i+1}) и соответствующих им частот n_i (n_i — сумма частот, которые попали в *i-й* интервал):

(x_1, x_2)	(x_2, x_3) …	(x_s, x_{s+1})
n_1	n_2	n_s

Требуется, используя критерий Пирсона, проверить гипотезу о том, что генеральная совокупность *X* распределена нормально.

Правило 2. Для того чтобы при уровне значимости α, проверить гипотезу о нормальном распределении генеральной совокупности, надо:

1) вычислить, например методом произведений, выборочную среднюю $\bar{x}$ и выборочное среднее квадратическое отклонение σ_B, причем в качестве вариант x_i^* принимают среднее арифметическое концов интервала:

$$x_i^* = (x_i + x_{i+1})/2;$$

2) пронормировать X, т. е. перейти к случайной величине

Z = (X$-\overline{x^*}$)$/\sigma^*$, и вычислить концы интервалов: $z_i = (x_i - \overline{x^*})/\sigma^*$,

$z_{i+1} = (x_i - \overline{x^*})/\sigma^*$, причем наименьшее значение Z, т. е. z_1, полагают равным – ∞, а наибольшее, т. е. z_{s+1}, полагают равным ∞;

3) вычислить теоретические частоты

$$n_i' = n \cdot P_i,$$

где n — объем выборки (сумма всех частот); $P_i = \Phi(z_{i+1}) - \Phi(z_i)$ — вероятности попадания X в интервалы (x_i, x_{i+1}); Ф (Z) — функция Лапласа;

4) сравнить эмпирические и теоретические частоты с помощью критерия Пирсона. Для этого:

а) составляют расчетную таблицу по которой находят наблюдаемое значение критерия Пирсона

$$\chi^2_{\text{набл}} = \sum (n_i - n_i')^2 / n_i'$$

б) по таблице критических точек распределения χ^2, по заданному уровню значимости α и числу степеней свободы k = s — 3 (s — число интервалов выборки) находят критическую точку правосторонней критической области $\chi^2_{\text{кр}}(\alpha; k)$.

Пример 1. Используя критерий Пирсона, при уровне значимости 0,05, проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X с эмпирическим распределением выборки объема n=200(таблица 1).

Таблица 1–Эмпирическое распределение выборки объема n=200

x_i	4	6	8	10	12	14	16	18	20
n_i	15	20	31	20	30	27	24	20	13

Решение. Используя метод произведений, найдем $\bar{x}_B$ и σ_B. Пусть C = 12, h=2.

Таблица 2 – Эмпирическое распределение выборки объема n=200

x_i	n_i	u_i	$n_i \cdot u_i$	$n_i \cdot u_i^2$	$n_i \cdot (u_i + 1)^2$
4	15	−4	−60	240	135
6	20	−3	−60	180	80
8	31	−2	−62	124	31
10	20	−1	−20	20	0
12	30	0	0	0	30
11	27	1	27	27	108
16	24	2	48	96	216
18	20	3	60	180	320
20	13	4	52	208	325
	200		15	1075	1245

Контроль:

$$\sum_{i=1}^{9} n_i \cdot (x_i + 1)^2 = 1245;$$

$$\sum_{i=1}^{9} n_i \cdot x_i^2 + 2\sum_{i=1}^{9} n_i \cdot x_i + n = 1075 - 30 + 200 = 1245;$$

$$M_1^* = \frac{-15}{100} = -0{,}15,\ M_2^* = \frac{1075}{100} = 10{,}75;$$

$$\bar{x}_B = M_1^* \cdot h + c,\ \bar{x}_B = -0{,}15 \cdot 2 + 12 = -0{,}3 + 12 = 11{,}7;$$

$$D_B = (M_2^* - (M_1^*)^2) \cdot h^2,$$

$$D_B = (10{,}75 - (-0{,}15)^2) \cdot 4 = (10{,}75 - 0{,}0225) \cdot 4 = 42{,}91;$$

$$\sigma_B = 6{,}6, \text{по формуле } n_i^1 = \frac{n \cdot h}{\sigma_B} \cdot \varphi(u_i) = 60{,}61 \cdot \varphi(u_i).$$

Составим расчетную таблицу. Значение функции $\varphi(u_i)$ найдем в приложении А (Таблица А1).

Таблица 3 – Эмпирическое распределение выборки объема n=200

i	x_i	$u_i = \frac{x_i - \bar{x}_B}{\sigma_B}$	$\varphi(u_i)$	$n_i^1 = 60{,}61 \cdot \varphi(u_i)$
1	4	−1,17	0,2012	12,2
2	6	−0,87	0,2756	16,7
3	8	−0,56	0,3410	20,7
4	10	−0,26	0,3857	23,4
5	12	0,05	0,3984	24,1
6	14	0,35	0,3867	23,4
7	16	0,65	0,3230	19,6
8	18	0,95	0,2541	15,4
9	20	1,26	0,1804	10,9

Сравним эмпирические и теоретические частоты:

а) Составим таблицу, из которой найдем наблюдаемое значение критерия

$$x^2_{набл} = \sum_i \frac{(n_i - n_i^1)^2}{n_i^1};$$

Таблица 4 – Эмпирическое распределение выборки объема n=200

i	n_i	n_i^1	$n_i - n_i^1$	$(n_i - n_i^1)^2$	$\frac{(n_i - n_i^1)^2}{n_i^1}$
1	15	12,2	2,8	7,84	0,64
2	20	16,7	3,3	10,89	0,65
3	31	20,7	10,3	106,09	5,1
4	20	23,4	−3,4	11,56	0,49
5	30	24,1	5,9	34,81	1,4
6	27	23,4	3,6	12,96	0,6
7	24	19,6	4,4	19,36	0,39
8	20	15,4	4,6	21,16	1,37
9	13	10,9	2,1	4,41	0,4
					$x^2_{набл} = 11{,}04$

б) по таблице критических точек распределения x^2 (приложение А4) по уровню значимости $\alpha = 0{,}05$ и числу степеней свободы $k = s - 3 = 9 - 3 = 6$ находим критическую точку правосторонней критической области $x^2_{кр}(0{,}05/6) = 12{,}6$. Так как $x^2_{набл} < x^2_{кр}$ – гипотезу о нормальном распределении генеральной совокупности нет оснований отвергать.

Найдем по таблице критических точек распределения x^2 (приложение А4) по уровню значимости $\alpha = 0{,}05$ и числу степеней свободы k = 7 критическую точку правосторонней критической области:

$$x^2_{кр.}(0{,}05;7)=14{,}1.$$

Так как $x^2_{набл.} > x^2_{кр.}$, гипотезу о равномерном распределении X отвергаем.

Вопросы для повторения

1 Статистическая гипотеза. Примеры.

2 Основная и альтернативная гипотезы. Примеры.

3 Простые и сложные гипотезы. Примеры.

4 Статистический критерий проверки гипотезы. Наблюдаемое значение критерия. Критерий согласия.

5 Ошибки первого и второго рода.

6 Теоретические и выравнивающие частоты.

7 Проверки гипотезы о нормальном распределении признака по критерию χ^2 Пирсона.

Задачи для решения в аудитории

1 При измерении роста 359 школьников получили результаты, представленные в таблице 5.

Таблица 5 – Рост 359 школьников

x_i	154-158	158-162	162-166	166-170	170-174	174-178	178-182
n_i	45	60	36	27	111	50	30

Можно ли предполагать, что распределение школьников по росту мало отличается от нормального?

2 В результате выборочного обследования стажа работы рабочих фермерского хозяйства получены следующие данные (таблица 6).

Таблица 6 – Стаж работы рабочих фермерского хозяйства

Стаж работы, год	0-4	4-8	8-12	12-16	16-20	20-24	24-28	28-32
Число раб.	1	4	13	20	23	15	3	1

Выяснить, является ли распределение стажа работы нормальным. Найти процент рабочих со стажем работы от 15 до 20 лет.

3 Предполагается, что случайная величина, эмпирическое распределение которой задано в таблице 7, обладает нормальным законом распределения. Приняв за математическое ожидание и дисперсию случайной величины, соответственно, среднее арифметическое и дисперсию признака, требуется:

1) вычислить для всех имеющихся в таблице 7 интервалов значений соответствующие вероятности и теоретические частоты;

2) построить в одной системе координат по данному эмпирическому распределению полигон и гистограмму, а также теоретическую кривую нормального распределения частот;

3) по критерию Пирсона оценить согласованность данного эмпирического распределения случайной величины с соответствующим теоретическим распределением, построенным по нормальному закону;

4) построить в одной системе координат графики эмпирической и теоретической функций распределения.

Таблица 7 – Группировка промышленных предприятий области по размеру роста валовой продукции

Валовая продукция в отчетном году в % к предыдущему году	80-90	90-100	100-110	110-120	120-130	130-140
Число предприятий	6	15	34	29	18	15

Задачи для самостоятельного решения

В таблице 8 приведены данные о суточной погрузке вагонов на железнодорожной сети за 80 дней июня-августа 2010 г.

Таблица 8 – Данные о суточной погрузке вагонов на железнодорожной сети

Погрузка в %	Число дней
91-93	2
93-95	7
95-97	17
97-99	22
99-101	19
101-103	10
103-105	2

Предполагая, что случайная величина имеет нормальное распределение, требуется найти:

1) общее выражение плотности вероятности и интегральной функции распределения;

2) с помощью критерия χ^2 Пирсона установить, согласуются ли эмпирические данные обследования с предположением о нормальном распределении рассматриваемой случайной величины;

3) вычислить вероятности, соответствующие каждому интервалу.

Раздел 4. Корреляционно-регрессионный анализ

Тема 7- 8 Корреляционная зависимость. Линейная корреляция

Теоретическая справка

Во многих задачах требуется установить и оценить зависимость изучаемой случайной величины Y от одной или нескольких других величин.

