Автономная некоммерческая образовательная организация
высшего образования
«Воронежский экономико-правовой институт»
(АНОО ВО «ВЭПИ»)

УТВЕРЖДАЮ
Проректор
по учебно-методической работе
_____ А.Ю. Жильников
«__» _____ 2018 г.

МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ ПО ВЫПОЛНЕНИЮ ЛАБОРАТОРНЫХ РАБОТ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

Б1.Б.22 Математические методы в психологии
(наименование дисциплины (модуля))

37.03.01 Психология
(код и наименование направления подготовки)

Направленность (профиль) Социальная психология
(наименование направленности (профиля))

Квалификация выпускника Бакалавр
(наименование квалификации)

Форма обучения Очная, заочная
(очная, очно-заочная, заочная)

Воронеж
2018

Методические рекомендации по выполнению лабораторных работ по дисциплине (модулю) рассмотрены и одобрены на заседании кафедры прикладной информатики, год начала подготовки – 2018.

Протокол от «17» января 2018 г. № 6

Заведующий кафедрой А.Г. Курина

Разработчики:

Профессор А.Г. Курина

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

В педагогике, психологии и других смежных науках о человеке, подавляющее большинство изучаемых явлений не поддается прямому измерению. Умственные способности, компетентность, академическая успеваемость, личностные качества, толерантность, мобильность и другие абстракции есть понятия, выделяемые исследователями и позволяющие описывать отношения между наблюдаемыми переменными. Эти феномены могут быть описаны не только семантически, их существование может быть подтверждено эмпирически, а в частном случае и количественно.

В психолого-педагогических исследованиях системное описание явлений указывает на необходимость осваивания и применения соответствующего математического аппарата. В исследованиях также используются математические методы многомерного анализа. Одним из современных обобщений методов моделирования причинно-следственных связей и латентных (скрытых) структур является структурное моделирование, базирующееся на методах математической статистики и становящееся все более популярным инструментом в работе психологов, педагогов, социологов и гуманитариев.

Лабораторный практикум предназначен для обучающихся всех форм обучения и ориентирован на освоение начального курса математических методов в психологических исследованиях и получение навыков статистического анализа и построения математических моделей с использованием пакета прикладных программ. Выбор MS Excel обусловлен доступностью данного пакета, наличием в нем точного набора средств статистического анализа и математических операций для решения задач, входящих в курс изучаемой дисциплины.

Все лабораторные работы содержат краткие теоретические сведения; содержание и этапы выполнения работы; примеры решения типовых задач в MS Excel с необходимыми пояснениями порядка действий и диалоговых окон; контрольные вопросы по теме работы и данные для самостоятельного исследования.

В практикуме используется двойная нумерация рисунков и формул, первое число указывает номер лабораторной работы, второе – порядковый номер рисунка или формулы в рамках данной лабораторной работы.

Практикум полезен также и преподавателям при организации и проведении практических занятий по дисциплинам, использующим аппарат математической статистики для решения практических задач.

Лабораторная работа № 1

Анализ выборочных данных. Базовые показатели распределения случайной величины (4 часа)

Цель: Изучить базовые показатели распределения случайных величин на примере блока описательных статистик.

Теоретические сведения

Генеральная совокупность – это совокупность всех возможных значений изучаемого признака, которые могли быть получены при данном комплексе условий. Генеральная совокупность рассматривается как случайная величина X с неизвестным законом распределения.

Выборка (выборочная совокупность) – часть объектов генеральной совокупности, на которой произведены измерения изучаемого признака. Совокупность X_1, X_2, $\cdots$, X_n измеренных на выборке значений изучаемого признака также называют *выборкой*.

Количество n произведенных измерений (наблюдений) признака, называется *объемом выборки*.

Сущность выборочного метода состоит в оценке по выборке X_1, X_2, $\cdots$, X_n свойств генеральной совокупности (свойств распределения случайной величины X).

Выборочные данные, упорядоченные по возрастанию или убыванию, называются *вариационным рядом*. Различные значения исследуемого признака в выборке называются *вариантами*.

В большинстве случаев данные концентрируются вокруг некоей центральной точки. Таким образом, чтобы описать любой набор данных, достаточно указать среднее значение. Другими числовыми характеристиками, которые используются для оценки среднего значения распределения, являются среднее арифметическое, медиана и мода.

Среднее арифметическое (часто называемое просто средним) — наиболее распространенная оценка среднего значения распределения. Она является результатом деления суммы всех наблюдаемых числовых величин на их количество. Для выборки, состоящей из чисел X_1, X_2, ..., X_n, выборочное среднее (обозначаемое символом $\bar{X}$) равно $\bar{X} = (X_1 + X_2 + \ldots + X_n) / n$, *или*

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} \ , \qquad (1.1)$$

где $\bar{x}$ — выборочное среднее, n — объем выборки, X_i – i-й элемент выборки.

Рассмотрим вычисление среднего арифметического значения на примере следующей выборки (табл.1.1).

Excel есть специальная функция =МЕДИАНА(), которая работает и с неупорядоченными массивами тоже.

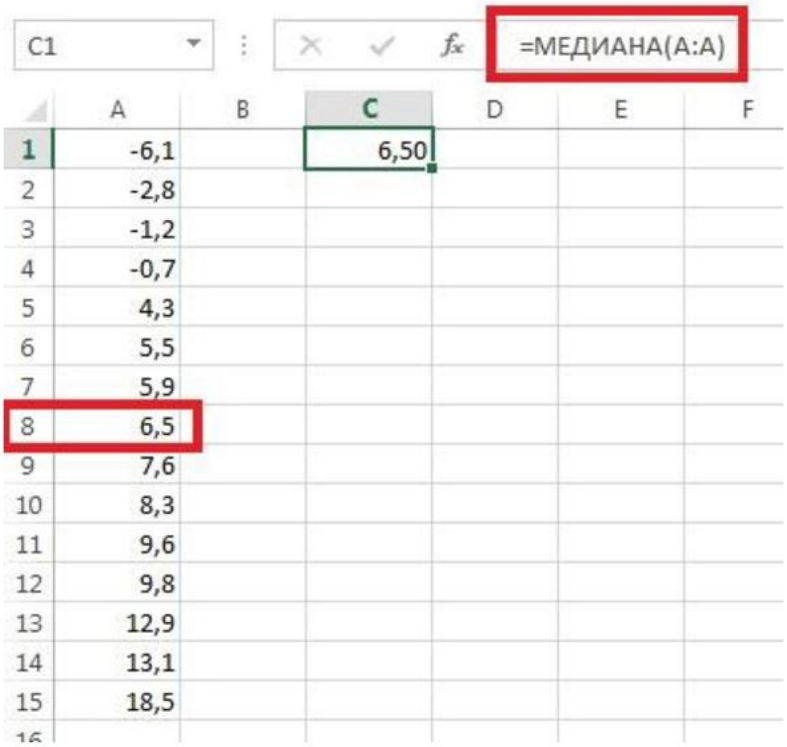


Рис. 1.2. Медиана упорядоченной выборки из нечетного числа элементов

Если удалить из выборки один элемент (последний), то медиана для оставшихся 14 переменных уменьшится до 6,2%, то есть не так значительно, как среднее арифметическое (рис. 1.3).

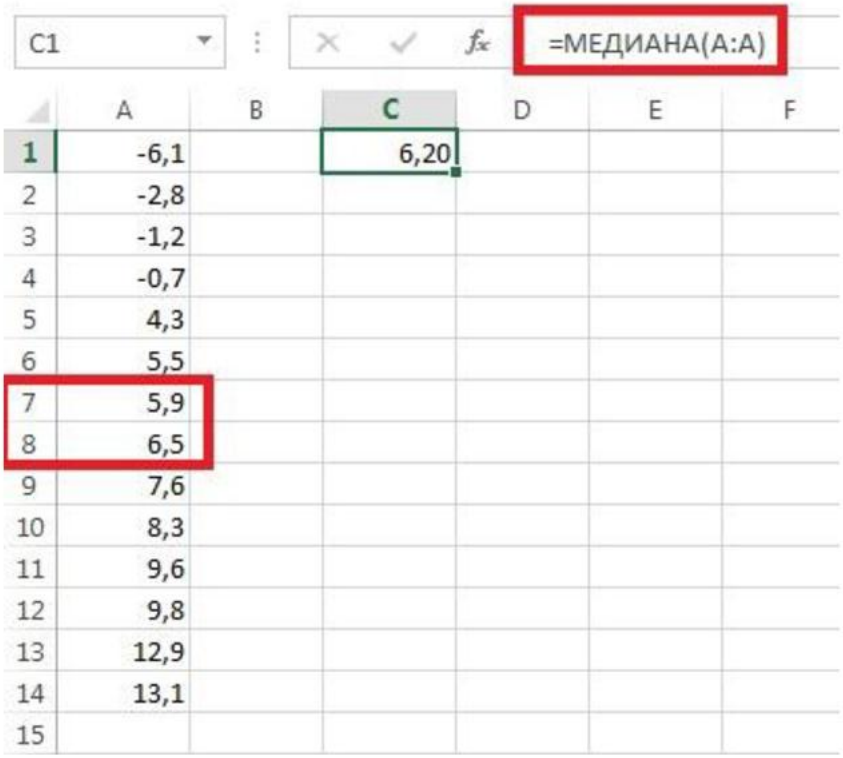
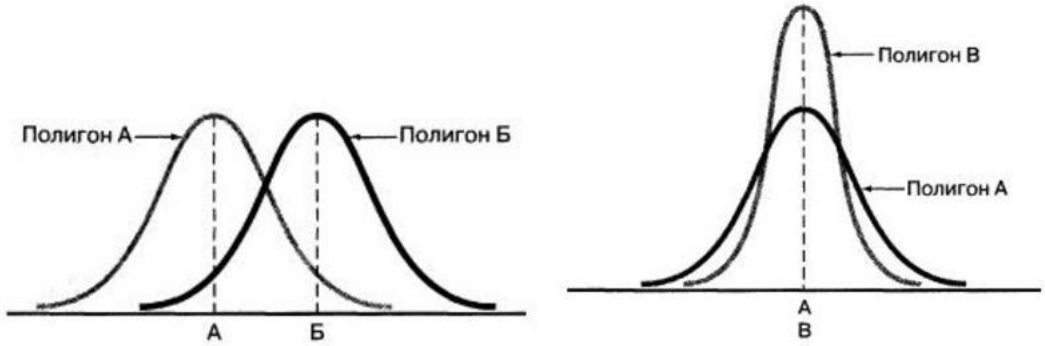


Рис. 1.3. Медиана упорядоченной выборки из четного числа элементов

Эти две функции для одномерного *дискретного ряда* дают различные значения (рис. 1.4). Например, при вычислении квартилей нашей выборки Q_1 = 1,8 для КВАРТИЛЬ.ВКЛ и –0,7 для КВАРТИЛЬ.ИСКЛ. Для расчета квартилей в Excel с помощью вышеприведенных формул массив данных можно не упорядочивать.

Расчет квартилей можно проводить для распределения на основе частот.

Вариация данных характеризует степень их дисперсии. Две разные выборки могут отличаться как средними значениями, так и вариациями, а могут иметь одинаковые вариации, но разные средние значения, либо одинаковые средние значения и совершенно разные вариации (рис. 1.6). Данные, которым соответствует полигон В на рис. 1.6. б, изменяются намного меньше, чем данные, по которым построен полигон А.



а) б)

Рис. 1.6. Два симметричных распределения колоколообразной формы с одинаковым разбросом и разными средними значениями

Существует пять оценок вариации данных:

- размах,
- межквартильный размах,
- дисперсия,
- стандартное отклонение,
- коэффициент вариации.

Размахом называется разность между наибольшим и наименьшим элементами выборки:

$$R = X_{max} - X_{min}. \qquad (1.6)$$

Размах нашей выборки вычислим, используя упорядоченный массив (рис. 1.4):

P = 18,5 – (–6,1) = 24,6.

Т.е. разница между наибольшим и наименьшим значением выборки равна 24,6% .

Размах позволяет измерить общий разброс данных, но не учитывает, как именно распределены данные между минимальным и максимальным

$$S = \sqrt{\frac{\sum_{i=1}^{n}(X_1 - \overline{X})^2}{n-1}}, \qquad (1.10)$$

в Excel используется функция =СТАНДОТКЛОН.В(). Для расчета этих функций массив данных может быть неупорядоченным.

Стандартное отклонение для нашей выборки равно 6,6 (рис. 1.8). Это значит, что основная масса данных отличается от среднего значения не более чем на 6,6% (т.е. колеблется в интервале от $\overline{X} - S = 6{,}2 - 6{,}6 = -0{,}4$ до $\overline{X} + S = 12{,}8$).

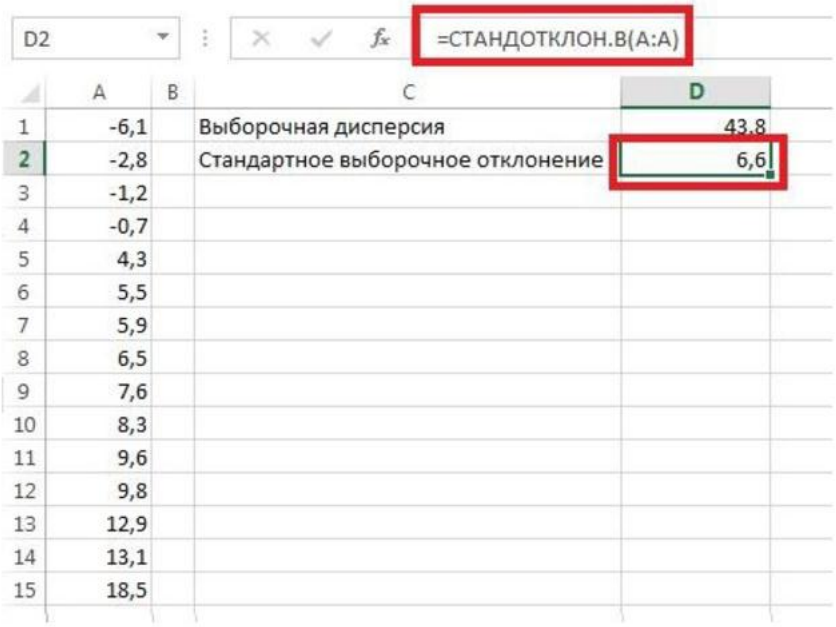


Рис. 1.8. Стандартное выборочное отклонение

Ни выборочная дисперсия, ни стандартное выборочное отклонение не могут быть отрицательными. Единственная ситуация, в которой показатели S^2 и S могут быть нулевыми, — если все элементы выборки равны между собой. В этом совершенно невероятном случае размах и межквартильный размах также равны нулю.

Коэффициент вариации измеряет рассеивание данных относительно среднего значения и измеряется в процентах (является относительной оценкой). Коэффициент вариации равен стандартному отклонению, деленному на среднее арифметическое и умноженному на 100%:

$$V = \frac{S}{\overline{x}} \cdot 100\%, \qquad (1.11)$$

где S — стандартное выборочное отклонение, $\overline{x}$ — выборочное среднее.

Коэффициент вариации позволяет сравнить две выборки, элементы которых выражаются в разных единицах измерения.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_1-\mu)^2}{N}, \quad (1.13)$$

где σ^2 – дисперсия генеральной совокупности. В Excel для вычисления дисперсии генеральной совокупности используется функция =ДИСП.Г().

Стандартное отклонение генеральной совокупности равно квадратному корню, извлеченному из дисперсии генеральной совокупности:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_1-\mu)^2}{N}}, \quad (1.14)$$

В Excel для вычисления стандартного отклонения генеральной совокупности используется функция =СТАНДОТКЛОН.Г().

Формулы для дисперсии и стандартного отклонения генеральной совокупности отличаются от формул для вычисления выборочной дисперсии и стандартного отклонения. При вычислении выборочных статистик S^2 и S знаменатель дроби может быть равен $n - 1$, а при вычислении параметров σ^2 и σ — объему генеральной совокупности N.

