Министерство образования Республики Беларусь
Белорусский государственный университет
Механико-математический факультет
Кафедра общей математики и информатики

О. А. Велько, М. В. Мартон, Н. А. Моисеева

Основы математической статистики и их применение в социологических исследованиях

Учебно-методическое пособие

Минск
БГУ
2023

УДК 316:519.2(075.8)
В 282

Решение о депонировании вынес:
Совет механико-математического факультета
28 февраля 2023 года, протокол №6

Авторы:
Велько Оксана Александровна, старший преподаватель;
Мартон Марина Владимировна, кандидат физико-математических наук, доцент;
Моисеева Наталья Александровна, старший преподаватель.

Рецензенты:
Барвенов С.А., доцент кафедры веб-технологий и компьютерного моделирования, кандидат физико-математических наук, доцент;
Гулина О.В., заместитель декана факультета экономики и менеджмента учреждения образования «Белорусский государственный экономический университет», кандидат физико-математических наук, доцент.

Велько, О. А. Основы математической статистики и их применение в социологических исследованиях : учебно-методическое пособие / О. А. Велько, М. В. Мартон, Н. А. Моисеева ; БГУ, Механико-математический фак., Каф. общей математики и информатики. – Минск : БГУ, 2023. – 110 с. : ил., табл.– Библиогр.: с. 102–104.

Учебно-методическое пособие предназначено для студентов факультета философии и социальных наук БГУ специальностей «Социология», «Социальные коммуникации». Каждая тема содержит исторические сведения, теоретический материал, примеры решения, задачи для самостоятельного решения и вопросы для самоконтроля. В каждой главе приведены лабораторные работы с использованием MS Excel. Многие математические понятия иллюстрируются примерами из социологии и экономики.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Понятие корреляции и регрессии появились в середине XIX в. благодаря работам английских статистиков Ф. Гальтона и К. Пирсона. Первый термин произошел от латинского «correlation» – соотношение, взаимосвязь. Второй термин (от лат. «regression» – движение назад) введен Ф. Гамильтоном, который, изучая зависимость между ростом родителей и их детей, обнаружил явление «регрессии к среднему» – у детей, родившихся у очень высоких родителей, рост имел тенденцию быть ближе к средней величине.

У истоков корреляционного анализа стоял Фрэнсис Гальтон (1822−1911). Первоначально Гальтон готовился стать врачом. Однако, обучаясь в Кембриджском университете, он увлекся естествознанием, метеорологией, антропологией, наследственностью и теорией эволюции. В его книге, посвященной природной наследственности, изданной в 1889 году им впервые были разработаны основы корреляционного анализа. Гальтон заложил основы новой науки и дал ей имя.

Однако превратил её в стройную научную дисциплину математик Карл Пирсон (1857−1936). В 1884 году Пирсон получает кафедру прикладной математики в Лондонском университете, а в 1889 году знакомится с Гальтоном и его работами. Большую роль в жизни Пирсона сыграл зоолог Ф.Велдон. Помогая ему в анализе реальных зоологических данных, Пирсон ввел в 1893 г. понятие среднего квадратического отклонения и коэффициента вариации. Пытаясь математически оформить теорию наследственности Гальтона, Пирсон в 1898 г. разрабатывает основы множественной регрессии. В 1903 г. Пирсон разработал основы теории сопряженности признаков, а в 1905 г. опубликовал основы нелинейной корреляции и регрессии.

Следующий этап развития связан с именем великого английского статистика Рональда Фишера (1890−1962). Во время обучения в Кембриджском университете Фишер знакомится с трудами Менделя и Пирсона. В 1913−1915 годах Фишер работает статистиком на одном из предприятий, а в 1915−1919 годах преподает физику и математику в средней школе. С 1919 года Фишер начинает работу статистиком на опытной сельскохозяйственной станции в Ротамстеде, где он проработал до 1933 года. Затем с 1933 года по 1943 год Фишер работает профессором в Лондонском университете, а с 1943 года по 1957 год заведует кафедрой генетики в Кембридже. За эти годы им были разработаны теория выборочных распределений, методы дисперсионного и дискриминантного анализа, теории планирования экспериментов, метод максимального правдоподобия и многое другое, что составляет основу современной прикладной математической статистики.

1. ПЕРВОНАЧАЛЬНАЯ СТАТИСТИЧЕСКАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

1.1. Вариационные ряды и их характеристики

Математическая статистика – раздел математики, посвященный математическим методам сбора, систематизации, обработки и использования статистических данных для научных и практических выводов.

Измерение – это приписывание чисел объектам или событиям, в соответствии с определёнными правилами. Эти правила устанавливают соответствие между некоторыми свойствами рассматриваемых объектов с одной стороны и ряда чисел с другой стороны.

Измерение является опытной, или экспериментальной процедурой, результатом активного взаимодействия исследователя с объектом познания. Переход от описания объекта познания к его измерению всегда означал переход к точному знанию. Можно сказать, что измерение сделало естественные науки такими, какими они существуют сегодня, и проникновение измерительных процедур в гуманитарные области знания приблизит их к точным наукам. Измерение позволяет перевести различия между объектами в известные, понятные любому взрослому человеку категории, называемые числами, и любая измерительная процедура, в конечном счете, обязательно должна закончиться числом. Однако, число, приписанное объекту, еще ни о чем не говорит, если не известны правила, по которым происходило это приписывание. Число приобретает смысл только в том случае, если известна шкала, в которой происходило измерение.

Существует 4 типа измерительных шкал: шкала наименований (номинальная шкала); порядковая шкала (порядковая или ординальная шкала); шкала интервалов и шкала отношений (абсолютная или пропорциональная шкала). Числа в этих шкалах обладают разными свойствами: они могут говорить о степени выраженности измеряемого признака, о количественных различиях между объектами и т.д. В зависимости от типа шкалы к числам могут быть применимы, а могут быть и неприменимы те или иные математические операции.

Шкала наименований (номинальная шкала). Измерение в этой шкале состоит в группировке предметов в классы, при условии что объекты, принадлежащие одному классу, аналогичны в отношении какого-либо признака или свойства. Числа, присвоенные объектам, говорят лишь о том, что эти объекты различны. Эта шкала определят, что разные свойства или признаки качественно отличаются друг от друга, но количественных операций производить нельзя. Чисел столько, сколько и классов. Наиболее часто используется дихотомическая шкала, где измеряемые признаки кодируются двумя символами или цифрами и разбиваются на два непересекающихся класса.

Приведем **примеры** из социологии и психологии.

1. Пол: мужчина (0) или женщина (1). Это дихотомическая шкала.
2. Группировка по темпераменту: сангвиник (1), холерик (2), флегматик (3), меланхолик (4).

3. Группировка по мотиву увольнения с работы:
   a) не устраивал заработок;
   b) неудобная сменность;
   c) плохие условия труда;
   d) конфликт с начальством.
4. Группировка по старшинству:
   a) старший ребёнок в семье;
   b) средний;
   c) младший;
   d) единственный.

В номинальной шкале можно подсчитать частоты значений встречающихся признаков, а затем работать с этими частотами. Единицей измерения является количество измерений. Не имеет значения, в каком порядке расположим классифицируемые объекты.

Порядковая шкала. Это шкала, классифицируемая по принципу больше – меньше. Все признаки располагаются в определённой последовательности: от большего к меньшему или наоборот. В порядковой шкале должно быть не менее трёх классов. От классов переходят к числам (низший класс – 1, средний – 2, высший – 3). Единицей измерения является расстояние в один класс, но расстояние между классами может быть разным.

Приведем **примеры** из социологии и психологии.

1. Рейтинг испытуемых;
2. Школьные оценки.

Пример. Пять испытуемых работали с заданием некоторого теста. Были получены следующие данные:

Испытуемый	Время решения в минутах
А	13
Б	12
В	18
Г	9
Д	15

Измерьте полученные данные в ранговой шкале.

Решение. Дополним таблицу еще одним столбцом и запишем соответствующие ранги:

Испытуемый	Время решения в минутах	Ранг
А	13	3
Б	12	2
В	18	5
Г	9	1
Д	15	4

Пример. Испытуемому предлагается задание, в котором 7 личностных качеств необходимо упорядочить (проранжировать) по степени значимости: в

левом столбце в соответствии с особенностями его «Я реального», а в правом с особенностями «Я идеального». Получилась следующая таблица:

Я реальное	Качества личности	Я идеальное
7	Ответственность	1
1	Общительность	5
3	Настойчивость	7
2	Энергичность	6
5	Жизнерадостность	4
4	Терпеливость	3
6	Решительность	2

Таким образом, ранжировать можно как качественные признаки, так и количественные.

При ранжировании следует следовать следующим **принципам ранжирования**:

a) Меньшему значению начисляется меньший ранг;

b) Наименьшему значению начисляется ранг 1. Следующему по величине назначается ранг 2 и т. д.;

c) Наибольшему значению начисляется ранг, соответствующий количеству ранжируемых значений.

d) Случай одинаковых рангов: если несколько значений равны, то им начисляется одинаковый средний ранг, представляющий собой среднее значение из тех рангов, которые эти величины получили бы, если бы стояли по порядку друг за другом и не были бы равны.

e) Общая сумма рангов должна быть равна $\frac{n\cdot(n+1)}{2}$, где n – общее число ранжируемых значений.

f) Не рекомендуется ранжировать более чем 20 величин (признаков, качеств, свойств и т.п.), так как в этом случае ранжирование в целом оказывается малоустойчивым.

g) При ранжировании достаточно большого числа объектов их следует объединить по какому-либо признаку в достаточно однородные классы (группы), а затем уже ранжировать полученные классы (группы).

Пример. Психолог получил у 12 испытуемых следующие значения показателя невербального интеллекта: 108, 108, 107, 123, 122, 117, 105, 117, 108, 114, 102, 104. Необходимо проранжировать эти показатели.

Решение. Запишем данные значения в таблицу:

№ испытуемого	1	2	3	4	5	6	7	8	9	10	11	12
показатель интеллекта	108	108	107	123	122	117	105	117	108	114	102	104
ранг	(5) 6	(6) 6	4	12	11	(9) 9,5	3	(10) 9,5	(7) 6	8	1	2

Испытуемый № 6 и № 8 имеют одинаковые показатели 117. Значит, каждому из них начисляется следующий средний ранг: $\frac{9+10}{2}=9{,}5$.

Испытуемые № 1, № 2 и № 9 также имеют одинаковые значения показателя. Их средний ранг равен: $\frac{5+6+7}{3}=\frac{18}{3}=6$.

Проверка правильности ранжирования:

1) сумма полученных рангов равна 6+6+4+12+11+9,5+3+9,5+6+8+1+2=78;

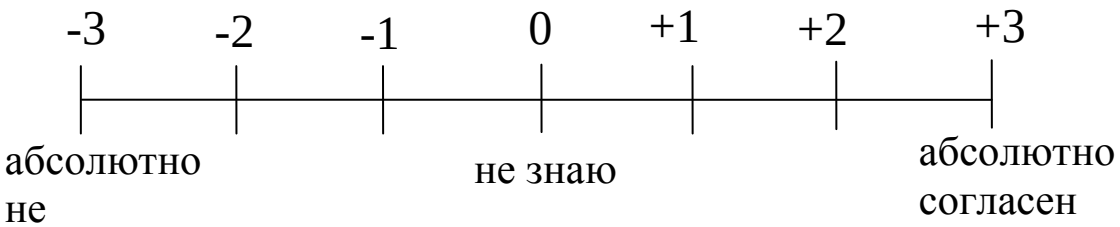
2) расчетная сумма рангов равна $\frac{n\cdot(n+1)}{2}=\frac{12\cdot13}{2}=78$.

Поскольку эти суммы равны, следовательно, ранжирование проведено верно.

Шкала интервалов. Это такое измерение, при котором числа отражают не только различия между объектами в уровне выраженности свойства (характеристика порядковой шкалы), но и то, насколько больше или меньше выражено это свойство. В этой шкале каждое из возможных значений признака отстоит от другого на равном расстоянии. Особенностью этой шкалы является то, что у неё нет точки отсчёта – нулевая точка часто является произвольной (либо отсутствует). Для интервального измерения устанавливаются специальные единицы измерения (в психологии это стены).

Приведем **примеры**.

1. Исчисление лет от Рождества Христова;
2. Температура по Цельсию;
3. Тестовые шкалы *IQ* Векслера, *T*-шкалы и т. д.;
4. В психологии часто используется семантический дифференциал Осгуда, который является примером измерения в шкале интервалов различных психологических особенностей личности, социальных установок, ценностных ориентаций субъективно-личностного смысла, различных аспектов самооценки и т.п.



Шкала отношений. Измерения в этой шкале отличаются от интервального тем, что нулевая точка не произвольна, а указывает на полное отсутствие измеряемого признака. В силу абсолютности нулевой точки, при сравнении объектов мы можем сказать не только о том, насколько больше или меньше выражено свойство, но и том во сколько раз (на сколько % и т. д.) больше или меньше оно выражено.

К числам в этой шкале применимы все математические операции, а значит, отношения между числами соответствуют (пропорциональны) отношениям между количествами измеряемых свойств у разных объектов.

Приведем **примеры.**

1. Шкалы длинны, массы, времени.
2. Температура по Кельвину.
3. Шкала «сырых баллов» в психологии (количество решённых задач, количество ошибок).

Между самими шкалами тоже существуют отношения порядка. Каждая из перечисленных шкал является шкалой более высокого порядка по отношению к предыдущей шкале. Так, например, измерения произведенные в шкале отношений можно перевести в шкалу интервалов, из шкалы интервалов – в шкалу порядка и т.д., но обратная процедура будет невозможна, т.к. при переходе к шкалам более низкого порядка часть информации (о единицах измерения, количествах свойств) теряется.

Тем не менее, это не всегда означает, что шкалы более высокого порядка предпочтительней по отношению к шкалам более низкого порядка, а в ряде случаев – даже, наоборот. Например, количество правильно выполненных заданий в тесте интеллекта (шкала отношений) гораздо выгодней представить в стандартизированной шкале IQ (шкала интервалов), а множество разнообразных поведенческих реакций в виде типа личности (шкала наименований). Наконец, существуют такие признаки объектов, которые можно измерить в любой шкале, как возраст, и такие, к измерению которых подходит только одна шкала, как, например, пол. На выбор измерительной шкалы, таким образом, могут оказывать влияние многие факторы, как достоинства самой шкалы, так и специфика самого объекта измерения.

Упражнение. Отнесите каждое из следующих измерений к одному из классов:

1. Числа, кодирующие темперамент;
2. Академический ранг: ассистент, старший преподаватель, доцент, профессор;
3. Метрическая система измерения расстояний;
4. Телефонные номера;
5. Результаты экспертной оценки испытуемых;
6. Количество положительных ответов;
7. Соответствие норме: норма – патология;
8. Психологический тип: экстраверт – интроверт.

Пусть требуется изучить данную совокупность объектов относительно некоторого признака. Например, исследуя работу продуктового магазина, можно определить его загруженность, скорость обслуживания, тип клиентов и т. п. Каждый такой признак образует случайную величину, над которой мы осуществляем наблюдения.

Вся подлежащая изучению совокупность объектов (наблюдений) называется ***генеральной совокупностью.*** В математической статистике понятие ***генеральной совокупности*** трактуется как совокупность всех мыслимых

наблюдений, которые могли бы быть произведены при данном реальном комплексе условий, и в этом смысле его не следует смешивать с реальными совокупностями, подлежащими статистическому изучению. Так, обследовав даже все предприятия подотрасли по определенным технико-экономическим показателям, мы можем рассматривать обследованную совокупность лишь как представителя гипотетически возможной более широкой совокупности предприятий, которые могли бы функционировать в рамках того же реального комплекса условий.

Состав генеральной совокупности зависит от целей исследования. Иногда генеральная совокупность – это все население определенного региона (например, когда изучается отношение потенциальных избирателей к кандидату), чаще всего задается несколько критериев, определяющих объект исследования. Например, женщины 25 – 80 лет, использующие крем для лица определенных марок не реже раза в неделю, и имеющие доход не ниже 100 у. е. на одного члена семьи.

Понятие генеральной совокупности в определенном смысле аналогично понятию случайной величины (закону распределения вероятностей, вероятностному пространству), так как полностью обусловлено определенным комплексом условий.

Как правило, проводить сплошное обследование, когда изучаются все объекты (например, перепись населения) дорого, трудно, экономически нецелесообразно (к примеру, не вскрывать же каждую упаковку конфет для проверки качества партии продукции), а иногда и невозможно (если объектов очень много или доступ к ним ограничен). В этих случаях наилучшим способом исследования является *выборочный* метод: из всего множества исследуемых объектов выбирают лишь часть и подвергают ее изучению.

Та часть объектов, которая отобрана случайным образом для непосредственного изучения из генеральной совокупности, называется ***выборочной совокупностью*** или ***выборкой***. Если говорить более строго, то выборка – это последовательность независимых одинаково распределенных случайных величин $X_1, X_2, \ldots, X_n$, распределение каждой из которых совпадает с распределением генеральной совокупности. Выборочными наблюдениями являются, например, проводимые социологические исследования, охватывающие часть населения страны, области, района и т. д.

Выборку можно рассматривать как некий эмпирический аналог генеральной совокупности. Сущность выборочного метода состоит в том, чтобы по некоторой части генеральной совокупности (по выборке) выносить суждение о ее свойствах в целом.

Число объектов, (наблюдений) в генеральной или выборочной совокупности называется её ***объемом*** *и обозначается через* $\boldsymbol{N}$ *(или* $\boldsymbol{n}$ *соответственно)*.

Используют два способа образования выборки:

– ***повторный отбор***, когда каждый элемент, случайно отобранный и обследованный, возвращается в общую совокупность и может быть повторно отобран;

– ***бесповторный отбор***, когда отобранный элемент не возвращается в общую совокупность.

1.2 Вариационные ряды и их графическое изображение

Установление статистических закономерностей, присущих массовым случайным явлениям, основано на изучении статистических данных – сведений о том, какие значения принял в результате наблюдений интересующий нас признак (случайная величина ***X***).

Пусть изучается некоторая случайная величина X, т. е. над ней проводится серия независимых наблюдений. В каждом из этих наблюдений случайная величина X принимает то или иное значение. Пусть она приняла n_1 раз некоторое значение x_1, n_2 раз значение x_2, …, n_k раз значение x_k. При этом $n_1 + n_2 + \ldots + n_k = n$ – объем выборки. Значения $x_1, x_2, \ldots, x_k$ называются вариантами случайной величины ***X***.

Числа n_i, показывающие, сколько раз встречаются варианты x_i в ряде наблюдений, называются частотами, а их отношение к общему числу наблюдений n – ***частостями*** или ***относительными частотами***, т.е.

$$w_i = \frac{n_i}{n}.$$

Вся совокупность значений случайной величины ***X*** представляет собой первичный статистический материал, который подлежит дальнейшей обработке и, прежде всего, упорядочиванию. Операция расположения значений случайной величины по неубыванию (иногда по невозрастанию) называется ранжированием статистических данных.

Вариационным рядом называется ранжированный в порядке возрастания или убывания ряд вариант с соответствующими им частотами n_i или частостями w_i.

При изучении вариационных рядов наряду с понятием частоты используется понятие ***накопленной частоты*** (обозначаем $n_i^{нак}$). Накопленная частота показывает, сколько наблюдалось вариант со значением признака, меньше x. Отношение накопленной частоты $n_i^{нак}$ к общему числу наблюдений n назовем ***накопленной частостью*** $w_i^{нак}$:

$$w_i^{нак} = \frac{n_i^{нак}}{n}.$$

Набор вариант и соответствующих им частот или частостей называется дискретным вариационным рядом. Дискретный вариационный ряд можно представить в виде таблицы. Первая ее строка содержит варианты, а вторая – их частоты n_i или частости w_i (табл. 1).

Таблица 1

Варианта x_i	x_1	x_2	…	x_k
Частота n_i	n_1	n_2	…	n_k

где $n_1 + n_2 + \ldots + n_k = n$ – объем выборки.

Примером дискретного вариационного ряда является распределение 50 рабочих механического цеха по тарифному разряду (табл. 2).

Таблица 2

Тарифный разряд x_i	1	2	3	4	5	6	Σ
Частота (количество рабочих) n_i	2	3	6	8	22	9	50

В случае, когда количество значений признака случайной величины ***X*** велико или признак является непрерывным, составляют ***интервальный вариационный ряд.***

Его также можно представить в виде таблицы. В первую строку вписывают частичные промежутки $[x_0, x_1)$, $[x_1, x_2)$, …, $[x_{k-1}, x_k]$. Обычно промежутки берут равными по длине. Для определения величины интервал k можно использовать формулу Стерджеса:

$$k = \frac{x_{\max} - x_{\min}}{1 + [3{,}322 \lg n]}.$$

где $x_{\max} - x_{\min}$ – размах вариации – разность между наибольшим и наименьшим значениями признака,

$[a]$ – операция округления числа a до ближайшего целого в большую сторону.

За начало первого интервала рекомендуется брать величину $x_{нач} = x_{\min} - \frac{k}{2}$.

Во второй строке таблицы интервального вариационного ряда записывают количества наблюдений n_i, попавших в каждый интервал.

Отметим, что во многих практических задачах интервальный вариационный ряд превращают в дискретный вариационный ряд, заменяя интервалы их серединами.

Возникает вопрос, как быть с теми значениями, которые лежат границе двух интервалов. В этом случае пользуются одним из правил:

1. Все такие значения относят к предыдущему интервалу.
2. Все такие значения относят к последующему интервалу.
3. Такие значения делят пополам: из них половину относим к предыдущему интервалу, а вторую к последующему.

На протяжении всего исследования используют только одно из этих правил. Иногда следует сдвинуть $[x_{\min}; x_{\max}]$, чтобы крайние точки не попали на границу.

Пример. Получены следующие проранжированные данные о распределении 100 рабочих цеха по выработке в отчётном году (в процентах к предыдущему году):

$$x_{\min} = \underbrace{97{,}0; 97{,}8; \ldots\ldots; 105{,}7; 112{,}5; 121{,}5; 142{,}0}_{n=100\ значений} = x_{\max}.$$

Рассчитаем величину интервала по формуле Стерджеса

$$k = \frac{x_{\max} - x_{\min}}{1 + [3{,}322 \lg n]} = \frac{142{,}0 - 97{,}0}{1 + [3{,}322 \lg 100]} = 5{,}89(\%).$$

Примем k=6,0(%). За начало первого интервала рекомендуется брать величину $x_{нач} = x_{\min} - \frac{k}{2}$. В данном случае $x_{нач} = 97{,}0 - \frac{6{,}0}{2} = 94{,}0(\%)$.