Две случайные величины могут быть связаны либо функциональной зависимостью, либо зависимостью другого рода, называемой статистической, либо быть независимыми.

Строгая функциональная зависимость реализуется редко, так как обе величины или одна из них подвержены еще действию случайных факторов, причем среди них могут быть и общие для обеих величин (под общими здесь подразумеваются такие факторы, которые воздействуют и на Y и на X). В этом случае возникает статистическая зависимость.

Например, если Y зависит от случайных факторов Z_1, Z_2, V_1, V_2, а X зависит от случайных факторов Z_1, Z_2, U_1, то между Y и X имеется статистическая зависимость, так как среди случайных факторов есть общие, а именно Z_1 и Z_2.

Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения другой. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой; в этом случае статистическую зависимость называют *корреляционной.*

Приведем пример случайной величины Y, которая не связана с величиной X функционально, а связана корреляционно. Пусть Y – урожай зерна,
X – количество удобрений. С одинаковых по площади участков земли при равных количествах внесенных удобрений снимают различный урожай, т. е. Y не является функцией от X. Это объясняется влиянием случайных факторов (осадки, температура воздуха и др.). Вместе с тем, как показывает опыт, средний урожай является функцией от количества удобрений, т.е. Y связан с X корреляционной зависимостью.

Определение 1. Две случайные величины X и Y находятся в *корреляционной* зависимости, если каждому значению любой из этих величин соответствует определенное распределение вероятностей другой величины.

Определение 2. *Условным математическим ожиданием* (кратко *УМО)* дискретной случайной величины X при $Y = y$ *(y* – определенное возможное значение *Y)* называется сумма произведений возможных значений величины X на их условные вероятности.

$$M_y(X) = \sum_{i=1}^{n} x_i P_y(X = x_i),$$

где $P_y(X = x_i)$ – условная вероятность равенства $X = x_i$ при условии, что $Y = y$. Для непрерывных величин $M_y(X) = \int_{-\infty}^{+\infty} x\varphi_y(x)dx,$

где $\varphi_y(x)$ – плотность вероятности случайной непрерывной величины X при условии $Y=y$.

УМО $M_y(X)$ есть функция от y : $M_y(X) = f(y)$, которую называют функцией *регрессии* величины X на величину Y.

Аналогично определяется *УМО* случайной величины Y и функция регрессии Y на X:

$$M_x(Y) = g(x).$$

Уравнение $x = f(y)\,(y = g(x))$ называется *уравнением регрессии* X на Y *(Y* на X), а линия на плоскости, соответствующая этому уравнению, называется *линией регрессии.*

Линия регрессии Y на X (X на Y) показывает, как в среднем зависит Y от X *(* X от Y*).*

Для характеристики корреляционной зависимости между случайными величинами вводится понятие коэффициента корреляции.

Определение 3. Если X и Y – независимые случайные величины, то

$$M(XY) = M(X) \cdot M(Y).$$

Если же X и Y не являются независимыми случайными величинами, то, вообще говоря, $M(XY) \neq M(X) \cdot M(Y)$.

Условились за меру связи (зависимости) двух случайных величин X и Y принять безразмерную величину r, определяемую соотношением:

$$r = \frac{M(XY) - M(X)M(Y)}{\sigma(X)\cdot\sigma(Y)}$$

или более кратно соотношением:

$$r = \frac{\mu}{\sigma_1\sigma_2},$$

где $\mu = M(XY) - M(X) \cdot M(Y), \sigma_1 = \sigma(X), \sigma_2 = \sigma(Y),$

и называемую *коэффициентом корреляции.*

Легко видеть, что

$$\mu = M(XY) - M(X) \cdot M(Y) = M[(X - M(X))(Y - M(Y))].$$

О п р е д е л е н и е 4. Случайные величины X и Y называются *некоррелированными,* если $r = 0$, и *коррелированными,* если $r \neq 0$.

Отметим некоторые свойства коэффициента корреляции.

1 Если X и Y – независимые случайные величины, то коэффициент корреляции равен нулю. Заметим, что обратное утверждение, вообще говоря, неверно.

2 Укажем без доказательства, что $|r| < 1$. При этом если $|r| = 1$, то между случайными величинами X и Y имеет место функциональная, а именно линейная зависимость.

3 Коэффициент корреляции характеризует относительную величину отклонения математического ожидания произведения $M(XY)$ от произведения математических ожиданий $M(X)M(Y)$ величин X и Y. Так как это отклонение имеет место только для зависимых величин, то можно сказать, что коэффициент корреляции характеризует тесноту зависимости между X и Y (таблица 9).

Таблица 9 – Количественный критерий оценки тесноты зависимости

Величина коэффициента корреляции	Характер связи
До $\|\pm 0{,}3\|$	Практически отсутствует
$\|\pm 0{,}3\| - \|\pm 0{,}5\|$	Слабая
$\|\pm 0{,}5\| - \|\pm 0{,}7\|$	Умеренная
$\|\pm 0{,}7\| - \|\pm 1\|$	Сильная

О п р е д е л е н и е 5. Корреляционная зависимость между случайными величинами X и Y называется *линейной корреляцией,* если обе функции регрессии $f(y)$ и $g(x)$ являются линейными. В этом случае обе линии регрессии являются прямыми; они называются *прямыми регрессии.*

Запишем уравнение прямой регрессии Y на X, т. е. найдем коэффициенты линейной функции $g(x) = Ax + B$.

Обозначим $M(X) = a$, $M(Y) = b$, $M[(X-a)^2] = \sigma_1^2$, $M[(Y-b)^2] = \sigma_2^2$, $A = \frac{\mu}{\sigma_1^2}$

Полученный коэффициент называется *коэффициентом регрессии* Y на X и обозначается через $p(Y/X)$:

$$p(Y/X) = \frac{\mu}{\sigma_1^2}.$$

Таким образом, уравнение прямой регрессии Y на X имеет вид:

$$y = p(Y/X)(x - a) + b$$

Аналогично можно получить уравнение прямой регрессии X на Y:

$$y = p(X/Y)(y - b) + a,$$

где коэффициент регрессии X на Y.

$$p(X/Y) = \frac{\mu}{\sigma_2^2}$$

Уравнения прямых регрессии можно записать в более симметричном виде, если воспользоваться коэффициентом корреляции. С учетом этого коэффициента

$$p\left(\frac{Y}{X}\right) = r\frac{\sigma_2}{\sigma_1},\ p\left(\frac{X}{Y}\right) = r\frac{\sigma_1}{\sigma_2},$$

и поэтому уравнения прямых регрессии принимают вид

$$y - b = r\frac{\sigma_2}{\sigma_1}(x - a), x - a = r\frac{\sigma_1}{\sigma_2}(y - b).$$

Из уравнений прямых регрессии видно, что обе эти прямые проходят через точку $(a;b)$; угловые коэффициенты прямых регрессии равны, соответственно, обозначения углов представлены на рисунке 1.

$$\mathrm{tg}\alpha = r\frac{\sigma_2}{\sigma_1},\ \mathrm{tg}\beta = \frac{1}{r}\cdot\frac{\sigma_2}{\sigma_1},$$

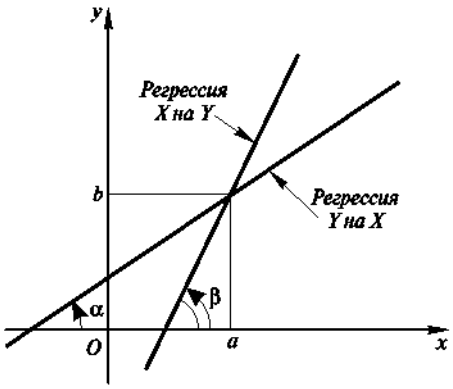

Рисунок 1 – Прямые линии регрессии

Так как $|r| \leq 1$, то $|\mathrm{tg}\alpha| \leq |\mathrm{tg}\beta|$. Это означает, что прямая регрессии Y на X имеет меньший наклон к оси абсцисс, чем прямая регрессии X на Y. Чем ближе $|r|$ к 1, тем меньше угол между прямыми регрессии. Эти прямые сливаются тогда и только тогда, когда $|r| = 1$.

При $r = 0$ прямые регрессии имеют уравнения $y = b$; $x = a$.

В этом случае $M_x(Y) = b = M(Y)$; $M_y(X) = a = M(X)$.

Коэффициенты регрессии имеют тот же знак, что и коэффициент корреляции r, и связаны соотношением:

$$p(Y/X)p(X/Y) = r^2.$$

Пусть проведено n опытов, в результате которых получены следующие значения системы величин $(X; Y)$: (x_i, y_i), $i = 1, 2, ..., n$. За приближенные значения $M(X)$, $M(Y)$, $D(X)$ и $D(Y)$ принимают их выборочные значения:

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i, \ \bar{y}_в = \frac{1}{n}\sum_{i=1}^{n} y_i, \ s_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_в)^2,$$

$$s_2^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y}_в)^2.$$

Оценкой для μ служит величина

$$\mu_в = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_в)(y_i - \bar{y}_в).$$

Заменяя в соотношениях величины μ, σ_1, σ_2 их выборочными значениями $\mu_в$, s_1, s_2, получим приближенные значения коэффициента корреляции и коэффициентов регрессии:

$$r \approx \frac{\mu_в}{s_1 s_2}, \ p(Y/X) \approx \frac{\mu_в}{s_1^2}, \ p(X/Y) \approx \frac{\mu_в}{s_2^2}$$

($\frac{\mu_в}{s_1 s_2}$ и $\frac{\mu_в}{s_1^2}, \frac{\mu_в}{s_2^2}$ – выборочные коэффициенты соответственно корреляции и регрессии).