Эмпирическое правило. Если распределение не имеет ярко выраженной асимметрии, а данные концентрируются вокруг некоего центра тяжести, для оценки изменчивости можно применять эмпирическое правило, которое гласит: если данные имеют колоколообразное распределение, то приблизительно 68% наблюдений отстоят от математического ожидания не более чем на одно стандартное отклонение, приблизительно 95% наблюдений отстоят от математического ожидания не более чем на два стандартных отклонения и 99,7% наблюдений отстоят от математического ожидания не более чем на три стандартных отклонения. Следовательно, для колоколообразных распределений значения, лежащие за пределами интервала $\mu \pm 2\sigma$, можно считать выбросами. Кроме того, только три из 1000 наблюдений отличаются от математического ожидания больше чем на три стандартных отклонения. Таким образом, значения, лежащие за пределами интервала $\mu \pm 3\sigma$ практически всегда являются выбросами.

Для распределений, имеющих сильную асимметрию или не имеющих колоколообразной формы, можно применять эмпирическое правило Бьенамэ-Чебышева: для любого набора данных, независимо от формы распределения, процент наблюдений, лежащих на расстоянии не превышающем k стандартных отклонений от математического ожидания, не меньше $(1 - 1/k^2)*100\%$.

Вычисление описательных статистик для распределения на основе частот. Если исходные данные недоступны, единственным источником информации становится распределение частот. В таких ситуациях можно вычислить приближенные значения количественных показателей

В процентах	2010	2011	2012	2013	Накопленная сумма процентов за 2013
до 5000	9,4	7,3	5,9	6,0	6,0
5000,1 – 7000,0	9,4	8,1	7,0	7,4	13,4
7000,1 – 10 000,0	14,6	13,4	12,1	13,0	26,4
10 000,1 – 14 000,0	16,6	16,2	15,4	16,4	42,8
14000,1 – 19 000,0	15,2	15,6	15,5	16,2	59,0
19 000,1 – 27 000,0	14,7	15,9	16,6	16,7	75,7
27 000,1 – 45 000,0	13,3	15,1	17,0	15,9	91,6
свыше 45 000	6,8	8,4	10,5	8,4	100,0
Всего, %	100	100	100	100	
Данные 2013 года относятся к первому кварталу и являются предварительными					

Рис. 1.10. Пример расчета нижнего квартиля

В нашем примере (рис. 1.10) нижний квартиль находится в интервале 7000,1 – 10 000, накопленная частота которого равна 26,4%. Нижняя граница этого интервала – 7000 руб., величина интервала – 3000 руб., накопленная частота интервала, предшествующего интервалу, содержащему нижний квартиль – 13,4%, частота интервала, содержащего нижний квартиль – 13,0%. Таким образом: $Q_1 = 7000 + 3000 * (¼ * 100 – 13,4) / 13 = 9677$ руб.

В Excel **описательные статистики** можно получить с помощью надстройки *Пакет анализа*. Если на вкладе *Данные* в области *Анализ* у вас не отображается пиктограмма *Анализ данных*, нужно предварительно установить надстройку *Пакет анализа*. В меню *Данные → Анализ данных* в открывшемся окне выберите строку *Описательная статистика* и кликните *Ok*. В окне *Описательная статистика* обязательно укажите *Входной интервал* (рис. 11).

Если вы хотите увидеть описательные статистики на том же листе, что и исходные данные, выберите переключатель *Выходной интервал* и укажите ячейку, куда следует поместить левый верхний угол выводимых статистик (в нашем примере C1). Если вы хотите вывести данные на новый лист или в новую книгу, достаточно просто выбрать соответствующий переключатель. Поставьте галочку напротив *Итоговая статистика*. По желанию также можно выбрать *Уровень сложности, k-ый наименьший и k-й наибольший*.

Excel вычисляет целый ряд статистик, рассмотренных выше: среднее, медиану, моду, стандартное отклонение, дисперсию, размах (*интервал*), минимум, максимум и объем выборки (*счет*). Кроме того, Excel вычисляет некоторые новые для нас статистики: стандартную ошибку, эксцесс и асимметричность.

– расстояние от X_{min} до Q_1 равно расстоянию от Q_3 до X_{max}.

– расстояние от Q_1 до медианы равно расстоянию от медианы до Q_3.

Когда данные распределены несимметрично, между элементами пятерки показателей возникают следующие зависимости:

– если распределение имеет положительную асимметрию, расстояние от X_{min} до медианы меньше расстояния от медианы до X_{max}.

– если распределение имеет положительную асимметрию, расстояние от Q_3 до X_{max} больше, чем от X_{min} до Q_1.

– если распределение имеет отрицательную асимметрию, расстояние от X_{min} до медианы больше расстояния от медианы до X_{max}.

– если распределение имеет отрицательную асимметрию, расстояние от Q3 до X_{min} меньше, чем от X_{max} до Q_1.

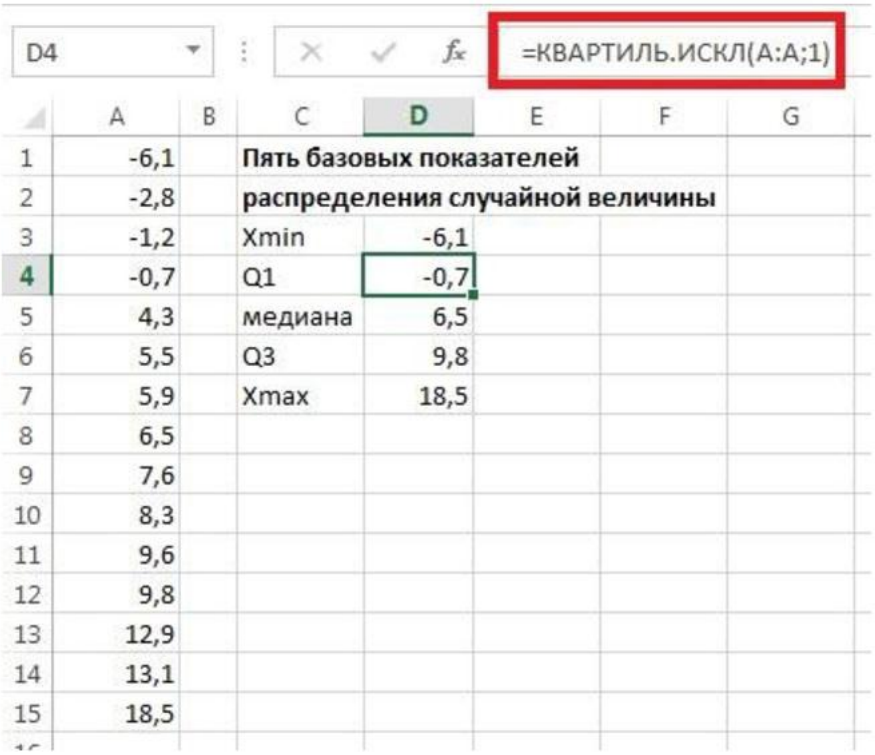


Рис. 1.12. Пятерка базовых показателей, характеризующих распределение случайной выборки

Исследуем на их основе симметричность распределения. Расстояние от медианы до X_{max} (18,5 – 6,5 = 12) приблизительно равно расстоянию от X_{min} до медианы (6,5 – (–6,1) = 12,6). Однако расстояние от Q_3 до X_{max} (18,5 – 9,8 = 8,7) превышает расстояние от X_{min} до Q_1 (–0,7 – (–6,1) = 5,4). Следовательно, распределение имеет слабую положительную асимметрию.

Точечная диаграмма позволяет наглядно представить саму выборку, пятерку базовых показателей и интервалы $\bar{X} \pm S$, $\bar{X} \pm 2S$, где $\bar{X}$ – среднее арифметическое выборки, S – стандартное отклонение выборки (рис. 1.13).

Таблица 1.2 – Основные характеристики выборки

Название	Обозначение	Название в сводной таблице	Метод вычисления	Формула Excel
Показатели вариации и центральной тенденции				
Размах вариации	R	Интервал	Разница max и min значений	МАКС (интервал) - МИН (интервал)
Объем выборки	n	Счет	Кол-во статистических единиц	СЧЕТ (интервал)
Медиана	Me	Медиана	Центральное значение отсортированной выборки	МЕДИАНА (интервал)
Мода	Mo	Мода	Наиболее часто встречающееся значение выборки	МОДА (интервал)
Среднее	$\bar{x}$	Среднее	Среднее арифметическое	СРЗНАЧ (интервал)
Среднее линейное отклонение	d	-	Средний модуль отклонения от среднего значения	СРОТКЛ (интервал)
Коэффициент осцилляции	V_R	-	$V_R = \frac{R}{\bar{x}} \cdot 100\%$	-
Линейный коэффициент вариации	V_d	-	$V_d = \frac{d}{\bar{x}} \cdot 100\%$	-
Коэффициент вариации	V_σ	-	$V_\sigma = \frac{\sigma}{\bar{x}} \cdot 100\%$	-
Показатели разброса данных (изменчивости)				
Дисперсия	D	Дисперсия	Средний квадрат отклонения от среднего значения	ДИСП (интервал)
Среднее квадратичное отклонение	σ	-	Среднее квадратичное отклонение от среднего значения	СТАНДОТКЛОНП (интервал)
Среднее квадратичное отклонение (несмещенное)	σ	Стандартное отклонение	Среднее квадратичное отклонение от среднего значения с поправкой на объем выборки	СТАНДОТКЛОН (интервал) – несмещенная оценка
Асимметрия	A	Асимметричность	$A = \frac{\sum (x_i - \bar{x})^3}{n\sigma^3}$	-
Эксцесс	E	Эксцесс	$E = \frac{\sum (x_i - \bar{x})^4}{n\sigma^4} - 3$	-

Лабораторная работа № 2

Построение доверительного интервала для математического ожидания генеральной совокупности (4 часа)

Цель: Изучить процедуру построения доверительного интервала для математического ожидания генеральной совокупности.

Теоретические сведения

Существует два вида оценок статистических данных: точечные и интервальные. *Точечная оценка* представляет собой отдельную выборочную статистику, которая используется для оценки параметра генеральной совокупности. Например, выборочное среднее — это точечная оценка математического ожидания генеральной совокупности, а выборочная дисперсия S^2 — точечная оценка дисперсии генеральной совокупности σ^2. Выборочное среднее является несмещенной оценкой математического ожидания генеральной совокупности, поскольку среднее значение всех выборочных средних (при одном и том же объеме выборки n) равно математическому ожиданию генеральной совокупности.

Для того чтобы выборочная дисперсия S^2 стала несмещенной оценкой дисперсии генеральной совокупности σ^2, знаменатель выборочной дисперсии следует положить равным $n - 1$, а не n. Иначе говоря, дисперсия генеральной совокупности является средним значением всевозможных выборочных дисперсий.

Для получения *интервальной оценки* математического ожидания генеральной совокупности анализируют распределение выборочных средних. Построенный интервал характеризуется определенным доверительным уровнем, который представляет собой вероятность того, что истинный параметр генеральной совокупности оценен правильно. Аналогичные доверительные интервалы можно применять для оценки доли признака p и основной распределенной массы генеральной совокупности.

Построение доверительного интервала для математического ожидания генеральной совокупности при известном стандартном отклонении. *Интервальная оценка*, доверительный уровень которой равен 95%, интерпретируется следующим образом: если из генеральной совокупности извлечь все выборки, имеющие объем n, и вычислить их выборочные средние, то 95% доверительных интервалов, построенных на их основе, будут содержать математическое ожидание генеральной совокупности, а 5% — нет. На практике, как правило, из генеральной совокупности извлекается только одна выборка, а математическое ожидание генеральной совокупности μ не известно. По этой причине невозможно гарантировать, что некий конкретный доверительный интервал содержит величину μ. Можно лишь утверждать, что вероятность этого события равна 95%.

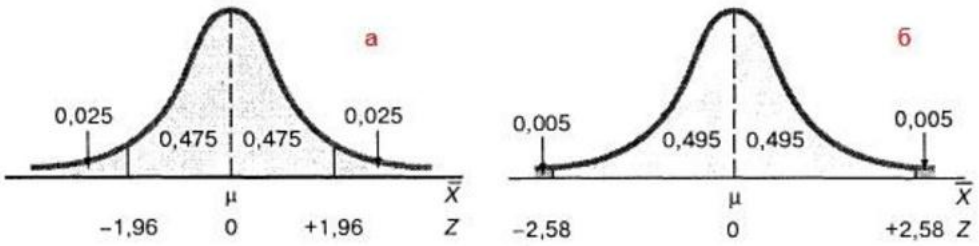


Рис. 2.1. Гауссова кривая для определения критического значения Z, соответствующего доверительному уровню, равному: (а) 95%; (б) 99%

Пример 1. При производстве бумаги средняя длина листа должна быть равной 11 дюймам, а ее стандартное отклонение — 0,02 дюйма. Периодически из произведенной продукции, чтобы оценить ее качество, извлекаются выборки. Допустим, выборка состоит из 100 листов, а ее выборочное среднее — 10,998 дюйма. Постройте интервал, содержащий математическое ожидание генеральной совокупности, доверительный уровень которого равен 95%.

Решение. Подставим в формулу (1) величину Z = 1,96, соответствующую доверительному уровню, равному 95%:

$$\overline{X} \pm Z\frac{\sigma}{\sqrt{n}} \leq 10.998 \pm 1.96\frac{0.02}{\sqrt{100}} = 10.998 \pm 0.00392$$

$$10.99408 \leq \mu \leq 11.00192$$

Таким образом, вероятность того, что математическое ожидание генеральной совокупности лежит в интервале от 10,99408 до 11,00192, равна 95%. Поскольку номинальная длина бумаги — 11 дюймов, она попадает в построенный интервал. Следовательно, производственный процесс выполняется правильно.

В Excel используется функция =ДОВЕРИТ.НОРМ(), возвращающая доверительный интервал для среднего генеральной совокупности с использованием нормального распределения. Для приведенного выше примера 1 вычисления в Excel показаны на рис. 2.3.

Построение доверительного интервала для математического ожидания генеральной совокупности при неизвестной дисперсии. На практике как математическое значение генеральной совокупности, так и его стандартное отклонение часто бывают неизвестными. Следовательно, необходимо построить доверительный интервал, содержащий математическое значение генеральной совокупности, используя лишь выборочные статистики $\overline{X}$ и S.