Сгруппированный ряд представим в виде таблицы 3.

Таблица 3

i	Выработка в отчетном году в процентах к предыдущему	Частота (количество рабочих) n_i	Частность (доля рабочих) $w_i = \frac{n_i}{n}$	Накопленная частота $n_i^{нак}$	Накопленная частность $w_i^{нак} = \frac{n_i^{нак}}{n}$
1	94,0-100,0	3	0,03	3	0,03
2	100,0-106,0	7	0,07	10	0,1
3	106,0-112,0	11	0,11	21	0,21
4	112,0-118,0	20	0,20	41	0,4
5	118,0-124,0	28	0,28	69	0,69
6	124,0-130,0	19	0,19	8	0,88
7	130,0-136,0	10	0,10	98	0,98
8	136,0-142,0	2	0,02	100	1,00
Σ		100	1,00	-	-

Если просмотр первичных, несгруппированных данных делал затруднительным представление об изменчивости значений признака, то полученный теперь вариационный ряд позволяет выявить закономерности распределения рабочих по интервалам выработки. Мы видим, например, что выработка колеблется от 94,0% до 142,0%, наибольшее число рабочих (48 или 0,48 от общего числа) увеличили выработку до 112,0%–124,0%, уменьшили выработку (в пределах от 94,0% до 100%) 3 рабочих и т.п.

Накопленные частоты (накопленные частости) для каждого интервала находятся последовательным суммированием частот (накопленных частостей) всех предшествующих интервалов, включая данный (см. табл. 3). Например, для x=124% накопленная частота $n_i^{нак} = 3 + 7 + 11 + 20 + 28 = 69$, т.е. 69 рабочих имели выработку, меньшую 124%.

Для графического изображения вариационных рядов наиболее часто используются полигон, гистограмма, кумулятивная кривая.

Полигон, как правило, служит для изображения дискретного вариационного ряда и представляет собой ломанную, в которой концы отрезков прямой имеют координаты (x_i, n_i), ***i***=1,2,…*m*.

Гистограмма служит только для изображения интервальных вариационных рядов и представляет собой ступенчатую фигуру из прямоугольников с основаниями, равными интервалам значений признака $k_i = x_{i+1} - x_i$, ***i***=1,2,….,*m*, и высотами, равными частотам (частостям) n_i

интервалов. Если соединить середины верхних оснований прямоугольников отрезками прямой, то можно получить полигон того же распределения.

Кумулятивная кривая (кумулята) – кривая накопленных частот (накопленных частостей). Для дискретного вариационного ряда кумулята представляет ломаную, соединяющую точки $\left(x_i, n_i^{нак}\right)$ или $\left(x_i, w_i^{нак}\right)$, i=1,2,…, m.

Для интервального вариационного ряда ломаная начинается с точки, абсцисса которой равна началу первого интервала, а ордината – накопленной частоте (накопленной частости), равной нулю. Другие точки этой ломанной соответствуют концам интервала.

Весьма важным является понятие эмпирической функции распределения.

Эмпирической функцией распределения $F_n(x)$ называется относительная частота того, что признак (случайная величина ***X***) примет значение, меньшее заданного ***x***, т.е.

$$F_n(x) = w\left(X < x\right) = w_i^{нак} \qquad (1)$$

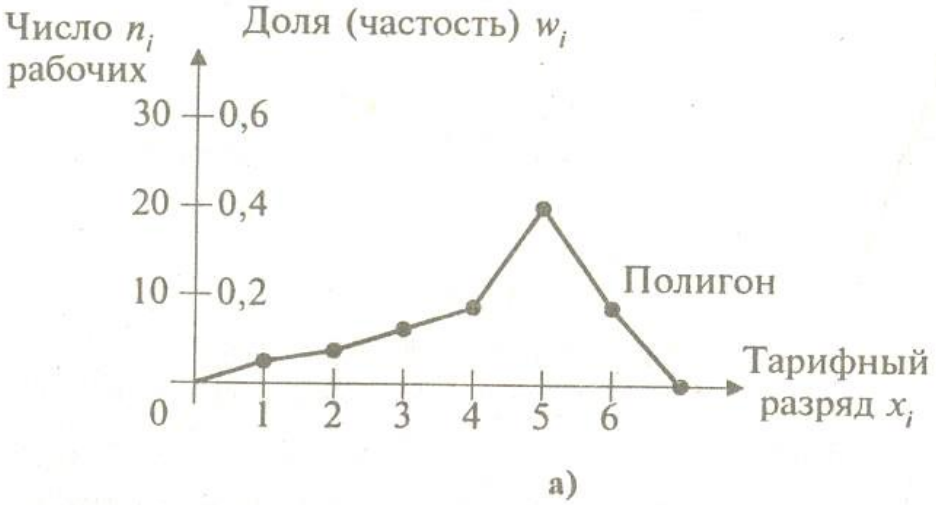
Другими словами, для данного X эмпирическая функция распределения представляет накопленную частость $w_i^{нак} = \dfrac{n_i^{нак}}{n}$.

Пример. Построить полигон, гистограмму, кумуляту и эмпирическую функцию распределения рабочих:

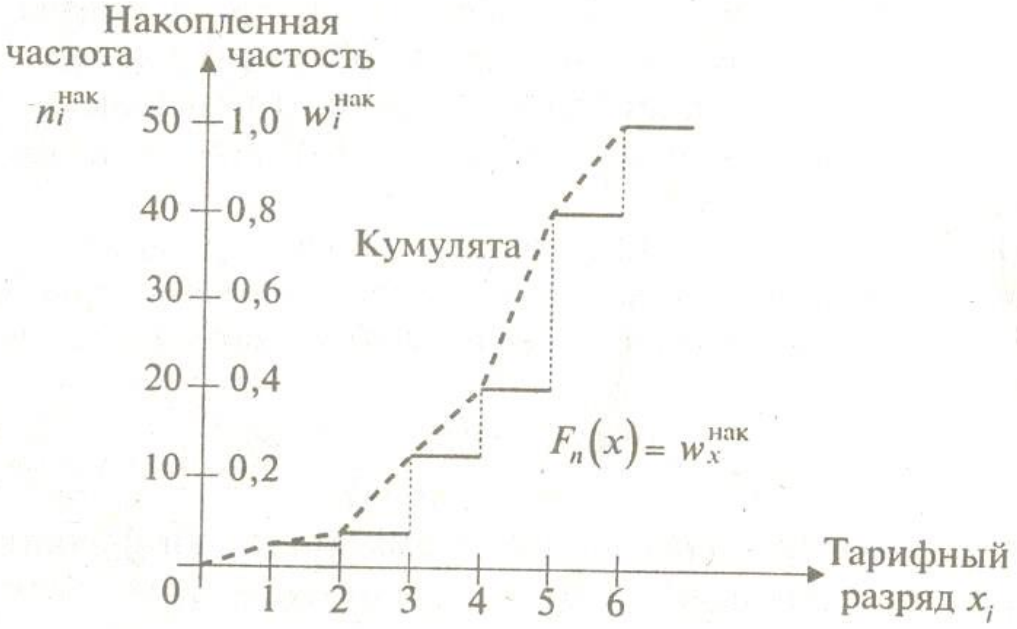
а) по тарифному разряду по данным табл. 2;

б) по выработке по данным табл. 3.

Решение. На рис. 1 и 2 изображены полигон, кумулята и эмпирическая функция распределения соответственно для дискретного (табл. 2) и интервального (табл. 3) вариационных рядов. Обращаем внимание на то, что для дискретного вариационного ряда эмпирическая функция распределения представляет собой разрывную ступенчатую функцию по аналогии с функцией распределения для дискретной случайной величины с той лишь разницей, что теперь по оси ординат вместо вероятностей располагаются частости (см. рис. 1).



а)



б)

Рис. 1

Для интервального вариационного ряда (табл. 3) имеем лишь значение функции распределения $F_n(x)$ на концах интервала (см. последнюю графу табл. 3). Поэтому для графического изображения этой функции целесообразно ее доопределить, соединив точки графика, соответствующие концам интервалов, отрезкам прямой. В результате получения ломаная совпадает с кумулятой (смотрите рис. 2.б.)

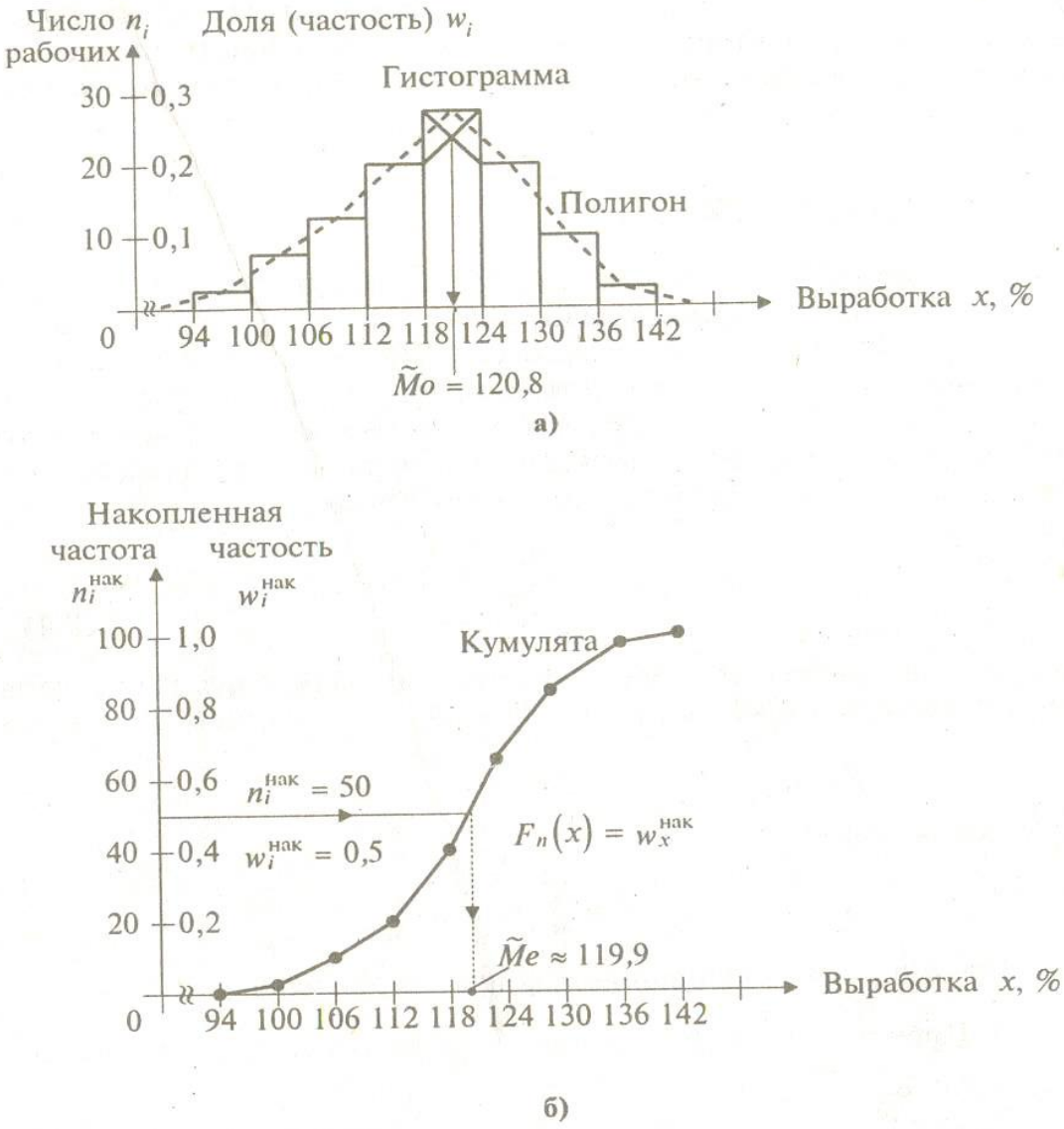


Рис. 2

Вариационный ряд является статистическим аналогом (реализацией) распределения признака (случайной величины ***X***). В этом смысле полигон аналогичен кривой распределения, а эмпирическая функция распределения – функция распределения случайной величины ***X***.

Вариационный ряд содержит достаточно полную информацию об изменчивости признака. Однако обилие числовых данных, с помощью которых он задается, усложняет их использование. В то же время на практике часто оказывается достаточным знание лишь сводных характеристик вариационных рядов: средних или характеристик центральной тенденции; характеристик изменчивости и др. Расчет статистических характеристик представляет собой второй после группировки этап обработки данных наблюдений.

1.3. Средние величины

Средние величины характеризуют значения признака, вокруг которого концентрируются наблюдения или, как говорят, центральную тенденцию распределения. Наиболее распространенной из средних величин является средняя арифметическая, как правило, ее называют выборочным средним.

Выборочным средним вариационного ряда называется сумма произведений всех вариант на соответствующие частоты, деленная на сумму частот:

$$\bar{x} = \frac{\sum_{i=1}^{m} x_i n_i}{n}, \qquad (2)$$

где x_i – варианты дискретного вариационного ряда или середины интервалов интервального вариационного ряда; n_i – соответствующие им частоты; $n = \sum_{i=1}^{m} n_i$.

Очевидно, что $\bar{x} = \sum_{i=1}^{m} x_i w_i$, где $w_i = \frac{n_i}{n}$ – частости вариант или интервалов.

Пример. Найти среднюю выработку рабочих по данным табл. 3.

Решение. По формуле (2) для интервального вариационного ряда

$$\bar{x} = \frac{97 \cdot 3 + 103 \cdot 7 + ... + 133 \cdot 10 + 139 \cdot 2}{100} = 119,2(\%),$$

где числа 97, 103, …, 133, 139 – середины соответствующих интервалов.

Для несгруппированного ряда все частоты $n_i = 1$ (i=1, 2, …, n), а

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (3)$$

есть «невзвешенная» выборочная средняя.

Отметим ***основные свойства*** выборочной средней, аналогичные свойствам математического ожидания случайной величины:

1. Выборочная средняя постоянной равна самой постоянной.
2. Если все варианты увеличить в одно и то же число раз, то выборочная средняя увеличиться во столько же раз:

$$\overline{kx} = k\bar{x} \text{ или } \frac{\sum_{i=1}^{m} (kx_i) n_i}{n} = k \frac{\sum_{i=1}^{m} x_i n_i}{n}.$$

3. Если все варианты увеличить (уменьшить) на одно, и то же число, то выборочная средняя увеличиться (уменьшится) на то же число:

$$\overline{x + c} = \bar{x} + c \text{ или } \frac{\sum_{i=1}^{m} (x_i + c) n_i}{n} = \frac{\sum_{i=1}^{m} x_i n_i}{n} + c.$$

4. Выборочная средняя отклонений вариант от выборочной средней равна нулю:

$$\overline{x-\bar{x}}=0 \text{ или } \sum_{i=1}^{m}\left(x_i-\bar{x}\right)n_i=0 \qquad (4)$$

При $c=\bar{x}$ имеем $\overline{x-c}=\bar{x}-c=\bar{x}-\bar{x}=0$.

5. Выборочная средняя алгебраической суммы нескольких признаков равна такой же сумме средних арифметических этих признаков:

$$\overline{x+y}=\bar{x}+\bar{y}.$$

6. Если ряд состоит из нескольких групп, общая выборочная средняя равна средней арифметической групповых средних:

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

где $\bar{x}$ – общая выборочная средняя (выборочная средняя всего ряда);

$\overline{x_i}$ – групповая средняя i-й группы, объем которой равен n_i;

l – число групп.

Медианой M_e вариационного ряда называется значение признака, приходящиеся на середину ранжированного ряда наблюдений.

Для дискретного вариационного ряда с нечетным числом членов медиана равна серединной варианте, а для ряда с четным числом членов – полусумме двух серединных вариант.

Пример. Найти медиану распределения рабочих по тарифному разряду по данным табл. 2.

Решение. Поскольку n=50 – четное, следовательно, серединных вариант две: $x_{25}=5$ и $x_{26}=5$. Поэтому $M_e=\frac{x_{25}+x_{26}}{2}=\frac{5+5}{2}=5\%$.

Для интервального вариационного ряда находиться медианный интервал, на который приходится середина ряда, а значение медианы на этом интервале находят с помощью линейного интерполирования. Не приводя соответствующей формулы, отметим, что медиана может быть приближенно найдена с помощью кумуляты как значение признака, для которого $n_x^{нак}=n/2$ или $w_x^{нак}=1/2$.

В случае интервального вариационного ряда находим медианный интервал (накопленные частоты до этого интервала меньше чем n/2, а после этого интервала больше n/2). Т.е. номер медианного интервала определяется на основании неравенства

$$\sum_{i=1}^{l}n_i\le\frac{n}{2}<\sum_{i=1}^{l+1}n_i.$$

Медиана находится по формуле

$$Me = x_l + h \cdot \frac{\frac{n}{2} - \sum_{i=1}^{l-1} n_i}{n_{Me}}.$$

где x_l – фактическая нижняя граница интервала медианы,

h – ширина медианного интервала,

$\sum_{i=1}^{l-1} n_i$ – накопленная к этому интервалу частота,

n_{Me} – частота на медианном интервале.

Достоинство медианы как меры центральной тенденции заключается в том, что на неё не влияет изменение крайних членов вариационного ряда, если любой из них, меньшей медианы, остается меньше ее, а любой, больший медианы, продолжает быть больше ее. Медиана предпочтительнее выборочной средней для ряда, у которого крайние варианты по сравнению с остальными оказались чрезмерно большими или малыми.

Модой $\boldsymbol{M_o}$ вариационного ряда называется варианта, которой соответствует наибольшая частота.

Например, для дискретного вариационного ряда табл. 2 мода $\boldsymbol{M_o}$=5, так как этой варианте соответствует наибольшая частота $n_i = 22$. Для интервального вариационного ряда находится модальный интервал, имеющий наибольшую частоту, а значение моды на этом интервале определяют с помощью формулы:

$$Mo = x_i + h \cdot \frac{n_i - n_{i-1}}{(n_i - n_{i-1}) + (n_i - n_{i+1})}.$$

Рассмотрим различные случаи:

1. Если все значения в группе встречаются одинаково часто, принято считать, что группа оценок не имеет моды. Таким образом, в группе (0,5; 0,5; 1,6; 1,6; 3,9; 3,9) моды нет.

2. Если два соседних значения имеют одинаковую частоту и они больше частоты любого другого значения, мода есть среднее этих двух значений. Итак, мода группы значений (0, 1, 1, 2, 2, 2, 3, 3, 3, 4) равна 2,5, т.к. (2+3)/2=2,5.

3. Если два несмежных значения в группе имеют равные частоты и они больше частот любого значения, то существуют две моды. В группе значений (10, 11, 11, 11, 12, 13, 14, 14, 14, 17) модами являются и 11 и 14; в таком случае говорят, что группа оценок является бимодальной.

Однако моду можно найти графическим путем с помощью полигона распределения и гистограммы. Если у нас дискретный вариационный ряд то мода – это абсцисса наиболее высокой точки полигона распределения.

Если распределения непрерывное, то находим модальный интервал (имеющий наибольшую частоту), тогда мода – это середина этого интервала.

Особенность моды как меры центральной тенденции заключается в том, что она не изменяется при изменении крайних членов ряда, т.е. обладает определенной устойчивостью к вариации признака.

Пример. Найти медиану и моду распределения рабочих по выработке по данным табл. 1.

Решение. На рис. 2б проведем горизонтальную прямую y=0.5 (или накопленной частоте $n_x^{нак} = 50$), до пересечения с графиком эмпирической функции распределения (или кумулятой). Абсцисса точки пересечения и будет медианой вариационного ряда: M_e=119,9 (%).

На гистограмме распределения (рис. 2а) находим прямоугольник с наибольшей частотой. Соединяя отрезками прямых вершины этого прямоугольника с соответствующими вершинами двух соседних прямоугольников (см. рис. 2а), получим точку пересечения этих отрезков, абсцисса которой и будет модой вариационного ряда: M_o=120,8%.

1.4. Показатели вариаций

Наибольший интерес представляют меры вариаций (рассеяние) наблюдений вокруг средних величин, в частности, вокруг выборочной средней.

Средним линейным отклонением вариационного ряда называется среднее арифметическая абсолютных величин отклонений вариант от их выборочной средней:

$$d = \frac{\sum_{i=1}^{m} |x_i - \bar{x}| \cdot n_i}{n}. \qquad (6)$$

Замечание. *Заметим, что «простая» сумма отклонений* $\sum_{i=1}^{m} (x_i - \bar{x}) n_i$ *не может характеризовать вариацию признака, ибо согласно свойству 4 выборочной средней эта сумма равна нулю для любого вариационного ряда).*

Дисперсией s^2 вариационного ряда называется выборочная средняя квадратов отклонений вариант от их выборочной средней:

$$s^2 = \frac{\sum_{i=1}^{m} \left(x_i - \bar{x}\right)^2 n_i}{n} \qquad (7)$$

Формулу для дисперсии вариационного ряда можно записать в виде:

$$s^2 = \sum_{i=1}^{m} (x_i - \bar{x})^2 w_i,$$

где $w_i = \frac{n_i}{n}$.

Для несгруппированного ряда $(n_i = 1)$ по формуле (7) имеем:

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

Дисперсию s^2 часто называют ***эмпирической*** или ***выборочной***, подчеркивая, что она (в отличие от дисперсии случайной величины) находится по опытным или статистическим данным.

Желательно в качестве меры вариации (рассеяния) иметь характеристику, выраженную в тех же единицах, что и значения признака. Такой характеристикой является ***выборочное среднее квадратическое отклонение s*** – арифметическое значение корня квадратного из дисперсии:

$$s = \sqrt{\frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2 n_i}{n}}. \qquad (8)$$

Рассматривается также безразмерная характеристика – ***коэффициент вариации,*** равный процентному отношению выборочного среднего квадратического отклонения к выборочной средней:

$$v = \frac{s}{\bar{x}} \cdot 100\% \ \left(\bar{x} \neq 0\right). \qquad (9)$$

Если коэффициент вариации признака, принимающего только положительные значения, высок (например, более 100%), то, как правило, это свидетельствует о неоднородности значений признака.