Подставляя в уравнения вместо a, b, $p(Y/X)$ и $p(X/Y)$ их приближенные значения, получим выборочные уравнения прямых регрессии: $y - \bar{y}_в = \frac{\mu_в}{s_1^2}(x - \bar{x}_в)$, $x - \bar{x}_в = \frac{\mu_в}{s_2^2}(y - \bar{y}_в)$

При большом числе наблюдений одно и то же значение x может встретиться n_x раз, одно и то же значение $y - n_y$ раз, одна и та же пара чисел (x, y) может наблюдаться n_{xy} раз. Поэтому данные наблюдений группируют,
т. е. подсчитывают частоты n_x, n_y, n_{xy}. Все сгруппированные данные записывают в виде таблицы, которую называют *корреляционной* (таблица 10).

Таблица 10 – Корреляционная таблица

Y	X				
	10	20	30	40	n_v
0.4	5	-	7	14	26
0.6	-	2	6	4	12
0.8	3	19	-	-	22
n_x	8	21	13	18	n= 60

В первой строке таблицы 10 указаны наблюдаемые значения (10; 20; 30; 40) признака *X*, а в первом столбце – наблюдаемые значения (0,4; 0,6; 0,8) признака *Y*. На пересечении строк и столбцов находятся частоты n_{xy} наблюдаемых пар значений признаков. Например, частота 5 указывает, что пара чисел (10; 0,4) наблюдалась 5 раз. Все частоты помещены в прямоугольнике. «Черточка» означает, что соответственная пара чисел, например (20; 0,4), не наблюдалась.

В последнем столбце записаны суммы частот строк. Например, сумма частот первой строки «жирного» прямоугольника равна $n_y = 5 + 7 + 14 = 26$; это число указывает, что значение признака *Y*, равное 0,4 (в сочетании с различными значениями признака *X*), наблюдалось 26 раз.

В последней строке записаны суммы частот столбцов. Например, число 8 указывает, что значение признака *X*, равное 10 (в сочетании с различными значениями признака *Y*), наблюдалось 8 раз.

В клетке, расположенной в нижнем правом углу таблицы, помещена сумма всех частот (общее число всех наблюдений n). Очевидно, $\sum n_x = \sum n_y = n$.

В нашем примере

$$\sum n_x = 8 + 21 + 13 + 18 = 60 \text{ и } \sum n_y = 26 + 12 + 22 = 60.$$

Выборочный коэффициент корреляции определяется равенством

$$r_в = \frac{\sum n_{xy} xy - n\bar{x}\bar{y}}{n\bar{\sigma}_x\bar{\sigma}_y},$$

где x, y – варианты (наблюдавшиеся значения) признаков *X* и *Y*; n_{xy} – частота пары вариант (x, y); n – объем выборки (сумма всех частот); $\bar{\sigma}_x$, $\bar{\sigma}_y$ – выборочные средние квадратические отклонения; $\bar{x}$, $\bar{y}$ – выборочные средние.

Выборочный коэффициент корреляции $r_в$ является оценкой коэффициента корреляции r генеральной совокупности и поэтому также служит для измере-

ния линейной связи между величинами – количественными признаками Y и X. Допустим, что выборочный коэффициент корреляции, найденный по выборке, оказался отличным от нуля. Так как выборка отобрана случайно, то отсюда еще нельзя заключить, что коэффициент корреляции генеральной совокупности также отличен от нуля. Возникает необходимость проверить гипотезу о значимости (существенности) выборочного коэффициента корреляции (или, что то же, о равенстве нулю коэффициента корреляции генеральной совокупности). Если гипотеза о равенстве нулю генерального коэффициента корреляции будет отвергнута, то выборочный коэффициент корреляции значим, а величины X и Y коррелированы; если гипотеза принята, то выборочный коэффициент корреляции незначим, а величины X и Y не коррелированы.

Если выборка имеет достаточно большой объем и хорошо представляет генеральную совокупность (репрезентативна), то заключение о тесноте линейной зависимости между признаками, полученное по данным *выборки,* в известной степени может быть распространено и на *генеральную совокупность.* Например, для оценки коэффициента корреляции $r_г$ нормально распределенной генеральной совокупности (при $n \geq 50$) можно воспользоваться формулой:

$$r_в - 3\frac{1-r_в^2}{\sqrt{n}} \leq r_г \leq r_в + 3\frac{1+r_в^2}{\sqrt{n}}$$

Замечание 1. Знак выборочного коэффициента корреляции совпадает со знаком выборочных коэффициентов регрессии, что следует из формул:

$$\rho_{yx} = r_в\frac{\bar{\sigma}_y}{\bar{\sigma}_x};\ \rho_{yx} = r_в\frac{\bar{\sigma}_x}{\bar{\sigma}_y}.$$

Требуется по данным корреляционной таблицы вычислить выборочный коэффициент корреляции. Можно значительно упростить расчет, если перейти к условным вариантам (при этом величина $r_в$ не изменится):

$$u_i = (x_i - C_1)/h_1 \text{ и } v_j = (y_j - C_2)/h_2.$$

В этом случае выборочный коэффициент корреляции вычисляют по формуле:

$$r_в = (\sum n_{uv}uv - n\bar{u}\,\bar{v})/(n\bar{\sigma}_u\bar{\sigma}_v).$$

Величины $\bar{u}$, $\bar{v}$, $\bar{\sigma}_u$ и $\bar{\sigma}_v$ можно найти методом произведений, а при малом числе данных – непосредственно, исходя из определений этих величин. Остается указать способ вычисления $\sum n_{uv}uv$, где n_{uv} – частота пары условных вариант (u, v).

Можно доказать, что справедливы формулы:

$$\sum n_{uv}uv = \sum vU, \text{ где } U = \sum n_{uv}u,$$

$$\sum n_{uv}uv = \sum uV, \text{ где } V = \sum n_{uv}v,$$

Для контроля целесообразно выполнить расчеты по обеим формулам и сравнить результаты; их совпадение свидетельствует о правильности вычислений.

Покажем на примере, как пользоваться приведенными формулами.

Пример 1. Вычислить $\sum n_{uv}uv$ по данным корреляционной таблицы 11.

$u_i = \frac{x_i - C_1}{h_1} = (x_i - 40)/10$ (в качестве ложного нуля C_1 взята варианта $x = 40$, расположенная примерно в середине вариационного ряда; шаг h_1 равен разности между двумя соседними вариантами: $20 - 10 = 10$) и $v_j = (y_j - C_2)/h_2 = (y_j - 35)/10$ (в качестве ложного нуля C_2 взята варианта $y = 35$, расположенная в середине вариационного ряда; шаг h_2 равен разности между двумя соседними вариантами: $25 - 15 = 10$).

Таблица 11 – Корреляционная таблица

Y	X						n_y
	10	20	30	40	50	60	
15	5	7	—	—	—	—	12
25	—	20	23	—	—	—	43
35	—	—	30	47	2	—	79
45	—	—	10	11	20	6	47
55	—	—	—	9	7	3	19
n_x	5	27	63	67	29	9	$n = 200$

Решение. Перейдем к условным вариантам:

Составим корреляционную таблицу в условных вариантах. Практически это делают так: в первом столбце вместо ложного нуля C_2 (варианты 35) пишут 0; над нулем последовательно записывают -1, -2; под нулем пишут 1, 2. В первой строке вместо ложного нуля C_1 (варианты 40) пишут 0; слева от нуля последовательно записывают -1, -2, -3; справа от нуля пишут 1, 2. Все остальные данные переписывают из первоначальной корреляционной таблицы. В итоге получим корреляционную таблицу 12 в условных вариантах.

Таблица 12 – Таблица условных вариант

v	u						n_v
	−3	−2	−1	0	1	2	
−2	5	7	—	—	—	—	12
−1	—	20	23	—	—	—	43
0	—	—	30	47	2	—	79
1	—	—	10	11	20	6	47
2	—	—	—	9	7	3	19
n_u	5	27	63	67	29	9	$n=200$

Теперь для вычисления искомой суммы $\sum n_{uv}uv$ составим расчетную таблицу 13.

Таблица – 13

v	u						$U=\sum n_{uv}u$	vU
	−3	−2	−1	0	1	2		
−2	5 (−15) (−10)	7 (−14) (−14)	—	—	—	—	−29	58
−1	—	20 (−40) (−20)	23 (−23) (−23)	—	—	—	−63	63
0	—	—	30 (−30) (0)	47 (0) (0)	2 (2) (0)	—	−28	0
1	—	—	10 (−10) (10)	11 (0) (11)	20 (20) (20)	6 (12) (6)	22	22
2	—	—	—	9 (0) (18)	7 (7) (14)	3 (6) (6)	13	26
$V=\sum n_{uv}v$	−10	−34	−13	29	34	12		$\sum_v vU=169$
uV	30	68	13	0	34	24	$\sum_u uV=169$	←Контроль↑

1 В каждой клетке, в которой частота $n_{uv} \neq 0$, записывают в правом верхнем углу произведение частоты n_{uv} на варианту u. Например, в правых верхних углах клеток первой строки записаны произведения: $5 \cdot (-3) = -15$; $7 \cdot (-2) = -14$.

2 Складывают все числа, помещенные в правых верхних углах клеток одной строки и их сумму записывают в клетку этой же строки столбца U. Например, для первой строки $U = -15 + (-14) = -29$.

3 Умножают варианту v на U и полученное произведение записывают в последнюю клетку той же строки, т. е в клетку столбца vU.

Например, в первой строке таблицы $v = -2$, $U = -29$; следовательно, $vU = (-2)\cdot(-29) = 58$.

4 Наконец, сложив все числа столбца vU, получают сумму $\sum_v vU$, которая равна искомой сумме $\sum n_{uv}uv$. Например, для табл. 16 имеем $\sum_v vU = 169$; следовательно, искомая сумма $\sum n_{uv}uv = 169$.