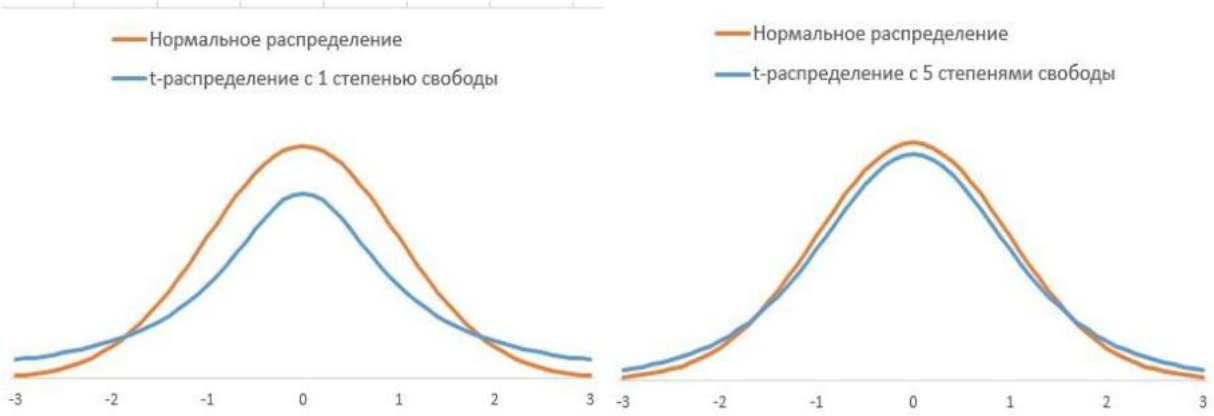
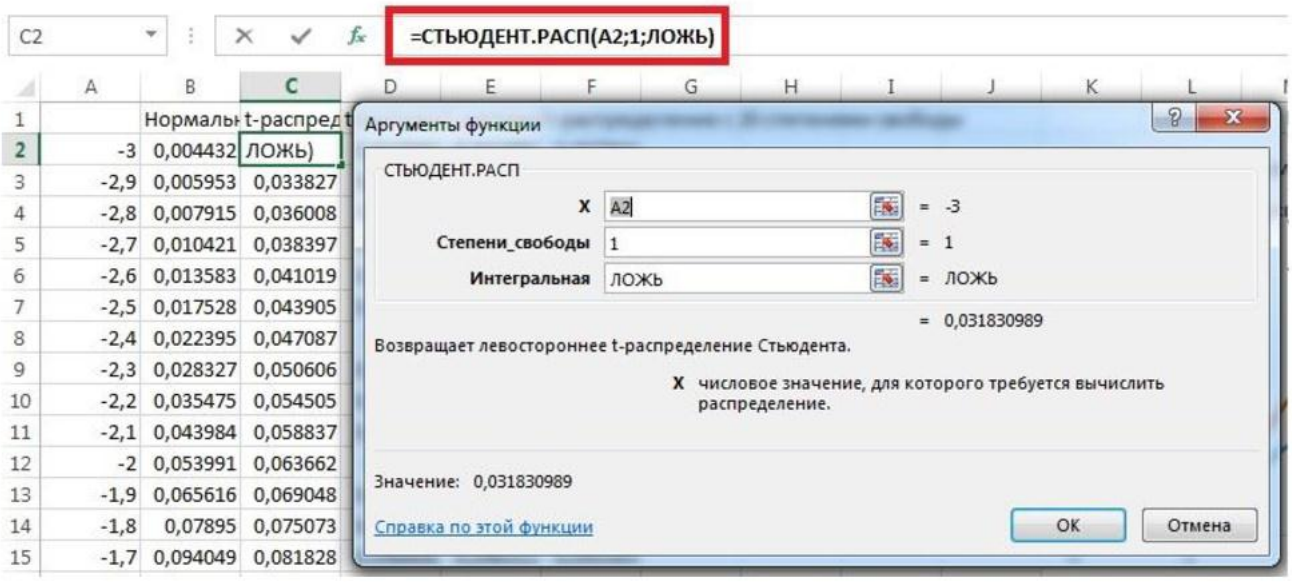


Рис. 2.4. Стандартизованное нормальное распределение и *t*-распределение Стьюдента с различным числом степеней свободы



	A	B	C
1		Нормальн	t-распред t
2	-3	0,004432	ЛОЖЬ)
3	-2,9	0,005953	0,033827
4	-2,8	0,007915	0,036008
5	-2,7	0,010421	0,038397
6	-2,6	0,013583	0,041019
7	-2,5	0,017528	0,043905
8	-2,4	0,022395	0,047087
9	-2,3	0,028327	0,050606
10	-2,2	0,035475	0,054505
11	-2,1	0,043984	0,058837
12	-2	0,053991	0,063662
13	-1,9	0,065616	0,069048
14	-1,8	0,07895	0,075073
15	-1,7	0,094049	0,081828

Рис. 2.5. Функция Excel =СТЬЮДЕНТ.РАСП() для построения плотности вероятности

Напомним, что *t*-распределение основано на предположении, что изучаемая случайная величина X является нормально распределенной. Однако на практике *t*-распределение можно применять для оценки неизвестного математического ожидания генеральной совокупности при неизвестном стандартном отклонении при достаточно большом объеме выборки и не слишком асимметричном распределении. При работе с небольшими выборками эти условия уже не выполняются автоматически, поэтому их следует проверять. Для этого необходимо строить гистограмму, диаграмму «ствол и листья», блочную диаграмму или график нормального распределения.

Степени свободы. Для вычисления выборочной дисперсии S^2 необходимо вычислить величину:

$$\sum_{i=1}^{n}(X_i - \bar{X})^2. \qquad (2.5)$$

значение, равное 1,9842. Поскольку *t*-распределение является симметричным и его математическое ожидание равно 0, площади, ограниченной правым хвостом, соответствует величина +1,9842, а площади, ограниченной левым хвостом, соответствует величина –1,9842. Величина 1,9842 означает следующее: вероятность того, что величина *t* превосходит +1,9842, равна 0,025, т.е. 2,5% (рис. 2.7). Вместо таблиц можно использовать функцию =СТЬЮДЕНТ.ОБР.2Х() (рис. 2.8).

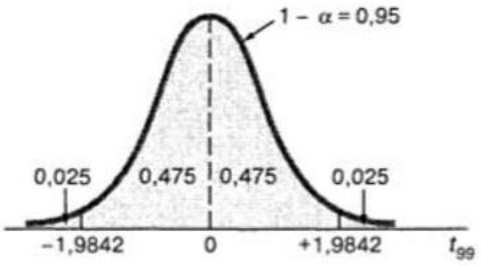


Рис. 2.7. Распределение Стьюдента с 99 степенями свободы

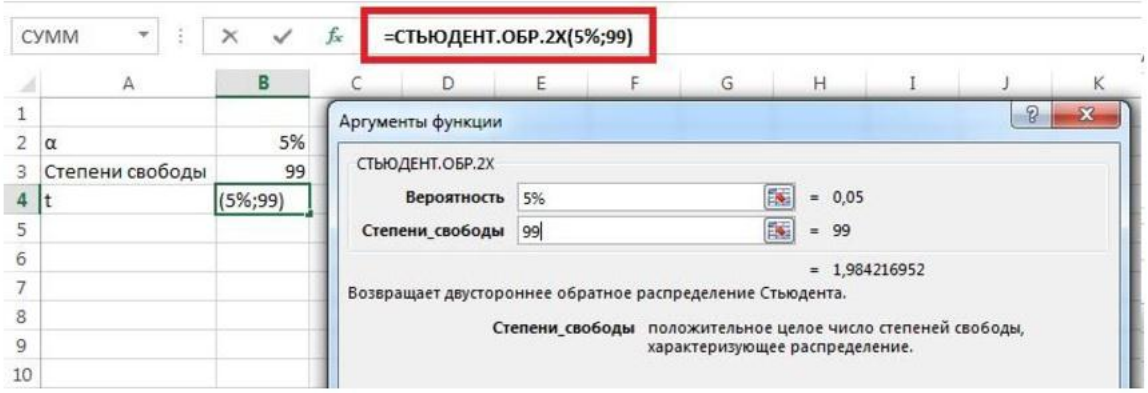


Рис. 2.8. Использование функции =СТЬЮДЕНТ.ОБР.2Х() для определения *t*-параметра по известной вероятности (обратная задача построению плотности распределения)

Доверительный интервал рассчитывается по формуле, содержащей математическое ожидание при неизвестном стандартном отклонении с вероятностью $(1 - \alpha)\cdot 100\%$:

$$\overline{X} \pm t_{n-1}\frac{\sigma}{\sqrt{n}}\ , \tag{2.6}$$

$$\overline{X} - t_{n-1}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + t_{n-1}\frac{\sigma}{\sqrt{n}}, \tag{2.7}$$

где t_{n-1} — критическое значение *t*-распределения с $n - 1$ степенями свободы, соответствующее площади, ограниченной правым хвостом и равной $\alpha/2$.

Проиллюстрируем применение этой формулы (рис. 2.9).

мода и медиана — совпадают друг с другом. Межквартильный размах нормального распределения равен 1,33 стандартного отклонения. Нормальное распределение является непрерывным, причем нормально распределенная случайная величина принимает произвольные значения, лежащие на всей числовой оси.

Многие непрерывные случайные величины не являются ни точно, ни приближенно нормальными. Свойства таких величин довольно сильно отличаются от свойств нормального распределения, перечисленных выше. Рассмотрим, например, оценки, полученные студентами при сдаче четырех тестов (рис. 2.10). Excel справляется с обработкой данных, не требуя их упорядочения. Вычислим описательные статистики результатов каждого теста в отдельности с помощью надстройки *Анализ данных*.

	A	B	C	D	E	F	G	H	I	J
1	Тест 1	Тест 2	Тест 3	Тест 4						
2	70	77	64	77			*Тест 1*	*Тест 2*	*Тест 3*	*Тест 4*
3	69	77	62	74						
4	64	73	55	62		Среднее (математическое ожидание)	65,0	70,7	59,3	65,0
5	55	58	50	44		Стандартная ошибка	2,1	2,3	2,3	3,9
6	72	78	66	80		Медиана	65,0	74,0	56,0	65,0
7	68	76	59	71		~~Мода~~	~~#Н/Д~~	~~77,0~~	~~53,0~~	~~#Н/Д~~
8	52	54	48	41		Стандартное отклонение	9,0	10,0	10,0	16,9
9	60	66	53	53		Дисперсия выборки	80,2	100,0	100,0	285,0
10	62	71	54	59		Эксцесс	-0,4	0,2	0,2	-1,2
11	58	64	52	50		Асимметричность	0,0	-1,0	1,0	0,0
12	78	82	76	89		Интервал (размах)	34	36	36	54
13	82	83	83	92		Минимум	48	47	47	38
14	66	75	57	68		Максимум	82	83	83	92
15	65	74	56	65		Сумма	1235	1343	1127	1235
16	57	61	51	47		Счет (объем выборки)	19	19	19	19
17	75	80	72	86						
18	48	47	47	38		Межквартильный размах	12,0	12,5	12,5	27,0
19	61	68	53	56		1,33σ	11,9	13,3	13,3	22,5
20	73	79	69	83		Размах	34	36	36	54
21						6σ	53,7	60,0	60,0	101,3
22						Доля наблюдений в окрестности ±σ	68%	74%	74%	58%
23						Доля наблюдений в окрестности ±2σ	100%	95%	95%	100%
24										

Рис. 2.10. Оценки, полученные студентами при сдаче четырех тестов (мода зачеркнута, так как не имеет смысла)

Приблизительно нормальным является распределение оценок только по первому тесту: математическое ожидание равно медиане, доля наблюдений в пределах окрестности ±1σ от математического ожидания составляет 68% (в точности, как и для нормального распределения), асимметричность = 0.

Второй подход к проверке гипотезы о нормальном распределении использует график. Для оценки смещения распределения были введены квартили. Кроме квартилей, для оценки нормальности распределения можно вычислять децили (разбивающие диапазон изменения данных на десятые доли), процентили (разбивающие диапазон изменения данных на сотые доли) и квантили (от слова *квант*), разбивающие всю совокупность данных на *n* диапазонов.

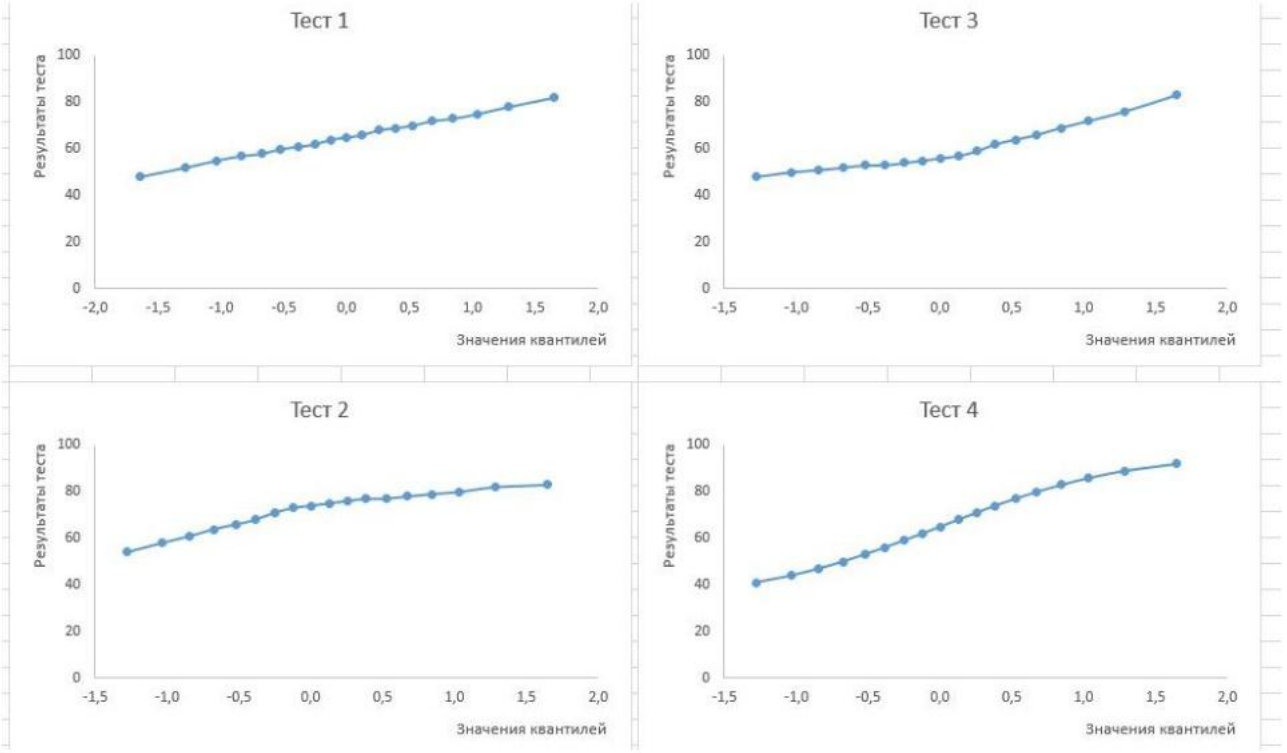


Рис. 2.13. Графики распределений для четырех тестов

График «Тест 1» свидетельствует, что наблюдаемые точки лежат очень близко к прямой линии, поэтому можно считать, что оценки, полученные студентами при сдаче первого теста, распределены практически нормально. «Тест 2» соответствует распределению с отрицательной асимметрией, о чем свидетельствует более длинный левый хвост распределения. «Тест 3»: наблюдается противоположная картина, соответствует распределению с положительной асимметрией, о чем свидетельствует более длинный правый хвост распределения. «Тест 4»: изображен симметричный график, средняя часть которого почти линейна. Значения случайной переменной сначала довольно медленно возрастают, затем их рост прекращается, а в третьей части — ускоряется. Этот рисунок не совпадает ни с панелью Б, ни с панелью В. Это распределение не имеет хвостов. Следовательно, оно является равномерным (или прямоугольным).

Оценка математического ожидания. Поправочный коэффициент для конечной генеральной совокупности используется для уменьшения стандартной ошибки в $\sqrt{\frac{N-n}{N-1}}$. Таким образом, доверительный интервал для математического ожидания, имеющий доверительный уровень, равный *(1 – α)x100%*, вычисляется по формуле:

$$\bar{X} \pm t_{n-1} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}. \qquad (2.8)$$

Как правило, доверительный уровень равен 95%. Он делится на две равные части: первая часть лежит слева от математического ожидания генеральной совокупности, а вторая — справа. Значение величины Z, соответствующей вероятности 2,5% (площади 0,025), равно –1,96, а значение величины Z, соответствующей суммарной площади 0,975, равно +1,96. Для расчета удобно использовать функцию Excel *Z*=НОРМ.СТ.ОБР(p), где p – вероятность, подставляя значения $p_1 = 2{,}5\%$ и $p_2 = 97{,}5\%$.

Если требуется поднять доверительный уровень, обычно выбирают величину, равную 99%. Если можно ограничиться более низким доверительным уровнем, выбирают 90%. Определяя ошибку выборочного исследования, не стоит думать о ее величине (в принципе, любая ошибка нежелательна). Следует задать такую ошибку, чтобы полученные результаты допускали разумную интерпретацию.

В нашем примере при $e = 5$, $\sigma = 25$ и $Z = 1{,}96$ (что соответствует 95%-ному доверительному уровню). По формуле (2.10) получаем:

$$n = \left(\frac{1.96 * 25}{5}\right)^2 = 96$$

Следовательно, $n = 96$. Таким образом, объем выборки, равный 100, был выбран удачно.

Порядок выполнения работы

3. По данным варианта построить доверительный интервал для математического ожидания генеральной совокупности.
4. Рассчитать t-распределение Стьюдента и построить график (полигон). Определить критические значения t-распределения.
5. Вычислить доверительный интервал с помощью функции Excel =ДОВЕРИТ.СТЬЮДЕНТ().
6. Проверить гипотезу о нормальности распределения выборки с помощью надстройки MS Excel *Анализ данных* и построить график.