Отметим ***основные свойства выборочной дисперсии,*** аналогичные свойствам дисперсии случайной величины:

1. Выборочная дисперсия постоянной равна нулю.
2. Если все варианты увеличить (уменьшить) в одно и то же число k раз, то выборочная дисперсия увеличится (уменьшится) в k^2 раз:

$$s_{kx}^2 = k^2 s_x^2 \text{ или } \frac{\sum_{i=1}^{m}\left(kx_i - k\bar{x}\right)^2 n_i}{n} = k^2 \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2 n_i}{n}.$$

3. Если все варианты увеличить (уменьшить) на одно и то же число, то выборочная дисперсия не изменится:

$$s_{x+c}^2 = s_x^2 = s^2 \text{ или } \frac{\sum_{i=1}^{m}\left((x_i + c) - (\bar{x} + c)\right)^2 n_i}{n} = \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2 n_i}{n}.$$

4. Выборочная дисперсия равна разности между средней арифметической квадратов вариантов и квадратом средней арифметической:

$$s^2 = \overline{x^2} - \left(\bar{x}\right)^2, \qquad (10)$$

где $\overline{x^2} = \frac{\sum_{i=1}^{m} x_i^2 n_i}{n}$. (11)

$$s^2 = \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2 n_i}{n} = \frac{\sum_{i=1}^{m} x_i^2 n_i}{n} - 2\bar{x}\frac{\sum_{i=1}^{m} x_i n_i}{n} + \bar{x}^2 \frac{\sum_{i=1}^{m} n_i}{n} = \overline{x^2} - 2\bar{x}\cdot\bar{x} + \bar{x}^2 = \overline{x^2} - \bar{x}^2, \text{ ибо} \sum_{i=1}^{m} n_i = n.$$

5. Если ряд состоит из нескольких групп наблюдений, то общая выборочная дисперсия равна сумме выборочной средней групповых дисперсий и межгрупповых дисперсий.

$$s^2 = \overline{s_i^2} + \delta^2, \qquad (12)$$

где s^2 – общая дисперсия (дисперсия всего ряда);

$$\overline{s_i^2} = \frac{\sum_{i=1}^{l} s_i^2 n_i}{n}$$ – выборочная средняя групповых дисперсий (13)

$$s_i^2 = \frac{\sum_{j=1}^{m} \left(x_j - \overline{x_i}\right)^2 n_j}{n_i};$$ (14)

$$\delta^2 = \frac{\sum_{i=1}^{l} \left(\overline{x_i} - \overline{x}\right)^2 n_i}{n}$$ – межгрупповая дисперсия. (15)

Формула (12), известная в статистике как "***правило сложения дисперсий***", имеет важное значение в статистическом анализе.

Пример. Вычислить выборочную дисперсию, выборочное среднее квадратическое отклонение и коэффициент вариации распределения рабочих по выработке по данным табл. 3.

Решение. Ранее было получено, что $\overline{x} = 119,2(\%)$. По определению выборочная дисперсия равна

$$s^2 = \frac{(97-119,2)^2 \cdot 3 + (103-119,2)^2 \cdot 7 + ... + (133-119,2)^2 \cdot 10 + (139-119,2)^2 \cdot 2}{100} = 87,48.$$

Выборочное среднее квадратическое отклонение $s = \sqrt{87,48} = 9,35\%$; коэффициент вариации по формуле (9) $v = \frac{s}{\overline{x}} \cdot 100\% = \frac{9,35}{119,2} \cdot 100\% = 7,85\%.$

Следует отметить, что вычисление выборочной дисперсии (особенно в случае, когда отклонения от выборочной средней $\left(x_i - \overline{x}\right)^2$ выражаются нецелыми числами) бывает удобнее проводить по формуле (10).

Например, в данном примере вначале по (11) найдем:

$$\overline{x^2} = \frac{97^2 \cdot 3 + 103^2 \cdot 7 + ... + 133^2 \cdot 10 + 139^2 \cdot 2}{100} = 14296,12.$$

Теперь по (10) $s^2 = \overline{x^2} - \left(\overline{x}\right)^2 = 14296,12 - 119,2^2 = 87,48.$

Пример. Имеются следующие данные о выборочных средних и выборочных дисперсиях заработной платы двух групп рабочих (табл. 4):

Таблица 4

Группа рабочих	Число рабочих	Средняя зар.плата одного рабочего в группе (руб.)	Дисперсия заработной платы
Работающие на одном станке	40	2400	180000
Работающие на двух станках	60	3200	200000

Найти общую выборочную дисперсию распределения рабочих по заработной плате и его коэффициент вариации.

Решение. Найдем общую выборочную среднюю по формуле (5):

$$\bar{x} = \frac{2400 \cdot 40 + 3200 \cdot 60}{100} = 2880(\textit{руб.}).$$

Найдем выборочную среднюю групповых дисперсий по формуле (13):

$$\overline{s_i^2} = \frac{180000 \cdot 40 + 200000 \cdot 60}{100} = 192000.$$

Найдем межгрупповую дисперсию по формуле (15):

$$\delta^2 = \frac{(2400 - 2880)^2 \cdot 40 + (3200 - 2880)^2 \cdot 60}{100} = 153600.$$

Используя правило сложения дисперсий (12), найдем общую выборочную дисперсию заработной платы и ее выборочное среднее квадратическое отклонение:

$$s^2 = 192000 + 153600 = 345600;\ s = \sqrt{345600} = 588(\textit{руб.})$$

По формуле (9) коэффициент вариации

$$v = \frac{588}{2880} \cdot 100\% = 20,41\%.$$

1.5. Упрощенный способ расчета средней арифметической и дисперсии

Вычисление выборочной средней $\bar{x}$ и выборочной дисперсии s^2 вариационного ряда можно упростить, если использовать не первоначальные варианты x_i $(i = 1,2,...m)$, а новые варианты

$$u_i = \frac{x_i - c}{k}, \tag{16}$$

где c и k – специально подобранные постоянные.

Согласно свойствам 2 и 3 выборочной средней и выборочной дисперсии

$$\bar{u} = \overline{\left(\frac{x - c}{k}\right)} = \frac{\bar{x} - c}{k}, \tag{17}$$

$$s_u{}^2 = s_{\frac{x-c}{k}}{}^2 = \frac{s_{x-c}^2}{k^2} = \frac{s_x^2}{k^2},$$

откуда

$$\bar{x} = \bar{u}k + c \tag{18}$$

и

$$s_x^2 = k^2 s_u^2. \tag{19}$$

Учитывая (10), а затем (17), получим

$$s_x^2 = k^2\left(\overline{u^2} - \bar{u}^2\right) = k^2\overline{u^2} - k^2\bar{u}^2 = k^2\overline{u^2} - k^2\left(\frac{\bar{x} - c}{k}\right)^2 = k^2\overline{u^2} - (\bar{x} - c)^2.$$

Теперь, заменяя в (18) и (19) $\bar{u}$ и $\overline{u^2}$ их выражениями (2) и (11) через варианты u_i, получим

$$\bar{x} = \frac{\sum_{i=1}^{m} u_i n_i}{n} \cdot k + c, \qquad (20)$$

$$s_x^2 = \frac{\sum_{i=1}^{m} u_i{}^2 n_i}{n} \cdot k^2 - \left(\bar{x} - c\right)^2 \qquad (21)$$

где u_i определяются по (16).

Формулы (20) и (21) дадут заметное упрощение расчетов, если в качестве постоянной k взять величину интервала по x, а в качестве c – середину серединного интервала. Если серединных интервалов два (при четном числе интервалов), то в качестве c рекомендуется взять середину одного из этих интервалов, например, имеющего большую частоту.

Замечание. *Формулы (20) и (21) для* $\bar{x}$ *и* s_x^2 *носят технический, вспомогательный характер и позволяют рассчитать характеристики ряда по новым, условным вариантам. Основными же формулами, вытекающими из определения средней арифметической и дисперсии вариационного ряда и отражающими их сущность, остаются соответственно формулы (3) и (7).*

Пример. Вычислить упрощенным способом выборочную среднюю и выборочную дисперсию распределения рабочих по выработке по данным табл. 3.

Решение. Возьмем постоянную k, равную величине интервала, т.е. k=6, и постоянную c, равную середине пятого (одного из двух серединных) интервала, т.е. c=121. По формуле (16) новые варианты $u_i = \left(\frac{x_i - 121}{6}\right)$.

Благодаря такому переходу получим вместо варианты x_i=97, 103, 109, 115, 121, 127, 133, 139 «простые» варианты u_i =-4, -3, -2, -1, 0, 1, 2, 3.

Теперь для расчета $\bar{x}$ и s_x^2 по (20) и (21) необходимо найти суммы $\sum_{i=1}^{m} u_i n_i$ и $\sum_{i=1}^{m} u_i{}^2 n_i$. Их вычисление представим в табл. 5.

Таблица 5

i	Интервалы x	Середина интервала x_i	$u_i = \left(\frac{x_i - 121}{6}\right)$	n_i	$u_i n_i$	$u_i{}^2 n_i$	$u_i + 1$	$(u_i + 1)^2 n_i$
1	94,0-100,0	97	-4	2	-12	48	-3	27
2	100,0-106,0	103	-3	7	-21	63	-2	28
3	106,0-112,0	109	-2	11	-22	44	-1	11
4	112,0-118,0	115	-1	20	20	20	0	0
5	118,0-124,0	121	0	28	0	0	1	28
6	124,0-130,0	127	1	19	19	19	2	76
7	130,0-136,0	133	2	10	20	40	3	90
8	136,0-142,0	139	3	2	6	18	4	32
	$\sum$			100	-30	252		292

В итоговой строке табл. 5 находим $\sum_{i=1}^{8} u_i n_i = -30$, $\sum_{i=1}^{m} u_i^2 n_i = 252$.

Последний столбец – контрольный. Если таблица составлена верно, то

$$\sum_{i=1}^{m} (u_i + 1)^2 n_i = \sum_{i=1}^{m} u_i^2 n_i + 2\sum_{i=1}^{m} u_i n_i + n, \text{ где } n = \sum_{i=1}^{m} n_i .$$

В данном случае $\sum_{i=1}^{8} (u_i + 1) n_i = 292 = 252 + 2 \cdot (-30) + 100$, т.е. расчеты проведены верно.

Теперь по формуле (20) получаем $\bar{x} = \frac{-30}{100} \cdot 6 + 121 = 119,2\%$, а по формуле (21) получаем $s_x^2 = \frac{252}{100} \cdot 6^2 - (119,2 - 121)^2 = 87,48$.

1.6. Начальные и центральные моменты вариационного ряда

Выборочная средняя и выборочная дисперсия вариационного ряда являются частными случаями более общего понятия – моментов вариационного ряда.

Начальный момент ν_k k-го порядка вариационного ряда определяется по формуле:

$$\nu_k = \frac{\sum_{i=1}^{m} x_i^k n_i}{n}. \qquad (22)$$

Очевидно, что $\nu_1 = \bar{x}$, т.е. выборочная средняя является начальным моментом первого порядка вариационного ряда.

Центральный момент μ_k k-го порядка вариационного ряда определяется по формуле:

$$\mu_k = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^k n_i}{n}. \qquad (23)$$

С помощью моментов распределения можно описать не только среднюю тенденцию, рассеяние, но и другие особенности вариации признака.

Очевидно, в силу (4), что $\mu_1 = 0$, а $\mu_2 = s^2$, т.е. центральный момент первого порядка для любого распределения равен нулю, а второго порядка является выборочной дисперсией вариационного ряда.

Коэффициентом асимметрии вариационного ряда называется число

$$A_s = \frac{\mu_3}{s^3} = \frac{\sum_{i=1}^{m} (x_i - \bar{x})^3 n_i}{n s^3}. \qquad (24)$$

Если $A_s = 0$, то распределение имеет симметричную форму, т.е. варианты, равноудаленные от $\bar{x}$, имеют одинаковую частоту. При $A_s > 0$ $(A_s < 0)$ говорят

о положительной (правосторонней) или отрицательной (левосторонней) асимметрии.

Эксцессом вариационного ряда называется число

$$E_k = \frac{\mu_4}{s^4} = \frac{\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^4 n_i}{ns^4} - 3. \qquad (25)$$

Эксцесс является показателем «крутости» вариационного ряда по сравнению с нормальным распределением, эксцесс нормально распределенной случайной величины равен нулю.

Если $E_k > 0$ $\left(E_k < 0\right)$, то полигон вариационного ряда имеет более крутую (пологую) вершину по сравнению с нормальной кривой.

Пример. Вычислить коэффициент асимметрии и эксцесс вариационного ряда, приведенного в табл. 3.

Решение. Коэффициент асимметрии и эксцесс вариационного ряда найдем по формулам (24) и (25):

$$A_s = \frac{\left(97-119,2\right)^3 \cdot 3 + \left(103-119,2\right)^3 \cdot 7 + ... + \left(139-119,2\right)^3 \cdot 2}{100 \cdot 9,35^3} = -0,302;$$

$$E_k = \frac{\left(97-119,2\right)^4 \cdot 3 + \left(103-119,2\right)^4 \cdot 7 + ... + \left(139-119,2\right)^4 \cdot 2}{100 \cdot 9,35^4} - 3 = -0,286.$$

В силу того, что коэффициент асимметрии A_s отрицателен и близок нулю, распределение рабочих по выработке обладает незначительной левосторонней асимметрией, а поскольку эксцесс E_k близок нулю, рассматриваемое распределение по крутости приближается к нормальной кривой.

Замечание. *Выборочная средняя $\bar{x}$, выборочная дисперсия s^2 и другие характеристики вариационного ряда являются статистическими аналогами математического ожидания* ***M(X)****, дисперсии σ^2 и соответствующих характеристик случайной величины* ***X.***

1.7 Лабораторная работа 1. Статистический анализ данных в Microsoft Excel

Цель: *научиться обрабатывать статистические данные в Excel.*

Задание 1. Построить эмпирическое распределение результатов тестирования 55 студентов в баллах: 64, 57, 63, 62, 58, 61, 63, 60, 60, 61, 65, 62, 62, 60, 64, 61, 59, 59, 63, 61, 62, 58, 58, 63, 61, 59, 62, 60, 60, 58, 61, 60, 63, 63, 58, 60, 59, 60, 59, 61, 62, 62, 63, 57, 61, 58, 60, 64, 60, 59, 61, 64, 62, 59, 65.

Решение.

1. В ячейку A1 введите слово *Наблюдения,* а в диапазоне A2:E12 – результаты тестирования студентов.

2. Найдите минимальное и максимальное значения, для этого в ячейке A14 вставьте статистическую функцию МИН и укажите в качестве аргумента интервал ячеек A2:E12, а в ячейку A15 вставьте статистическую функцию МАКС и укажите в качестве аргумента тот же интервал ячеек.

3. Выберите ширину интервала 1 балл. Тогда при крайних значениях 57 и 65 получится 9 интервалов. В ячейки G1 и G2 введите названия интервалов *Результаты тестирования* и *Баллы*, соответственно. В диапазон G4:G12 введите граничные значения интервалов (*57, 58, 59, 60, 61, 62, 63, 64, 65).*

4. Введите заголовки создаваемой таблицы: в ячейки H1:H2 – *Абсолютные частоты,* в ячейки I1:I2 *–Относительные частоты,* в ячейки J1:J2 – *Накопленные частоты*.

5. Заполните столбец абсолютных частот. Для этого выделите для них блок ячеек H4:H12. На панели инструментов *Стандартная* вызовите *Мастер функций* (кнопка ). В появившемся диалоговом окне *Мастер функций* выберите категорию *Статистические* и функцию ЧАСТОТА, после чего нажмите кнопку ОК. В рабочее поле *Массив_данных* мышью введите диапазон интервалов A2:E12. В рабочее поле *Массив_интервалов* мышью введите диапазон интервалов G4:G12. Последовательно нажмите комбинацию клавиш *Ctrl+Shift+Enter*. В столбце H4:H12 появится массив абсолютных частот.

6. В ячейке H13 найдите общее количество наблюдений. Для этого установите курсор в ячейку H13, на панели *Стандартная* нажмите кнопку *Автосумма*. Убедитесь, что диапазон суммирования указан правильно *(H4:H12)*, и нажмите клавишу *Enter.*

7. Заполните столбец относительных частот. В ячейку I4 введите формулу для вычисления относительной частоты: *= H4/H$13.* Нажмите клавишу *Enter.* Протягиванием (за правый нижний угол при нажатой левой кнопке мыши) скопируйте введенную формулу в диапазон I5:I12. Получим массив относительных частот.

8. Заполните столбец накопленных частот. В ячейку J4 скопируйте значение относительной частоты из ячейки I4. В ячейку J5 введите формулу: *=J4+I5*. Нажмите клавишу *Enter*. Протягиванием скопируйте введенную формулу в диапазон J6:J12. Получим массив накопленных частот.

	A	B	C	D	E	F	G	H	I	J
1	Наблюдения						Баллы	Абсолютные	Относительные	Накопленные
2	64	62	58	63	61			частоты	частоты	частоты
3	57	62	63	58	58					
4	63	60	61	60	60		57	2	0,036363636	0,036363636
5	62	64	59	59	64		58	6	0,109090909	0,145454545
6	58	61	62	60	60		59	7	0,127272727	0,272727273
7	61	59	60	59	59		60	10	0,181818182	0,454545455
8	63	59	60	61	61		61	9	0,163636364	0,618181818
9	60	63	58	62	64		62	8	0,145454545	0,763636364
10	60	61	61	62	62		63	7	0,127272727	0,890909091
11	61	62	60	63	59		64	4	0,072727273	0,963636364
12	65	58	63	57	65		65	2	0,036363636	1
13								55		
14	57									
15	65									

9. Постройте диаграмму относительных и накопленных частот. Щелчком указателя мыши по кнопке на панели инструментов вызовите *Мастер диаграмм*. В появившемся диалоговом окне выберите вкладку *Нестандартные* и тип диаграммы *График/гистограмма 2*. После нажатия кнопки *Далее* укажите

диапазон данных – I1:J12 (с помощью мыши). Проверьте положение переключателя *Ряды* в: столбцах. Выберите вкладку *Ряд* и с помощью мыши введите в рабочее поле *Подписи оси X* диапазон подписей оси X: *G4:G12*. Нажав кнопку *Далее,* введите названия осей X и Y: в рабочее поле *Ось X* (категорий) – Баллы; *Ось Y* (значений) – Относительная частота; *Вторая ось Y* (значений) – Накопленная частота. Нажмите кнопку *Готово*.

10. Сохраните документ под именем *Статистика.*

Задание 2. Рассматриваются ежемесячные количества реализованных туристической фирмой путевок за периоды до и после начала активной рекламной компании. Ниже приведены количества реализованных путевок по месяцам.

Месяц	Количество реализованных путевок	
	С рекламой	Без рекламы
Январь	125	110
Февраль	120	115
Март	130	125
Апрель	162	135
Май	156	136
Июнь	144	125
Июль	145	147
Август	168	156
Сентябрь	156	145
Октябрь	137	130
Ноябрь	115	111
Декабрь	155	120

Требуется определить для данных показателей: среднее значение, медиану, моду, дисперсию, ранги, стандартные отклонения, эксцесс и асимметрию.

Решение.

1. Перейдите на второй лист рабочей книги и создайте следующую таблицу:

	A	B	C	D	E
1	Месяц	Количество реализованных путевок		Ранги	
2		С рекламой	Без рекламы		
3	Январь	125	110		
4	Февраль	120	115		
5	Март	130	125		
6	Апрель	162	135		
7	Май	156	136		
8	Июнь	144	125		
9	Июль	145	147		
10	Август	168	156		
11	Сентябрь	156	145		
12	Октябрь	137	130		
13	Ноябрь	115	111		
14	Декабрь	155	120		
15	Среднее				
16	Медиана				
17	Мода				
18	Дисперсия				
19	Стандартное отклонение				
20	Эксцесс				
21	Асимметрия				

2. Для вычисления среднего в ячейку B15 вставьте статистическую функцию СРЗНАЧ и укажите в качестве ее аргумента интервал ячеек B3:B14. Скопируйте полученную формулу в ячейку C15.

3. Для вычисления медианы в ячейку B16 вставьте статистическую функцию МЕДИАНА и укажите в качестве ее аргумента интервал ячеек B3:B14. Скопируйте полученную формулу в ячейку C16.

4. Для вычисления моды в ячейку B17 вставьте статистическую функцию МОДА и укажите в качестве ее аргумента интервал ячеек B3:B14. Скопируйте полученную формулу в ячейку C17.

5. Для вычисления дисперсии в ячейку B18 вставьте статистическую функцию ДИСП и укажите в качестве ее аргумента интервал ячеек B3:B14. Скопируйте полученную формулу в ячейку C18.

6. Для определения ранга реализации путевок по месяцам после начала рекламной компании в ячейку D3 вставьте статистическую функцию РАНГ и укажите в качестве аргумента *Число* ячейку B3, а в качестве аргумента *Ссылка* интервал ячеек B3:B14 (также воспользуйтесь клавишей F4 для задания абсолютной ссылки на интервал ячеек). Скопируйте полученную формулу в ячейки D4:D14. Определите месяцы, в которых было наибольшее и наименьшее количество реализованных путевок.

7. Аналогично определите ранг реализации путевок до начала рекламной компании в ячейках E3:E14.

8. Для вычисления стандартного отклонения в ячейку B19 вставьте статистическую функцию СТАНДОТКЛОН и укажите в качестве ее аргумента интервал ячеек B3:B14. Скопируйте полученную формулу в ячейку C19.

9. Для вычисления эксцесса в ячейку B20 вставьте статистическую функцию ЭКСЦЕСС и укажите в качестве ее аргумента интервал ячеек B3:B14. Скопируйте полученную формулу в ячейку C20.

10. Для вычисления асимметрии в ячейку B21 вставьте статистическую функцию СКОС и укажите в качестве ее аргумента интервал ячеек B3:B14. Скопируйте полученную формулу в ячейку C21.

11. Сохраните документ.

Задание 3. Психолог исследует эффективность четырех различных методик обучения математике. Для этой цели из всех учащихся СШ были выбраны 4 параллельных класса, обучавшихся четырьмя этими методами.

Эффективность обучения оценивалась по итогам контрольной работы (по 20-балльной системе).

Получены следующие данные:

№	Методика A	Методика B	Методика C	Методика D
1	11	15	8	16
2	9	16	8	14
3	10	15	9	12
4	15	14	15	14
5	13	12	14	17
6	20	18	12	18
7	19	17	17	20
8	17	13	10	15
9	11	10	9	11
10	16	11	15	12
11	14	12	12	13
12	10	15	16	10
13	8	16	13	10
14	9	14	18	15
15	15	20	19	16
16	12	19	14	16
17	18	20	8	14
18	16	18	9	18
19	20	15	9	17
20	19	14	7	19
21	17	17	10	20
22	13	12	10	19
23	10	11		
24		11		

Необходимо определить основные статистические характеристики в группах данных.

Решение.