Для контроля аналогичные вычисления производят по столбцам: произведения $n_{uv}v$ записывают в левый нижний угол клетки, содержащей частоту $n_{uv} \neq 0$; все числа, помещенные в левых нижних углах клеток одного столбца, складывают и их сумму записывают в строку V; далее умножают каждую варианту u на V и результат записывают в клетках последней строки.

Наконец, сложив все числа последней строки, получают сумму $\sum_u uV$, которая также равна искомой сумме $\sum n_{uv}uv$. Например, для таблицы 13 имеем $\sum_u uV = 169$; следовательно, $\sum n_{uv}uv = 169$.

Теперь, когда мы научились вычислять $\sum n_{uv}uv$, приведем пример на отыскание выборочного коэффициента корреляции.

Пример 2. Вычислить выборочный коэффициент корреляции $r_в = (\sum n_{uv}uv - n\bar{u}\,\bar{v})/(n\bar{\sigma}_u\bar{\sigma}_v)$ по данным корреляционной таблицы 10

Решение. Перейдя к условным вариантам, получим корреляционную таблицу 12. Величины $\bar{u}$, $\bar{v}$, $\bar{\sigma}_u$ и $\bar{\sigma}_v$ можно вычислить методом произведений; однако, поскольку числа u_i, v_i малы, вычислим $\bar{u}$ и $\bar{v}$, исходя из определения средней, а $\bar{\sigma}_u$ и $\bar{\sigma}_v$ – используя формулы:

$$\bar{\sigma}_u = \sqrt{\bar{u}^2 - (\bar{u})^2},\ \bar{\sigma}_v = \sqrt{\bar{v}^2 - (\bar{v})^2}.$$

Найдем $\bar{u}$ и $\bar{v}$:

$$\bar{u} = (\sum n_u u)/n = [5\cdot(-3) + 27\cdot(-2) + 63\cdot(-1) + 29\cdot 1 + 9\cdot 2]/200 =$$
$$-0{,}425\bar{v} = (\sum n_v v)/n = [12\cdot(-2) + 43\cdot(-1) + 47\cdot 1 + 19\cdot 2]/200 = 0{,}09$$

Вычислим вспомогательную величину $\bar{u}^2$, а затем σ_u:

$$\bar{u}^2 = \frac{\sum n_u u^2}{n} = (5\cdot 9 + 27\cdot 4 + 63\cdot 1 + 29\cdot 1 + 9\cdot 2)/200 = 1{,}405.$$

$$\bar{\sigma}_u = \sqrt{\bar{u}^2 - (\bar{u})^2} = \sqrt{1{,}405 - (0{,}425)^2} = 1{,}106.$$

Аналогично получим $\bar{\sigma}_в = 1{,}209$.

Найдем искомый выборочный коэффициент корреляции, учитывая, что ранее уже вычислена $\sum n_{uv}uv = 169$:

$$r_в = (\sum n_{uv}uv - n\bar{u}\,\bar{v})/(n\bar{\sigma}_u\bar{\sigma}_v) = [169 - 200\cdot(-0{,}425)\cdot 0{,}09]/(200\cdot 1{,}106\cdot 1{,}209) = 0{,}603$$

Итак, $r_в = 0{,}603$.

Задача

Найти выборочное уравнение прямой линии регрессии Y на X по данным корреляционной таблицы 14.

Таблица 14 – Корреляционная таблица

X Y	20	25	30	35	40	n_y
16	4	6	-	-	-	10
26	-	8	10	-	-	18
36	-	-	32	3	9	44
46	-	-	4	12	6	22
56	-	-	-	1	5	6
n_x	4	14	46	16	20	n=100

Решение. Выборочное уравнение прямой линии регрессии Y на X имеет вид:

$$\overline{y_x} - \bar{y} = r_B \frac{\sigma_y}{\sigma_x}(x - \bar{x}), \text{где}$$

r_B – выборочный коэффициент корреляции, причем

$$r_B = \frac{\sum_{i=1}^{k} n_{xy} \cdot x \cdot y - n \cdot \bar{x} \cdot \bar{y}}{n \cdot \sigma_x \cdot \sigma_y},$$

$\sigma_x \cdot \sigma_y$ – выборочные средние квадратические отклонения;

$\bar{x} \cdot \bar{y}$ –выборочные средние признаков X и Y;

$\overline{y_x}$ –условная средняя.

Выборочные средние признаков X и Y найдем методом произведений:

$$\bar{x} = M_{1,x}^{*} \cdot h_1 + C_1$$

$$\bar{y} = M_{1,y}^{*} \cdot h_2 + C_2$$

Пусть $C_1 = 30, C_2 = 36.$ Составим вспомогательные таблицы(таблицы 15, 16).

Таблица 15 – Вспомогательная таблица

x_i	n_i	u_i	$n_i \cdot u_i$	$n_i \cdot u_i^2$	$n_i \cdot (u_i + 1)^2$
20	*4*	*-2*	*-8*	*16*	*4*
25	*14*	*-1*	*-14*	*14*	*0*
30	*46*	*0*	*0*	*0*	*46*
35	*16*	*1*	*16*	*16*	*64*
40	*20*	*2*	*40*	*80*	*180*
	100		*34*	*126*	*294*

Контроль:

$$\sum_{i=1}^{k} n_i \cdot (u_i + 1)^2 = 294,$$

$$\sum_{i=1}^{k} n_i \cdot u_i^2 + 2\sum_{i=1}^{k} n_i \cdot u_i + n = 126 + 2 \cdot 34 + 100 = 294.$$

$$M_{1,x}^* = \frac{\sum_{i=1}^{k} n_i \cdot u_i}{n}, \quad M_{1,x}^* = \frac{34}{100} = 0{,}34, h_1 = 5, \quad \bar{x} = 0{,}34 \cdot 5 + 30 = 31{,}7.$$

Таблица 16 – Вспомогательная таблица

y_i	n_i	v_i	$h_i \cdot v_i$	$h_i \cdot v_i^2$	$n_i \cdot (v_i + 1)^2$
16	10	-2	-20	40	10
26	18	-1	-18	18	0
36	44	0	0	0	44
46	22	1	22	22	88
56	6	2	12	24	54
	100		-4	104	196

Контроль:

$$\sum_{i=1}^{k} n_i \cdot (v_1 + 1)^2 = 196,$$

$$\sum_{i=1}^{k} n_i \cdot v_i^2 + 2\sum_{i=1}^{k} n_i \cdot v_i + n = 104 + 2 \cdot (-4) + 100 = 196$$

$$M_{1,y}^* = \frac{\sum_{i=1}^{k} n_i \cdot v_i}{n}, M_{1,y}^* = \frac{-4}{100} = -0{,}04, \bar{y} = -0{,}04 \cdot 10 + 36 = -0{,}4 + 36 = 35{,}6$$

Вычислим σ_x, σ_y:

$$\sigma_x = \sqrt{D_{Bx}}, \sigma_y = \sqrt{D_{By}}.$$

Найдем D_{Bx} и D_{By}. Воспользуемся методом произведений.

$$D_B = (M_2^* - (M_1^*)^2) \cdot h^2;$$

$$M_{2,x}^* = \frac{\sum_{i=1}^{k} n_i \cdot u_i^2}{n}, M_{2,x}^* = \frac{126}{100} = 1{,}26.$$

$$D_{Bx} - (1{,}26 - (0{,}34)^2) \cdot 25 - 28{,}61.$$

$$\sigma_x = \sqrt{28{,}61} = 5{,}35.$$

Аналогично найдем σ_y, $\sigma_y - 10{,}2$.

Вычислим $\sum_{i=1}^{n} n_{x,y} \cdot x \cdot y$.

$$\sum_{i=1}^{n} n_{x,y} = 16 \cdot (20 \cdot 4 + 25 \cdot 6) + 26 \cdot (8 \cdot 25 + 10 \cdot 30) + 36 \cdot (32 \cdot 30 + 3 \cdot 35 + 9 \cdot 40) + 46 \cdot (4 \cdot 30 + 12 \cdot 35 + 6 \cdot 40) + 56 \cdot (35 + 5 \cdot 40) = 3680 + 13000 + 36 \cdot (960 + 105 + 300) + 35880 + 13160 = 117020.$$

Найдем $r_B = \frac{\sum_{i=1}^{n} n_{x,y} \cdot x \cdot y - n \cdot \bar{x} \cdot \bar{y}}{n \cdot \sigma_x \cdot \sigma_y}$,

$$r_B = \frac{117020 - 100 \cdot 31{,}7 \cdot 35{,}6}{100 \cdot 5{,}35 \cdot 10{,}2} = \frac{4168}{5457} \approx 0{,}76;$$

$$\bar{y}_x - 35{,}6 = 0{,}76 \cdot \frac{10{,}2}{5{,}35} \cdot (x - 31{,}7);$$

$$\bar{y}_x = 1{,}45 \cdot x - 10{,}365.$$

Проверка гипотезы о значимости выборочного коэффициента корреляции

Пусть двумерная генеральная совокупность *(X, Y)* распределена нормально. Из этой совокупности извлечена выборка объема *n* и по ней найден выборочный коэффициент корреляции $r_в$, который оказался отличным от нуля. Так как выборка отобрана случайно, то еще нельзя заключить, что коэффициент корреляции генеральной совокупности также отличен от нуля. Поэтому возникает необходимость при заданном уровне значимости α проверить нулевую гипотезу H_0: $r_г = 0$ о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе H_1: $r_г \neq 0$.

Если нулевая гипотеза отвергается, то это означает, что выборочный коэффициент корреляции значимо отличается от нуля (кратко говоря, значим), а *X* и *Y* — коррелированные случайные величины, т. е. связаны линейной зависимостью.

Если нулевая гипотеза будет принята, то выборочный коэффициент корреляции незначим, а *X* и У не коррелированные случайные величины, т. е. не связаны линейной зависимостью.