Данные: 64, 57, 63, 53, 62, 58, 61, 63, 47, 70, 53, 60, 61, 65, 62, 51, 62, 40, 64, 61, 59, 56, 59, 63, 61, 55, 57, 60, 52, 64.

Контрольные вопросы

1. В чем заключается суть выборочного метода? Какая выборка называется репрезентативной?
2. Что представляет собой точечный вариационный ряд?
3. Как строится интервальный вариационный ряд?
4. Что понимается под эмпирическим распределением, с помощью чего оно может быть представлено?
5. Как определяется эмпирическая функция распределения?
6. Статистическим аналогом чего является полигон относительных частот?

Лабораторная работа № 3

Точечный и интервальный вариационные ряды. Графическое представление вариационного ряда (4 часа)

Цель работы. Оценка свойств генеральной совокупности по эмпирическим (наблюдаемым) данным (выборке) путем построения эмпирического распределения, нахождения числовых характеристик выборки, нахождения точечных и интервальных оценок параметров нормального распределения, проверки гипотезы о нормальном распределении генеральной совокупности.

Теоретические сведения

Выборочные данные, упорядоченные по возрастанию или убыванию, называются *вариационным рядом*. Различные значения исследуемого признака в выборке называются *вариантами*.

Упорядоченная по возрастанию или убыванию последовательность вариант x_i с указанием частот f_i (или относительной частоты w_i) их повторения в выборке называется *точечным вариационным рядом*. Точечный вариационный ряд представляется таблицей, в первой строке (столбце) которой приводятся упорядоченные по возрастанию варианты x_i, в последующих строках (столбцах) соответствующие им частоты f_i и относительные частоты w_i (табл. 3.1).

Таблица 3.1– Пример точечного вариационного ряда

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k
$\frac{n_i}{n}$	$\frac{n_1}{n}$	$\frac{n_2}{n}$	…	$\frac{n_k}{n}$
m_i	$m_1 = n_1$	$m_2 = n_1 + n_2$	…	$m_i = n$

Здесь k – количество различных вариант в выборке, m_i – накопленные частоты вариант x_i.

Точечный вариационный ряд является статистическим аналогом ряда распределения дискретной случайной величины.

При большом количестве вариант или при непрерывном характере исследуемого признака от точечного вариационного ряда переходят к *интервальному вариационному ряду* – выборочным данным сгруппированным по k последовательным интервалам числовой оси. Количество интервалов k определяется как натуральное число $k \approx 1 + \log_2 n$. Длины h интервалов группирования находятся как $h = \frac{x_{max} - x_{min}}{k}$, где x_{max} и x_{min} наибольшее и

На начальном этапе проведения группировки необходимо определить целесообразное число групп и величину интервала. Необходимо самостоятельно выбрать порядок образования интервалов, для этого можно воспользоваться формулой Стерджесса – (3.1).

Формируя группы, нужно помнить, что распределение единиц совокупности внутри групп должно быть как можно более равномерным (в каждую группу должно входить не менее 2-3-х значений), при этом распределение должно иметь только один модальный интервал (имеющий максимальную частоту). В целях достижения соответствия этим критериям фактическая величина интервала и число групп могут отличаться от расчётных значений.

Для удобства восприятия и анализа группировки рекомендуется брать величину интервала, кратную пяти, десяти, ста, и т.д. в зависимости от величины и степени вариации признака.

Построение интервального вариационного ряда и точечного вариационного ряда по серединам интервалов. Необходимо найти максимальное x_{max} и минимальное x_{min} значение в приведенных данных (функции МАКС(число1;[число2];...) и МИН(число1;[число2];...), в скобках указываем диапазон ячеек).

Количество интервалов k определяем по формуле Стерджесса :

$$k = 1 + 3{,}322 * lg\ n, \qquad (3.1)$$

где n – объем выборки.

Расчетное значение k округляем до ближайшего целого.

Определяем шаг интервала:

$$h = \frac{x_{\max} - x_{\min}}{k}. \qquad (3.2)$$

За начало C_0 первого интервала примем x_{min}. Последующие границы интервалов найдем как $Ci=Ci\text{-}1+h,\ i=1,\ 2\ldots$.

Частоты f_i интервалов определим по точечному вариационному ряду, середины интервалов вычислим как $x_i'=(C_{i-1}+ C_i)\ /2,\ i=1,\ 2\ldots$.

Если в какие-либо интервалы попадает менее 3 значений, количество интервалов рекомендуется сократить и дальнейшую обработку данных производить по скорректированное число интервалов.

В интервальном вариационном ряду столбцы с четвертого по седьмой (рис. 3.1) образуют точечный вариационный ряд, построенный по серединам интервалов. В последней строке интервального вариационного ряда вычислены контрольные суммы: сумма частот интервалов, которая должна равняться объему выборки, и сумма относительных частот интервалов, которая должна равняться единице. Для вычисления этих сумм выделяются ячейки, содержащие суммируемые величины, во вкладке «Формулы» выбирается

Интервал, имеющий наибольшую частоту, будет являться *модальным*, а конкретное (дискретное) значение моды будет находиться внутри него. Рассчитать конкретное, значение моды в интервальном ряду можно по следующей формуле:

$$M_0 = x_{M0} + i\frac{f_{M0} - f_{M0-1}}{(f_{M0} - f_{M0-1}) + (f_{M0} - f_{M0+1})}, \quad (3.4)$$

где x_{Mo} — нижняя граница модального интервала,
i — длина модального интервала;
f_{Mo} — частота модального интервала;
f_{Mo-1} — частота, соответствующая предшествующему интервалу;
f_{Mo+1} — частота, соответствующая последующему интервалу.

Медиана применяется для количественной характеристики структуры и равна такому варианту, который делит ранжированную совокупность на две равные части. У одной половины совокупности признаки не больше медианы (меньше или равны), у второй — не меньше медианы (больше или равны). Если рассматриваемый ряд интервальный, то накопленные частоты покажут нам медианный интервал.

Конкретное значение медианы рассчитывается по формуле:

$$Me = x_{Me} + i\frac{\frac{\sum f}{2} - f'_{Me-1}}{f_{Me}}, \quad (3.4)$$

где i — длина медианного интервала;
f'_{Me-1} — накопленная частота в интервале, предшествующем медианному;
f_{Me} — частота медианного интервала.

Размах вариации составит (абсолютная вариация):

$$R = x_{max} — x_{min}. \quad (3.5)$$

Для **графического представления свойств выборки** наиболее часто используются гистограмма и полигон частот (или относительных частот), а также кумулята (рис. 3.2).

Гистограмма частот – это графическое представление выборки, где по оси абсцисс (OX) отложены величины интервалов, а по оси ординат (OУ) – величины частот f_i, попадающих в данный i-й интервал. При увеличении до бесконечности размера выборки выборочные функции распределения превращаются в теоретические: гистограмма превращается в график плотности распределения.

равной нулю. Другие точки этой ломаной соответствуют концам интервалов и накопленным частотам.

Кумулята относительных частот отличается тем, что накапливаются относительные частоты: $w_i^{нак} = \frac{m_i^{нак}}{n}$.

По виду полигона и рассчитанным характеристикам можно сделать вывод о характере распределения. Так, если мода, медиана и средняя примерно равны, то распределение данных близко к нормальному.

В MS Excel **для построения выборочных функций распределения** используются специальная функция ЧАСТОТА и процедура *Гистограмма* из пакета анализа (*Надстройки → Анализ данных*).

Функция ЧАСТОТА (*массив_данных, двоичный_массив*) вычисляет частоты появления случайной величины в интервалах значений и выводит их как массив цифр:

- *массив_данных* – это массив или ссылка на множество данных, для которых вычисляются частоты;
- *двоичный_массив* — это массив интервалов, по которым группируются значения выборки.

Процедура *Гистограмма* выводит результаты выборочного распределения в виде таблицы и графика. Параметры диалогового окна *Гистограмма*:

- *входной диапазон* – диапазон исследуемых данных (выборка);
- *интервал карманов* – диапазон ячеек или набор граничных значений, определяющих выбранные интервалы (карманы). Эти значения должны быть введены в возрастающем порядке.

Если диапазон карманов не был введен, то набор интервалов, равномерно распределенных между минимальным и максимальным значениями данных, будет создан автоматически.

- *выходной диапазон* – предназначен для ввода ссылки на левую верхнюю ячейку выходного диапазона;
- переключатель *Интегральный процент* – параметр, который позволяет установить режим включения в гистограмму графика интегральных процентов;
- переключатель *Вывод графика* – параметр, который позволяет установить режим автоматического создания встроенной диаграммы на листе, содержащем выходной диапазон.

Для построения отдельных графиков используют *Мастер диаграмм* закладку *Стандартные*: режимы «гистограмма» и «график» (для полигона). С помощью диалогового меню *Мастера диаграмм* подписывают графические данные, называют оси и наносят числовые параметры и линии сетки, выводят величины самих значений.

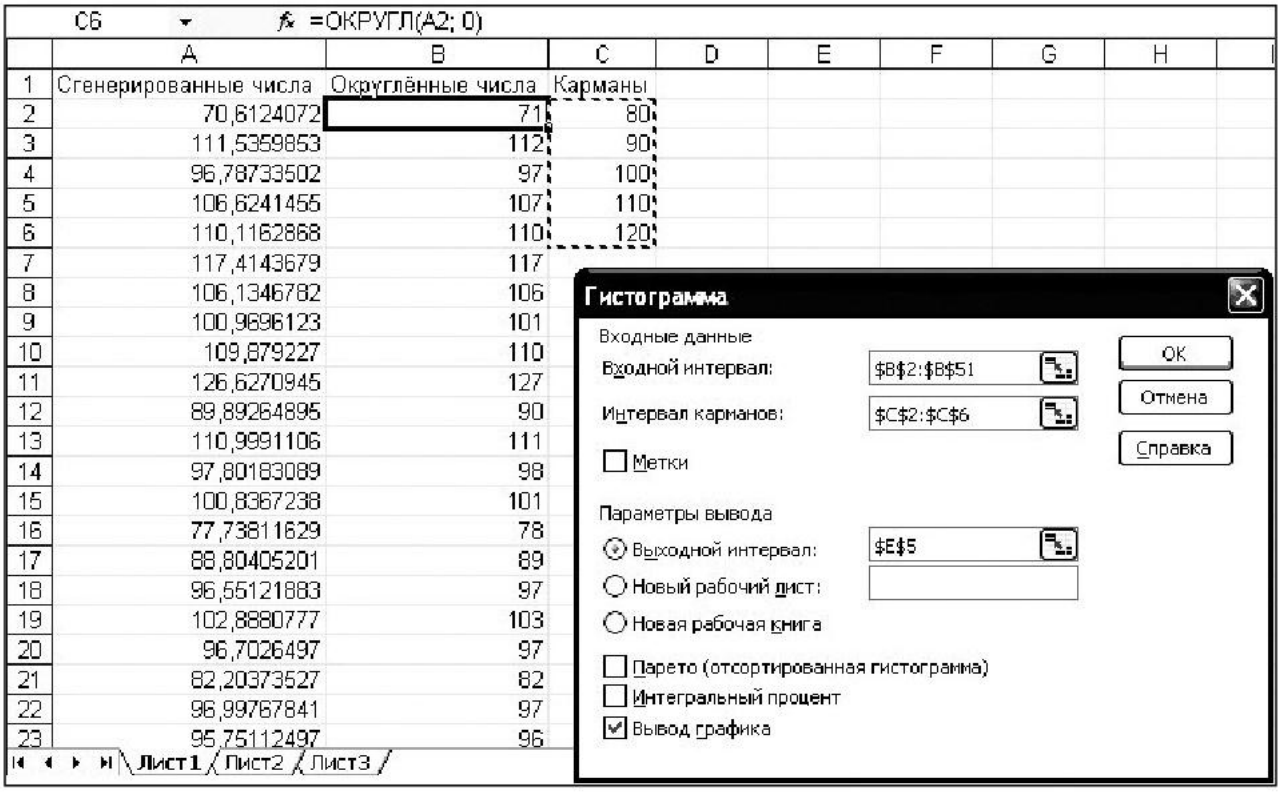


Рисунок 3.5. Настройка функции «Гистограмма»

Для группировки данных с помощью статистической надстройки выбираем меню **Сервис - Анализ данных - Гистограмма**. Указываем следующие параметры:

Входные данные:

– *Входной интервал* – выборка исходных данных;

– *Интервал карманов* – нижние границы интервалов группировки. Этот параметр можно не задавать – тогда программа сама выберет необходимое количество интервалов группировки и определит их границы.

Параметры вывода:

– *Выходной интервал* – расположение результатов группировки на листе;

– *Вывод графика* – построение гистограммы.

– *Интегральный процент* – вычисление накопленных частностей.

Результат работы функции **Гистограмма** представлен на рис. 3.6.

Порядок выполнения работы

По заданной выборке значений изучаемого признака генеральной совокупности в соответствии с номером варианта:

1. Построить точечный вариационный ряд. Выборочные данные отредактировать (вкладка «*Редактирование*» → «*Сортировка и фильтр*» → «*Сортировка по возрастанию*»).

2. От точечного вариационного ряда перейти к интервальному вариационному ряду.

3. Построить точечный вариационный ряд по серединам интервалов.

4. Найти интервальные выборочные среднюю, дисперсию, среднеквадратическое отклонение, моду, медиану, коэффициент вариации, размах вариации. Промежуточные данные анализа вариационного ряда представить в виде расчетной таблицы.

Номер интервала	Диапазон значений	f	x'_i	$x'_i \cdot f$	$\left(x_i' - \bar{x}\right)^2 f_i$

5. По серединам интервалов построить графики вариационного ряда: полигон частот, гистограмму, кумуляту.

6. На основании произведенных расчетов сделать вывод о характере выборочного распределения.

Пример. В качестве примера выборки возьмем распределение веса студентов (в кг): 64, 57, 63, 53, 62, 58, 61, 63, 47, 70, 53, 60, 61, 65, 62, 51, 62, 40, 64, 61, 59, 56, 59, 63, 61, 55, 57, 60, 52, 64.

Алгоритм построения гистограммы и кумуляты относительных частот.

1. В ячейку A1 введите слово *Наблюдения*, в ячейку B1 – *Вес, кг*.

2. В диапазон A2:A21 введите значения веса студентов (см. рис. 1). В диапазон B2:B8 введите граничные значения интервалов (40, 45, 50, 55, 60, 65, 70).

3. Введите заголовки создаваемой таблицы: в ячейки C1 – *Абсолютные частоты*, в D1 – *Относительные частоты*, в E1 – *Накопленные относительные частоты,* в *F1 – Накопленные абсолютные частоты.*

4. С помощью функции **ЧАСТОТА** заполните столбец абсолютных частот, для этого выделите блок ячеек C2:C8. С панели инструментов **Стандартная** вызовите **Мастер функций** (кнопка ***fx***). Для Excel 2007 на панели инструментов выберите вкладку **Функции** и нажмите кнопку **Вставить функцию**. В появившемся диалоговом окне выберите категорию **Статистические** и функцию **ЧАСТОТА**, после чего нажмите кнопку **ОК**. Указателем мыши в рабочее поле **Массив_данных** введите диапазон данных наблюдений (A2:A21). В рабочее поле **Двоичный_массив** мышью введите диапазон интервалов (B2:B8).