1. Перейдите на третий лист рабочей книги и создайте следующую таблицу:

	A	B	C	D	E
1	№	Методика A	Методика B	Методика C	Методика D
2	1	11	15	8	16
3	2	9	16	8	14
4	3	10	15	9	12
5	4	15	14	15	14
6	5	13	12	14	17
7	6	20	18	12	18
8	7	19	17	17	20
9	8	17	13	10	15
10	9	11	10	9	11
11	10	16	11	15	12
12	11	14	12	12	13
13	12	10	15	16	10
14	13	8	16	13	10
15	14	9	14	18	15
16	15	15	20	19	16
17	16	12	19	14	16
18	17	18	20	8	14
19	18	16	18	9	18
20	19	20	15	9	17
21	20	19	14	7	19
22	21	17	17	10	20
23	22	13	12	10	19
24	23	10	11		
25	24		11		

2. Для использования инструментов пакета анализа в меню *Сервис*, выберите команду *Анализ данных.* Затем в появившемся списке *Инструменты анализа* выберите строку *Описательная статистика*.

3. В появившемся диалоговом окне в рабочем поле *Входной интервал*, укажите входной диапазон B2:E25. Активировав рабочее поле *Выходной интервал*, укажите выходной диапазон – ячейку A27. В разделе *Группировка* переключатель установите в положение по столбцам. Установите флажок в поле *Итоговая статистика* и нажмите кнопку ОК.

4. Проанализируйте полученный результат.

5. Сохраните документ.

Примеры решения задач

1. Составить вариационный ряд оценок, полученных студентами одной группы по высшей математике: 3,4,5,4,3,2,5,5,3,3,4,3,4,2,3,3,4,4,5,3.

Решение. Составим вариационный ряд и запишем его в виде таблицы

Варианты (x_i)	2	3	4	5
Частоты (n_i)	2	8	6	4

Общая сумма частот вариационного ряда равна объему выборки, т.е.

$$n = \sum_{i=1}^{4} n_i = 2 + 8 + 6 + 4 = 20 .$$

Построим полигон частот по данному распределению. Для этого отложим на оси абсцисс варианты x_i, а на оси ординат – соответствующие им частоты n_i, соединив точки (x_i, n_i) отрезками прямых, получим искомый полигон частот.

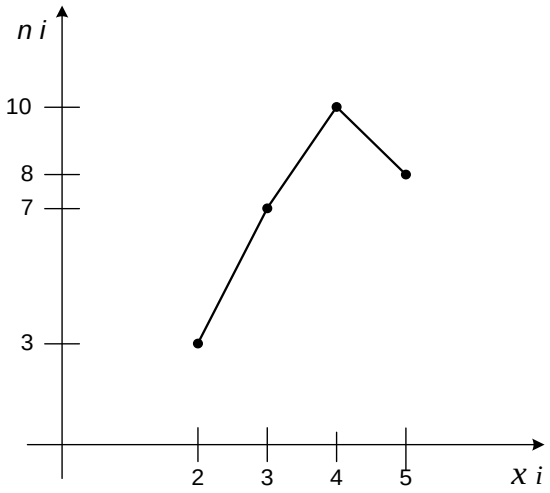


Рис. 3.

Частоты могут накапливаться. Накопленные частоты получаются последовательным суммированием значений частот от первой до последней.

Варианты (x_i)	2	3	4	5
Частоты (n_i)	2	8	6	4
Кумуляты частот	2	10	16	20

2. Приведены результаты контрольной работы по высшей математике у студентов одной группы: 2,4,4,2,3,3,5,3,5,5,5,2,3,3,4,4,4,5,4,3,5,5,3,4,4,4,5,4.

Найти распределение относительных частот. Построить полигон частот и полигон относительных частот.

Решение. Построим соответствующий вариационный ряд 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5.

Выборка содержит 4 различные значения, опишем ее статистическим рядом

x_i	2	3	4	5
n_i	3	7	10	8

Объем выборки: n=3+7+10+8=28. Найдем относительные частоты по формуле $w_i = \frac{m_i}{n}$: $w_1 = \frac{3}{28} = 0{,}11$; $w_2 = \frac{7}{28} = 0{,}25$; $w_3 = \frac{10}{28} = 0{,}36$; $w_4 = \frac{8}{28} = 0{,}28$. Напишем искомое распределение относительных частот:

x_i	2	3	4	5
w_i	0,11	0,25	0,36	0,28

Построим полигон частот по данному распределению. Для этого отложим на оси абсцисс варианты x_i, а на оси ординат – соответствующие им частоты n_i соединив точки (x_i, n_i) отрезками прямых, получим искомый полигон частот рис. 4.

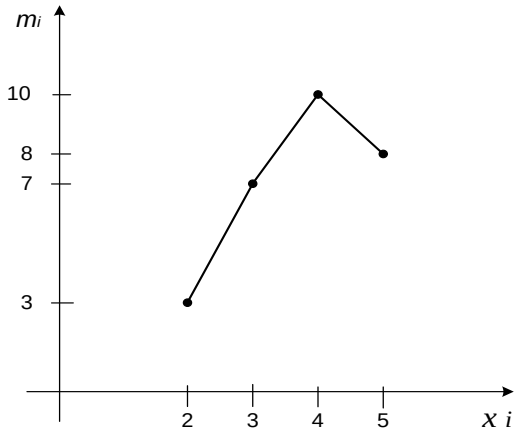


Рис. 4.

Чтобы построить полигон относительных частот, воспользуемся распределением относительных частот. Отложим на оси абсцисс варианты x_i, а на оси ординат w_i – соответствующие относительные частоты, соединим точки (x_i, w_i) отрезками прямых и получим искомый полигон относительных частот на рис. 5.

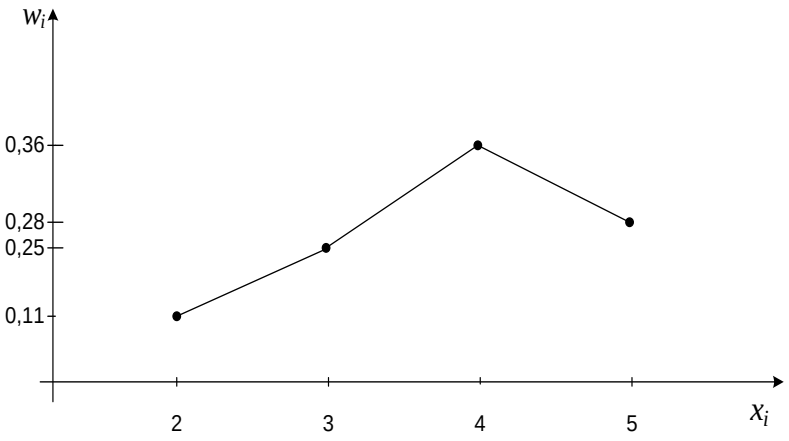


Рис. 5

3. В лыжной команде вариационным рядом размера обуви является для юношей 39,39,40,40,40,40,41,41,41, а для девушек – 35,35,35,35,36,39. Найти соответствующие медианы.

Решение. В первом случае нечетное число членов ряда N=9, поэтому M_e=x_5=40. Второй ряд имеет четное число членов N=6, тогда $M_e = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}(35+35) = 35$.

4. Рабочая бригада из 11 человек имеет следующие тарифные разряды: 9,5,9,6,6,8,7,6,7,8,6. Найти моду выборки.

Решение. Упорядочим вариационный ряд по возрастанию 5,6,6,6,6,7,7,8,8,9,9. Мода отражает наиболее распространенную варианту рассматриваемого признака (т.е. варианту, которая имеет наибольшую частоту). В данном случае M_o=6, т.е. это будет рабочий шестого разряда.

5. По данным социологического исследования был получен ряд распределения рабочих цеха по числу изготавливаемых за смену деталей.

x_i	8	10	12	14
n_i	1	4	10	6

Вычислить выборочное среднее, выборочную дисперсию, выборочное среднеквадратическое отклонение, коэффициент вариации и размах вариации.

Решение. Для нахождения выборочного среднего воспользуемся формулой:

$$\overline{x} = \frac{x_1 \cdot n_1 + x_2 \cdot n_2 + x_3 \cdot n_3 + x_4 \cdot n_4}{n},\ \overline{x} = \frac{8 \cdot 1 + 10 \cdot 4 + 12 \cdot 10 + 14 \cdot 6}{1+4+10+6} = \frac{8+40+120+84}{21} = 12.$$

Найдем выборочную дисперсию

$$s^2 = \frac{(x_1 - \overline{x})^2 n_1 + (x_2 - \overline{x})^2 n_2 + (x_3 - \overline{x})^2 n_3 + (x_4 - \overline{x})^2 n_4}{n} =$$

$$= \frac{(8-12)^2 \cdot 1 + (10-12)^2 \cdot 4 + (12-12)^2 \cdot 10 + (14-12)^2 \cdot 6}{21} = \frac{4^2 \cdot 1 + 2^2 \cdot 4 + 0^2 \cdot 10 + 2^2 \cdot 6}{21} =$$

$$= \frac{16+16+0+24}{21} = \frac{56}{21}.$$

Выборочное среднеквадратическое отклонение $s=\sqrt{s^2}=\sqrt{\frac{56}{21}}\approx 1,63$. По формуле $v=\frac{s}{\overline{x}}\cdot 100\%$ вычисляем коэффициент вариации: $v=\frac{1,63}{12}\cdot 100\%=13,58\%.$

Размах вариации: $R=x_{\max}-x_{\min}=14-8=6$.

6. Обследование качества пряжи на прочность дало следующие результаты

Прочность нити	Число случаев (частота)	Накопленная частота
120-140	1	1
140-160	6	7
160-180	19	26
180-200	58	84
200-220	53	137
220-240	24	161
240-260	16	177
260-280	3	180

Найти моду, медиану.

Решение. Так как наибольшая частота $n_i=58$ отвечает интервалу 180-200, то $x_i=180,\ n_{i-1}=19,\ n_{i+1}=53,\ h=20$. По формуле

$$Mo=x_i+h\cdot\frac{n_i-n_{i-1}}{(n_i-n_{i-1})+(n_i-n_{i+1})}$$

получаем:

$$Mo=180+20\cdot\frac{58-19}{(58-19)+(58-52)}=197,73.$$

Номер медианного интервала определяется на основании неравенства:

$$\sum_{i=1}^{l}m_i\le\frac{n}{2}<\sum_{i=1}^{l+1}m_i\ .$$

$$\sum_{i=1}^{l}m_i\le\frac{180}{2}<\sum_{i=1}^{l+1}m_i\Rightarrow\sum_{i=1}^{4}m_i\le 90<\sum_{i=1}^{5}m_i\Rightarrow 84\le 90<137.$$

Следовательно, номер медианного интервала $m_{Me}=5$, а сам интервал 200-220. По формуле:

$$Me=x_l+h\cdot\frac{\frac{n}{2}-\sum_{i=1}^{l-1}m_i}{m_{Me}}.$$

получаем

$$Me=200+20\cdot\frac{90-84}{53}=202,26.$$

7. Был проведен следующий эксперимент: книгу «Основы высшей математики и теории вероятностей» открывали на случайной странице и искали слово «вероятность». При этом фиксировалось, сколько раз встречается

искомое слово. В результате 20 опытов получена следующая выборка 4, 1, 4, 5, 1, 13, 4, 10, 2, 4, 7, 2, 2, 4, 6, 4, 5, 6, 2, 4. Найти эмпирическую функцию распределения и построить ее график.

Решение. Соответствующий вариационный ряд 1, 1, 2, 2, 2, 2, 4, 4, 4, 4, 4, 4, 4, 5, 5, 6, 6, 7, 10, 13. Выборка содержит 8 различных значений и описывается следующим статистическим рядом:

x_i	1	2	4	5	6	7	10	13
m_i	2	4	7	2	2	1	1	1

Объем выборки n=2+4+7+2+2+1+1+1=20. Наименьшая варианта равна 1, тогда при $x \le 1$ $F_n(x)$=0, значение x<2, а именно x_1=1 наблюдалось 2 раза, тогда, $F_n(x)$=2/20=0,1 при $1 < x \le 2$. Значения x<4, а именно x_1=1 и x_2=2 принимались 2+4=6 раз, следовательно $F_n(x)$=6/20=0,3 при $2 < x \le 4$. Значения x<5 принимались 2+4+7=13 раз, тогда $F_n(x)$=13/20=0,65 при $4 < x \le 5$. Значения x<6 принимались 2+4+7+2=15 раз, тогда $F_n(x)$=15/20=0,75 при $5 < x \le 6$. Значения x<7 принимались 2+4+7+2+2=17 раз, тогда $F_n(x)$=17/20=0,85 при $6 < x \le 7$. Значения x<10 наблюдались 17+1=18 раз, тогда $F_n(x)$=18/20=0,9 при $7 < x \le 10$. Значения x<13 принимались 18+1=19 раз, тогда $F_n(x)$=19/20=0,95 при $10 < x \le 13$. Так как x=13 – наибольшая варианта, то $F_n(x)$=1, при x>13.Аналитическое выражение искомой эмпирической функции распределения:

$$F_n(x) = \begin{cases} 0, & x \le 1 \\ 0{,}1, & 1 < x \le 2 \\ 0{,}3, & 2 < x \le 4 \\ 0{,}65, & 4 < x \le 5 \\ 0{,}75, & 5 < x \le 7 \\ 0{,}85, & 6 < x \le 7 \\ 0{,}9, & 7 < x \le 10 \\ 0{,}95, & 10 < x \le 13 \\ 1, & x > 13 \end{cases}$$

Построим график этой функции

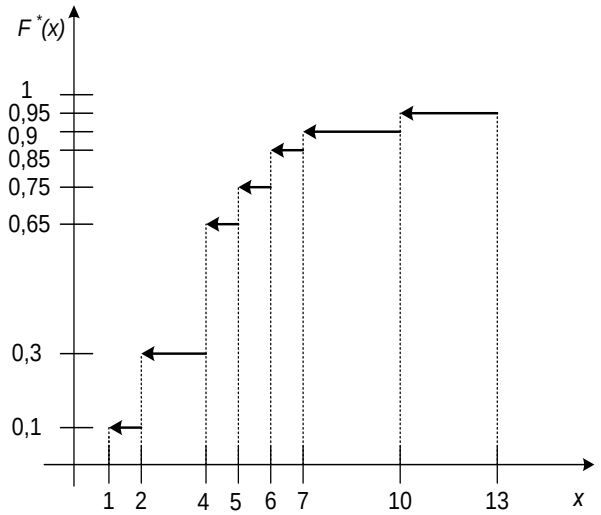


Рис. 6.

8. Требуется найти выборочную среднюю и выборочную дисперсию для интервального вариационного ряда заданного таблицей:

Рост	150-154	154-158	158-162	162-166	166-170	170-174	174-178	178-182	182-186
Число студентов, n_i	10	30	105	230	250	220	115	30	10

Решение. Для расчета выборочной средней переходим от интервального вариационного ряда к дискретному вариационному ряду, заменяя каждый интервал его серединой. Далее перейдем к условным вариантам по формуле: $u_i = \frac{x_i - c}{k}$, где c = 168, а k = 4, т.е. $u_i = \frac{x_i - 168}{4}$. Все необходимые вычисления приведем в таблице:

Рост, (см)	Число студентов, n_i	Середина интервала, x_i	$u_i = \frac{x_i - c}{k}$	$u_i \cdot n_i$	$u_i^2 \cdot n_i$
150-154	10	152	-4	-40	160
154-158	30	156	-3	-90	270
158-162	105	160	-2	-210	420
162-166	230	164	-1	-230	230
166-170	250	168	0	0	0
170-174	220	172	1	220	220
174-178	115	176	2	230	460
178-182	30	180	3	90	270
182-186	10	184	4	40	160
Сумма	1000	-	-	10	2190

Тогда $\bar{u} = \frac{\sum u_i \cdot n_i}{n} = \frac{10}{1000}$.

Отсюда найдем выборочную среднюю $\bar{x} = \bar{u} \cdot k + c = \frac{10}{1000} \cdot 4 + 168 = 168,04$.

Найдем выборочную дисперсию по формуле

$$s_x^2 = \frac{\sum_{i=1}^{m} u_i^2 n_i}{n} \cdot k^2 - \left(\bar{x} - c\right)^2 = \frac{2190}{1000} \cdot 4^2 - \left(168,04 - 168\right)^2 = 35,0384 \approx 35,04.$$

9. Дано распределение признака X. Найти асимметрию и эксцесс.

x_i	-2	-1	0	1	2	3
m_i	2	4	6	5	2	1

Решение. Сначала находим начальные моменты первого, второго, третьего и четвёртого порядков данного распределения.

x_i	m_i	$m_i x_i$	$m_i x_i^2$	$m_i x_i^3$	$m_i x_i^4$
-2	2	-4	8	-16	32
-1	4	-4	4	-4	4
0	6	0	0	0	0
1	5	5	5	5	5
2	2	4	8	16	32
3	1	3	9	27	81
$\sum$	20	4	34	28	154

Отсюда $\nu_1 = \frac{4}{20} = 0,2$, $\nu_2 = \frac{34}{20} = 1,7$, $\nu_3 = \frac{28}{20} = 1,4$, $\nu_4 = \frac{154}{20} = 7,7$, следовательно,

$$\mu_2 = \nu_2 - \nu_1^2 = 1,7 - (0,2)^2 = 1,66,$$

$$\mu_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3 = 1,4 - 3 \cdot 0,2 \cdot 1,7 + 2(0,2)^3 = 0,40,$$

$$\mu_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4 = 7,7 - 4 \cdot 0,2 \cdot 1,4 + 6(0,2)^2 \cdot 1,7 - 3(0,2)^4 = 6,17.$$

Найдем асимметрию $A_s = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{(\sqrt{\mu_2})^3} = \frac{\mu_3}{\mu_2\sqrt{\mu_2}} = \frac{0,40}{1,66 \cdot 1,29} = 0,19$. Имеем правостороннюю асимметрию.

Найдем эксцесс $E_k = \frac{\mu_4}{\sigma^4} - 3 = \frac{\mu_4}{\mu_2^2} - 3 = \frac{6,17}{(1,66)^2} - 3 = 0,77.$ Такое значение указывает на более островершинное распределение, чем нормальное.

Задачи для самостоятельного решения

1. Измерение веса прямоугольных плиток дало следующие результаты: 35; 32; 33; 34; 31; 34; 30; 34; 32; 31; 31; 31; 35; 32; 31. Найти распределение относительных частот. Построить интервальный вариационный ряд частот.

2. Начертить полигон распределения 20 абитуриентов по числу баллов, полученных ими на приемных экзаменах 9, 10, 9, 9, 7, 8, 7, 7, 8, 9, 6, 6, 8, 8, 8, 7, 10, 6, 7, 8. Построить полигон относительных частот.

3. Опрос 20 сотрудников фирмы дал следующие результаты. Стаж работы около 5 лет имеют 2 сотрудника, стаж работы 7 лет у 3 сотрудников, стаж работы 10 лет имеют 8 сотрудников и стаж работы 15 лет у 7 сотрудников. Эти данные сведены в таблицу:

Стаж работы	5	7	10	15
Количество сотрудников	2	3	8	7

Построить эмпирическую функцию распределения по данному распределению выборки. Найти моду и медиану.

4. Дан интервальный вариационный ряд распределения 250 рабочих цеха по возрасту

Возрастная группа	16-18	18-20	20-22	22-24	24-26	26-28	28-30	30-32	32-34	34-36
Число рабочих	6	12	24	30	35	42	35	28	20	18

Построить гистограммы частот и относительных частот. Найти моду и медиану.

5. В таблице сгруппированы отчетные данные строительных организаций по фонду заработной платы за определенный отрезок времени.

Интервал (млн. руб.)	Частота (число организаций)
0,5-1,0	59
1,0-1,5	29
1,5-2,0	22
2,0-2,5	53

Построить гистограммы частот и относительных частот.

6. Найти моду и медиану оценок контрольной работы по высшей математике по данным следующей таблицы:

x_i	2	3	4	5
n_i	3	7	10	8

7. Дано распределение 446 ткачей по числу обслуживаемых ими станков:

Число станков	2	4	6	8	10	12
Число ткачей	2	64	154	128	78	20

Вычислить выборочную среднюю и выборочную дисперсию по данному распределению.

8. Анализируются объемы ежедневных продаж некоторого товара за 60 дней. Получены следующие результаты: 5, 6, 3, 1, 7, 7, 6, 6, 10, 11, 6, 4, 5, 6, 3, 12, 9, 10, 7, 4, 6, 7, 8, 8, 10, 5, 5, 4, 3, 6, 6, 7, 7, 8, 8, 10, 6, 4, 5, 6, 12, 7, 7, 8, 11, 9, 10, 5, 6, 4, 2, 7, 11, 8, 7, 9, 5,6, 9, 5, 6, 9, 5.

a. Постройте дискретный вариационный ряд, интервальный вариационный ряд.
b. Найдите эмпирическую функцию распределения выборки и постройте ее график.
c. Постройте гистограмму частостей.
d. Найдите выборочное среднее, выборочную дисперсию, выборочное среднеквадратическое отклонение, моду и медиану.

9. Количество дорожно-транспортных происшествий в регионе за 30 дней составило: 40, 21, 26, 29, 28, 28, 20, 31, 39, 20, 21, 39, 22, 22, 32, 36, 38, 29, 22, 34, 20, 30, 22, 28, 29, 29, 31, 37, 39, 35.

a. Постройте дискретный вариационный ряд, интервальный вариационный ряд.
b. Найдите эмпирическую функцию распределения выборки и постройте ее график.
c. Постройте гистограмму частостей.
d. Найдите выборочное среднее, выборочную дисперсию, выборочное среднеквадратическое отклонение, моду и медиану.

10. Данные о заработной плате двадцати сотрудников фирмы приведены в следующей таблице:

Заработная плата, руб.	2560	2600	2620	2650	2700
Количество сотрудников	2	3	10	4	1

Найдите выборочное среднее (среднюю заработную плату), выборочную дисперсию, выборочное среднеквадратическое отклонение, моду и медиану по заданному распределению выборки.

11. Опрос 10 студентов факультета философии и социальных наук БГУ показал:

Средний балл	4–6	6–8	8–10
Количество студентов	2	3	5

Найдите выборочное среднее, выборочную дисперсию, выборочное среднеквадратическое отклонение, моду и медиану по заданному распределению выборки.

12. Данные об урожайности ржи на различных участках колхозного поля в ц/га приведены в следующей таблице:

Урожайность ржи (ц/га)	9-12	12-15	15-18	18-21	21-24	24-27
Доля участка (в % к общей площади)	5	15	33	23	17	7

Вычислить среднюю урожайность, коэффициент вариации и размах вариации.

13. Анализируются объемы ежегодных продаж туристических услуг двадцатью турфирмами. Данные сведены в таблицу:

Ежегодные продажи, млн. руб.	1	5	7	9
Количество фирм	6	12	1	1

Найти выборочное среднее, выборочную дисперсию, выборочное среднеквадратическое отклонение, моду и медиану по данному распределению выборки.