В качестве критерия проверки нулевой гипотезы примем случайную величину $T = \frac{r_в \cdot \sqrt{n-2}}{\sqrt{1-r_в^2}}$.

Величина *T* при справедливости нулевой гипотезы имеет распределение Стьюдента с $k = n - 2$ степенями свободы.

Поскольку конкурирующая гипотеза имеет вид $r_г \neq 0$, то критическая область — двусторонняя.

Обозначим значение критерия, вычисленное по данным наблюдений, через $T_{набл}$ и сформулируем правило проверки нулевой гипотезы.

Правило. Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу H_0: $r_г = 0$ *о* равенстве нулю генерального коэффициента корреляции нормальной двумерной случайной величины при конкурирующей гипотезе $H_1: r_г \neq 0$, надо вычислить наблюдаемое значение критерия: $T_{набл} = \frac{r_в \cdot \sqrt{n-2}}{\sqrt{1-r_в^2}}$ и по таблице критических точек распределения Стьюдента, по заданному уровню значимости и числу степеней свободы $k = n - 2$ найти критическую точку $t_{кр}$ $(\alpha; k)$ для двусторонней критической области.

Если $|T_{набл}| < t_{кр}$ — нет оснований отвергнуть нулевую гипотезу.

Если $|T_{набл}| > t_{кр}$ — нулевую гипотезу отвергают.

Пример. По выборке объема n =122, извлеченной нз нормальной двумерной совокупности, найден выборочный коэффициент корреляции $r_в$ = 0,4. При уровне значимости α =0,05 проверить нулевую гипотезу о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе $H_1: r_г \neq 0$.

Р е ш е н и е . Найдем наблюдаемое значение критерия:

$$T_{набл} = \frac{r_в\sqrt{n-2}}{\sqrt{1-r_в^2}} = \frac{0.4\cdot\sqrt{122-2}}{\sqrt{1-0.4^2}} = 4.78.$$

По условию, конкурирующая гипотеза имеет вид $r_г \neq 0$, поэтому критическая область — двусторонняя.

По уровню значимости 0,05 и числу степеней свободы $k = 122 - 2 = 120$ находим по таблице критических точек для двусторонней критической области критическую точку $t_{кр} = t_{кр}$ (0,05; 120)= 1,98.

Поскольку $T_{набл} > t_{кр}$ — нулевую гипотезу отвергаем. Другими словами, выборочный коэффициент корреляции значимо отличается от нуля,

т. е. *X* и *Y* – коррелированны.

Выборочный коэффициент ранговой корреляции Кендалла и проверка гипотезы о его значимости

Можно оценивать связь между двумя качественными признаками, используя коэффициент ранговой корреляции Кендалла. Пусть ранги объектов выборки объема n :

по признаку А $x_1, x_2, \dots, x_n$

по признаку В $y_1, y_2, \dots, y_n$

Допустим, что правее y_1 имеется R_1 рангов, больших y_1; правее y_2 имеется R_2 рангов, больших y_2 ; ... ; правее y_{n-1} имеется R_{n-1} рангов, больших y_{n-1} . Введем обозначение суммы рангов R_i $(i = 1, 2, \dots, n-1)$:

$$R = R_1 + R_2 + \cdots + R_{n-1} .$$

Выборочный коэффициент ранговой корреляции Кендалла определяется формулой :

$$\tau_в = \left[\frac{4R}{n(n-1)}\right] - 1 \quad (*),$$

где n — объем выборки, $R = \sum_{i-1}^{n-1} R_i$.

Убедимся, что коэффициент Кендалла имеет те же свойства, что и коэффициент Спирмена.

1 В случае «полной прямой зависимости» признаков

$$x_1 = 1, x_2 = 2, \dots, x_n = n,$$

$$y_1 - 1, y_2 - 2, \dots, y_n - n$$

Правее y_1 имеется n-1 рангов, больших y_1, поэтому $R_1 = n - 1$. Очевидно, что $R_2 = n - 2, \dots, R_{n-1} = 1$. Следовательно,

$$R = (n-1) + (n-2) + \cdots + 1 = \frac{n(n-1)}{2}. \quad (**)$$

Подставив **(**)** в **(*)**, получим:

$$\tau_в = 1.$$

2 В случае «противоположной зависимости»

$$x_1 = 1, x_2 = 2, \dots, x_n = n$$

$$y_1 = n, y_2 = n - 1, \dots, y_n = 1$$

Правее y_1 нет рангов, больших y_1; поэтому $R_1 = 0$. Очевидно, что $R_2 = R_3 = \cdots = R_{n-1} = 0$. Следовательно,

$$R = 0. \quad (***)$$

Подставив **(***)** в **(*)**, получим:

$$\tau_в = -1$$

Замечание. При достаточно большом объеме выборки и при значениях коэффициентов ранговой корреляции, не близких к единице, имеет место приближенное равенство:

$$\rho_в = \frac{3}{2}\tau_в$$

Приведем правило, позволяющее установить значимость или незначимость ранговой корреляционной связи Кендалла.

Для того чтобы при уровне значимости α, проверить нулевую гипотезу о равенстве нулю генерального коэффициента ранговой корреляции $\tau_г$ Кендалла при конкурирующей гипотезе $H_1: \tau_г \neq 0$, надо вычислить критическую точку:

$$T_{кр} = z_{кр}\sqrt{\frac{2(2n+5)}{9n(n-1)}},$$

где n — объем выборки; $z_{кр}$ — критическая точка двусторонней критической области, которую находят по таблице функции Лапласа по равенству $\Phi(z_{кр}) = \frac{(1-\alpha)}{2}$.

Если $|\tau_в| < T_{кр}$ — нет оснований отвергнуть нулевую гипотезу. Ранговая корреляционная связь между качественными признаками незначимая.

Если $|\tau_в| > T_{кр}$ — нулевую гипотезу отвергают. Между качественными признаками существует значимая ранговая корреляционная связь.

Приведем примеры расчета коэффициента ранговой корреляции $\tau_в$-Кендалла

Пример 1. Найти выборочный коэффициент ранговой корреляции Кендалла по данным рангам объектов выборки объема n=10. При уровне значимости α =0,05 проверить, является ли ранговая корреляционная связь значимой:

по признаку A ... x_i 1 2 3 4 5 6 7 8 9 10;

по признаку B ... y_i 6 4 8 1 2 5 10 3 7 9

Правее $y_1 = 6$ имеется 4 ранга (8, 10, 7, 9), больших y_1 поэтому $R_1 = 4$. Аналогично найдем. $R_2 = 5$, $R_3 = 2$, $R_4 = 6$, $R_5 = 5$, $R_6 = 3$, $R_7 = 0$, $R_8 = 2$, $R_9 = 1$. Следовательно, сумма рангов $R = 28$.

Найдем искомый коэффициент ранговой корреляции Кендалла, учитывая, что n=10:

$$\tau_в = \left[\frac{4R}{n(n-1)}\right] - 1 = \left[\frac{4*28}{10*9}\right] - 1 = 0,24.$$

Найдем критическую точку $z_{кр}$:

$$\Phi(z_{кр}) = \frac{(1-\alpha)}{2} = \frac{(1-0,05)}{2} = 0,475.$$

По таблице функции Лапласа находим $z_{кр} = 1,96$.

Найдем критическую точку:

$$T_{кр} = z_{кр}\sqrt{\frac{2(2n+5)}{9n(n-1)}},$$

Подставив $z_{кр} = 1,96$ и n=10, получим $T_{кр} = 0,487$.

Так как $\tau_в < T_{кр}$ — нет оснований отвергнуть нулевую гипотезу; ранговая корреляционная связь между признаками незначимая.

Допустим, что объекты генеральной совокупности обладают двумя качественными признаками. Под *качественным* подразумевается признак, который

невозможно измерить точно, но он позволяет сравнивать объекты между собой и, следовательно, расположить их в порядке убывания или возрастания качества. Для определенности *будем всегда располагать объекты в порядке ухудшения качества*. При таком «ранжировании» на первом месте находится объект наилучшего качества по сравнению с остальными; на втором месте окажется объект «хуже» первого, но «лучше» других, и т. д.

Пусть выборка объема n содержит независимые объекты, которые обладают двумя качественными признаками A и B. Для оценки *степени связи признаков* вводят, в частности, коэффициенты ранговой корреляции Спирмена и Кендалла.

Для практических целей использование ранговой корреляции весьма полезно. Например, если установлена высокая ранговая корреляция между двумя качественными признаками изделий, то достаточно контролировать изделия только по одному из признаков, что удешевляет и ускоряет контроль.

Расположим сначала объекты выборки в порядке ухудшения качества по признаку A при допущении, что *все объекты имеют различное качество по обоим признакам* (случай, когда это допущение не выполняется, рассмотрим ниже). Припишем объекту, стоящему на - м месте, число-ранг x_i, равный порядковому номеру объекта. Например, ранг объекта, занимающего первое место, $x_1 = 1$; объект, расположенный на втором месте, имеет ранг $x_2 = 2$, и т. д. В итоге получим последовательность рангов по признаку A: $x_1 - 1$, $x_2 = 2$,…, $x_n = n$.

Расположим теперь объекты в порядке убывания качества по признаку B и припишем каждому из них ранг y_i, однако (для удобства сравнения рангов) *индекс i при y будет по-прежнему равен порядковому номеру объекта по признаку A*. Например, запись $y_2 = 5$ означает, что по признаку A объект стоит на втором месте, а по признаку B – на пятом.

В итоге получим две последовательности рангов:

по признаку A: … $x_1, x_2, \ldots, x_n$;

по признаку B: … $y_1, y_2, \ldots, y_n$

Заметим, что в первой строке индекс i совпадает с порядковым номером объекта, а во второй, вообще говоря, не совпадает.

Итак, в общем случае $x_i \neq y_i$.