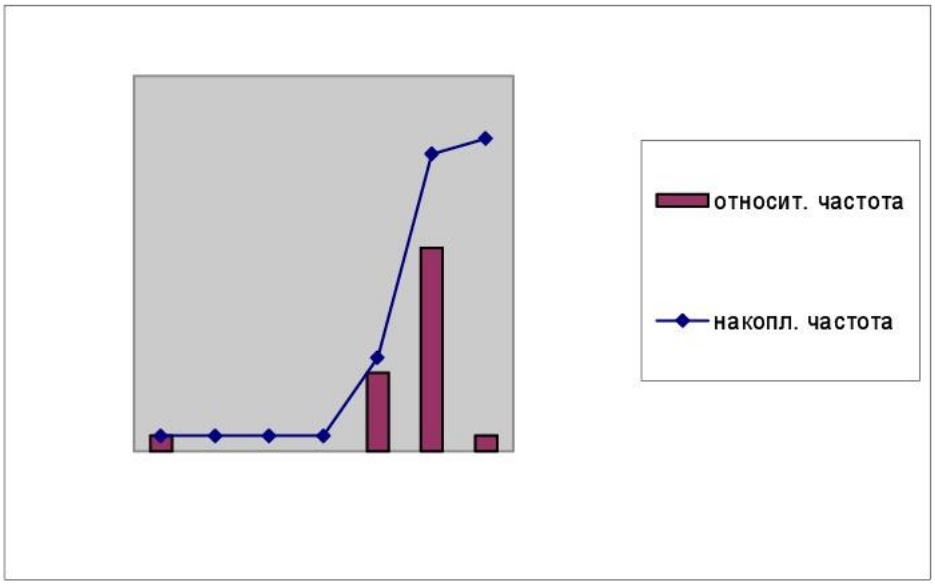


Рисунок 3.9. Гистограмма и кумулята относительных частот

Контрольные вопросы

1. Сформулируйте понятия генеральной совокупности и выборки.
2. В чем заключается суть выборочного метода?
3. Какие методы формирования выборки существуют?
4. Какая выборка называется репрезентативной (представительной)?
5. Что представляет собой точечный вариационный ряд?
6. Как строится интервальный вариационный ряд?
7. Что понимается под эмпирическим распределением и с помощью чего оно может быть представлено?
8. Как определяется эмпирическая функция распределения?
9. Статистическим аналогом чего является полигон относительных частот?
10. Приведите числовые характеристики положения выборки.
11. Сформулируйте понятия выборочной моды и медианы.
12. Как находится выборочная средняя по точечному вариационному ряду?
13. Как определяются выборочные мода и медиана по интервальному вариационному ряду?
14. Приведите числовые характеристики вариации выборки.
15. Как находятся выборочные дисперсия и среднее квадратическое отклонение?

Таблица 4.1 – Значения выборочных совокупностей и расчеты для нахождения эмпирического значения *t*

№	X_1	X_2	d	$(d - M_d)$	$(d - M_d)^2$
1	5	6	-1	0,333	0,111
2	8	7	1	2,333	5,444
3	9	10	-1	0,333	0,111
4	4	6	-2	-0,667	0,444
5	5	5	0	1,333	1,778
6	7	9	-2	-0,667	0,444
7	5	8	-3	-1,667	2,778
8	6	9	-3	-1,667	2,778
9	7	8	-1	0,333	0,111
∑			-12	0	14

В MS Excel расчет критерия *t* Стьюдента для зависимых выборок проводится с помощью опции *Парный двухвыборочный t-тест для средних*.

Данная опция находится в Анализе данных… Заходим в меню ***Сервис***, выбираем ***Анализ данных…*** Раскроется окно со списком *Инструментованализа*. В этом списке находим средство ***Парный двухвыборочный t-тест для средних***.

В появившимся диалоговом окне в поле *Интервал переменной 1* указывается диапазон ячеек первой выборки. В поле *Интервал переменной 2* указывается диапазон значений второй выборки. Если диапазон указывается вместе с заголовком, то напротив *Метки* ставим флажок. В поле *Альфа* указываем уровень значимости (по умолчанию стоит 0,05).

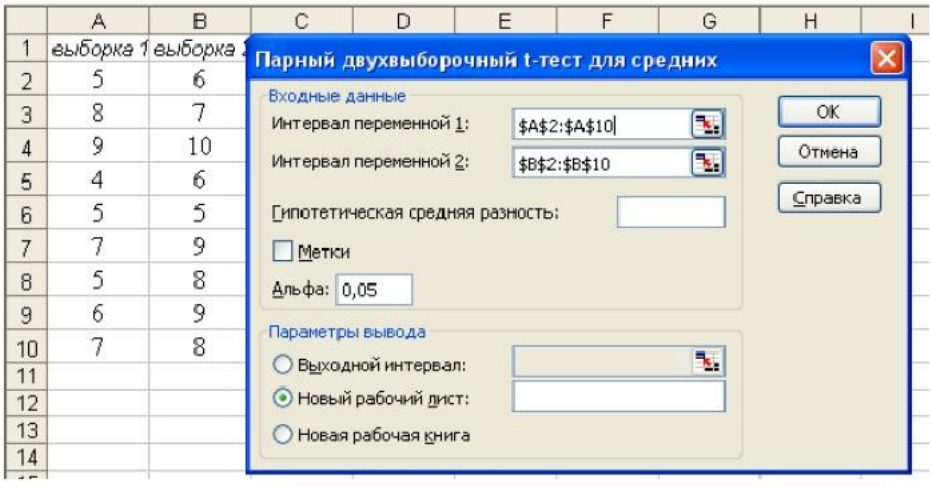


Рис. 4.1. Диалоговое окно для критерия *t* Стьюдента для зависимых выборок

Выходные результаты представлены в виде таблицы, где приводятся статистические характеристики выборочных значений, соответствующие

- Хвосты Обязательный. Число хвостов распределения. Если значение "хвосты" = 1, функция СТЬЮДЕНТ.ТЕСТ возвращает одностороннее распределение. Если значение "хвосты" = 2, функция СТЬЮДЕНТ.ТЕСТ возвращает двустороннее распределение.
- Тип Обязательный. Вид выполняемого t-теста: парный, гомоскедастический (двухвыборочный с равными дисперсиями) или гетероскедастический (Двухвыборочный с неравными дисперсиями).

Порядок выполнения работы

Задание. Надо сравнить между собой результаты выполнения тестов на внимание в двух группах. Чтобы узнать различаются ли группы между собой необходимо вычислить **t-критерий Стьюдента для независимых выборок.**

Алгоритм.

1. Внесем данные по группам в таблицу:
2. Проверить распределения на нормальность.
3. Рассчитать среднее арифметическое, стандартное отклонение и количество человек в каждой группе.
4. Вычисляем эмпирическое значения по формуле t-критерия Стьюдента для независмых выброк
5. Вычисляем степени свободы.
6. Определяем по таблице критическое значение t-Стьюдента для заданного уровня значимости и применяем правило статистического вывода:

Данные

№	Результаты группы №1 (сек.)	Результаты группы №2 (сек.)	№	Результаты группы №1 (сек.)	Результаты группы №2 (сек.)
1	30	46	16	34	42
2	45	49	17	33	40
3	41	52	18	49	58
4	38	55	19	32	54
5	34	56	20	46	53
6	36	40	21	41	51
7	31	47	22	44	57
8	30	51	23	38	56
9	49	58	24	50	44
10	50	46	25	37	42
11	51	46	26	39	49
12	46	56	27	40	50
13	41	53	28	46	55
14	37	57	29	42	43
15	36	44			

критическому значению 39,36 или превысит его, равна 0,025. Таким образом, вероятность того, что тестовая χ^2-статистика лежит между критическими значениями 12,40 и 39,36, равна 0,95. Т.е., задав уровень значимости и определив количество степеней свободы, мы можем найти критическое значение тестовой χ^2-статистики для любого конкретного распределения χ^2.

Пусть для анализа создана выборка, состоящая из 25элементов. Чтобы определить, отличается ли стандартное отклонение от заданной величины, равной, например, 15, можно применить двусторонний критерий. Нулевая и альтернативная гипотезы формулируются следующим образом: H_0: $\sigma = 15$; H_1: $\sigma \neq 15$.Нулевая гипотеза отклоняется, если тестовая χ^2-статистика попадает в область отклонения гипотезы, ограниченную левым или правым хвостами кривой распределения χ^2 с (25 — 1) = 24 степенями свободы. Следовательно, решающее правило формулируется следующим образом: гипотеза H_0 отклоняется, если $\chi^2 > \chi_U^2 = 39{,}364$ или $\chi^2 < \chi_L^2 = 12{,}401$, в противном случае гипотеза не отклоняется.

Предположим, что выборочное стандартное отклонение S, вычисленное для выборки $n = 25$, составляет 17,7. Для того чтобы проверить нулевую гипотезу при уровне значимости, равном 0,05, воспользуемся формулой (5.2):

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{(25-1)(17{,}7)^2}{15^2} = 33{,}42$$

Расчетное значение тестовой χ^2-статистики 33,42 лежит в интервале, ограниченном нижним и верхним критическими значениями, т.е. между 12,40 и 39,36, следовательно, гипотезу H_0 отклонять нельзя и нет оснований утверждать, что стандартное отклонение генеральной совокупности отличается от 15.

Критерий χ^2 для проверки гипотезы от дисперсии или стандартном отклонении считается классической параметрической процедурой. При проверке гипотезы о дисперсии генеральной совокупности или стандартном отклонении предполагается, что исходные данные имеют нормальное распределение. Как и большинство параметрических критериев, χ^2-критерий довольно чувствителен к нарушению этих предположений (т.е. этот критерий не является устойчивым). Следовательно, если генеральная совокупность сильно отличается от нормального распределения, особенно, когда объем выборки невелик, точность критерия значительно снижается.

Использование χ^2-критерия согласия для распределения Пуассона. Распределение Пуассона может быть использовано, например, для моделирования количества клиентов, прибывающих в отделение банка в течение минуты. Предположим, что в течение недели фактическое количество клиентов, приходящих в отделение банка в течение минуты, измерялось 200 раз (рис. 5.3).

Для того чтобы определить, имеет ли количество прибытий в минуту распределение Пуассона, формулируются нулевая и альтернативная гипотеза. H_0: количество прибытий в минуту подчиняется распределению Пуассона, H_1: количество прибытий в минуту не подчиняется распределению Пуассона. Поскольку распределение Пуассона имеет один параметр —математическое ожидание λ, в нулевую и альтернативную гипотезы можно включать либо величину λ, либо ее выборочную оценку. В нашем примере для оценки среднего количества прибытий клиентов необходимо воспользоваться формулой:

$$\overline{X} = \frac{\sum_{j=1}^{c} m_j f_j}{n}. \qquad (5.3)$$

Для расчета по этой формуле в Excel удобно воспользоваться функцией =СУММПРОИЗВ() (рис. 5.3).

	A	B	C	D	E	F	G
1	Прибытия	Частота					
2	0	14					
3	1	31					
4	2	47					
5	3	41					
6	4	29					
7	5	21					
8	6	10					
9	7	5					
10	8	2					
11		200					
12							
13	Среднее	$\overline{X}$	2,90	=СУММПРОИЗВ(A2:A10;B2:B10)/B11			

Рис. 5.3. Распределение частоты прибытий в минуту во время ланча

Для оценки параметра λ можно воспользоваться оценкой $\overline{X}$. Теоретическую частоту X успехов (X = 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 и более), соответствующую параметру λ = 2,9, можно определить с помощью функции =ПУАССОН.РАСП(X;X̄;ЛОЖЬ). Умножив пуассоновскую вероятность на объем выборки n, получим теоретическую частоту f_e (рис. 5.4).

Как следует из рис. 5.4, теоретическая частота девяти и более прибытий не превосходит 1,0. Для того чтобы каждая категория содержала частоту, равную 1,0 или большему числу, категорию «9 и более» следует объединить с категорией «8». То есть, остается девять категорий (0, 1, 2, 3, 4, 5, 6, 7, 8 и более). Поскольку математическое ожидание распределения Пуассона определяется на основе выборочных данных, количество степеней свободы

	A	B	C	D
1	Прибытия	Наблюдаемая частота f_0	Теоретическая частота $f_e = nP(X)$	$(f_0 - f_e)^2/f_e$
2	0	14	11,00	0,8153
3	1	31	31,91	0,0261
4	2	47	46,27	0,0114
5	3	41	44,73	0,3114
6	4	29	32,43	0,3629
7	5	21	18,81	0,2550
8	6	10	9,09	0,0908
9	7	5	3,77	0,4040
10	8 и более	2	1,98	0,0003
11	Итого	200	χ^2-критерий согласия	2,2772
12				

Рис. 5.5. Расчет χ^2-критерия согласия для распределения Пуассона

	A	B	C	D	E	F	G	H	I	J
1	-6,1	-2,8	-1,2	-0,7	0,5	1,8	1,9	2,5	2,8	3,3
2	3,5	3,8	3,8	4,0	4,2	4,3	4,5	4,6	5,0	5,1
3	5,2	5,4	5,5	5,8	5,9	6,0	6,2	6,3	6,5	6,5
4	7,0	7,1	7,1	7,2	7,2	7,3	7,5	7,6	7,6	7,8
5	7,8	7,8	7,9	8,1	8,1	8,2	8,3	8,3	8,4	8,5
6	8,5	8,5	8,6	8,8	8,8	8,8	9,0	9,0	9,1	9,1
7	9,1	9,2	9,3	9,3	9,5	9,5	9,5	9,5	9,6	9,6
8	9,7	9,8	9,9	9,9	9,9	9,9	10,1	10,1	10,1	10,1
9	10,2	10,3	10,3	10,4	10,5	10,5	10,5	10,5	10,5	10,5
10	10,6	10,7	10,7	10,8	10,9	11,0	11,0	11,1	11,1	11,1
11	11,2	11,2	11,3	11,3	11,3	11,3	11,4	11,5	11,5	11,5
12	11,6	11,7	11,7	11,9	11,9	12,2	12,2	12,3	12,3	12,4
13	12,5	12,7	12,9	12,9	12,9	13,0	13,1	13,2	13,4	13,4
14	13,7	13,7	13,9	14,1	14,7	14,8	14,9	15,0	15,7	15,8
15	15,8	16,0	16,9	17,0	17,0	17,6	17,8	18,1	18,1	18,2
16	18,5	18,5	18,7	18,9	21,4	22,0	22,9	26,3		
17										
18	Среднее значение				$\bar{X}$	10,149				
19	Стандартное отклонение				S	4,773				
20										

Рис. 5.6. Упорядоченный массив выборочных данных

Выборочные данные можно сгруппировать, разбив, например, на классы (интервалы) шириной 5% (рис. 5.7).

	A	B	C	D	E	F	G	H	I
1	Пятилетняя среднегодовая доходность				X	Z	Площадь в классе "меньше"	Площадь в классе	$f_e = nP(X)$
2			менее	-10,0	-10,0	-4,22	0,00001	0,00001	0,00
3	от	-10,0	до	-5,0	-5,0	-3,17	0,00075	0,00074	0,12
4	от	-5,0	до	0,0	0,0	-2,13	0,01673	0,01598	2,52
5	от	0,0	до	5,0	5,0	-1,08	0,14030	0,12358	19,53
6	от	5,0	до	10,0	10,0	-0,03	0,48752	0,34721	54,86
7	от	10,0	до	15,0	15,0	1,02	0,84527	0,35775	56,53
8	от	15,0	до	20,0	20,0	2,06	0,98049	0,13522	21,36
9	от	20,0	до	25,0	25,0	3,11	0,99907	0,01858	2,94
10	от	25,0	до	30,0	30,0	4,16	0,99998	0,00091	0,14
11			более	30,0	–	+∞	1,00000	0,00001	0,00
12									
13	Среднее значение				$\bar{X}$	10,149			
14	Стандартное отклонение				S	4,773			
15									

Рис. 5.8. Площади и ожидаемые частоты для каждого класса выборочных данных

	A	B	C	D	E	F	G	H
1	Пятилетняя среднегодовая доходность				Наблюдаемая частота f_0	Теоретическая частота $f_e = nP(X)$	$(f_e - f_0)^2/f_e$	
2			менее	0,0	4,0	2,64	0,699	
3	от	0,0	до	5,0	14,0	19,53	1,564	
4	от	5,0	до	10,0	58,0	54,86	0,180	
5	от	10,0	до	15,0	61,0	56,53	0,354	
6	от	15,0	до	20,0	17,0	21,36	0,892	
7			более	20,0	4,0	3,08	0,275	
8						χ^2-критерий согласия	3,964	
9								
10	Число степеней свободы					3		
11	Уровень значимости					0,05		
12	Критическое значение χ^2-статистики					7,815	=ХИ2.ОБР(1-F11;F10)	
13								

Рис. 5.9. Вычисления, связанные с применением χ^2-критерия согласия для нормального распределения

Видно, что χ^2-статистика = 3,964 < $\chi_U^2$7,815, следовательно гипотезу H_0 отклонять нельзя. Иначе говоря, у нас нет оснований утверждать, что данные выборки не подчиняются нормальному распределению.