14. Дано распределения товарооборота сельских магазинов одного из районов за год:

Товарооборот (тыс. руб.)	65-75	75-85	85-95	95-105	105-115	115-125	125-135
Число сельмагов	3	5	2	20	10	0	4

Найти выборочное среднее, выборочную дисперсию, выборочное среднеквадратическое отклонение, моду и медиану.

15. Анализируются объемы ежегодного товарооборота десятью крупнейшими магазинами города. Данные сведены в таблицу:

Объемы ежегодного товарооборота, млн. руб.	12,5	17,5	22,5	27,5	32,5
Количество магазинов	1	2	4	2	1

По данным выборки найти выборочное среднее, выборочную дисперсию, выборочное среднеквадратическое отклонение, моду и медиану.

Вопросы для самоконтроля

1. Что такое генеральная совокупность и выборка?
2. Назовите основные виды выборок и способы отбора элементов в них.
3. Что такое вариационный ряд?
4. Что такое дискретный вариационный и интервальный вариационный ряды?

5. Дать определение эмпирической функции распределения.
6. Что такое полигон частот и гистограмма? Для чего они используются?
7. Что такое выборочная средняя вариационного ряда? Основные свойства выборочной средней.
8. Что такое выборочная дисперсия вариационного ряда? Основные свойства выборочной дисперсии.
9. Начальные и центральные моменты вариационного ряда.
10. Что такое коэффициент асимметрии вариационного ряда? Что этот коэффициент показывает?
11. Что такое эксцесс вариационного ряда? Что этот коэффициент показывает?
12. Как находятся мода и медиана?

2. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ

2.1. Понятие оценки параметров. Состоятельность, несмещенность, эффективность оценок.

Сформулируем задачу оценки параметров в общем виде. Пусть распределение признака X генеральной совокупности задается функцией вероятностей $\varphi(x_i,\theta) = P(X = x_i)$ (для дискретной случайной величины X) или плотностью вероятности $\varphi(x,\theta) = p(x)$ (для непрерывной случайной величины X), которая содержит неизвестный параметр θ. Например, это параметр λ в распределении Пуассона или параметры a и σ^2 для нормального закона распределения и т.д.

Оценкой $\hat{\theta}$ параметра θ называют всякую функцию результатов наблюдений над случайной величиной X (иначе – статистику), с помощью которой судят о значении параметра θ:

$$\hat{\theta} = \hat{\theta}(X_1, X_2, ..., X_n).$$

Поскольку $X_1, X_2, \ldots, X_n$ – случайные величины, то и оценка $\hat{\theta}$ (в отличие от оцениваемого параметра θ – величины неслучайной, детерминированной) является случайной величиной, зависящей от закона распределения случайной величины X и числа n.

Всегда существует множество функций от результатов наблюдений $X_1, X_2, \ldots, X_n$ (от n «экземпляров» случайной величины X), которые можно предложить в качестве оценки параметра θ. Например, если параметр θ является математическим ожиданием случайной величины X, т.е. генеральной средней $\bar{x}_0$, то в качестве его оценки $\hat{\theta}$ по выборке можно взять: среднюю арифметическую результатов наблюдений – выборочную среднюю $\bar{x}$, моду M_o, медиану M_e и полусумму наименьшего и наибольшего значений по выборке, т.е. $(x_{\min} + x_{\max})/2$, и т.д. Какими свойствами должна обладать оценка $\hat{\theta}$, чтобы в каком-то смысле быть «доброкачественной» оценкой?

Назвать «наилучшей» оценкой такую, которая наиболее близка к истинному значению оцениваемого параметра, невозможно, так как выше отмечено, что $\hat{\theta}$ случайная величина, поэтому невозможно предсказать индивидуальное значение оценки в данном частном случае. Так что о качестве оценки следует судить не по индивидуальным ее значениям, а лишь по распределению ее значений в большой серии испытаний, т.е. по выборочному распределению оценки. Если значения оценки $\hat{\theta}$ концентрируются около истинного значения параметра θ, т.е. основная часть массы выборочного распределения оценки сосредоточена в малой окрестности оцениваемого параметра θ, то с большой вероятностью можно считать, что оценка $\hat{\theta}$ отличается от параметра θ лишь на малую величину. Поэтому, чтобы значение $\hat{\theta}$ было близко к θ, надо, очевидно, потребовать, чтобы рассеяние случайной величины $\hat{\theta}$ относительно θ, выражаемое, например, математическим ожиданием квадрата отклонения оценки от оцениваемого параметра $M(\hat{\theta} - \theta)^2$,

было по возможности меньшим. Таково основное условие, которому должна удовлетворять «наилучшая» оценка.

Рассмотрим наиболее **важные свойства оценок.**

Оценка $\hat{\theta}$ параметра θ называется ***несмещенной,*** если ее математическое ожидание равно оцениваемому параметру, т.е.

$$M(\hat{\theta}) = \theta .$$

В противном случае оценка называется ***смещенной.***

Если это равенство не выполняется, то оценка $\hat{\theta}$, полученная по разным выборкам, будет в среднем либо завышать значение θ (если $M(\hat{\theta}) > \theta$), либо занижать его (если $M(\hat{\theta}) < \theta$).

Таким образом, требование несмещенности гарантирует отсутствие систематических ошибок при оценивании.

Оценка $\hat{\theta}$ параметра θ называется ***состоятельной,*** если она сходится по вероятности к оцениваемому параметру, т.е.:

$$\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| < \varepsilon\right) = 1.$$

В случае использования состоятельных оценок оправдывается увеличение объема выборки, так как при этом становятся маловероятными значительные ошибки при оценивании. Поэтому практический смысл имеют только состоятельные оценки. Если оценка состоятельна, то практически достоверно, что при достаточно большом n $\hat{\theta} \approx \theta$.

Несмещенная оценка $\hat{\theta}$ параметра θ называется ***эффективной,*** если она имеет наименьшую дисперсию среди всех возможных несмещенных оценок параметра θ, вычисленных по выборкам одного и того же объема n.

В качестве статистических оценок параметров генеральной совокупности желательно использовать оценки, удовлетворяющие одновременно требованиям несмещенности, состоятельности и эффективности. Однако достичь этого удается не всегда. Может оказаться, что для простоты расчетов целесообразно использовать незначительно смещенные оценки или оценки, обладающие большей дисперсией по сравнению с эффективными оценками, и т.п.

2.2. Метод моментов

Согласно методу моментов, предложенному К. Пирсоном, определенное количество выборочных моментов (начальных ν_k или центральных μ_k, или тех и других) приравнивается к соответствующим теоретическим моментам распределения (ν_k или μ_k) случайной величины X.

Пример. Найти методом моментов оценку для параметра λ закона Пуассона.

Решение. В данном случае для нахождения единственного параметра λ достаточно приравнять теоретический ν_1 и эмпирический $\hat{\nu}_1$ начальные моменты первого порядка, ν_1 – математическое ожидание случайной величины X, распределенной по закону Пуассона, $M(X) = \lambda$. Момент ν_1 равен $\bar{x}$.

Следовательно, оценка методом моментов параметра λ закона Пуассона есть выборочная средняя $\bar{x}$.

Оценки методом моментов обычно состоятельны, однако по эффективности они не являются «наилучшими». Метод моментов часто используется на практике, так как он приводит к сравнительно простым вычислениям.

2.3 Метод максимального правдоподобия

Основным методом получения оценок параметров генеральной совокупности по данным выборки является метод максимального правдоподобия, предложенный Р. Фишером.

Основу метода составляет ***функция правдоподобия***, выражающая плотность распределения вероятностей совместного появления результатов выборки $x_1, x_2, \ldots, x_n$.

$$L(x_1, x_2, ..., x_n, \theta) = \phi(x_1, \theta)\phi(x_2, \theta)\ldots\phi(x_n, \theta).$$

Согласно методу максимального правдоподобия в качестве оценки неизвестного параметра θ принимается такое значение $\hat{\theta}$, которое максимизирует функцию L. Естественность подобного подхода к определению статистических оценок вытекает из смысла функции правдоподобия, которая при каждом фиксированном значении параметра θ является мерой правдоподобности получения наблюдений $x_1, x_2, \ldots, x_n$. И оценка $\hat{\theta}$ такова, что имеющиеся у нас наблюдения $x_1, x_2, \ldots, x_n$ являются наиболее правдоподобными.

Нахождение оценки $\hat{\theta}$ упрощается, если максимизировать не саму функцию L, а $\ln L$, поскольку максимум обеих функций достигается при одном и том же значении θ. Поэтому для отыскания оценки параметра θ (одного или нескольких) надо решить уравнение (систему уравнений) правдоподобия, получаемое приравниванием производной (частных производных) нулю по параметру (параметрам) θ:

$$\frac{d \ln L}{d\theta} = 0 \text{ или } \frac{1}{L} \cdot \frac{dL}{d\theta} = 0,$$

а затем отобрать то решение, которое доставляет функции $\ln L$ максимум.

Пример. Найти методом максимального правдоподобия оценку для вероятности p наступления некоторого события A по данному числу m появлений этого события в n независимых испытаниях.

Решение. Составим функцию правдоподобия:

$$L(x_1, x_2, ..., x_n; p) = \underbrace{p \cdot p \cdot \ldots \cdot p}_{m} \cdot \underbrace{(1-p) \cdot (1-p) \cdot \ldots \cdot (1-p)}_{n-m} \text{ или } L = p^m \cdot (1-p)^{n-m}.$$

Тогда $\ln L = m \ln p + (n-m)\ln(1-p)$ и $\frac{d \ln L}{dp} = \frac{m}{p} - \frac{n-m}{1-p}$, откуда $\hat{p} = \frac{m}{n}$.

2.4 Оценки параметров генеральной совокупности

Оценка генеральной средней. Пусть из генеральной совокупности объема N отобрана случайная выборка $X_1, X_2, \ldots, X_n$, где X_k – случайная величина, выражающая значение признака у k-го элемента выборки ($k = 1,2,\ldots,n$). Следует найти «наилучшую» оценку для генеральной средней.

Рассмотрим в качестве такой возможной оценки выборочную среднюю $\bar{x} = \frac{1}{n}\sum_{k=1}^{n} X_k$.

а) Выборка повторная.

Закон распределения для каждой случайной величины X_k $(k = 1,2,\ldots,n)$ имеет вид:

x_i	x_1	x_2	…	x_i	…	x_m
p_i	$\frac{N_1}{N}$	$\frac{N_2}{N}$	…	$\frac{N_i}{N}$	…	$\frac{N_m}{N}$

Действительно, вероятность того, что 1-й отобранный в выборку элемент имеет значение признака x_1, согласно классическому определению вероятности равна $P(X_1 = x_1) = \frac{N_1}{N}$, так как из общего числа N элементов генеральной совокупности N_1 элементов имеют значение признака x_1. Так как выборка повторная и каждый отобранный и обследованный элемент возвращается в исходную совокупность, восстанавливая всякий раз ее первоначальные состав и объем, то вероятность $p_1 = P(X_k = x_1) = \frac{N_1}{N}$ для любого элемента выборки, т.е. для $k = 1,2,\ldots,n$. Аналогично можно определить $P(X_k = x_i) = \frac{N_i}{N}$ для $k = 1,2,\ldots,n$; $i = 1,2,\ldots,m$ и убедиться в том, что закон распределения каждой случайной величины X_k один и тот же**.**

Случайные величины $X_1, X_2, \ldots, X_n$ независимы, так как независимы любые события $\{X_k = x_i\}$ $(k = 1,2,\ldots,n; i = 1,2,\ldots,m)$ и их комбинации. Например, независимы события $\{X_2 = x_1\}$ и $\{X_1 = x_1\}$, ибо $P(X_2 = x_1 \mid X_1 = x_1) = P(X_2 = x_1 \mid X_1 \neq x_1) = P(X_2 = x_1) = \frac{N_1}{N}$, т.е. вероятность того, что значение признака у 2-го отобранного в выборку элемента равно x_1, не меняется в зависимости от того, какое значение признака у 1-го элемента, и т.д.

Найдем числовые характеристики случайной величины X_k:

$$M(X_k) = \sum_{i=1}^{m} x_i p_i = \frac{1}{N}\sum_{i=1}^{m} x_i N_i = \bar{x}_0, \quad D(X_k) = \sum_{i=1}^{m} (x_i - \bar{x}_0)^2 p_i = \frac{1}{N}\sum_{i=1}^{m} (x_i - \bar{x}_0)^2 N_i = \sigma^2,$$

т.е. математическое ожидание и дисперсия каждой случайной величины X_k – это соответственно генеральная средняя и генеральная дисперсия.

Теорема. *Выборочная средняя* $\bar{x}$ *есть несмещенная и состоятельная оценка генеральной средней* $\bar{x}_0$*, причем* $\sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n}$**.**

б) Выборка бесповторная.

В этом случае случайные величины X_k $(k = 1,2,\ldots,n)$ будут зависимыми. Рассмотрим, например, события $\{X_1 = x_1\}$ и $\{X_2 = x_1\}$. Теперь вероятности $P(X_2 = x_1 | X_1 = x_1) = \dfrac{N_1 - 1}{N - 1}$, $P(X_2 = x_1 | X_1 \neq x_1) = \dfrac{N_1}{N - 1}$, так как отобранный элемент (в случае бесповторной выборки) в исходную совокупность не возвращается, то в ней остается всего N–1 элемент, которых со значением признака $N_1 - 1$ или N_1 в зависимости от того, принимает ли первый отобранный элемент значение признака x_1 .

Однако и для бесповторной выборки выборочная средняя является «хорошей» оценкой. Об этом свидетельствует следующая теорема, которую приводим без доказательства.

Теорема. *Выборочная средняя* $\bar{x}$ *бесповторной выборки есть несмещенная и состоятельная оценка генеральной средней* $\bar{x}_0$*, причем* $\sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n} \cdot \dfrac{N - n}{N - 1} \approx \dfrac{\sigma^2}{n} \cdot \left(1 - \dfrac{n}{N}\right)$**.**

Пример. Найти несмещенную и состоятельную оценку средней выработки рабочих цеха по данным выборки, представленной в табл. 3.

Решение. Несмещенная и состоятельная оценка генеральной средней $\bar{x}_0$ есть выборочная средняя $\bar{x}$, т.е. $\bar{x}$ =119,2(%).

Оценка генеральной дисперсии. На первый взгляд, наиболее подходящей оценкой для генеральной дисперсии σ^2 является выборочная дисперсия s^2. Следующая теорема свидетельствует о том, что s^2 не является «наилучшей» оценкой.

Теорема. Выборочная дисперсия s^2 есть смещенная и состоятельная оценка генеральной дисперсии σ^2.

Принимая без доказательства состоятельность оценки s^2, докажем, что она – смещенная оценка. Имеем $s^2 = \overline{x^2} - \bar{x}^2$.

Если все значения признака уменьшить на одно и то же число c, то средняя уменьшится на это число, т.е. $\overline{x - c} = \bar{x} - c$, а дисперсия не изменится:

$$s^2 = s_x^2 = s_{x-c}^2 = \overline{(x-c)^2} - \overline{(x-c)}^2 = \overline{(x-c)^2} - (\bar{x} - c)^2.$$

Полагая $c = \bar{x}_0$, получим $s^2 = \overline{(x - \bar{x}_0)^2} - (\bar{x} - \bar{x}_0)^2$.

а) Выборка повторная.

Для повторной выборки выборочные значения рассматриваем как *независимые* случайные величины $X_1, X_2, \ldots, X_n$, каждая из которых имеет один и тот же закон распределения с числовыми характеристиками $M(X_k) = \bar{x}_0$, $D(X_k) = \sigma^2$, $k = 1,2,\ldots,n$.

Найдем математическое ожидание оценки s^2:

$$M(s^2) = M\left(\frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{x}_0)^2\right) - M(\bar{x} - \bar{x}_0)^2.$$

Первый член в правой части

$$M\left(\frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{x}_0)^2\right) = \frac{1}{n}\sum_{k=1}^{n}M(X_k - \bar{x}_0)^2 = \frac{1}{n}\sum_{k=1}^{n}D(X_k) = \frac{1}{n}\sum_{k=1}^{n}\sigma^2 = \sigma^2.$$

Второй член с учетом того, что $\bar{x}$ есть несмещенная оценка $\bar{x}_0$, т.е. $M(\bar{x}) = \bar{x}_0$,

$$M(\bar{x} - \bar{x}_0)^2 = D(\bar{x}) = \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}.$$

Поэтому

$$M(s^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

б) Выборка бесповторная.

Как уже было рассмотрено выше, для бесповторной выборки $X_1, X_2, ...X_n$ – зависимые случайные величины. Можно показать, что

$$M\left(s^2\right) = \frac{n-1}{n}\cdot\frac{N}{N-1}\sigma^2 \approx \frac{n-1}{n}\sigma^2$$

(так как объем генеральной совокупности *N*, как правило, большой и $N \approx N-1$).

Итак, и для повторной выборки, и для бесповторной

$$M\left(s^2\right) = \frac{n-1}{n}\sigma^2,$$

т.е. s^2 – смещенная оценка.

Так как $\frac{n-1}{n} < 1$ и $M\left(s^2\right) < \sigma^2$, то выборочная дисперсия занижает генеральную дисперсию. Поэтому, заменяя σ^2 на s^2, мы допускаем систематическую погрешность в меньшую сторону. Чтобы ее ликвидировать, достаточно ввести поправку, умножив s^2 на $\frac{n-1}{n}$. Тогда получим «исправленную» выборочную дисперсию

$$\bar{s}^2 = \frac{n-1}{n}s^2 = \frac{1}{n}\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^2 n_i.$$

Очевидно, что

$$M\left(\bar{s}^2\right) = M\left(\frac{n-1}{n}s^2\right) = \frac{n}{n-1}M\left(s^2\right) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2,$$

т.е. $\bar{s}^2$ является несмещенной и состоятельной оценкой генеральной дисперсии σ^2.

Примет. Найти несмещенную и состоятельную оценку дисперсии случайной величины *X* – выработки рабочих цеха по данным выборки, представленной в таблице 3.

Решение. Несмещенной и состоятельной оценкой дисперсии случайной величины (генеральной дисперсии) σ^2 является «исправленная» выборочная дисперсия $\bar{s}^2$. Ранее была вычислена выборочная дисперсия $s^2 = 87,48$. При $n = 100$ имеем

$$\bar{s}^2 = \frac{100}{99} \cdot 87,48 = 88,36.$$

Разница между s^2 и $\bar{s}^2$ заметна при небольшом числе наблюдений. При $n > 30$ $\bar{s}^2 \approx s^2$, т.е. в качестве оценки для σ^2 вполне можно использовать выборочную дисперсию s^2.

2.5. Интервальные оценки. Доверительная вероятность и доверительный интервал. Надежность и точность оценки.

Вычисленная на основе выборки оценка $\hat{\theta}$ является лишь приближением к неизвестному значению параметра θ даже в том случае, когда эта оценка состоятельная, несмещенная и эффективная. Чтобы получить представление о точности и надежности оценки $\hat{\theta}$ параметра θ, используют интервальную оценку параметра.

Интервальной оценкой параметра θ называется числовой интервал (θ_1, θ_2), который с заданной вероятностью γ накрывает неизвестное значение параметра θ: $P(\theta_1 < \theta < \theta_2) = \gamma$. Такой интервал (θ_1, θ_2) называется ***доверительным***, а вероятность γ – ***доверительной вероятностью*** или ***надежностью оценки***. Граничные точки доверительного интервала называются соответственно ***нижним*** (θ_1) и ***верхним*** (θ_2) ***доверительными пределами***.

Заданному γ соответствует не единственный доверительный интервал. Доверительные интервалы могут изменяться от выборки к выборке. Более того, для данной выборки различные методы построения доверительных интервалов могут привести к различным интервалам. Величина доверительного интервала существенно зависит от объема выборки n (уменьшается с ростом n) и от значения доверительной вероятности γ (увеличивается с приближением γ к единице). Очень часто (но не всегда) доверительный интервал выбирается симметричным относительно параметра θ, т.е. $(\hat{\theta} - \Delta, \hat{\theta} + \Delta)$, где $\hat{\theta}$ – оценка параметра θ. При этом Δ называют ***ошибкой оценки*** $\hat{\theta}$, а наибольшее значение Δ – ***предельной ошибкой***.

2.6. Доверительный интервал для математического ожидания при известной генеральной дисперсии в случае нормального распределения.

Пусть $x_1, x_2, \ldots, x_n$ – выборка из нормальной совокупности с математическим ожиданием a и σ, причем значение параметра a неизвестно, а σ^2 известно. В качестве оценки математического ожидания a возьмем выборочное среднее:

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

Величина $\bar{x}$ имеет нормальное распределение с математическим ожиданием a и дисперсией $\sigma_{\bar{x}}^2 = \sigma^2 / n$. Поэтому вероятность того, что разность $\bar{x} - a$ не превзойдет по абсолютной величине наперед заданного числа Δ, равна

$$P\left(|\bar{x} - a| < \Delta\right) = 2\Phi\left(\frac{\Delta}{\sigma_{\bar{x}}}\right) = 2\Phi\left(\frac{\Delta\sqrt{n}}{\sigma}\right), \text{ где } \Phi(x) - \text{функция Лапласа.}$$

Зададим доверительную вероятность γ (в качестве доверительной вероятности, иначе уровня доверия, обычно полагают $\gamma = 0{,}95;\ 0{,}99$). Пусть t_γ – корень уравнения $2\Phi(t) = \gamma$, который можно найти по таблице функции Лапласа. Тогда

$$\frac{\Delta\sqrt{n}}{\sigma} = t_\gamma \text{ или } \Delta = \frac{t_\gamma \sigma}{\sqrt{n}} \,.$$

Таким образом, практически достоверно (точнее, с вероятностью γ), что

$$|\bar{x} - a| < \Delta = \frac{t_\gamma \sigma}{\sqrt{n}} \,.$$

Последнее неравенство запишем в виде:

$$\bar{x} - t_\gamma \frac{\sigma}{\sqrt{n}} < a < \bar{x} + t_\gamma \frac{\sigma}{\sqrt{n}} \text{ или } \bar{x} - t_\gamma \sigma_{\bar{x}} < a < \bar{x} + t_\gamma \sigma_{\bar{x}}$$

– это и есть искомый доверительный интервал.

Выборочная средняя $\bar{x}$ и выборочная доля w выборки представляют собой сумму независимых случайных величин, имеющих один и тот же закон распределения с конечным математическим ожиданием и дисперсией. При $n \to \infty$ распределения $\bar{x}$ и w неограниченно приближаются к нормальным (практически при $n > 30$ распределения $\bar{x}$ и w можно считать приближенно нормальными).