Рассмотрим два «крайних случая».

1 Пусть ранги по признакам A и B совпадают при всех значениях индекса i: $x_i = y_i$. В этом случае ухудшение качества по одному признаку влечет ухудшение качества по другому. Очевидно, признаки связаны – имеет место «полная прямая зависимость».

2 Пусть ранги по признакам А и В противоположны в том смысле, что если $x_1 = 1$, то $y_1 = n$; если $x_2 = 2$, то $y_2 = n - 1$; … если $x_n = n$, то $y_n = 1$. В

этом случае ухудшение качества по одному признаку влечет улучшение по другому. Очевидно, признаки связаны – имеет место «противоположная зависимость».

На практике чаще будет встречаться промежуточный случай, когда ухудшение качества по одному признаку влечет для некоторых объектов ухудшение, а для других – улучшение качества. Задача состоит в том, чтобы оценить связь между признаками. Для ее решения рассмотрим ранги $x_1, x_2, \dots, x_n$ как возможные значения случайной величины X, а $y_1, y_2, \dots, y_n$ – как возможные значения случайной величины Y. Таким образом, о связи между качественными признаками A и B можно судить по связи между случайными величинами X и Y, для оценки которой используем коэффициент корреляции.

Вычислим выборочный коэффициент корреляции случайных величин X и Y в условных вариантах:

$$r_в = \frac{\sum n_{uv} uv - n\bar{u}\bar{v}}{n\sigma_u \sigma_v},$$

приняв в качестве условных вариант отклонения $u_i = x_i - \bar{x}$, $v_i = y_i - \bar{y}$. Каждому рангу x_i соответствует только один ранг y_i, поэтому частота любой пары рангов с одинаковыми индексами, а следовательно, и любой пары условных вариант с одинаковыми индексами, равна единице: $n_{u_i v_i} = 1$. Очевидно, что частота любой пары вариант с разными индексами равна нулю. Учитывая, кроме того, что среднее значение отклонения равно нулю, т.е. $\bar{u} = \bar{v} = 0$, получим более простую формулу вычисления выборочного коэффициента корреляции:

$$r_в = \frac{\sum u_i v_i}{n\sigma_u \sigma_v}. \qquad (*)$$

Таким образом, надо найти $\sum u_i v_i$, σ_u и σ_v.

Выразим $\sum u_i v_i$ через известные числа – объем выборки n и разности рангов $d_i = x_i - y_i$. Заметим, что поскольку средние значения рангов $\bar{x} = (1 + 2 + \dots + n)/n$ и $\bar{y} = (1 + 2 + \dots + n)/n$ равны между собой, то $\bar{y} - \bar{x} = 0$.

Используем последнее равенство:

$$d_i = x_i - y_i = x_i - y_i + (\bar{y} - \bar{x}) = (x_i - \bar{x}) - (y_i - \bar{y}) = u_i - v_i.$$

Следовательно,

$$d_i^2 = (u_i - v_i)^2.$$

Учитывая, что (смотрите далее пояснение)

$$\sum u_i^2 = \sum v_i^2 = (n^3 - n)/12, \qquad (**)$$

имеем:

$$\sum d_i^2 = \sum (u_i^2 - v_i^2)^2 = \sum u_i^2 - 2\sum u_i v_i + \sum v_i^2 = [(n^3 - n)/6] - 2\sum u_i v_i.$$

Отсюда

$$\sum u_i v_i = [(n^3 - n)/12] - \sum d_i^2/2\,. \qquad (***)$$

Остается найти σ_u и σ_v. По определению выборочной дисперсии, учитывая, что $\bar{u} = 0$, и используя (**), получим

$$D_u = \sum (u_i - \bar{u})^2/n = \sum u_i^2/n = (n^3 - n)/12n = (n^2 - 1)/12\,.$$

Отсюда среднее квадратическое отклонение

$$\sigma_u = \sqrt{(n^2 - 1)/12}\,.$$

Аналогично найдем

$$\sigma_v = \sqrt{(n^2 - 1)/12}\,.$$

Следовательно,

$$n\sigma_u\sigma_v = (n^2 - n)/12\,.$$

Подставив правые части этого равенства и соотношения (***) в (*), окончательно получим *выборочный коэффициент ранговой корреляции Спирмена*

$$\rho_в = 1 - \left[\left(6\sum d_i^2\right)/(n^3 - n)\right], \qquad (****)$$

где $d_i = x_i - y_i$.

Пояснение. Покажем, что $\sum u_i^2 = (n^3 - n)/12$. Действительно, учитывая, что

$$\sum x_i = 1 + 2 + \cdots + n = (1 + n)n/2\,,$$

$$\bar{x} = \sum x_i/n = (1 + n)/2\,,$$

$$\sum x_i^2 = 1^2 + 2^2 + \cdots + n^2 = [n(n+1)(2n+1)]/6,$$

$$\sum u_i^2 = \sum (x_i - \bar{x})^2 = \sum x_i^2 - 2\bar{x}\sum x_i + n(\bar{x})^2,$$

после элементарных выкладок получим

$$\sum u_i^2 = (n^3 - n)/12.$$

Аналогично можно показать, что

$$\sum v_i^2 = (n^3 - n)/12.$$

Пример : Найти выборочный коэффициент ранговой корреляции Кендалла по данным рангам объектов выборки объема n=10. При уровне значимости $\alpha = 0{,}05$ проверить, является ли ранговая корреляционная связь значимой?:

по признаку A ... x_i 1 2 3 4 5 6 7 8 9 10
по признаку B ... y_i 6 4 8 1 2 5 10 3 7 9

Правее $y_1 = 6$ имеется 4 ранга (8, 10, 7, 9), больших y_1 поэтому $R_1 = 4$. Аналогично найдем. $R_2 = 5$, $R_3 = 2$, $R_4 = 6$, $R_5 = 5$, $R_6 = 3$, $R_7 = 0$, $R_8 = 2$, $R_9 = 1$. Следовательно, сумма рангов $R = 28$.

Найдем искомый коэффициент ранговой корреляции Кендалла, учитывая, что n=10:

$$\tau_{\text{в}} = \left[\frac{4R}{n(n-1)}\right] - 1 = \left[\frac{4*28}{10*9}\right] - 1 = 0{,}24.$$

Найдем критическую точку $z_{\text{кр}}$:

$$\Phi(z_{\text{кр}}) = \frac{(1-\alpha)}{2} = \frac{(1-0{,}05)}{2} = 0{,}475\,.$$

По таблице функции Лапласа находим $z_{\text{кр}} = 1{,}96$.

Найдем критическую точку:

$$T_{\text{кр}} = z_{\text{кр}}\sqrt{\frac{2(2n+5)}{9n(n-1)}}\,,$$

Подставив $z_{\text{кр}} = 1{,}96$ и n=10, получим $T_{\text{кр}} = 0{,}487$.

Так как $\tau_{\text{в}} < T_{\text{кр}}$ — нет оснований отвергнуть нулевую гипотезу; ранговая корреляционная связь между признаками незначимая.

Вопросы для повторения

1 Корреляционная зависимость.
2 Условное математическое ожидание дискретных и непрерывных случайных величин.
3 Регрессия Y на X (X на Y). Линия регрессии Y на X (X на Y).
4 Коэффициент корреляции и его свойства.
5 Линейная корреляция.
6 Уравнение линейной регрессии Y на X (X на Y).
7 Расчет прямых линий регрессии Y на X (X на Y) по опытным данным.
8 Метод четырех полей вычисления выборочного коэффициента корреляции.
9 Проверка гипотезы о значимости выборочного коэффициента корреляции.
10.Выборочный коэффициент ранговой корреляции Спирмена и Кендалла.

Задачи для решения в аудитории

1 Найти выборочное уравнение прямой регрессии Y на X по данным таблицы.

Таблица 17 – Опытные данные

X	22	23	24	24,5	24,5	25	25,5	26	26,5	27
Y	0,46	0,48	0,5	0,49	0,5	0,51	0,52	0,51	0,52	0,54

2 Средняя температура в г. Саратове (X) и в г. Алатыре (Y) измерялась в течение 13 лет и данные приведены в таблице18.

Таблица 18 – Средняя температура в г. Саратове (X) и в г. Алатыре (Y)

Год	1891	1892	1893	1894	1895	1896	1897
X	-19,2	-14,8	-19,6	-11,1	-9,4	-16,9	-13,7
Y	-21,8	-15,4	-20,8	-11,3	-11,6	-19,2	-13

Таблица 18 (продолжение)

Год	1899	1911	1912	1913	1914	1915
X	-4,9	-13,9	-9,4	-8,3	-7,9	-5,3
Y	-7,4	-15,1	-14,4	-11,1	-10,5	-7,2

Найти выборочный коэффициент корреляции средних январских температур в Саратове и Алатыре, написать выборочное уравнение линейной регрессии Y на X и оценить характер связи Y с X.

3 В таблице 19 приведены опытные данные, характеризующие потери зерна в зависимости от сроков уборки.

Таблица 19– Данные, характеризующие потери зерна в зависимости от сроков уборки

Срок уборки	Собрано (ц. с 1 га)	Потери (ц. с 1 га)
В период полной спелости зерна	29,5	0
После наступления полной спелости зерна	28,4	1,1
Через 5 дней	28,4	1,1
Через 10 дней	23,4	6,1
Через 15 дней	21,6	7,9
Через 20 дней	18,5	11,0

На основании этих данных требуется составить выборочное уравнение прямых линий регрессии.

4 На основе выборочных данных о деятельности аудиторско-консультационных фирм Москвы в 2001 г. найти выборочный коэффициент прогрессии r_{xy}, оценить тесноту связи между совокупной выручкой этих фирм и общей численностью профессионалов. Записать уравнение связи. Установить будет ли связь существенной (таблица 20).