Анализ категорий данных может проводиться с помощью непараметрических процедур: рангового критерия Уилкоксона, используемого в ситуациях, когда не выполняются условия применения *t*-критерия для поверки гипотезы о равенстве математических ожиданий двух независимых групп, а также критерия Крускала-Уоллиса, который является альтернативой однофакторному дисперсионному анализу.

Лабораторная работа № 6

Меры связи выборочных данных. Ковариация. Корреляция (4 часа)

Цель: Изучить количественные показатели, характеризующие силу зависимости между двумя случайными величинами.

Теоретические сведения

Первичную оценку меры связи между выборочными переменными проводят с помощью диаграммы рассеяния (разброса данных) – рис. 6.1. Она показывает взаимосвязь между двумя видами выборочных данных и характеризует степень их взаимозависимости.

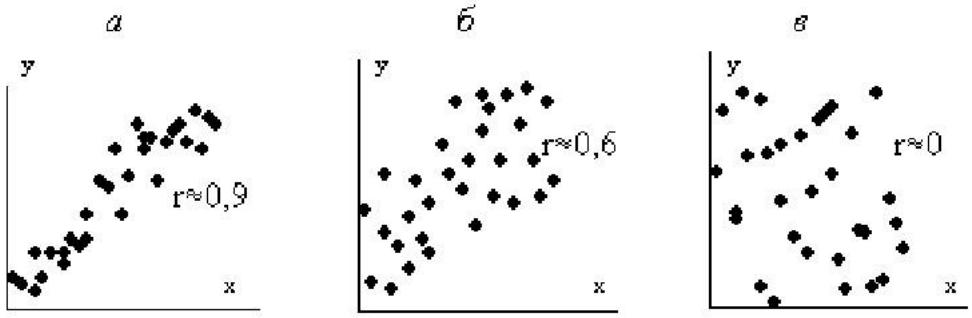


Рис 6.1. Характерные варианты скопления точек на диаграммах рассеяния

Для построения диаграммы рассеяния рекомендуется использовать не менее 30 пар данных *(x,y)*. Оси *x* и *y* строят так, чтобы длины рабочих частей были примерно одинаковы. Точки, далеко отстоящие от основной группы, являются выбросами, и их исключают (по правилу «трех сигм»).

Ковариация оценивает силу линейной зависимости между двумя числовыми переменными X и Y. Выборочная ковариация:

$$\operatorname{cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_1-\overline{X})(Y_1-\overline{Y})}{n-1}. \qquad (6.1)$$

При этом ковариация случайной величины с собой равна дисперсии:

$$\operatorname{cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}. \qquad (6.2)$$

Для расчета ковариации двух выборок в Excel используется функция КОВАРИАЦИЯ.В().

$$m_{pxy} = \sqrt{\frac{1 - p_{xy}^2}{n-2}}\,. \qquad (6.5)$$

При $r/m_r \geq 3$ коэффициент корреляции считается достоверным, т.е. связь доказана. При $r/m_r < 3$. связь недостоверна.

Линейность корреляции означает, что все точки, изображенные на диаграмме разброса, лежат на прямой (рис 6.2). В реальных ситуациях коэффициент корреляции редко принимает точные значения -1, 0 и +1 (рис. 6.1).

Рис. 6.2. Примеры зависимости между двумя переменными

Зависимости на рис. 6.1 и 6.2 говорят о *тенденциях*, поскольку наличие корреляции между переменными X и Y не означает наличия причинно-следственных связей между ними, т.е. изменение значения одной из переменных не обязательно приводит к изменению значения другой. Сильная корреляция может быть случайной (иметь низкую статистическую значимость) или объясняться третьей переменной, оставшейся за рамками анализа. В таких ситуациях необходимо проводить дополнительное исследование. Таким образом, можно утверждать, что причинно-следственные связи порождают корреляцию, но корреляция не означает наличия причинно-следственных связей.

Выборочный коэффициент корреляции можно вычислить через ковариацию:

$$r = \frac{\mathrm{cov}(X,Y)}{S_x S_r}, \qquad (6.6)$$

где

$$\mathrm{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n-1} \quad S_x = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} \quad S_y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}}$$

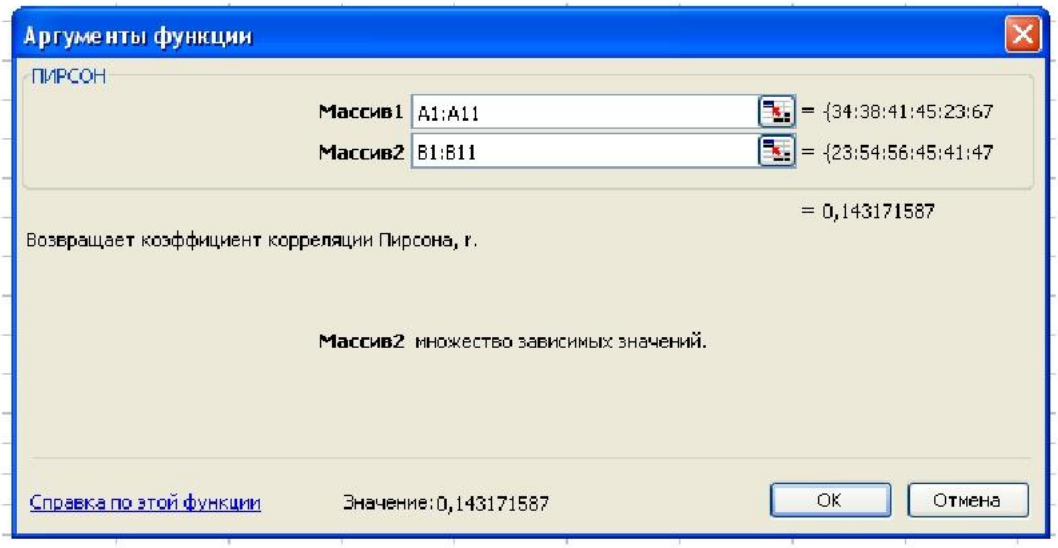


Рис. 6.4. Диалоговое окно Аргумента функции ПИРСОН

Оба аргумента должны содержать одинаковое количество значений. Если они имеют различные объемы данных, то функция возвращает значение ошибки # Н/Д, если хотя бы один аргумент не задан, то функция возвращает значение ошибки # ДЕЛ/0!.

В ячейке, которую мы выделили для записи функции, появится значение коэффициента корреляции Пирсона. Полученные результаты можно приводить к стандартным значениям, проводить *Z*-преобразования:

$$Z = \frac{X_i - \overline{X}}{\sigma}. \qquad (6.7)$$

В таком случае коэффициент корреляции Пирсона выглядит проще:

$$R_{xy} = \frac{\sum_{i=1}^{N} Z_{xi} Z_{yi}}{N-1}. \qquad (6.8)$$

На величину коэффициента корреляции не влияет то, в каких единицах представлены признаки. Любые линейные преобразования признаков(умножение на константу, прибавление константы) не меняют значения коэффициента корреляции. Исключением является умножение одного из признаков на отрицательную константу, в таком случае знак коэффициента корреляции меняется на противоположный.

Вычислить коэффициенты корреляции также можно с помощью функции КОРРЕЛ со следующими характеристиками**:** КОРРЕЛ (массив 1; массив 2), где: массив 1 = диапазон данных для первой переменной, массив 2 = диапазон данных для второй переменной.

Контрольные вопросы

1. Сформулируйте понятия функциональной и стохастической зависимостей случайных переменных.
2. Какая взаимосвязь случайных величин называется корреляционной?
3. В чем заключается основная задача корреляционного анализа?
4. Что такое корреляционного поле? Приведите примеры с разными значениями корреляционных зависимостей между переменными.
5. Что называется ковариацией?
6. Для оценки какой корреляционной зависимости используется выборочный коэффициент корреляции? Каковы его свойства?
7. Какие параметры взаимосвязи выборочных данных характеризует коэффициент корреляции?
8. Какие виды корреляционной зависимости Вам известны?
9. Как проверяется значимость коэффициента корреляции?
10. Что показывает интервальная оценка коэффициента корреляции?
11. Что характеризует эмпирическое корреляционное отношение? Каковы его свойства?
12. Что характеризует эмпирический коэффициент детерминации?
13. В каких диапазонах значений находятся коэффициент корреляции и коэффициент детерминации? Приведите примеры и интерпретируйте их значение.
14. Что такое множественная корреляция? Частная корреляция? Определяет ли корреляция причинно-следственные связи между исследуемыми параметрами?

в которой r_{ij} – выборочные коэффициенты корреляции между величинами X_i и X_j,. Матрицы Q_p и q_p симметричные, поэтому при вычислении матрицы q_p приводятся только элементы, расположенные на главной диагонали и под ней.

Теснота линейной корреляционной связи одной из величин X_i с совокупностью остальных p - 1 величин X_1, X_2, X_3, ... , X_p оценивается *выборочным коэффициентом множественной корреляции*

$$R_{i/1...p} = \sqrt{1 - \frac{|q_p|}{q_{ii}}}, \qquad (7.1)$$

где $|\, q_p \,|$ – определитель матрицы q_p, q_{ii} – алгебраическое дополнение элемента r_{ii} матрицы q_p,. В частности, для трех величин X_1, X_2, X_3 выборочный коэффициент множественной корреляции $R_{i/jk}$ вычисляется по формуле

$$R_{i/jk} = \sqrt{\frac{r_{ij}^2 + r_{ik}^2 - 2r_{ij}\,r_{ik}\,r_{jk}}{1 - r_{jk}^2}}\,. \qquad (7.2)$$

Выборочный коэффициент множественной корреляции принимает значения от 0 до 1. Чем ближе значение $R_{i/1...p}$ к единице тем теснее линейная корреляционная связь X_i с остальными величинами X_1, X_2, X_3, ... , X_p.

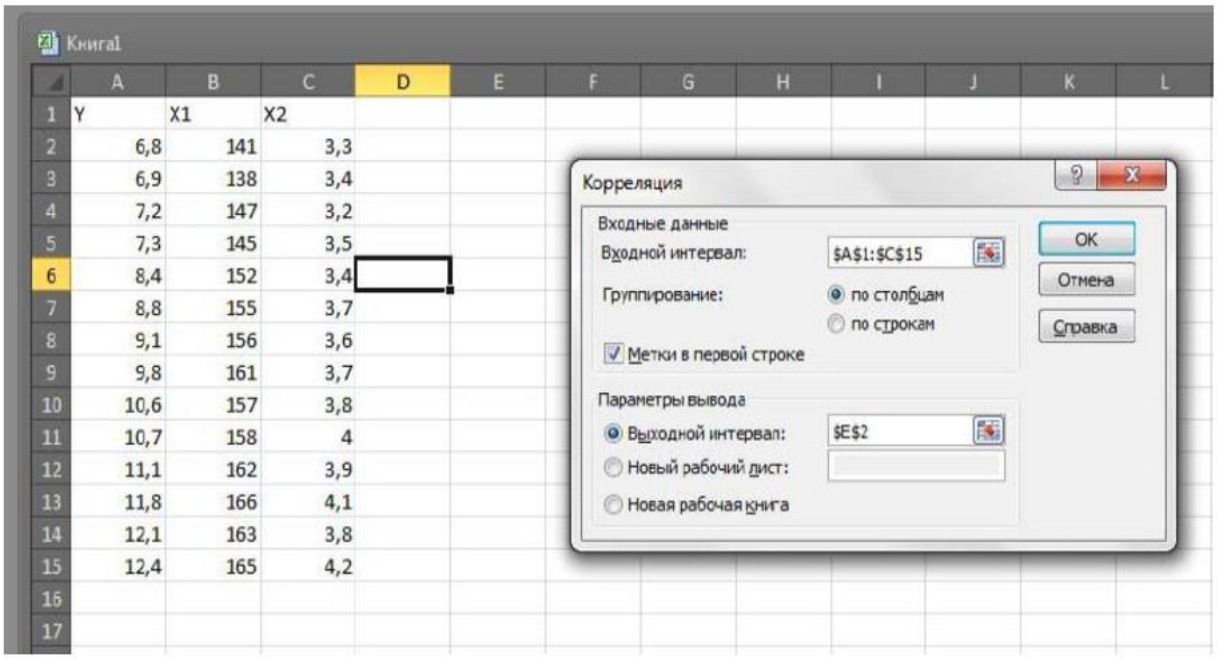
Величина $R^2 = \left(R_{i/1...p}\right)^2$ называется *выборочным множественным коэффициентом детерминации*, которая показывает долю вариации переменной X_i объясняемую вариацией остальных переменных. Множественный коэффициент корреляции $R_{i/1...p}$ значим при уровне значимости α, если вычисленное значение F-статистики:

$$F = \frac{R^2(n-p)}{(1-R)^2(p-1)} > F(\alpha, \mathrm{p}\text{-}1, \mathrm{n}\text{-}\mathrm{p}), \qquad (7.3)$$

где $F(\alpha, \mathrm{p}\text{-}1, \mathrm{n}\text{-}\mathrm{p})$ значение F-критерия на уровне значимости α при числе степеней свободы $k_1 = p - 1$ и $k_2 = n - p$.

Частные коэффициенты корреляции. Если величины из совокупности X_1, X_2, X_3, ... , X_p коррелируют друг с другом, то на величинах парных коэффициентов корреляции r_{ij} переменных X_i и X_j, сказывается влияние и других переменных совокупности, что приводит к искажению значений коэффициентов корреляции r_{ij}. Для оценки линейной корреляционной зависимости между величинами X_i и X_j, очищенной от влияния других величин совокупности, используется *выборочный частный коэффициент корреляции* $r_{ij/1...p}$. Он определяется соотношением

переменных расположены по столбцам, если значения переменных расположены по строкам, то выбирается «по строкам»; поставим флажок в поле «Метки в первой строке (столбце)», что указывает на то, что в первой строке (столбце) сгруппированных по столбцам (строкам) данных находятся имена переменных. В части «Параметры вывода» выбирается место расположения результатов выполнения функции «**Корреляция**»: «Выходной интервал» – указывается ячейка текущего листа, с которого (вправо и вниз) будет расположена корреляционная матрица q_p; «Новый рабочий лист» – вывод корреляционной матрицы на новый рабочий лист; «Новая рабочая книга» – вывод корреляционной матрицы в новую рабочую книгу. Выберем «Выходной интервал» и ячейку **E2**, с которой будет расположена корреляционная матрица. По «ОК» получим в ячейках **E2-H5** корреляционную матрицу. Заполнение окна «Корреляция» приведено на рис. 7.1, результаты – на рис. 7.2.