Среднее квадратическое отклонение выборочной средней $\sigma_{\bar{x}}$ и выборочной доли σ_w называется ***средней квадратической (стандартной) ошибкой*** выборки (для бесповторной выборки обозначаем соответственно $\sigma'_{\bar{x}}$ и σ'_w).

Итак, при заданной доверительной вероятности γ предельная ошибка выборки равна t-кратной величине средней квадратической ошибки, где $2\Phi(t) = \Phi^*(t) = \gamma$, т.е.

$$\Delta = t\sigma_{\bar{x}} \qquad (\Delta = t\sigma_w).$$

Интервальные оценки (доверительные интервалы) для генеральной средней $\bar{x}_0$ и генеральной доли p находятся по формулам:

$$\bar{x} - \Delta \le x_0 \le \bar{x} + \Delta \qquad (w - \Delta \le p \le w + \Delta).$$

Напомним, что для

- повторной выборки $\sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n}$, $\sigma_w^2 = \dfrac{pq}{n} \approx \dfrac{w(1-w)}{n}$;
- бесповторной выборки $\sigma_{\bar{x}}'^2 \approx \dfrac{\sigma^2}{n}\left(1 - \dfrac{n}{N}\right)$, $\sigma_w'^2 \approx \dfrac{pq}{n}\left(1 - \dfrac{n}{N}\right) \approx \dfrac{w(1-w)}{n}\left(1 - \dfrac{n}{N}\right)$.

Замечание. *При определении средней квадратической ошибки выборки для доли, если даже* w *неизвестна, в качестве* pq *можно взять его максимально возможное значение* $(pq)_{\max} = 0{,}25$ *(так как* $pq = p(1-p) = -(p^2 - p) = 0{,}25 - (p - 0{,}5)^2$, *то* pq *максимально при* $p=0{,}5$*).*

Пример. При обследовании выборки 1000 рабочих цеха в отчетном году по сравнению с предыдущим было отобрано 100 рабочих. Получены следующие данные (см. первые две графы табл. 3). Необходимо определить: а) вероятность того, что средняя выработка рабочих цеха отличается от средней выборочной не более, чем на 1% (по абсолютной величине); б) границы, в которых с вероятностью 0,9545 заключена средняя выработка рабочих цеха. Рассмотреть случаи повторной и бесповторной выборки.

Решение. а) Имеем $N = 1000$, $n = 100$. Ранее были вычислены $\bar{x} = 119{,}2\ (\%)$, $s^2 = 87{,}48$.

а) Найдем среднюю квадратическую ошибку выборки для средней:

1) для повторной выборки: $\sigma_{\bar{x}} = \sqrt{\dfrac{87{,}48}{100}} = 0{,}935\ (\%)$;

2) для бесповторной выборки: $\sigma'_{\bar{x}} = \sqrt{\dfrac{87{,}48}{100}\left(1 - \dfrac{100}{1000}\right)} = 0{,}887\ (\%)$.

Теперь находим искомую доверительную вероятность:

$$1)\ P(|\bar{x} - x_0| \le 1) = 2\Phi\left(\frac{1}{0{,}935}\right) = 2\Phi(1{,}07) = 2 \cdot 0{,}3577 = 0{,}7154.$$

$$2)\ P(|\bar{x} - x_0| \le 1) = 2\Phi\left(\frac{1}{0{,}887}\right) = 2\Phi(1{,}13) = 2 \cdot 0{,}3708 = 0{,}7416.$$

Итак, вероятность того, что выборочная средняя отличается от генеральной средней не более чем на 1% (по абсолютной величине), равна 0,715 для повторной и 0,741 для бесповторной выборки.

б) Находим предельные ошибки повторной и бесповторной выборок при $t = 2{,}00$:

$$1)\ \Delta = 2{,}00 \cdot 0{,}935 = 1{,}870\ (\%), \qquad 2)\ \Delta' = 2{,}00 \cdot 0{,}887 = 1{,}774\ (\%).$$

Теперь определяем искомый доверительный интервал:

1) $119{,}2 - 1{,}870 \le x_0 \le 119{,}2 + 1{,}870$ или $117{,}33 \le x_0 \le 121{,}07$;

2) $119{,}2 - 1{,}774 \le x_0 \le 119{,}2 + 1{,}774$ или $117{,}43 \le x_0 \le 120{,}97$.

Таким образом, с надежностью 0,9545 средняя выработка рабочих цеха заключена в границах от 117,43 до 121,07%, если выборка повторная, и от 117,43 до 120,97%, если выборка бесповторная.

2.7. Распределения χ^2 (хи-квадрат), Стьюдента, Фишера-Снедекора.

В статистических исследованиях применяются следующие распределения, основанные на нормальном распределении.

Распределением χ^2 *(хи-квадрат)* ***(Пирсона) с*** k ***степенями свободы*** называется распределение суммы квадратов k независимых случайных величин, распределенных по стандартному нормальному закону, т.е.

$$\chi^2 = \sum_{i=1}^{k} Z_i^2 ,$$

где $Z_i \in N(0;1)$, $i = 1,2,...,k$.

Распределением Стьюдента (или t***-распределением***) называется распределение случайной величины

$$t = \frac{Z}{\sqrt{\frac{1}{k}\chi^2}},$$

где $Z \in N(0;1)$, χ^2 – независимая от Z случайная величина, имеющая χ^2-распределение с k степенями свободы. Практически при $k > 30$ t-распределение можно считать приближенно нормальным.

Распределением Фишера-Снедекора (или F ***-распределением***) называется распределение случайной величины

$$F = \frac{\frac{1}{k_1}\chi^2(k_1)}{\frac{1}{k_2}\chi^2(k_2)},$$

где $\chi_1^2(k_1)$ и $\chi_2^2(k_2)$ – случайные величины, имеющие распределение соответственно с k_1 и k_2 степенями свободы.

2.8. Построение доверительного интервала для генерального среднего при неизвестной генеральной дисперсии (случай нормального распределения).

Стандартное отклонение выборочной средней $Z = \frac{\bar{x} - \bar{x}_0}{\sigma_{\bar{x}}} = \frac{\bar{x} - \bar{x}_0}{\sigma / \sqrt{n}}$ имеет стандартное нормальное распределение $N(0; 1)$, т.е. нормальное распределение с нулевым математическим ожиданием и единичной дисперсией. Действительно, сумма нормально распределенных величин есть нормально распределенная величина и ее математическое ожидание и дисперсия:

$$M(Z) = M\left(\frac{\bar{x} - \bar{x}_0}{\sigma}\sqrt{n}\right) = \frac{\sqrt{n}}{\sigma}\left[M(\bar{x}) - M(\bar{x}_0)\right] = \frac{\sqrt{n}}{\sigma}(\bar{x}_0 - \bar{x}_0) = 0,$$

$$\sigma_z^2 = D(Z) = D\left(\frac{\bar{x} - \bar{x}_0}{\sigma}\sqrt{n}\right) = \left(\frac{\sqrt{n}}{\sigma}\right)^2\left[D(\bar{x}) - D(\bar{x}_0)\right] = \frac{n}{\sigma^2}\left(\frac{\sigma^2}{n} - 0\right) = 1.$$

Однако на практике почти всегда генеральная дисперсия σ^2 (как и оцениваемая генеральная средняя $\bar{x}_0$) неизвестна. Если заменить σ^2 ее «наилучшей» оценкой по выборке, а именно «исправленной» выборочной дисперсией $\hat{s}^2$, то большой интерес представляет распределение выборочной характеристики (статистики) $t = \frac{\bar{x} - \bar{x}_0}{\hat{s} / \sqrt{n}}$ или, что то же с учетом $\hat{s}^2 = \frac{n}{n-1}s^2$, распределение статистики $t = \frac{\bar{x} - \bar{x}_0}{s / \sqrt{n-1}}$.

Представим статистику t в виде:

$$t = \frac{(\bar{x} - \bar{x}_0) / \dfrac{\sigma}{\sqrt{n}}}{\sqrt{\dfrac{n}{n-1} \cdot \dfrac{s^2}{\sigma^2}}}.$$

Числитель этого выражения, как показано выше, имеет стандартное нормальное распределение $N(0;\ 1)$. Случайная величина ns^2/σ^2 имеет χ^2-распределение с $k = n-1$ степенями свободы. Следовательно, статистика t имеет распределение Стьюдента с $k = n-1$ степенями свободы. Указанное распределение не зависит от неизвестных параметров распределения случайной величины X, а зависит лишь от числа степеней свободы k.

Число степеней свободы k определяется как разность между общим числом n наблюдений (вариант) случайной величины X и числом уравнений, связывающих эти наблюдения. Так, например, для распределения статистики t число степеней свободы $k = n-1$, ибо одна степень свободы «теряется» при определении выборочной средней $\bar{x}$ (n наблюдений связаны одним уравнением $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$).

При $n \le 30$ для заданной доверительной вероятности γ из таблицы распределения Стьюдента имеем

$$P(|t| < t_{\gamma,k}) = \int_{-t_{\gamma,k}}^{t_{\gamma,k}} p_k(y)dy = \gamma.$$

Таким образом, с доверительной вероятностью γ выполняется неравенство

$$|t| < t_{\gamma,k} \text{ или } \left|\frac{\bar{x} - \bar{x}_0}{s/\sqrt{n-1}}\right| < t_{\gamma,k}.$$

Преобразуя последнее неравенство, получаем

$$\bar{x} - t_{\gamma,k}\frac{s}{\sqrt{n-1}} < x_0 < \bar{x} + t_{\gamma,k}\frac{s}{\sqrt{n-1}}$$

– искомый доверительный интервал для генерального среднего $\bar{x}_0$.

При $n > 30$ статистика t имеет распределение, близкое к $N(0;1)$, поэтому с вероятностью $\approx \gamma$

$$\bar{x} - t_\gamma\frac{s}{\sqrt{n-1}} < \bar{x}_0 < \bar{x} + t_\gamma\frac{s}{\sqrt{n-1}}, \text{ где } \Phi(t_\gamma) = \gamma/2.$$

Пример. Для контроля срока службы электроламп из большой партии было отобрано 17 электроламп. В результате испытаний оказалось, что средний срок службы отобранных ламп равен 980 ч, а среднее квадратическое отклонение их срока службы – 18 ч. Необходимо определить границы, в которых с вероятностью 0,95 заключен средний срок службы ламп во всей партии.

Решение. Имеем по условию $n = 17$, $\bar{x} = 980$ (ч), $s = 18$ (ч). Учитывая, что $\gamma = 0{,}95$ и (приложение 3) $t_{0{,}95;16}$=2,12, находим предельную ошибку выборки $\Delta = \frac{2{,}12 \cdot 18}{\sqrt{16}} = 9{,}5$ (ч). Теперь искомый доверительный интервал

$980-9{,}5<x_0<980+9{,}5$ или $970{,}5<x_0<989{,}5$, т.е. с надежностью 0,95 средний срок службы электроламп в партии заключен от 970,5 до 989,5 ч.

Замечание. *Каков бы ни был закон распределения независимых одинаково распределенных случайных величин* $\xi_1,\xi_2,\ldots,\xi_n$, *имеющих конечную дисперсию, их сумма распределена приближенно нормально при достаточно больших* n *(согласно центральной предельной теореме). Поэтому при достаточно больших* n *и известной дисперсии* σ^2 *имеем*

$$P\left(\bar{x}-t_\gamma\frac{\sigma}{\sqrt{n}}<\bar{x}_0<\bar{x}+t_\gamma\frac{\sigma}{\sqrt{n}}\right)\approx\gamma\,.$$

Если же σ *неизвестно, то можно использовать ее оценку:* $\sigma\approx\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}=\hat{s}$. *В этом случае интервал*

$$\left(\bar{x}-t_\gamma\frac{\hat{s}}{\sqrt{n}};\bar{x}+t_\gamma\frac{\hat{s}}{\sqrt{n}}\right)$$

является доверительным интервалом для $\bar{x}_0$ *с доверительной вероятностью, близкой к* γ.

2.9. Построение доверительного интервала для среднего квадратического отклонения (случай нормального распределения)

Рассмотрим независимые центрированные случайные величины $X_1,X_2,\ldots,X_n$, распределенные нормально с нулевым математическим ожиданием и единичной дисперсией. Тогда случайной величины

$$\xi=\sum_{i=1}^{n}X_i^2\,,$$

имеет χ^2-распределение (распределением Пирсона) и $k=n$ степенями свободы. Очевидно, что $\xi\geq 0$ и $P(\xi<0)=0$.

χ^2-распределение зависит только от одного параметра – числа степеней свободы k. И чем больше k, тем более симметрично распределение χ^2, хотя некоторая правая асимметрия существует всегда.

Это распределение непосредственно связано с распределением статистической дисперсии s^2 выборки. Справедливо следующее утверждение: если $X_1,X_2,\ldots,X_n$ – независимые нормально распределенные случайные величины с одинаковыми математическими ожиданиями и дисперсиями, то случайная величина $V^2=ns^2/\sigma^2$ имеет χ^2-распределение с $k=n-1$ степенями свободы.

Зададим доверительную вероятность γ и определим числа α и β ($\alpha>0$, $\beta>0$) из условия $\int_{\alpha}^{\beta}p(t)dt=\gamma$, где $p(t)$ – плотность распределения вероятности закона χ^2 с $k=n-1$ степенями свободы. Очевидно, числа α и β, удовлетворяющие данному условию, можно выбрать бесчисленным

множеством способов. Потребуем дополнительно, чтобы $\int_0^{\alpha} p(t)dt = \frac{1-\gamma}{2}$, тогда $\int_{\beta}^{\infty} p(t)dt = \frac{1-\gamma}{2}$, и числа α и β однозначно определены. Их значения находятся из таблицы χ^2-распределения с k степенями свободы: по $k = n-1$ $(n \le 30)$ в соответствующей строке таблицы (см. прилож. 4) выбираем два числа α и β, одно из которых отвечает вероятности $p_1 = (1-\gamma)/2$, другое – вероятности $p_2 = (1+\gamma)/2$.

Итак, с вероятностью γ выполнены неравенства

$$\alpha < \frac{ns^2}{\sigma^2} < \beta, \text{ откуда } \frac{ns^2}{\beta} < \sigma^2 < \frac{ns^2}{\alpha} \text{ или } \frac{\sqrt{n}s}{\sqrt{\beta}} < \sigma < \frac{\sqrt{n}s}{\sqrt{\alpha}}$$

– это и есть доверительные интервалы для дисперсии и среднего квадратического отклонения, соответствующие доверительной вероятности γ в случае нормально распределенной генеральной совокупности.

Пример. На основании выборочных наблюдений производительности труда 20 работниц было установлено, что среднее квадратическое отклонение суточной выработки составляет 15 м ткани в час. Предполагая, что производительность труда работницы имеет нормальное распределение, найти границы, в которых с надежностью 0,9 заключены генеральные дисперсия и среднее квадратическое отклонение суточной выработки работниц.

Решение. Имеем $\gamma = 0{,}9$; $(1-\gamma)/2 = 0{,}05$, $(1+\gamma)/2 = 0{,}95$. При числе степеней свободы $k = n-1 = 20-1 = 19$ определим $\alpha = \chi_1^2$ и $\beta = \chi_2^2$ по таблице приложения 4 для вероятностей 0,95 и 0,05, т.е. $\chi_1^2 = 10{,}1$ и $\chi_2^2 = 30{,}1$. Тогда доверительный интервал для σ^2 можно записать в виде: $\frac{20}{30{,}1} \cdot 15^2 < \sigma^2 < \frac{20}{10{,}1} \cdot 15^2$ или $149{,}5 < \sigma^2 < 445{,}6$ и для σ: $\sqrt{149{,}5} < \sigma < \sqrt{445{,}6}$ или $12{,}2 < \sigma < 21{,}1$ (м/ч). Итак, с надежностью 0,9 дисперсия суточной выработки работниц заключена в границах от 149,5 до 445,6, а ее среднее квадратическое отклонение – от 12,2 до 21,1 метров ткани в час.

2.10. Лабораторная работа 2. Анализ выборок в Microsoft Excel

Цель: *научиться проводить совместный анализ нескольких выборок, познакомиться с основными статистическими критериями и их реализацией в Excel.*

Критерий Стьюдента (t) наиболее часто используется для проверки гипотезы: «Средние двух выборок относятся к одной и той же совокупности». Критерий позволяет найти вероятность того, что оба средних относятся к одной и той же совокупности. Если эта вероятность *p* ниже уровня значимости (*p<0.05)*, то принято считать, что выборки относятся к двум разным совокупностям.

Для оценки достоверности отличий по критерию Стьюдента принимается нулевая гипотеза, что средние выборок равны между собой. Затем вычисляется

значение вероятности того, что изучаемые события произошли случайным образом.

В MS Excel для оценки достоверности отличий по критерию Стьюдента используется функция ТТЕСТ.

Синтаксис. **ТТЕСТ (массив1; массив2; хвосты; тип)**

Массив1 – первое множество данных.

Массив2 – второе множество данных.

Хвосты – число хвостов распределения. Если хвосты = 1, то функция ТТЕСТ использует одностороннее распределение. Если хвосты = 2, то функция ТТЕСТ использует двустороннее распределение.

Тип – вид исполняемого t – теста. Если тип =1, то выполняется парный t – тест. Если тип =2, то выполняется двухвыборочный t – тест с равными дисперсиями. Если тип =3, то выполняется двухвыборочный t – тест с неравными дисперсиями.

Задание 1. Выяснить, достоверны ли отличия при сравнении данных реализации туристической фирмой путевок за периоды до и после начала рекламной компании (см. лабораторная работа 1, задание 2).

Решение.

1. Перейдите на новый лист рабочей книги.
2. Выделите данные в диапазоне A1:C14 на *листе 2* и скопируйте их на *лист 5*.
3. Для выявления достоверности отличий табличный курсор установите в свободную ячейку (A15). На панели инструментов необходимо нажать кнопку *Вставка функции*. В появившемся диалоговом окне *Мастер функций* выберите категорию *Статистические* и функцию ТТЕСТ, после чего нажмите кнопку ОК. Указателем мыши введите диапазон данных контрольной группы в поле *Массив1* (*B3:B14*). В поле *Массив2* введите диапазон данных исследуемой группы (*C3:C14*). В поле Х*восты* введите с клавиатуры цифру 2, а в поле *Тип* введите цифру 3. Нажмите ОК. В ячейке *A15* появится значение вероятности – 0,057893.
4. Поскольку величина вероятности случайного появления анализируемых выборок (0,057893) больше уровня значимости ($\alpha = 0{,}05$), то нулевая гипотеза принимается. Следовательно, различия между выборками могут быть случайными и средние выборок не считаются достоверно отличающимися друг от друга.
5. Сохраните документ.

Критерий согласия χ^2 применяется в том случае, если вам необходимо сравнить две относительные или выраженные в процентах величины (доли). Здесь, как и в случае с критерием Стьюдента, принимается нулевая гипотеза о том, что выборки принадлежат к одной генеральной совокупности. Кроме того, определяется ожидаемое значение результата. Затем оценивается вероятность того, что ожидаемые значения и наблюдаемые принадлежат к одной генеральной совокупности.

В MS Excel критерий χ^2 реализован в функции ХИ2ТЕСТ. Функции ХИ2ТЕСТ вычисляет вероятность совпадения наблюдаемых (фактических)

значений и теоретических (гипотетических) значений. Если вычисленная вероятность ниже уровня значимости (0,05), то нулевая гипотеза отвергается и утверждается, что наблюдаемые значения не соответствуют теоретическим (ожидаемым) значениям.

Синтаксис. **ХИ2ТЕСТ (фактический_интервал; ожидаемый_интервал).**

Фактический_интервал – интервал данных, которые содержат наблюдения, подлежащие сравнению с ожидаемыми значениями;

Ожидаемый_интервал – интервал данных, который содержит теоретические (ожидаемые) значения для соответствующих наблюдаемых значений.

Задание 2. В ходе социологического опроса на вопрос о перенесенном в детстве заболевании ответы распределились следующим образом:

	Да	Нет	Не помню
Мужчины	58	11	10
Женщины	35	25	23

Есть ли достоверные отличия в ответах женщин и мужчин?

Решение.

1. Перейдите на шестой лист рабочей книги и создайте следующую таблицу:

	A	B	C
1	**Фактические значения**		
2	**Ответ**	**Мужчины**	**Женщины**
3	**Да**	58	35
4	**Нет**	11	25
5	**Не помню**	10	23

2. Вычисляем объемы выборок. Для этого поставьте курсор в ячейку *B6* и на панели инструментов нажимаем кнопку  (*Автосумма*).

Аналогично вычислите объем второй выборки (ячейка *C*6).

3. Вычислите P – долю признака. Для этого в ячейку D2 введите P – доля, а в ячейку D3 введите =(B3+C3)/(B6+C6). Затем скопируйте полученную формулу в ячейки D4:D5.

4. Создайте следующую таблицу для ожидаемых значений:

	A	B	C
7			
8	**Ожидаемые значения**		
9	**Ответ**	**Мужчины**	**Женщины**
10	**Да**		
11	**Нет**		
12	**Не помню**		

5. Вычислите ожидаемые (теоретические) значения. Для этого в ячейку B10 введите формулу =B$6*$D3. Скопируйте (растяните) эту формулу в ячейки B10:C12.

6. Вычислите значение вероятности того, что нет достоверных отличий в ответах женщин и мужчин. Для этого установите курсор в свободную ячейку (A14). На панели инструментов нажмите кнопку Вставка функции и в категории Статистические выберите функцию ХИ2ТЕСТ. В поле фактический

интервал введите диапазон B3:C5, а в поле ожидаемый интервал введите диапазон B10:C12. Нажмите ОК. В ячейке A14 появится значение вероятности.

7. Проанализируйте полученный результат (сделайте вывод о том, есть ли достоверные отличия в ответах женщин и мужчин).
8. Сохраните документ.

Задание 3. В двух группах измерялось время сложной сенсомоторной реакции. В экспериментальную группу входили 15 спортсменов, а в контрольную группу 12 обычных человек. Проверить есть ли достоверные отличия между средней скоростью реакции спортсменов и людей, не занимающихся профессионально спортом.

№	Скорость реакции	
	спортсмены	обычные люди
1	504	580
2	560	692
3	420	700
4	600	621
5	580	640
6	530	561
7	490	680
8	580	630
9	470	708
10	550	652
11	460	500
12	540	515
13	544	---
14	485	---
15	500	---

Решение.

Для решения задач такого типа используется *t*-критерий Стьюдента.