Таблица 20 – Связь между совокупной выручкой фирм и общей численностью профессионалов

Общая численность профессионалов, чел., x	23	32	50	53	55	58	59	62	69	75
Совокупная выручка млн. руб., y	2,62	3,04	3,15	3,83	3,58	4,08	4,09	4,2	4,8	4,24

5 По результатам таблицы 21 вычислить выборочный коэффициент корреляции и проверить гипотезу о значимости выборочного коэффициента корреляции.

Таблица 21 – Связь между признаками

X	92	91	90	86	85	85	85	83	83	80	80	78
Y	84	85	84	81	76	77	75	79	75	78	78	78

6 В таблице 22 приведены результаты измерения силы звука самолета Y(Дб.) на различных расстояниях от точки взлета X(км). Запишите уравнение прямой линии регрессии Y на X. Найдите:

а) на каком расстоянии от точки взлета звук становится смертельно опасным для человека (свыше 120 Дб)

б) на каком расстоянии от аэродрома можно строить жилые помещения (менее 75 Дб)

Таблица 22– Сила звука самолета Y(Дб) на различных расстояниях от точки взлета X (км)

X	1	2,5	3	5,5	7	8,5	10	15	20	30
Y	115	108	102	98	93	89	87	72	65	60

7 На вопрос, который был задан студентам: «Какие качества вы больше всего цените в товарищах?», были получены следующие ответы (таблица 23).

Таблица 23 – Качества товарища

Качества	1 курс	5 курс
1 Взаимовыручка (А)	58	50
2 Коммуникабельность (Б)	53	27
3 Доброта (В)	48	22
4 Жизнелюбие (Г)	45	48
5 Преданность (Д)	36	45
6 Честность (Е)	28	15
7 Отзывчивость (Ж)	21	37
8 Понимание (З)	15	20
9 Справедливость (И)	13	18
10 Трудолюбие (К)	10	11

Выяснить, есть ли зависимость между отношением к данным качествам и возрастом респондентов, т. е найти:

а) выборочный коэффициент ранговой корреляции Спирмена (ρ_b) и проверить его значимость на уровне α = 0,05;

б) выборочный коэффициент ранговой корреляции Кендалла (τ_b) и проверить его значимость на уровне α = 0,05.

8 Знания 10 студентов проверены по двум тестам: А и В. Оценки по 100-балльной системе оказались следующими (таблица 24).

Таблица 24 – Оценка знаний студентов по двум тестам

A:	95	90	86	84	75	70	62	60	57	50
B:	92	93	83	80	55	60	45	72	62	70

Найти выборочные коэффициенты корреляции Спирмена и Кендалла между оценками по двум тестам и проверить их значимость: ρ_b на уровне α = 0,01; τ_b на уровне α = 0,05.

Задачи для самостоятельного решения

1 Средняя температура июня в Москве (X) и Ярославле (Y) измерялась в течение 40 лет. Эти данные приведены в таблице 25.

Таблица 25 – Значения температуры в городах Москва и Ярославль

X	У	X	У	X	У	X	У	X	У
12	10,8	13,9	10,1	15	13,8	17,2	13,9	18,1	16
12	11,3	14,2	10	15	16	16,9	14,8	18,4	17,8
12	12	14	10	15,5	13,9	16,9	15	19,2	15
12	13	14	12	15,9	14,7	17	16	19,3	16,1
12,8	10,9	13,9	12,4	16	13	16,8	17	20	17
13,8	10	15	11	15,9	15	17,5	16	20,1	17,7
13,1	11,5	14,9	13	16	16	18	14	14	14,8
13	13	14,9	14,2	16,9	12,9	18	14,8	14	15,3

Найти выборочные средние июньские температуры в Москве и Ярославле и их средние квадратические отклонения.

С помощью критерия χ^2 – Пирсона согласие опытных данных с гипотезой о нормальном распределении средних июньских температур. Найти выборочный коэффициент корреляции X и Y, написать выборочное уравнение линейной регрессии Y на X. Охарактеризовать зависимость Y и X.

Примерные задания для рубежного контроля 2 (проверочная самостоятельная работа)

1 В таблице 26 приведены результаты измерения силы звука самолета Y(Дб) на различных расстояниях от точки взлета X(км). Запишите уравнение прямой линии регрессии Y на X. Найдите:

а) на каком расстоянии от точки взлета звук становится смертельно опасным для человека (свыше 120 Дб)

б) на каком расстоянии от аэродрома можно строить жилые помещения (менее 75 Дб).

Таблица 26 – силы звука самолета на различных расстояниях от точки взлета

X	1	2,5	3	5,5	7	8,5	10	15	20	30
Y	115	108	102	98	93	89	87	72	65	60

2 Двумя методами проведены измерения одной и той же физической величины. Получены следующие результаты:

а) $x_1 = 9{,}6$; $x_2 = 10$; $x_3 = 9{,}8$; $x_4 = 10{,}2$; $x_5 = 10{,}6$;

б) $y_1 = 10{,}4$; $y_2 = 9{,}7$; $y_3 = 10$; $y_4 = 10{,}3$.

Можно ли считать, что оба метода обеспечивают одинаковую точность измерений, если принять уровень значимости $\alpha = 0{,}1$? Предполагается, что результаты измерений распределены нормально и выборки независимы.

3 Н. Хольмбергом наблюдалось распределение красных кровяных шариков по 169 отделениям прибора гемацитометра числа vi отделений, содержащих по ni красных кровяных шариков. Данные указанны в таблице 27.

Таблица 27 – распределение красных кровяных шариков по 169 отделениям прибора гемацитометра

vi	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
ni	1	3	5	8	12	14	15	15	21	18	17	16	9	6	3	2	2	1

При уровне значимости $\rho = 0{,}05$ проверить гипотезу о нормальном распределении.

4 Зависимость между объемом промышленной продукции и инвестициями в основной капитал по 10 областям одного из федеральных округов РФ в 2003 году характеризуется следующими данными, представленными в таблице 28.

Таблица 28 – Зависимость между объемом промышленной продукции и инвестициями

Область	Объем промышленной продукции, млрд руб.	Инвестиции в основной капитал, млрд руб.
Белгородская	64,6	10,22
Брянская	21,5	4,12
Владимирская	51,1	8,58
Воронежская	54,4	14,79
Ивановская	20,6	2,88
Калужская	35,7	7,24

Костромская	18,4	5,57
Курская	37,1	9,67
Липецкая	90,6	10,45
Смоленская	39,8	10,48

Вычислите ранговые коэффициенты корреляции Кендалла. Проверьте значимость при α=0,05. Сформулируйте вывод о зависимости между объемом промышленной продукции и инвестициями в основной капитал по рассматриваемым областям РФ.

Вопросы к зачету

1 Статистическое распределение выборки.
2 Статистическая оценка параметров распределения. Генеральная и выборочная средние.
3 Генеральная дисперсия. Выборочная дисперсия.
4 Оценка генеральной средней.
5 Интервальная оценка. Доверительная вероятность. Доверительный интервал.
6 Доверительный интервал для МО нормального распределения при известном σ.
7 Доверительный интервал для МО нормального распределения при неизвестном σ.
8 Доверительный интервал для среднего квадратического отклонения.
9 Проверка гипотезы о нормальном распределении генеральной совокупности по критерию X^2 Пирсона.
10 Корреляционная зависимость. Линейная корреляция. Расчет прямых линий регрессии по опытным данным.
11 Проверка гипотезы о равенстве дисперсий двух нормальных совокупностей.
12 Выборочные коэффициенты Кендалла и Спирмена. Проверка гипотезы о их значимости.

Список литературы

1 Баврин, И. И. Высшая математика [Текст] / И. И. Баврин. – Москва : Просвещение, 2004.
2 Владимирский, Б. М. Математика. Общий курс [Текст] : учебник для бакалавров / Б. М. Владимирский. – Москва : Просвещение, 2008.
3 Гмурман, В. Е. Руководство к решению задач по теории вероятностей и математической статистики: учебное пособие для студентов [Текст] /
В. Е. Гмурман. – Москва : Высшая школа, 2006.
4 Ильин, В. А. Высшая математика [Текст] / В. А. Ильин. – Москва : Проспект, 2005.
5 Карелина, И. Г. Математика [Текст] / И. Г. Карелина. – Воронеж : Воронежский государственный университет, 2004
6 Колемаев, В. А., Калинин В. Н. Теория вероятностей и математическая статистка [Текст] / В. А. Колемаев, В. Н. Калинин. – Москва : Инфра, 1997.
7 Лобоцкая, Н. Л. Основы высшей математики [Текст] / Н. Л. Лобоцкая. – Минск : Высшая школа, 2000.
8 Маркович, Э. С. Курс высшей математики с элементами теории вероятностей и математической статистики [Текст] / Э. С. Маркович. – Москва : Высшая школа, 1982.
9 Минорский, В. П. Сборник задач по высшей математике [Текст] /
В. П. Минорский. – Москва : Изд-во физ.-мат., 2001.
10 Крамер, Г. Математические методы статистики [Текст] / Г. Крамер – Москва : Наука, 1975.
11 Кремер, Н. М. Теория вероятностей и математическая статистика [Текст] /
Н. М. Кремер. – Москва: Наука, 2000.