	A	B	C
1	Y	X1	X2
2	6,8	141	3,3
3	6,9	138	3,4
4	7,2	147	3,2
5	7,3	145	3,5
6	8,4	152	3,4
7	8,8	155	3,7
8	9,1	156	3,6
9	9,8	161	3,7
10	10,6	157	3,8
11	10,7	158	4
12	11,1	162	3,9
13	11,8	166	4,1
14	12,1	163	3,8
15	12,4	165	4,2

Рис. 7.1. Заполнение окна «Корреляция»

	A	B	C	D	E	F	G	H	I	J	K	L
1	X1	X2	X3	Корреляционная матрица								
2	6,8	141	3,3			X1	X2	X3				
3	6,9	138	3,4		X1	1	0,944018	0,919392				
4	7,2	147	3,2		X2	0,944018	1	0,853917				
5	7,3	145	3,5		X3	0,919392	0,853917	1				
6	8,4	152	3,4									
7	8,8	155	3,7									
8	9,1	156	3,6									
9	9,8	161	3,7			t-статистики						
10	10,6	157	3,8		rX1X2	9,912769						
11	10,7	158	4		rX1X3	8,096928						
12	11,1	162	3,9		rX2X3	5,684104						
13	11,8	166	4,1									
14	12,1	163	3,8		t(1-α,n-2)=	2,178813		Доверительный интервал для ρ(X1,X2)				
15	12,4	165	4,2		z(X1,X2)=	1,773738		0,828328		0,982491		
16					t(1-α)=	1,959964						

Рис. 7.2. Результаты корреляционного анализа

Проведем Z-преобразование Фишера для выборочного коэффициента корреляции R_{21}= 0,944. Для этого выделим, например, ячейку **F15**. Во вкладке «**Формулы**» выберем «**Другие функции**», в группе «**Статистические**» выберем функцию «**ФИШЕР**». В окне этой функции в поле «x» введем значение коэффициента корреляции R_{21}. По «ОК» в ячейке **F15** получим значение z, равное в этом примере 1,7736 (см. рис. 7.2). Для вычисления значений $z-\frac{t_{1-\alpha}}{\sqrt{n-3}}$ u $z+\frac{t_{1-\alpha}}{\sqrt{n-3}}$ предварительно найдем значение $t_{1-\alpha}$. Выделим, например, ячейку **F16.** Во вкладке «**Формулы**» выберем «**Другие функции**», в группе«**Статистические**» выберем функцию «**НОРМ. СТ. ОБР.**». В окне этой функции в поле «Вероятность» введем значение $1-\frac{\alpha}{2}$, равное 0,975. По «ОК» в ячейке **F16** получим значение $t_{1-\alpha}$, равное 1,9599 (см. рис. 7.2).

Для получения нижней границы доверительного интервала для ρ_{12}используем функцию **ФИШЕРОБР** вычисления гиперболического тангенса $th(x)$. Выделим ячейку **H15** и в строке формул введем **=ФИШЕРОБР(F15-F16/(14-3)^0,5)** По «Enter» в ячейке **H15** получим искомую нижнюю границу доверительного интервала, в этом примере равную 0,8283.

Для получения верхней границы доверительного интервала для ρ_{12}выделим ячейку **J15** и в строке формул введем **=ФИШЕРОБР(F15+F16/(14-3)^0,5)** По «Enter» в ячейке **J15** получим искомую верхнюю границу доверительного интервала, равную в этом примере 0,9825 (см. рис. 7.2). Аналогичным образом могут быть построены доверительные интервалы для других генеральных коэффициентов корреляции.

Для нахождения выборочных коэффициентов множественной корреляции частных коэффициентов корреляции построим предварительно матрицу алгебраических дополнений q_{ij} элементов выборочной корреляционной матрицы, см. рис. 7.3. Для этого в ячейке **A19** вычислим определитель корреляционной матрицы: выделим эту ячейку и в строке формул, учитывая расположение выборочной корреляционной матрицы, см. рис. 7.2, введем **=МОПРЕД(F3:H5)**.По Enter в **A19** получим значение определителя. Выделим ячейки **A21 – C23** и введем в строке формул **МОБР(F3:H5).** Нажав**Ctrl+Shift+Enter**, вячейках**A21 – C23** получим матрицу обратную к корреляционной матрице. Для получения матрицы алгебраических дополнений q_{ij} элементов выборочной корреляционной матрицы необходимо умножить элементы полученной обратной матрицы на определитель корреляционной матрицы. Матрицу алгебраических дополнений q_{ij} элементов выборочной корреляционной матрицы разместим в ячейках **F21 – H23**, см. рис. 7.3. Выделим ячейку **F21** и введя в строке формул **=A21*A19** по «Enter» в ячейке **F21** получим значение q_{11}. Аналогично вычисляются другие алгебраические дополнения элементов корреляционной матрицы.

ячейке получим значение $R_{21/3}$. Выделив ячейку **B30** и введя в строкемул=-**F23/КОРЕНЬ(F21*H23),** по «ОК» в ячейке **B30** получим значение $R_{31/2}$.

Выделив ячейку **C30** и введя в строке формул =-**G23/КОРЕНЬ(G22*H23),** по «ОК» в ячейке **C30** получим значение $R_{32/1}$. Остальные элементы матрицы частных коэффициентов корреляции (ячейки **C28, D28, D29**) заполняются исходя из ее симметричности (см. рис. 7.3).

Общее заключение. Значения выборочных парных коэффициентов корреляции R_{12}= 0,944 и R_{13} = 0,9194 говорят о сильной линейной корреляционной зависимости переменной (X1) от переменных (X2) и (X3). Переменные (факторы) X2 и X3 также сильно коррелированы, R_{23}= 0,8539. Все коэффициенты парной корреляции значимы, о чем свидетельствуют значения их t-статистик t_{12}= 9,913,t_{13}= 8,097, t_{23}= 5,684, модули которых превышают критическое значение t-статистики t (0,95; 12) = 2,179. Для генерального коэффициента корреляции ρ_{12} = 95%-й доверительный интервал имеет вид (0,8283; 0,9824), что также говорит о сильной линейной корреляционной связи между переменными X1 и X2. Значение множественного коэффициента корреляции X1 с X2 и X3 равно 0,9688. Значение множественного коэффициента детерминации говорит о том, что 93,86% вариации производительности труда объясняется вариацией переменных X2 и X3. Значения частных коэффициенты корреляции $R_{12/3}$0,776 и $R_{13/2}$ 0,66 и проверка их значимости говорят о значимом влиянии переменных X2 и X3 на переменную X1. Проверка значимости частного коэффициента корреляции $R_{23/1}$ = - 0,107 говорит об отсутствии значимой линейной корреляционной зависимости переменных X2 и X3.

Порядок выполнения работы

1. Ввод выборочных данных для исследования корреляционной зависимости совокупности величин X1, X2, … , Xn.
2. Построение матрицы выборочных коэффициентов корреляции и оценка наличия и тесноты линейной корреляционной зависимости между парами величин.
3. Проверка значимости наибольшего по модулю коэффициента корреляции при уровне значимости $\alpha = 0,05$.
4. Построение доверительного интервала надежности $\gamma = 1 - \alpha$для генерального коэффициента корреляции ρ между наиболее тесно связанными величинами заданной совокупности.
5. Нахождение выборочного коэффициента множественной корреляции$R_{1/2,\dots,p}$и выборочного множественного коэффициента детерминации $R^2= R_{1/1,\dots,p}$.
6. Построение матрицы выборочных частных коэффициентов корреляции и оценка «очищенной» корреляционной зависимости X1с другими величинами совокупности.
7. Общее заключение о корреляционной зависимости исследуемых величин.

15. Определите выборочный множественный коэффициент детерминации $R^2 = (R_{1/2,3})^2$ по матрице выборочных коэффициентов корреляции, приведенной в 11-м вопросе.

16. Для характеристики, какой взаимосвязи используется частный коэффициент корреляции?

17. Определите выборочный частный коэффициент корреляции $R_{13/2}$ по матрице выборочных коэффициентов корреляции, приведенной в 11-м вопросе.

18. Проверьте значимость частного коэффициента корреляции $R_{13/2}$, найденного в предыдущем вопросе, при объеме выборки n=19 и уровне значимости $\alpha = 0{,}05$.

где L_i - число связок (видов повторяющихся элементов) в оценках i-го эксперта, t_l – количество элементов в l-й связке для i-го эксперта (количество повторяющихся элементов). Если нет связанных рангов, то T_i равно нулю.

В некотором смысле коэффициент конкордации W служит мерой общности. Описание коэффициента конкордации Кендалав литературе может ограничиваться формулами, в которых не учитываются связанные ранги, а критерий ограничивается требованием стремления W к единице, что существенно затрудняет его практическое использование. В абсолютном выражении W может оказаться очень малым, но его значение будет статистически значимым для проверки гипотезы о равномерном распределении рангов (согласии ранжировок). Вычисление коэффициента конкордации без введения поправочных коэффициентов и проверки на статистическую значимость может привести к существенным ошибкам.

В использовании коэффициент конкордации Кендала можно выделить два ограничения:

- невозможность рассчитать согласованность мнений экспертов по каждой переменной в отдельности;
- коэффициент измеряет согласованность мнений в смысле их коррелированности, но не совпадения.

Пример расчета коэффициента конкордации. Произвести экспертную оценку интегрального показателя, состоящего из n факторов, по степени выраженности признака. Исходные данные: число факторов $n = 6$; число экспертов $m = 4$.

Этап 1. Создание экспертной комиссии. В экспертную группу вошло 4 эксперта.

Этап 2. Сбор мнений специалистов путем анкетного опроса. Оценку степени значимости технических параметров холодильника для потребителей эксперты производят путем присвоения им рангового номера. Фактору, которому эксперт дает наивысшую оценку, присваивается ранг 1. Если эксперт признает несколько факторов равнозначными, то им присваивается одинаковый ранговый номер. На основе данных анкетного опроса составляется сводная матрица рангов.

Этап 3. Составление сводной матрицы рангов.

Таблица 8.1 – Матрица рангов

Исследуемый признак	Единицы измерения	Эксперты			
		1	2	3	4
X 1		3	4	5	4
X 2		5	6	3	4
X 3		4	3	1	3
X 4		1	1	3	1
X 5		6	5	6	6
X 6		2	2	2	2

$$\Delta = \sum_{i-1}^{m} x_{ij} - \frac{\sum_{i-1}^{n} \sum_{i-1}^{m} x_{ij}}{n} \quad . \qquad (8.4)$$

Проверка правильности составления матрицы на основе исчисления контрольной суммы:

$$\sum_{j-1}^{n} x_{ij} = \frac{(1+n) \times n}{2} = \frac{(1+6) \times 6}{2} = 21 .$$

Сумма по столбцам матрицы равны между собой и контрольной суммы, значит матрица составлена правильно.

Этап 4. Анализ значимости исследуемых факторов. В данном примере факторы по значимости распределились следующим образом (табл. 8.5).

Таблица 8.5 – Расположение факторов по значимости

Факторы	x_4	x_6	x_3	x_1	x_2	x_5
Сумма рангов	6,5	8	11	16,5	19	23

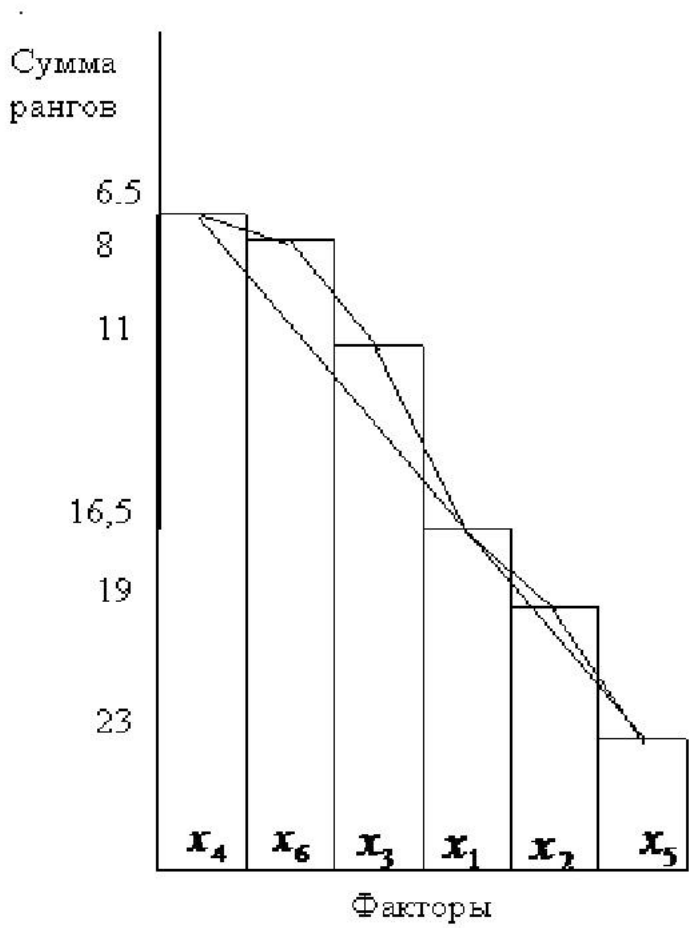


Рис. 8.1. Гистограмма и полигон распределения сумм рангов

$$x^2 = \frac{213{,}5}{\frac{1}{12} \times 4 \times 6(6+1) - \frac{1}{6-1} \times 1} = 15{,}471.$$

Вычисленный $x^2 = 15{,}471$ сравним с табличным значением для числа степеней свободы K = n-1 = 6-1 = 5 и при заданном уровне значимости $\alpha = 0{,}05$.

Так как x^2 расчетный 15,471 > x^2 кр = 11,07 (см. статистические таблицы), то W = 0.744 является статистически значимой оценкой, а потому полученные результаты для интегрального показателя имеют смысл и могут использоваться в дальнейших исследованиях.

Этап 7. Подготовка решения экспертной комиссии. В результате проведенного исследования на основе экспертных оценок выяснилось, что наиболее выраженными свойствами обладают факторы X4, X5 и X2. Следовательно, дальнейшие исследования необходимо проводить по изучению этих свойств. На основе получения суммы рангов (табл. 8.4) можно вычислить показатели весомости рассмотренных факторов с тем, чтобы их можно было учитывать при комплексной оценке. Для этого произведем следующие вычисления. Сначала по каждому параметру вычислим величины, обратные сумме рангов, то есть

$x_1 = \frac{1}{16{,}5} = 0{,}06$, $x_2 = \frac{1}{19} = 0{,}05$, $x_3 = \frac{1}{11} = 0.09$, $x_4 = \frac{1}{6{,}5} = 0{,}15$, $x_5 = \frac{1}{23} = 0{,}04$,

$x_6 = \frac{1}{8} = 0{,}12$.

Это делается для того, чтобы привести в соответствие содержание сумм рангов коэффициентам весомости. Расположим полученные числа по мере убывания, сложим их, взвесим каждое число в полученной сумме, которую примем равной 1 (табл. 8.6).

Таблица 8.6 – Взвешенные по экспертным заключениям переменные

Переменные (признаки)	Величины, обратные сумме рангов	Коэффициенты весомости параметров
x_4	0,15	0,29
x_6	0,12	0,23
x_3	0,09	0,18
x_1	0,06	0,12
x_2	0,05	0,10
x_5	0,04	0,08

Лабораторная работа № 9

Расчет социометрических критериев в MS Excel (4 часа)

Цель: Изучить проведение социометрического анализа малых групп

Теоретические сведения

Социометрический метод - вид опроса, направленный на количественное измерение и анализ структуры межличностных отношений в малых социальных группах путем фиксации среди членов малой группы связей предпочтения в ситуациях выбора. Относится к числу экспериментальных методов изучения малых групп.

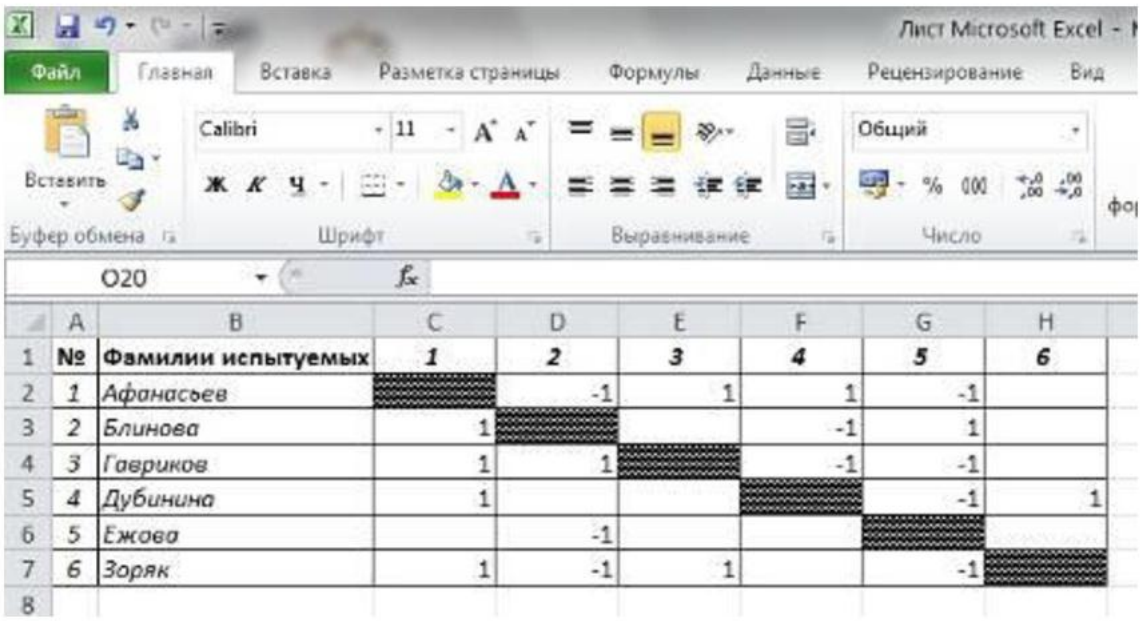
Основное назначение: служит диагностике состояния взаимоотношений в малых группах и коллективах. Предназначен для получения информации о структуре связей между членами группы относительно выделенного критерия. С помощью социометрического метода дается описание структуры взаимоотношений в группе, количественная оценка (измерение) характера эмоциональных отношений между ее членами (чувства симпатии, неприязни), устанавливается место в указанной структуре, занимаемое тем или иным членом группы (Табл. 1).