1. Перейдите на новый лист рабочей книги и создайте следующую таблицу:

	A	B	C
1	**№**	**Скорость реакции**	
2	**испытуемого**	**спортсмены**	**обычные люди**
3	1	504	580
4	2	560	692
5	3	420	700
6	4	600	621
7	5	580	640
8	6	530	561
9	7	490	680
10	8	580	630
11	9	470	708
12	10	550	652
13	11	460	500
14	12	540	515
15	13	544	
16	14	485	
17	15	500	

2. Для реализации t-критерий Стьюдента воспользуемся пакетом анализа. Для этого в пункте меню *Сервис* выберите строку *Анализ данных* и далее выберите *Двухвыборочный t-тест с различными дисперсиями.*

3. В появившемся диалоговом окне задайте *интервал переменной 1* диапазон ячеек B3:B17. Аналогично укажите *интервал переменной 2* диапазон C3:C14.

4. Далее укажите выходной диапазон. Для этого поставьте переключатель в положение *Выходной диапазон*, затем щелкните по полю ввода *Выходной диапазон* и, установите курсор на ячейку D1. Нажмите ОК.

5. *Интерпретация результатов.* Поскольку вероятность нулевой гипотезы (нет разницы между средней скоростью реакции) 0,000457726 меньше чем уровень значимости (0,05), то эта гипотеза отвергается. Таким образом, можно утверждать, что различия в скорости сенсомоторной реакции спортсменов и людей, не занимающихся профессионально спортом неслучайны, то есть различия достоверны.

6. Сохраните документ.

Задание 4. Проверить соответствие выборочных данных из задания 1 лабораторной работы 1 нормальному закону распределения.

Решение.

1. Перейдите на новый лист. Выделите данные в диапазоне A1:J13 на *листе 1* и скопируйте их на *лист 4*.

2. Найдите теоретические частоты нормального распределения. Для этого предварительно необходимо найти среднее значение и стандартное отклонение выборки. В ячейке I14 с помощью функции СРЗНАЧ найдите среднее значение для данных из диапазона A2:E12. В ячейке J14 с помощью функции СТАНДОТКЛОН найдите стандартное отклонение для этих же данных. В ячейки K1 и K2 введите название столбца – *Теоретические частости*. Затем с помощью функции НОРМРАСП найдите теоретические частости. Установите курсор в ячейку K4, вызовите указанную функцию и заполните ее рабочие поля: *x—G4; Среднее—I14; Стандартное_ откл—J14; Интегральный—0.* Далее скопируйте содержимое ячейки K4 в диапазон ячеек K5:K12. Затем в ячейки L1 и L2 введите название нового столбца – *Теоретические частоты*. Установите курсор в ячейку L4 и введите формулу *=H$13*K4*. Далее скопируйте содержимое ячейки L4 в диапазон ячеек L2:L12.

3. С помощью функции ХИ2ТЕСТ определите соответствие данных нормальному закону распределения. Для этого установите курсор в свободную ячейку L13. На панели инструментов *Стандартная* нажмите кнопку *Вставка функции* (кнопка ). В появившемся диалоговом окне выберите категорию *Статистические* и функцию ХИ2ТЕСТ, после чего нажмите ОК. В появившемся окне в рабочие поля введите фактический H4:H12 и ожидаемые L4:L12 диапазоны частот. В ячейке L13 появится значение вероятности того, что выборочные данные соответствуют нормальному закону распределения.

4. Проанализируйте полученный результат.

5. Сохраните документ.

Примеры решения задач

1. Из партии, содержащей 2000 деталей, для проверки по схеме бесповторной выборки было отобрано 200 деталей, среди которых оказалось 184 стандартных. Найти: а) вероятность того, что доля нестандартных деталей во всей партии отличается от полученной доли в выборке не более чем на 0,02 (по абсолютной величине); б) границы, в которых с надежностью 0,95 заключена доля нестандартных деталей во всей партии.

Решение. Имеем $N = 2000$, $n = 200$, $n = 200 - 184 = 16$ нестандартных деталей. Выборочная доля нестандартных деталей $w = \frac{m}{n} = \frac{16}{200} = 0,08$.

а) Находим среднюю квадратическую ошибку бесповторной выборки для доли:

$$\sigma'_w = \sqrt{\frac{0,08 \cdot 0,92\left(1 - \frac{200}{2000}\right)}{200}} = 0,0182.$$

Тогда искомая доверительная вероятность**:**

$$P(|w - p| \le 0,02) = 2\Phi\left(\frac{0,02}{0,0182}\right) = 2\Phi(1,10) = 0,729,$$

т.е. вероятность того, что выборочная доля нестандартных деталей будет отличаться от генеральной доли не более чем на 0,02, равна 0,729.

б) Учитывая, что $\gamma = 2\Phi(t) = 0,95$ и $t = 1,96$, найдем предельную ошибку выборки для доли**:** $\Delta = 1,96 \cdot 0,0182 = 0,0357$**.** Теперь искомый доверительный интервал:

$$0,08\text{-}0,0357 \le p \le 0,08+0,0357 \text{ или } 0,044 \le p \le 0,116.$$

Итак, с надежностью 0,95 доля нестандартных деталей во всей партии заключена от 0,044 до 0,116.

2. Найти с надежностью 0,95 доверительный интервал для неизвестного математического ожидания *a,* нормально распределенного признака *X*, если генеральная дисперсия $\sigma^2 = 7$, выборочная средняя $\bar{x} = 4,5$ и объем выборки n=20.

Решение. Требуется найти доверительный интервал по формуле

$$\bar{x} - t_\gamma \frac{\sigma}{\sqrt{n}} < a < \bar{x} + t_\gamma \frac{\sigma}{\sqrt{n}} \cdot$$

По условию $\bar{x} = 4,5$; n=20; $\sigma = \sqrt{7}$; т.е. все величины кроме t_γ известны. Найдем t_γ из соотношения $\Phi(t) = \frac{\gamma}{2}$, где $\gamma = 0,95$, $\Phi(t) = \frac{0,95}{2} = 0,475$. По таблице (Приложение 2) находим t_γ=1,96.

Подставим t_γ=1,96 в формулу $\bar{x} - t_\gamma \frac{\sigma}{\sqrt{n}} < a < \bar{x} + t_\gamma \frac{\sigma}{\sqrt{n}}$ и получим $4,5 - 1,96 \cdot \frac{\sqrt{7}}{\sqrt{20}} < a < 4,5 + 1,96 \cdot \frac{\sqrt{7}}{\sqrt{20}}$. Окончательно получим 95% доверительный интервал для математического ожидания

(4,5-1,16; 4,5+1,16)=(3,34; 5,66).

3. В некотором районе в личном владении находится 3500 коров. Выборочно обследовали 800 коров и установили, что среднегодовой удой у этой группы коров 4200 кг. Оценить вероятность того, что среднегодовой удой у всех коров отличается от среднего выборочного не более чем на 10 кг, если генеральное среднеквадратичное отклонение равно 250 кг.

Решение. Рассмотрим случаи повторного и бесповторного отбора. Из условия имеем n= 800; $\Delta = 10$; $\sigma = 250; a = 4200; N = 3500$.

Для повторного отбора

$$P(|\bar{x} - a| < \Delta) = P(|\bar{x} - 4200| < 10) = 2\Phi\left(\frac{\Delta\sqrt{n}}{\sigma}\right) = 2\Phi\left(\frac{10 \cdot \sqrt{800}}{250}\right) = 2\Phi(1{,}13) = 2 \cdot 0{,}3708 = 0{,}7416 \approx 74\%.$$

Для бесповторного отбора

$$P(|\bar{x} - a| < \Delta) = P(|\bar{x} - 4200| < 10) = 2\Phi\left(\frac{\Delta\sqrt{n}}{\sigma\sqrt{1 - \frac{n}{N}}}\right) = 2\Phi\left(\frac{10 \cdot \sqrt{800}}{250\sqrt{1 - \frac{800}{3500}}}\right) = 2\Phi(1{,}288) \approx 2\Phi(1{,}29) =$$

$$= 2 \cdot 0{,}4015 = 0{,}803 = 80\%$$

Мы видим, что бесповторная выборка дает более точный результат при одном и том же объеме.

4. Определить объем повторной и бесповторной выборок для определения средней продолжительности горения лампочки в партии из 5000 лампочек, чтобы с вероятностью 0,99 предельная ошибка выборки не превосходила 25 часов. Предельная ошибка распределена нормально со среднеквадратическим отклонением равным 150.

Решение. По условию $\sigma = 150$; Δ=25; γ=0,99; N=5000.

Из соотношения $\Phi(t) = \frac{\gamma}{2}$ найдем t: $\Phi(t) = \frac{0{,}99}{2} = 0{,}495$. По таблице (Приложение 2) находим t=2,58.

Следовательно, необходимый объем выборки в случае повтора

$$n = \frac{t^2\sigma^2}{\Delta^2} = \frac{(150)^2 \cdot (2{,}58)^2}{(25)^2} = 239{,}6 \approx 240.$$

В случае бесповторного отбора

$$n = \frac{Nt^2\sigma^2}{t^2\sigma^2 + \Delta^2 N} = \frac{5000 \cdot (150)^2 \cdot (2{,}58)^2}{(2{,}58)^2 \cdot (25)^2 + (150)^2 \cdot (2{,}58)^2} \approx 229.$$

5. Из 2000 деталей было отобрано 400, распределение которых по размеру задается следующей таблицей.

Размер детали, мм	7,975	8,025	8,075	8,125	8,175	8,225
Количество деталей	12	28	132	150	62	16

Найти среднюю ошибку выборки при повторном и бесповторном отборе.

Решение. Выборочную среднюю и “исправленное” среднеквадратическое отклонение найдем соответственно по формулам

$$n = \sum m_i = 12+28+132+150+62+16,$$

$$\bar{x} = \frac{\sum x_i m_i}{n} = \frac{1}{400}\left(12 \cdot 7{,}975 + 28 \cdot 8{,}025 + 132 \cdot 8{,}075 + 150 \cdot 8{,}125 + 62 \cdot 8{,}175 + 16 \cdot 8225\right) \approx 8{,}11,$$

$$s = \sqrt{\frac{\sum m_i (x_i - \bar{x})^2}{n-1}} =$$

$$= \sqrt{\frac{1}{399} \cdot (12 \cdot (7{,}975 - 8{,}11)^2 + 28 \cdot (8{,}025 - 8{,}11)^2 + 132 \cdot (8{,}075 - 8{,}11)^2 + 150 \cdot (8{,}125 - 8{,}11)^2 +}$$

$$\overline{62 \cdot (8{,}175 - 8{,}11)^2 + 16 \cdot (8{,}225 - 8{,}11)^2)} \approx 0{,}052.$$

Таким образом, средняя ошибка для повторной выборки

$$\mu = \frac{s}{\sqrt{n}} = \frac{0{,}052}{\sqrt{400}} = 0{,}0026,$$

а для бесповторной выборки $\mu = \frac{s}{\sqrt{n}} \cdot \sqrt{1 - \frac{n}{N}} = \frac{0{,}052}{\sqrt{400}} \cdot \sqrt{1 - \frac{400}{2000}} = \frac{0{,}052}{20} \cdot \frac{2}{\sqrt{5}} \approx 0{,}0023.$

6. Из генеральной совокупности, где признак распределен по нормальному закону, извлечена выборка объема n=12:

x_i	-0,5	-0,4	-0,2	0	0,2	0,6	0,8	1	1,2	1,5
n_i	1	2	1	1	1	1	1	1	2	1

Найти доверительный интервал для математического ожидания, который отвечает доверительной вероятности $\gamma = 0{,}95$.

Решение. В данном примере воспользуемся формулой

$$\bar{x} - \frac{st_\alpha}{\sqrt{n}} < a < \bar{x} + \frac{st_\alpha}{\sqrt{n}}$$

Сначала найдем

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{1}{12}\left(-0{,}5 \cdot 1 - 0{,}4 \cdot 2 - 0{,}2 \cdot 1 + 0 + 0{,}2 \cdot 1 + 0{,}6 \cdot 1 + 0{,}8 \cdot 1 + 1 + 1{,}2 \cdot 2 + 1{,}5 \cdot 1\right) \approx 0{,}417$$

$$s = \sqrt{\frac{\sum n_i (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1}{11}((-0{,}5 - 0{,}417)^2 \cdot 1 + (-0{,}4 - 0{,}417)^2 \cdot 2 + (-0{,}2 - 0{,}417)^2 \cdot 1 +}$$

$$+(0 - 0{,}417)^2 \cdot 1 + (0{,}2 - 0{,}417)^2 \cdot 1 + (0{,}6 - 0{,}417)^2 \cdot 1 + (0{,}8 - 0{,}417)^2 \cdot 1 + (1 - 0{,}417)^2 \cdot 1 +$$

$$\overline{+(1{,}2 - 0{,}417)^2 \cdot 2 + (1{,}5 - 0{,}417)^2 \cdot 1)} \approx 0{,}720.$$

В случае малой выборки (n<30) вместо нормального распределения пользуются распределением Стьюдента, которое зависит от одного параметра – числа степеней свободы этого распределения. В нашем случае $\nu = n - 1 = 12 - 1 = 11$. По соответствующей таблице распределения Стьюдента (см. приложение 3) находим $t_{кр}(11;0{,}05)$=2,2, где $\alpha = 1 - 0{,}95 = 0{,}05$. Тогда искомый интервал $\left(0{,}417 - \frac{0{,}720}{\sqrt{12}} \cdot 2{,}2; 0{,}417 + \frac{0{,}720}{\sqrt{12}} \cdot 2{,}2\right) = (-0{,}04; 0{,}88)$. Это означает, что если будет произведено достаточно большое число выборок данного объема, то 95% из них найденный доверительный интервал накроет математическое ожидание

и только в 5% случаев оцениваемое математическое ожидание может выйти за границы доверительного интервала.

7. Перед выборами было опрошено $n = 650$ человек. Из них $k = 350$ человек отдали предпочтение демократам. На сколько голосов могут рассчитывать демократы, если число избирателей в городе равно $N = 65000$. Вычисления произвести с доверительной вероятностью 0,95.

Решение. Обозначим через m число избирателей, которые проголосовали «за». Применим формулу для бесповторного отбора:

$$P\left(|w-p|<\varepsilon\right)\approx 2\Phi\left(\frac{\varepsilon}{\mu}\right),$$

где

$$\mu=\sqrt{\frac{pq}{n}\left(1-\frac{n}{N}\right)}.$$

Откуда $p-\varepsilon\le W\le p+\varepsilon$.

Так как $W=\frac{m}{n}$, то $\left(p-\varepsilon\right)N\le m\le\left(p+\varepsilon\right)N$. Найдем p и ε. По условию $N = 65000$; $p=\frac{k}{n}=\frac{350}{650}=\frac{7}{13}$; $q=1-\frac{7}{13}=\frac{6}{13}$.

Доверительная вероятность по условию $P = 0,95$, а значит $2\Phi\left(t\right)=0,95$ или $\Phi\left(t\right)=0,475$. Из приложения 2 находим, что $t = 1,96$. Тогда

$$\frac{\varepsilon}{\mu}=t=1,96\ \Rightarrow\ \varepsilon=1,96\cdot\mu=1,96\cdot\sqrt{\frac{\frac{7}{13}\cdot\frac{6}{13}}{650}\cdot\left(1-\frac{650}{65000}\right)}\approx 0,038.$$

Таким образом, $\frac{7}{13}-\varepsilon\le W\le\frac{7}{13}+\varepsilon\ \Rightarrow\ 65000\left(\frac{7}{13}-\varepsilon\right)\le m\le 65000\left(\frac{7}{13}+\varepsilon\right)$.

Следовательно, $65000\cdot 0,5\le m\le 65000\cdot 0,576$, т.е. $32500\le m\le 37440$.

Итак, демократы смогут рассчитывать на m голосов, где $32500\le m\le 37440$.

Задачи для самостоятельного решения

1. Для определения средней урожайности массива пшеницы площадью в 400 га был произведен случайный отбор 50 опытных участков, каждый площадью 0,25 га. Выборочная средняя урожайность оказалась равной 19 ц/га, а среднеквадратическое отклонение 1,5 ц/га. Найти с вероятностью 0,99 возможные пределы для определяемой средней урожайности.

2. Глубина моря измеряется прибором, систематическая ошибка которого равна 0, а случайная ошибка распределена нормально с $\sigma = 30$ м. Сколько надо сделать независимых измерений, чтобы определить глубину с ошибкой не более 15м при $\gamma = 0,9$.

3. При определении качества продукции в партии из 10000 штук была произведена повторная выборка объемом в 300 единиц. Определить доверительные границы доли изделий первого сорта в партии с доверительной

вероятностью, равной 0,999, если частость изделия первого сорта в выборке оказалась равной 0,6. Произвести расчет для бесповторной выборки.

4. Для определения процента вкладов, не превышающих 1000 ден. ед., произведена повторная выборка 900 лицевых счетов. Среди них оказалось 30% вкладов, не более 1000 ден. ед. каждый. С какой доверительной вероятностью можно утверждать, что процент таких вкладов в данной кассе будет отличаться от найденного не более чем на 2%?

5. Производится выборочное обследование доли лиц с высшим образованием в данной местности. Сколько нужно обследовать лиц, чтобы полученный результат гарантировать с вероятностью 0,95 при допустимой ошибке в определяемой доле 0,01?

6. Из генеральной совокупности с нормальным распределением извлечена выборка с объемом n=10

x_i	-2	0	1	2	3	4	5
n_i	1	1	2	1	2	2	1

Найти доверительный интервал для математического ожидания с доверительной вероятностью $\gamma = 0{,}99$.

7. Обследуется средняя продолжительность разговора. Сколько телефонных разговоров должно быть зафиксировано, чтобы с вероятностью 0,997, можно было бы утверждать, что отклонение выборочной средней от генеральной средней не превосходит 10 секунд, если среднеквадратическое отклонение равно 2,5 мин.

8. При изучении выборочным путем срока службы телевизоров получено следующее выборочное распределение:

Срок службы (в днях)	до 50	50-100	100-150	150-200	200-250
Число телевизоров	10	30	60	40	20

Определить с вероятностью 0,99 доверительные границы для среднего срока службы телевизоров данной партии.

9. Перед выборами было опрошено 30000 человек. Из них 200 человек отдали предпочтение демократам. На сколько голосов могут рассчитывать демократы, если число избирателей в городе равно 600. Вычисления произвести с доверительной вероятностью 0,95.

Вопросы для самоконтроля

1. Что такое интервальные оценки? Как они строятся?
2. Чем отличаются интервальные оценки для математического ожидания нормальной СВ при известной и неизвестной дисперсиях?
3. Построение доверительных интервалов для дисперсии и среднеквадратического отклонения нормальной СВ.

3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

3.1. Статистические гипотезы

Очень часто при решении практических задач, связанных с применением методов математической статистики, возникает вопрос: может ли на основании данных некоторой выборки быть принято или отвергнуто некоторое предположение (гипотеза) относительно генеральной совокупности. Например, испытана новая методика обучения. Можно ли по результатам испытания сделать обоснованный вывод о том, что новая методика по сравнению с предыдущей более эффективна (имеет лучшие оценочные показатели). Аналогичный вопрос возникает и при апробации новых лекарств, при внедрении новых технологии и т. п. Процедура сопоставления высказанного предположения (гипотезы) с данными выборки называется проверкой гипотез. Эта процедура состоит в следующем. Относительно генеральной совокупности высказывается некоторая гипотеза (или несколько гипотез). Из генеральной совокупности извлекается выборка. Нужно указать правило, которое бы давало ответ на вопрос: следует ли отклонить гипотезу (некоторые гипотезы) или принять ее (одну из них). Отметим, что статистическими методами доказать гипотезу нельзя. Есть лишь возможность опровергнуть или не опровергнуть ее.

Статистические гипотезы разделяются на два вида: *нулевые* и *альтернативные*. Нулевая гипотеза (H_0) утверждает об отсутствии различий между двумя распределениями (различия равны нулю), альтернативная (H_1) – о существовании или значимости различий. Нулевая и альтернативная гипотезы являются взаимоисключающими, и в этом плане, одна из них должна будет оказаться истинной, а другая – ложной. Для проверки статистических гипотез служат статистические критерии. Статистические гипотезы могут быть *направленные* и *ненаправленные*. Если гипотеза просто утверждает отсутствие или значимость различий, то она является ненаправленной, т.к. в ее формулировку не входит направление различий. Если гипотеза помимо отсутствия или значимости различий утверждает и то, что параметры одного распределения должны оказаться больше или меньше, чем параметры другого, то она является направленной.

Статистический критерий – это решающее правило, обеспечивающее надежное принятие истинной гипотезы и отклонение ложной с высокой вероятностью, а также метод расчета числа, говорящего о значимости различий между распределениями случайной величины и само это число.

Статистические критерии, сами по себе, не являются средством решения научных проблем, так как статистические методы не заменяют собой мышления ученого. Результат, полученный при помощи применения статистических критериев всегда носит вероятностный характер, т.к. исследователь в большинстве случаев имеет дело не только со случайной величиной, но и со случайной выборкой, и поэтому, выводы его также обладают определенной степенью достоверности или значимости.

Уровень значимости (α) – это вероятность того, что исследователь счел различия существенными, а они на самом деле случайны. Уровень значимости – это допустимая для данной задачи вероятность ошибки первого рода (ложноположительного решения), то есть вероятность отклонить нулевую гипотезу, когда на самом деле она верна. Другими словами, уровень значимости – это такое (достаточно малое) значение вероятности события, при котором событие уже можно считать неслучайным.

В социологических исследованиях обычно используется три уровня значимости: 5-процентный, 1-процентный и 0,1- процентный (хотя последний намного реже). Если указывают, что различия достоверны на 5%-ом уровне значимости ($p \leq 0,05$), то имеют ввиду, что вероятность ошибочного вывода составляет 0,05, если на 1%-ом – 0,01 ($p \leq 0,01$) и т.д. При этом, 5%-й уровень считается низшим, а 0,1%-й – высшим уровнем значимости.

Степенью свободы называется характеристика распределения, используемая при проверке статистических гипотез (обозначается *df* или *ν*). Число степеней свободы равно числу классов вариационного ряда минус число условий, при которых он был сформирован.

Предположим, что выборка из 100 человек была разбита на три класса в зависимости от степени выраженности какого-либо признака. В первый класс могут попасть те, у кого признак выражен максимально, во второй те, у кого он выражен в средней степени, но в третий могут попасть только оставшиеся, вне зависимости от того минимально выражен у них признак, или вовсе отсутствует. Можно, конечно, допустить и другое разбиение, но число степеней свободы в данном случае будет равно $df = 3–1= 2$. Если исследователь имеет дело с классификацией из 100 классов, то *df* будет равно 99 и т.д. Для двух распределений $df = c – 2$ (c – число классов), а при представлении переменных в таблице размером $a \times b$, $df = (a – 1)(b – 1)$, где a – число столбцов, а b – число строк.