Приложения

Приложение А

Статистические таблицы

Таблица А 1 – Значения функции $\varphi(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{x^2}{2}}$

	0	1	2	3	4	5	6	7	8	9
0,0	0,3989	3989	3989	3988	3986	3984	3982	3980	3977	3973
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925	3918
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836	3825
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712	3697
0,4	3683	3668	3652	3637	3621	3605	3589	3572	3555	3538
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372	3352
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166	3144
0,7	3123	3101	3079	3056	3034	3011	2989	2966	2943	2920
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709	2685
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468	2444
1,0	0,2420	2396	2371	2347	2323	2299	2275	2251	2227	2203
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989	1965
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758	1736
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539	1518
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334	1315
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145	1127
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973	0957
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818	0804
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681	0669
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562	0551
2,0	0,0540	0529	0519	0508	0498	0488	0478	0468	0459	0449
2,1	0440	0431	0422	0413	0404	0396	0387	0379	0371	0363
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297	0290
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235	0229
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184	0180
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143	0139
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110	0107
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084	0081
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063	0061
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047	0046
3,0	0,0044	0043	0042	0040	0039	0038	0037	0036	0035	0034
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013	0013
3,4	0012	0012	0012	0011	0011	0010	0010	0010	0009	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007	0006
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005	0004
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003	0003
3,8	0003	0003	0003	0003	0003	0002	0002	0002	0002	0002
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001	0001

Таблица А2 – Значения функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^x e^{-\frac{z^2}{2}} dz$

	Φ(*x*)	***x***	**Φ(*x*)**	***x***	**Φ(*x*)**	***x***	**Φ(*x*)**
0,00							
0,01	0,0000	0,32	0,1255	0,64	0,2389	0,96	0,3315
0,02	0,0040	0,33	0,1293	0,65	0,2422	0,97	0,3340
0,03	0,0080	0,34	0,1331	0,66	0,2454	0,98	0,3365
0,04	0,0120	0,35	0,1368	0,67	0,2486	0,99	0,3389
0,05	0,0160	0,36	0,1406	0,68	0,2517	1,00	0,3413
0,06	0,0199	0,37	0,1443	0,69	0,2549	1,01	0,3438
0,07	0,0239	0,38	0,1480	0,70	0,2580	1,02	0,3461
0,08	0,0279	0,39	0,1517	0,71	0,2611	1,03	0,3485
0,09	0,0319	0,40	0,1554	0,72	0,2642	1,04	0,3508
0,10	0,0359	0,41	0,1591	0,73	0,2673	1,05	0,3531
0,11	0,0398	0,42	0,1628	0,74	0,2703	1,06	0,3554
0,12	0,0438	0,43	0,1664	0,75	0,2734	1,07	0,3577
0,13	0,0478	0,44	0,1700	0,76	0,2764	1,08	0,3599
0,14	0,0517	0,45	0,1736	0,77	0,2794	1,09	0,3621
0,15	0,0557	0,46	0,1772	0,78	0,2823	1,10	0,3643
0,16	0,0596	0,47	0,1808	0,79	0,2852	1,11	0,3665
0,17	0,0636	0,48	0,1844	0,80	0,2881	1,12	0,3686
0,18	0,0675	0,49	0,1879	0,81	0,2910	1,13	0,3708
0,19	0,0714	0,50	0,1915	0,82	0,2939	1,14	0,3729
0,20	0,0753	0,51	0,1950	0,83	0,2967	1,15	0,3749
0,21	0,0793	0,52	0,1985	0,84	0,2995	1,16	0,3770
0,22	0,0832	0,53	0,2019	0,85	0,3023	1,17	0,3790
0,23	0,0871	0,54	0,2054	0,86	0,3051	1,18	0,3810
0,24	0,0910	0,55	0,2088	0,87	0,3078	1,19	0,3830
0,25	0,0948	0,56	0,2123	0,88	0,3106	1,20	0,3849
0,26	0,0987	0,57	0,2157	0,89	0,3133	1,21	0,3869
0,27	0,1026	0,58	0,2190	0,90	0,3159	1,22	0,3883
0,28	0,1064	0,59	0,2224	0,91	0,3186	1,23	0,3907
0,29	0,1103	0,60	0,2257	0,92	0,3212	1,24	0,3925
0,30	0,1141	0,61	0,2291	0,93	0,3238	1,25	0,3944
0,31	0,1179	0,62	0,2324	0,94	0,3264		
	0,1217	0,63	0,2357	0,95	0,3289		

Продолжение таблицы А2

1,26	0,3962	1,59	*0,4441*	1,92	0,4726	*2,50*	0,4938
1,27	0,3980	1,60	*0,4452*	1,93	0,4732	*2,52*	0,4941
1,28	0,3997	1,61	*0,4463*	1,94	0,4738	*2,54*	0,4945
1,29	0,4015	1,62	*0,4474*	1,95	0,4744	*2,56*	0,4948
1,30	0,4032	1,63	*0,4484*	1,96	0,4750	*2,58*	0,4951
1,31	0,4049	1,64	*0,4495*	1,97	0,4756	*2,60*	0,4953
1,32	0,4066	1,65	*0,4505*	1,98	0,4761	*2,62*	0,4956
1,33	0,4082	1,66	*0,4515*	1,99	0,4767	*2,64*	0,4959
1,34	0,4099	1,67	*0,4525*	2,00	0,4772	*2,66*	0,4961
1,35	0,4115	1,68	*0,4535*	2,02	0,4783	*2,68*	0,4963
1,36	0,4131	1,69	*0,4545*	2,04	0,47,93	*2,70*	0,4965
1,37	0,4147	1,70	*0,4554*	2,06	0,4803	*2,72*	0,4967
1,38	0,4162	1,71	*0,4564*	2,08	0,4812	*2,74*	0,4969
1,39	0,4177	1,72	*0,4573*	2,10	0,4821	*2,76*	0,4971
1,40	0,4192	1,73	*0,4582*	2,12	0,4830	*2,78*	0,4973
1,41	0,4207	1,74	*0,4591*	2,14	0,4838	*2,80*	0,4974
1,42	0,4222	1,75	*0,4599*	2,16	0,4846	*2,82*	0,4976
1,43	0,4236	1,76	*0,4608*	2,18	0,4854	*2,84*	0,4977
1,44	0,4251	1,77	*0,4616*	2,20	0,4861	*2,86*	0,4979
1,45	0,4265	1,78	*0,4625*	2,22	0,4868	*2,88*	0,4980
1,46	0,4279	1,79	*0,4633*	2,24	0,4875	*2,90*	0,4981
1,47	0,4292	1,80	*0,4641*	2,26	0,4881	*2,92*	0,4982
1,48	0,4306	1,81	*0,4649*	2,28	0,4887	*2,94*	0,4984
1,49	0,4319	1,82	*0,4656*	2,30	0,4893	*2,96*	0,4985
1,50	0,4332	1,83	*0,4664*	2,32	0,4898	*2,98*	0,4986
1,51	0,4345	1,84	*0,4671*	2,34	0,4904	*3,00*	0,49865
1,52	0,4357	1,85	*0,4678*	2,36	0,4909	*3,20*	0,49931
1,53	0,4370	1,86	*0,4686*	2,38	0,4913	*3,40*	0,49966
1,54	0,4382	1,87	*0,4693*	2,40	0,4918	*3,60*	0,499841
1,55	0,4394	1,88	*0,4699*	2,42	0,4922	*3,80*	0,499928
1,56	0,4406	1,89	*0,4706*	2,44	0,4927	*4,00*	0,499968
1,57	0,4418	1,90	*0,4713*	2,46	0,4931	*4,50*	0,499997
1,58	0,4429	1,91	*0,4719*	2,48	0,4934	*5,00*	0,499997

Таблица А 3 – Значения $q = q(\gamma, n)$

γ / n	**0,95**	**0,99**	**0,999**	γ / n	**0,95**	**0,99**	**0,999**
5	1,37	2,67	5,64	20	0,37	0,58	0,88
6	1,09	2,01	3,88	25	0,32	0,49	0,73
7	0,92	1,62	2,98	30	0,28	0,43	0,63
8	0,80	1,38	2,42	35	0,26	0,38	0,56
9	0,71	1,20	2,06	40	0,24	0,35	0,50
10	0,65	1,08	1,80	45	0,22	0,32	0,46
11	0,59	0,98	1,60	50	0,21	0,30	0,43
12	0,55	0,90	1,45	60	0,188	0,269	0,38
13	0,52	0,83	1,33	70	0,174	0,245	0,34
14	0,48	0,78	1,23	80	0,161	0,226	0,31
15	0,46	0,73	1,15	90	0,151	0,211	0,29
16	0,44	0,70	1,07	100	0,143	0,198	0,27
17	0,42	0,66	1,01	150	0,115	0,160	0,211
18	0,40	0,63	0,96	200	0,099	0,136	0,185
19	0,39	0,60	0,92	250	0,089	0,120	0,162

Таблица А 4 – Критические точки распределения χ^2

Число степеней свободы *k*	Уровень значимости α					
	0,01	**0,025**	**0,05**	**0,95**	**0,975**	**0,89**
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Таблица А 5 – Значения $t_\gamma = t(\gamma, n)$

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	2,36	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

γ / n	**0,95**	**0,99**	**0,999**
20	2,093	2,861	3,883
25	2,064	2,797	3,745
30	2,045	2,756	3,659
35	2,032	2,720	3,600
40	2,023	2,708	3,558
45	2,016	2,692	3,527
50	2,009	2,679	3,502
60	2,001	2,662	3,464
70	1,996	2,649	3,439
80	1,001	2,640	3,418
90	1,987	2,633	3,403
100	1,984	2,627	3,392
120	1,980	2,617	3,374
∞	1,960	2,576	3,291

Лукерьянова Елена Александровна

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

(Часть 2)

Материалы для практических занятий и самостоятельной работы
для студентов очной, очно-заочной и заочной форм обучения
37.03.01 «Психология», 37.05.02 « Психология служебной деятельности»

Редактор Л. П. Чукомина

Подписано в печать 28.01.19 Формат 60 х 84 1/16 Бумага 65 г/м^2
Печать цифровая Усл. печ. л 3,0 Уч.-изд . 3,0
Заказ №30 Тираж 25 Не для продажи

БИЦ Курганского государственного университета.
640020, г. Курган, ул. Советская, 63/4.
Курганский государственный университет.