Область применения: используется как средство активного управления групповой деятельностью. Применяется при изучении институциональных малых групп, коллективов для работы в экстремальных (и приближенных к экстремальным) условиях, а также тогда, когда характер эмоциональных взаимоотношений оказывает значительное влияние на результаты совместной деятельности. В прикладном социологическом исследовании может быть использован наряду с другими методами в качестве основного или вспомогательного. Используется также в медицине при исследовании и терапии неврозов.

Основные нормативные требования: обоснованность применения метода, его адекватность изучаемому объекту и характеру исследовательских задач. Надежность социометрической техники. Доверительность взаимоотношений исследователя с респондентами, их заинтересованное отношение к результатам опроса, тщательность интерпретации полученных данных. К социометрическому методу предъявляются все основные требования социального эксперимента.

Ограничения в применении: может применяться только в рамках анализа структуры взаимоотношений в малых (контактных) группах, достаточно сложившихся, имеющих опыт совместной деятельности не менее 6 месяцев. Данные, полученные с его использованием, нельзя рассматривать как полную картину внутригрупповых отношений; она ограничивается выделенными критериями; существует необходимость дополнения данных другими способами сбора информации, особенно при выработке практических рекомендаций.

Программа проведения социометрического опроса. Особое внимание уделяется предварительному знакомству с группой, точному определению



	A	B	C	D	E	F	G	H
1	№	Фамилии испытуемых	1	2	3	4	5	6
2	1	Афанасьев		-1	1	1	-1	
3	2	Блинова	1			-1	1	
4	3	Гавриков	1	1		-1	-1	
5	4	Дубинина	1				-1	1
6	5	Ежова		-1				
7	6	Зоряк	1	-1	1		-1	
8								

Рис. 9.1. Первый шаг обработки социометрических данных

Ниже под фамилиями подсчитываем число полученных положительных голосов, формула будет иметь вид:

=СЧЕТЕСЛИ(C2:C7;1)

где C2:C7 диапазон ячеек, значения в которых необходимо вычислить, 1 – показатель положительного отношения одного человека к другому в нашем исследовании.

Для того чтобы не дублировать формулу под каждым участником, ее можно «растянуть», выделив ячейку содержащую формулу, а появившейся черный маркер заполнения потянуть вправо. (см. рис. 9.2)

Далее подсчитывают количество полученных негативных голосов. При этом формула будет иметь вид:

=СЧЕТЕСЛИ(C2:C7;-1)

где C2:C7 диапазон ячеек, значения в которых необходимо вычислить, -1 – показатель отрицательного отношения одного человека к другому в нашем исследовании.

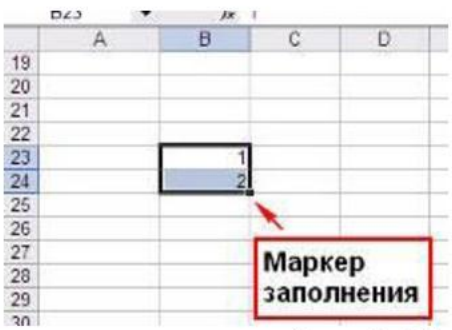


Рис. 9.2 Маркер автозаполнения ячеек

Аналогичным образом подсчитываем количество возможных не сделанных (проигнорированных) выборов, т.е. пустых ячеек. Формула приобретет следующий вид:

=СЧИТАТЬПУСТОТЫ(C2:H7)-6

где C2:H7 – диапазон ячеек, в которых необходимо произвести расчет, 6-количество участников социометрического опроса (поскольку участник не может выбрать сам себя).

Сумма абсолютно всех выборов может быть посчитана по формуле:

=СУММ(C11:C12)

где C11:C12 ячейки с количеством сделанных и проигнорированных выборов (см. рис. 9.5).

G31

	B	C	D	E	Г	G	H	I	J
1	Фамилии испытуемых	1	2	3	4	5	6	отданные "+" голоса	отданные "-" голоса
2	Афанасьев		-1	1	1	-1		2	2
3	Блинова	1			-1	1		2	1
4	Гавриков	1	1		-1	-1		2	2
5	Дубинина	1				1	1	2	1
6	Ежова		-1					0	1
7	Зорян	1	-1	1		-1		2	2
8	Число полученных "+" голосов	4	1	2	1	1	1		
9	Число полученных "-" голосов	0	3	0	2	4	0		
10	Статус участника	4	-2	2	-1	-3	1		
11	Сумма сделанных выборов	19							
12	Сумма несделанных выборов	11							
13	Сумма всех возможных выборов	30							

Рис. 9.5. Четвертый шаг обработки социометрических данных

A	B	C	D	E	F	G	H
№	Фамилии испытуемых	1	2	3	4	5	6
1	Афанасьев		-1	1	1	-1	
2	Блинова	1			-1	1	
3	Гавриков	1	1		-1	-1	
4	Дубинина	1				-1	1
5	Ежова		-1				
6	Зорян	1	-1	1		-1	
	Число полученных "+" голосов	4	1	2	1	1	1
	Число полученных "-" голосов	0	3	0	2	4	0
	Статус участника	4	-2	2	-1	-3	1
	Сумма сделанных выборов	19					
	Сумма несделанных выборов	11					
	Сумма всех возможных выборов	30					

Рис. 9.6. Первый шаг транспонирования таблицы

Далее нам необходимо посчитать количество взаимных положительных и отрицательных выборов. Для этой цели необходимо транспонировать таблицу. Выделяем диапазон ячеек, характеризующий выборы участников социометрии и копируем их (см. рис. 6).

Таблица в зеркальном отображении появится в указанном месте (рис 9.9)

№	Фамилии испытуемых	1	2	3	4	5	6	отданные "+" голоса	отданные "-" голоса
1	Афанасьев		-1	1	1	-1		2	2
2	Блинова	1			-1	1		2	1
3	Гаврилов	1	1		-1	-1		2	2
4	Дубинина	1				-1	1	2	1
5	Ежова		-1					0	1
6	Зорин	1	-1	1		-1		2	2
	Число полученных "+" голосов	4	1	2	1	1	1		
	Число полученных "-" голосов	0	3	0	2	4	0		
	Статус участника	4	-2	2	-1	-3	1		
	Сумма сделанных выборов	19							
	Сумма несделанных выборов	11							
	Сумма всех возможных выборов	30							

Фамилии испытуемых	Афанасьев	Блинова	Гаврилов	Дубинина	Ежова	Зорин
1		1	1	1		1
2	-1		1		-1	-1
3	1					1
4	1	-1	-1			
5	-1	1	-1	-1		-1
6				1		

Рис. 9.9. Вставка транспонированной таблицы

Далее нам необходимо, вычислить, сколько взаимных выборов было сделано в группе. Для этого используем формулу вида:

=ЕСЛИ(C2=C16;(ЕСЛИ(C2=0;0;1));0)

где C2 и C16 ячейки, которые необходимо сравнить между собой. Данные, полученные в ячейке «растягиваем» по количеству участников обследуемой группы. В получившейся таблице взаимные выборы отображаются цифрой 1 (см. рис 9.10).

K26

№	Фамилии испытуемых	1	2	3	4	5	6	отданные "+" голоса	отданные "-" голоса
1	Афанасьев		-1	1	1	-1		2	2
2	Блинова	1			-1	1		2	1
3	Гаврилов	1	1		-1	-1		2	2
4	Дубинина	1				-1	1	2	1
5	Ежова		-1					0	1
6	Зорин	1	-1	1		-1		2	2
	Число полученных "+" голосов	4	1	2	1	1	1		
	Число полученных "-" голосов	0	3	0	2	4	0		
	Статус участника	4	-2	2	-1	-3	1		
	Сумма сделанных выборов	19							
	Сумма несделанных выборов	11							
	Сумма всех возможных выборов	30							

Фамилии испытуемых	Афанасьев	Блинова	Гаврилов	Дубинина	Ежова	Зорин
1		1	1	1		1
2	1		1		1	1
3	1					1
4	1	1	1			
5	1	1	1	1		1
6				1		

Афанасьев	0	0	1	1	0	0
Блинова	0	0	0	0	0	0
Гаврилов	1	0	0	0	0	0
Дубинина	1	0	0	0	0	0
Ежова	0	0	0	0	0	0
Зорин	0	0	0	0	0	0

Рис. 9.10. Пятый шаг обработки социометрических данных

В нашем примере эмоциональная экспансивность будет иметь вид:

=(I2+J2)/(6-1)

Основными групповыми индексами являются индексы экспансивности группы и взаимности выборов (сплоченности группы).

Индекс групповой экспансивности подсчитывается по формуле:

$$Э_{гр} = \frac{V_N^+ + V_N^-}{N}, \qquad (9.3)$$

где V+N – сумма всех положительных выборов в группе, V-N – сумма всех отрицательных выборов, N – численность группы.

Для нашего исследования подсчет будет иметь вид:

=СУММ(I2:J7)/6

где I2:J7 – диапазон ячеек, в которых отражены выборы участников группы

Индекс психологической взаимности выборов характеризует степень групповой сплоченности. Он выражает «удельный вес» положительной взаимности отношений в коллективе по отношению к идеальной возможной взаимности (в %). Определяется по формуле:

$$ВВ_{гр} = \frac{100X^+}{\frac{1}{2}N(N-1)}, \qquad (9.4)$$

где X+ - число положительных взаимных связей в группе.

В нашем случае формула подсчета будет иметь вид:

= (100*C14)/(0,5*6*(6-1))

В целом вычисления по итогам социометрического опроса могут выглядеть следующим образом (см. рис. 9.12)

Порядок выполнения работы

По результатам опроса группы из 10 участников (табл. 9.1) провести социометрическую оценку предпочтений членов группы. Социоматрицу заполнить произвольно. Пример заполнения социоматрицы показан в табл.9.2.

Таблица 9.1 – Таблица соцопроса

№№ пп	Фамилии членов группы	Если бы коллектив заново формировался, с кем бы вы хотели снова быть в одной группе?		
		В первую очередь	Во вторую очередь	В третью очередь
1.	Алексеев			
2.	Борисов			
3.	Волков			
4.	Грибов			
5.	Дёмин			
6.	Ежов			
7.	Жуков			
8.	Зайцев			
9.	Исаев			
10.	Королёв			

Таблица 9.2 – Пример заполнения социоматрицы

№№ пп	Кто выбирает (фамилии)	Кого выбирают									
		1	2	3	4	5	6	7	8	9	10
1.	Алексеев			1			2			3	
2.	Борисов					2			1		3
3.	Волков	3								1	2
4.	Грибов	3					2			1	
5.	Дёмин		1	2						3	
6.	Ежов				1	2				3	
7.	Жуков		3		1		3				
8.	Зайцев		1							2	3
9.	Исаев	1			2		3				
10.	Королёв		2						1	3	

Контрольные вопросы

1. Опишите особенности социометрических видов исследований.
2. Назовите основные социометрические критерии.
3. Основные этапы построения социометрической матрицы.
4. Алгоритм составления социограммы.
5. Как рассчитать верхнюю и нижнюю границы страт.
6. Что такое социометрический статус?
7. Что такое индекс эмоциональной экспансивности?
8. Что такое индекс групповой сплоченности?

Критические значения критерия χ^2 Фридмана для четырех выборок численностью $n<5$

n=2		n=3		n=4			
χ^2	p	χ^2	p	χ^2	p	χ^2	p
0,0	1,000	0,0	1,000	0,0	1,000	5,7	0,141
0,6	0,958	0,6	0,958	0,3	0,992	6,0	0,105
1,2	0,834	1,0	0,910	0,6	0,928	6,3	0,094
1,8	0,792	1,8	0,727	0,9	0,900	6,6	0,077
2,4	0,625	2,2	0,608	1,2	0,800	6,9	0,068
3,0	0,542	2,6	0,524	1,5	0,754	7,2	0,054
3,6	0,458	3,4	0,446	1,8	0,677	7,5	0,052
4,2	0,375	3,8	0,342	2,1	0,649	7,8	0,036
4,8	0,208	4,2	0,300	2,4	0,524	8,1	0,033
5,4	0,167	5,0	0,207	2,7	0,508	8,4	0,019
6,0	0,042	5,4	0,175	3,0	0,432	8,7	0,014
		5,8	0,148	3,3	0,389	9,3	0,012
		6,6	0,075	3,6	0,355	9,6	0,0069
		7,0	0,054	3,9	0,324	9,9	0,0062
		7,4	0,033	4,5	0,242	10,2	0,0027
		8,2	0,017	4,8	0,200	10,8	0,0016
		9,0	0,0017	5,1	0,190	11,1	0,00094
				5,4	0,158	12,0	0,000072

Критические значения критерия t Стьюдента

df	p		df	p		df	p		df	p	
	0,05	0,01		0,05	0,01		0,05	0,01		0,05	0,01
1	12,706	63,656	24	2,064	2,797	48	2,011	2,682	71	1,994	2,647
2	4,302	9,924	25	2,059	2,787	49	2,010	2,680	72	1,993	2,646
3	3,182	5,840	26	2,059	2,778	50	2,009	2,678	73	1,993	2,645
4	2,776	4,604	27	2,052	2,771	51	2,008	2,676	74	1,993	2,644
5	2,570	4,032	28	2,048	2,763	52	2,007	2,674	75	1,992	2,643
6	2,446	3,707	29	2,045	2,756	53	2,006	2,672	76	1,992	2,642
7	2,365	3,499	30	2,042	2,750	54	2,005	2,670	77	1,992	2,641
8	2,306	3,355	31	2,040	2,744	55	2,004	2,668	78	1,991	2,640
9	2,262	3,250	32	2,036	2,738	56	2,003	2,667	79	1,990	2,639
10	2,228	3,169	33	2,035	2,733	57	2,002	2,665	80	1,990	2,638
11	2,201	3,105	34	2,032	2,728	58	2,002	2,663	90	1,987	2,632
12	2,179	3,084	35	2,030	2,724	59	2,001	2,662	100	1,984	2,626
13	2,160	3,112	36	2,028	2,719	60	2,000	2,660	110	1,982	2,621
14	2,145	2,976	37	2,026	2,715	61	2,000	2,659	120	1,980	2,617
15	2,131	2,947	38	2,024	2,712	62	1,999	2,657	130	1,978	2,614
16	2,119	2,920	39	2,023	2,708	63	1,998	2,656	140	1,977	2,611
17	2,110	2,898	40	2,021	2,704	64	1,998	2,655	150	1,976	2,609
18	2,101	2,878	41	2,020	2,701	65	1,997	2,654	200	1,972	2,601
19	2,093	2,861	42	2,018	2,698	66	1,997	2,652	250	1,969	2,597
20	2,086	2,845	43	2,017	2,695	67	1,996	2,651	300	1,968	2,592
21	2,079	2,831	44	2,015	2,692	68	1,995	2,650	350	1,967	2,592
22	2,074	2,819	45	2,014	2,690	69	1,995	2,649	400	1,9659	2,588
23	2,069	2,807	46	2,013	2,687	70	1,994	2,648	500	1,9640	2,785