Поскольку уже говорилось, что статистический критерий – не только метод расчета числа, говорящего о различиях между распределениями, но и само это число, в задачи исследователя входит и правильная интерпретация полученного значения статистического критерия. Для того чтобы определить, какая из двух гипотез верна, необходимо обратиться к таблицам значимости статистических критериев. В этих таблицах даются критические значения статистического критерия для соответствующего числа степеней свободы и уровня значимости. Например, если применялся *t*-критерий Стьюдента, а число степеней свободы было равно 20-ти, то необходимо найти значения *t*-критерия на 5%-ом и 1%-ом уровне значимости (2,09 и 2,85 соответственно). Если полученное эмпирическое значение окажется меньше, либо равняется критическому (табличному) значению на 5%-ом уровне, то необходимо признать верной нулевую гипотезу, если же выше, чем на 1%-ом уровне – альтернативную. В том случае, когда эмпирическое значение оказывается между двух критических, ни нулевую, ни альтернативную гипотезу принять нельзя, необходимо либо увеличить объем выборки, чтобы различия стали достоверны, либо воспользоваться другим критерием. Так обстоит дело с

большинством критериев – чем выше число, тем достоверней различия между распределениями, и лишь в отношении некоторых критериев картина обратная (см. описания критериев).

Мощность критерия (1–β) – это его способность выявлять различия, если они есть, т.е. его способность отклонить нулевую гипотезу об отсутствии различий, если она неверна.

Мощность критерия определяется эмпирическим путем. Для проверки одной и той же гипотезы можно использовать разные критерии, но при этом обнаруживается, что одни критерии выявляют различия, а другие – нет. Те критерии, которые обнаруживают различия, особенно на малых выборках, в то время как другие неспособны это сделать признаются более мощными, и это снижает, хотя это и не устраняет вероятности ошибочного вывода.

Статистический вывод связан с так называемыми ошибками I и II рода. Ошибка, состоящая в том, что была отклонена нулевая гипотеза, в то время, как она верна, называется ошибкой I рода. Ошибка, состоящая в том, что была принята нулевая гипотеза, в то время как она неверна, является ошибкой II рода. Иначе говоря, это ошибки отвержения истинной гипотезы и принятия ложной. Ниже представлено распределение истинных решений и возможных ошибок статистического вывода.

Таблица 6. Распределение ошибок и истинных решений в зависимости от верности гипотез и решений исследователя

	H_0 верна H_1 неверна	H_0 неверна H_1 верна
Отклонить H_0 Принять H_1	Ошибка I рода α	Истинное решение $1-\beta$
Принять H_0 Отклонить H_1	Истинное решение $1-\alpha$	Ошибка II рода β

Критерии принято делить на параметрические и непараметрические. Параметрическими критериями являются те, в формулу расчета которых входят параметры распределения – средние или дисперсии. Непараметрические критерии в отличии от параметрических основаны на использовании в их формулах частот, долей или рангов. Непараметрические критерии применимы к переменным выраженным в любых шкалах, а параметрические – только лишь к тем переменным, которые выражены в шкалах интервалов или отношений.

И те, и другие критерии имеют свои преимущества и недостатки. В тех случаях, когда переменная измерена в шкале интервалов и ее распределение близко к нормальному, лучше пользоваться параметрическими критериями, т.к. они оказываются более мощными, чем непараметрические. Но в том случае, если эти условия не выполняются, более эффективными окажутся непараметрические критерии, так как им "все равно" в каких шкалах измерены переменные и соответствует распределение нормальному или нет. В ряде случаев непараметрическим критериям нет замены, особенно если признак определялся не количественно, а качественно.

Правило отклонения нулевой и принятия альтернативной гипотезы.

Если эмпирическое значение критерия равняется критическому значению, соответствующему $\alpha \leq 0{,}05$, или превышает его, то H_0 отклоняется, но мы еще не можем определенно принять H_1.

Если эмпирическое значение критерия равняется критическому значению, соответствующему $\alpha \leq 0{,}01$, или превышает его, то H_0 отклоняется и принимается H_1.

Для облегчения процесса принятия решения можно всякий раз вычерчивать «ось значимости»:

Ось значимости представляет собой прямую, на левом конце которой располагается 0, хотя он, как правило, не отмечается на самой этой прямой, и слева направо идет увеличение числового ряда. По сути дела это привычная школьная ось абсцисс *Ox* декартовой системы координат. Однако особенность этой оси в том, что на ней выделено три участка, «зоны». Левая зона называется «зоной незначимости», правая – «зоной значимости», а промежуточная – «зоной неопределенности». Границами всех трех зон являются критическое значение, соответствующее α=0,05 (обозначается как $Ч_{0,05}$) и критическое значение, соответствующее α=0,01 (обозначается как $Ч_{0,05}$).

Вправо от критического значения $Ч_{0,01}$ простирается «зона значимости» – сюда попадают эмпирические значения, превышающие $Ч_{0,01}$, и, следовательно, значимые. В этом случае принимается альтернативная гипотеза H_1.

Влево от критического значения $Ч_{0,05}$ простирается «зона незначимости» – сюда попадают эмпирические значения, которые ниже $Ч_{0,05}$ следовательно, незначимы, и в этом случае принимается гипотеза H_0 об отсутствии различий.

Если эмпирическое значение попадает в «зону неопределенности», то отклоняется гипотеза о недостоверности различий (H_0), но гипотеза об их достоверности (H_1) не принимается.

До настоящего времени созданы десятки статистических критериев, которые существуют для решения довольно ограниченного круга задач. Создание статических критериев не является самоцелью, каждый из таких методов проверки гипотез имеет свои преимущества и недостатки, и в некоторых случаях может, а в некоторых – не может быть заменен другими критериями. Основанием для выбора критерия является не только его мощность, но и другие характеристики: простота вычисления, применимость к неравным по объему выборкам, применимость к нескольким выборкам сразу, возможность использования его для переменных, измеренных в разных шкалах, универсальность (возможность применения его к решению самых различных задач).

Все многообразие задач, с которыми приходится сталкиваться экспериментатору при проверке гипотез, можно свести к нескольким группам:

a) Выявление различий в распределении переменной в разных группах испытуемых;

b) Проверка совпадения эмпирических результатов с ожидаемыми теоретическими;

c) Обнаружение влияния фактора на распределение переменной;

d) Обнаружение интересующего исследователя эффекта в одной или разных выборках испытуемых.

3.2. Проверка гипотез о равенстве дисперсий (случай нормального распределения)

Гипотезы о дисперсиях возникают довольно часто, так как дисперсия характеризует такие исключительно важные показатели, как точность машин, приборов, технологических процессов, степень однородности совокупностей, риск, связанный с отклонением доходности активов от ожидаемого уровня, и т.д.

Сформулируем задачу. Пусть имеются две нормально распределенные совокупности, дисперсии которых равны $\sigma_1^{\,2}$ и σ_2^2. Необходимо проверить нулевую гипотезу о равенстве дисперсий, т.е. $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2$ относительно $\mathrm{H}_1 : \sigma_1^2 > \sigma_2^2$ или $\mathrm{H}_1' : \sigma_1^2 \neq \sigma_2^2$.

Для проверки гипотезы H_0 из этих совокупностей взяты две независимые выборки объемом n_1 и n_2. Для оценки дисперсий σ_1^2 и σ_2^2 используются «исправленные» выборочные дисперсии $\hat{s}_1^2$ и $\hat{s}_2^2$. Следовательно, задача проверки гипотезы сводится к сравнению дисперсий $\hat{s}_1^2$ и $\hat{s}_2^2$.

При справедливости гипотезы $\mathrm{H}_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$ в качестве оценки σ^2 можно взять те же дисперсии $\hat{s}_1^2$ и $\hat{s}_2^2$, рассчитанные по элементам первой и второй выборок.

Напомним, что выборочные характеристики $\dfrac{(n_1-1)\hat{s}_1^2}{\sigma^2}$ и $\dfrac{(n_2-1)\hat{s}_2^2}{\sigma^2}$ имеют распределение соответственно χ^2 с $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ степенями свободы, а их отношение $\dfrac{\frac{1}{k_1}\chi^2(k_1)}{\frac{1}{k_2}\chi^2(k_2)}$ имеет F-распределение Фишера–Снедекора с k_1 и k_2 степенями свободы. Следовательно, случайная величина F, определяемая отношением:

$$F = \frac{\dfrac{1}{n_1 - 1}\left[(n_1-1)\hat{s}_1^2 \Big/ \sigma^2\right]}{\dfrac{1}{n_2 - 1}\left[(n_2-1)\hat{s}_2^2 \Big/ \sigma^2\right]} = \frac{\hat{s}_1^2}{\hat{s}_2^2},$$

т.е. отношением «исправленных» выборочных дисперсий, имеет F-распределение Фишера–Снедекора с $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ степенями свободы. Вид некоторых кривых F-распределения показан на рис. 7.

При формировании критерия отклонения (принятия) гипотезы H_0 следует учесть, что распределение статистики F (в отличие от нормального или распределения Стьюдента) является несимметричным.

Поэтому гипотеза H_0 отвергается, если $F > F_{\alpha;k_1,k_2}$ (в случае правосторонней критической области – рис. 7, *а*), либо если $F < F_{1-\alpha;k_1,k_2}$ (в случае левосторонней – рис. 7, *б*), либо если $F < F_{1-\alpha/2;k_1,k_2}$ и $F > F_{\alpha/2;k_1,k_2}$ (в случае двусторонней критической области – рис. 7, *в*). В противном случае гипотеза H_0 не отвергается (принимается).

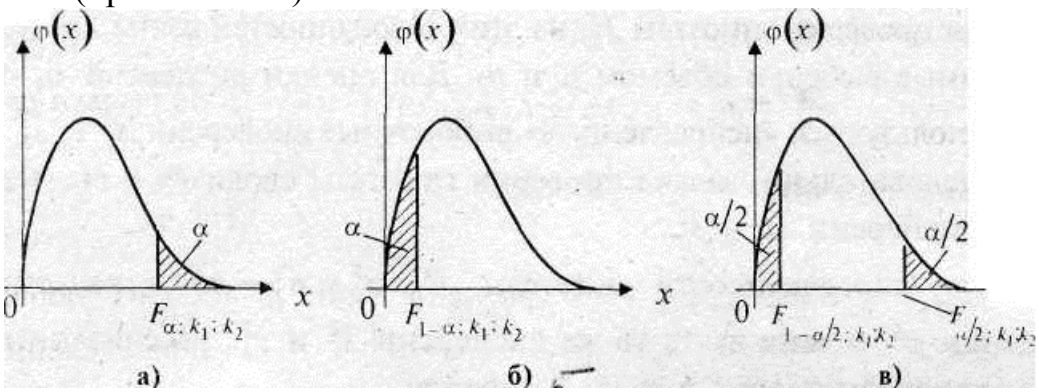


Рис. 7.

В приложении 5 приведены таблицы значений $F_{\alpha;k_1k_2}$ для $\alpha = 0{,}05$.

Пример. На двух токарных станках обрабатываются втулки. Отобраны две пробы: из втулок, сделанных на первом станке, $n_1 = 15$ шт., на втором станке – $n_2 = 18$ шт. По данным этих выборок рассчитаны выборочные дисперсии $s_1^2 = 8{,}5$ (для первого станка) и $s_2^2 = 6{,}3$ (для второго станка). Полагая, что размеры втулок подчиняются нормальному закону распределения, на уровне значимости $\alpha = 0{,}05$ выяснить, можно ли считать, что станки обладают различной точностью.

Решение. Имеем нулевую гипотезу $H_0 : \sigma_1^2 = \sigma_2^2$, т.е. дисперсии размера втулок, обрабатываемых на каждом станке, равны. Возьмем в качестве конкурирующей гипотезу $H_1 : \sigma_1^2 > \sigma_2^2$ (дисперсия больше для первого станка). Статистика критерия (в качестве дисперсии s_1^2, стоящей в числителе, берут большую из двух дисперсий – это дает возможность, учитывая свойства F-распределения, в два раза сократить объем его табличных значений):

$$F = \frac{\hat{s}_1^2}{\hat{s}_2^2} = \frac{\dfrac{n_1}{n_1 - 1} s_1^2}{\dfrac{n_2}{n_2 - 1} s_2^2} = \frac{(15/14) \cdot 8{,}5}{(18/17) \cdot 6{,}3} = 1{,}37$$

По приложению 5 критическое значение F-критерия на уровне значимости $\alpha = 0{,}05$ при числе степеней свободы $k_1 = n_1 - 1 = 14$ и $k_2 = n_2 - 1 = 17$, т.е. $F_{0{,}05;14;17} = 2{,}33$. Так как $F > F_{0{,}05;14;17}$, то гипотеза H_0 не отвергается, т.е. имеющиеся данные не позволяют считать, что станки обладают различной точностью.

3.3. Проверка гипотез о числовых значениях параметров

Гипотезы о числовых значениях встречаются в различных задачах. Пусть $x_i (i = 1,2,...,n)$ – значения некоторого параметра изделий, производящихся станком автоматической линии, и пусть a – заданное номинальное значение этого параметра. Каждое отдельное значение x_i может, естественно, как-то отклоняться от заданного номинала. Очевидно, для того, чтобы проверить правильность настройки этого станка, надо убедиться в том, что среднее значение параметра у производимых на нем изделий будет соответствовать номиналу, т.е. проверить гипотезу $H_0 : \bar{x}_0 = a$ против альтернативной:

$$H_1 : \bar{x}_0 \neq a \ \ (H_2' : \bar{x}_0 < a \ \text{ или } \ H_2'' : \bar{x}_0 > a).$$

При произвольной настройке станка может возникнуть необходимость проверки гипотезы о том, что точность изготовления изделий по данному параметру, задаваемая дисперсий σ^2, равна заданной величине ${\sigma_0}^2$, т.е. $H_0 : \sigma^2 = {\sigma_0}^2$ или, например, того, что доля бракованных изделий, производимых станком, равна заданной величине p_0, т.е. $H_0 : p = p_0$ и т.д.

Аналогичные задачи могут возникнуть, например, в финансовом анализе, когда по данным выборки надо установить, можно ли считать доходность актива определенного вида или портфеля ценных бумаг, либо ее риск равным заданному числу; или по результатам выборочной аудиторской проверки однотипных документов нужно убедиться, можно ли считать процент допущенных ошибок равным номиналу, и т.п.

В общем случае гипотезы подобного типа имеют вид $H_0 : \theta = \Delta_0$, где θ – некоторый параметр исследуемого распределения, а Δ_0 – область его конкретных значений, состоящая в частном случае из одного значения.

Соответствующие критерии проверки гипотез о числовых значениях параметров нормального закона приведены в табл. 7.

Таблица 7

Нулевая гипотеза	Предположения	Статистика критерия	Альтернативная гипотеза	Критерий отклонения гипотезы
$a = a_0$	σ^2 известна	$t = \dfrac{\bar{x} - a_0}{\sigma / \sqrt{n}}$	$a = a_1 > a_0$ $a = a_1 < a_0$ $a = a_1 \neq a_0$	$\lvert t \rvert > t_{1-2\alpha}$ $\lvert t \rvert > t_{1-\alpha}$
	σ^2 неизвестна	$t = \dfrac{\bar{x} - a_0}{s / \sqrt{n-1}}$	$a = a_1 > a_0$ $a = a_1 < a_0$ $a = a_1 \neq a_0$	$\lvert t \rvert > t_{1-2\alpha,n-1}$ $\lvert t \rvert > t_{1-\alpha,n-1}$
$\sigma^2 = \sigma_0^2$	a неизвестно	$\chi^2 = \dfrac{ns^2}{{\sigma_0}^2}$	$\sigma^2 = \sigma_1^2 > \sigma_0^2$ $\sigma^2 = \sigma_1^2 < \sigma_0^2$ $\sigma^2 = \sigma_1^2 \neq \sigma_0^2$	$\chi^2 < \chi^2_{\alpha;n-1}$ либо $\chi^2 > \chi^2_{1-\alpha;n-1}$ $\chi^2 > \chi^2_{\alpha/2;n-1}$ $\chi^2 < \chi^2_{1-\alpha/2;n-1}$

$p = p_0$	Достаточно большие n	$t = \dfrac{w - p_0}{\sqrt{p_0 q_0 / n}}$	$p = p_1 > p_0$ $p = p_1 < p_0$ $p = p_1 \neq p_0$	$\lvert t\rvert > t_{1-2\alpha}$ $\lvert t\rvert > t_{1-\alpha}$

Примечание. *Критические значения статистик на уровне значимости α определяют по соответствующим таблицам приложений исходя из соотношений:*

$$P\left(|t| < t_{1-\alpha}\right) = \Phi\left(t_{1-\alpha}\right) = 1-\alpha,$$
$$P\left(|t| < t_{1-\alpha,n-1}\right) = \theta\left(t_{1-\alpha,n-1}\right) = 1-\alpha,$$

$\theta(\lambda, n) = \dfrac{\lambda^n}{n!} e^{-\lambda}$ *– функция Пуассона,*

$$P\left(\chi^2 > \chi^2_{\alpha,n-1}\right) = \alpha.$$

Пример. На основании сделанного прогноза средняя дебиторская задолженность однотипных предприятий региона должна составить $a_0 = 120$ ден. ед. Выборочная проверка 10 предприятий дала среднюю задолженность $\bar{x} = 135$ ден. ед., а среднее квадратическое отклонение задолженности $s = 20$ ден. ед. На уровне значимости 0,05: а) выяснить, можно ли принять данный прогноз; б) найти мощность критерия, если в действительности средняя дебиторская задолженность всех предприятий региона равна 130 ден. ед.

Решение. а) Проверяемая гипотеза $H_0 : \bar{x}_0 = a = 120$. В качестве альтернативной возьмем гипотезу $H_1 : a > 120$. Так как генеральная дисперсия σ^2 неизвестна, то используем t-критерий Стьюдента. Статистика критерия в соответствии с табл. 7 равна $t = \dfrac{\bar{x} - a_0}{s/\sqrt{n-1}} = \dfrac{135-120}{20/\sqrt{10-1}} = 2{,}25.$ Критическое значение статистики находим из приложения 3

$$t_{1-2\cdot 0{,}05;10-1} = t_{0{,}9;9} = 1{,}83.$$

Так как $|t| > t_{0{,}9;9} (2{,}25 > 1{,}83)$,то гипотеза H_0 отвергается, т.е. на 5%-ном уровне значимости сделанный прогноз должен быть отвергнут.

б) Альтернативная гипотеза $H_1 : \bar{x}_0 = a_1 = 120.$ Так как $a_1 = 130 > a_0 = 120$, то критическая область правосторонняя и критическое значение выборочной средней

$$\bar{x}_{кр} = \bar{x}_0 + t_{1-2\alpha,n-1}\frac{s}{\sqrt{n-1}} = a + t_{0{,}9;9}\frac{s}{\sqrt{n-1}} = 120 + 1{,}83\frac{20}{\sqrt{10-1}} = 132{,}2(\textit{ден.ед.})$$

т.е. критическая область значений для $\bar{x}$ есть интервал $(132{,}2;+\infty)$. Мощность критерия равна вероятности P отвергнуть гипотезу H_0, когда верна гипотеза H_1 т.е.

$$P = P(132{,}2 < \bar{x} < +\infty) = \frac{1}{2} + \frac{1}{2}\theta(t, n-1),$$

где $t = \dfrac{\bar{x} - a_1}{s/\sqrt{n-1}} = \dfrac{132{,}2-130}{20/\sqrt{10-1}} = 0{,}33.$ Следовательно, $\theta(0{,}33;9) \approx 0{,}25.$

Итак, $P \approx \dfrac{1}{2}(1 + 0{,}25) = 0{,}62.$

Аналогично проверяются и другие гипотезы о числовых значениях параметров в соответствии с критериями проверки, приведенными в табл. 7.

3.4. Проверка гипотез о законе распределения с помощью критериев Пирсона и Колмогорова

Одной из важнейших задач математической статистики является установление теоретического закона распределения случайной величины, характеризующей изучаемый признак по опытному (эмпирическому) распределению, представляющему вариационный ряд.

Для решения этой задачи необходимо определить вид и параметры закона распределения.

Предположение о виде закона распределения может быть выдвинуто исходя из теоретических предпосылок (например, выполнение условий центральной предельной теоремы может свидетельствовать о нормальном законе распределения случайной величины), опыта аналогичных предшествующих исследований и, наконец, на основании графического изображения эмпирического распределения.

Параметры распределения, как правило, неизвестны, поэтому их заменяют наилучшими оценками по выборке.

Как бы хорошо ни был подобран теоретический закон распределения, между эмпирическим и теоретическим распределениями неизбежны расхождения. Естественно возникает вопрос: объясняются ли эти расхождения только случайными обстоятельствами, связанными с ограниченным числом наблюдений, или они являются существенными и связаны с тем, что теоретический закон распределения подобран неудачно. Для ответа на этот вопрос и служат **критерии согласия.**

Пусть необходимо проверить нулевую гипотезу H_0 о том, что исследуемая случайная величина X подчиняется определенному закону распределения. Для проверки гипотезы H_0 выбирают некоторую случайную величину U, характеризующую степень расхождения теоретического и эмпирического распределений, закон распределения которой при достаточно больших n известен и практически не зависит от закона распределения случайной величины X.

Зная закон распределения U, можно найти вероятность того, что приняла значение не меньше, чем фактически наблюдаемое в опыте u, т.е. $U \geq u$. Если $P(U \geq u) = \alpha$ мала, то это означает в соответствии с принципом практической уверенности, что такие, как в опыте, и большие отклонения практически невозможны. В этом случае гипотезу H_0 отвергают. Если же вероятность $P(U \geq u) = \alpha$ не мала, расхождение между эмпирическим и теоретическим распределениями несущественно и гипотезу H_0 можно считать правдоподобной или по крайней мере не противоречащей опытным данным.

Использование критерия согласия Пирсона. В наиболее часто используемом на практике критерии Пирсона в качестве меры расхождения U

берется величина χ^2, равная сумме квадратов отклонений частостей (статистических вероятностей) w_i, от гипотетически p_i, рассчитанных по предполагаемому распределению, взятых с некоторыми весами c_i.

$$U = \chi^2 = \sum_{i=1}^{m} c_i (w_i - p_i)^2.$$

Веса c_i вводятся таким образом, чтобы при одних и тех же отклонениях $(w_i - p_i)^2$ больший вес имели отклонения, при которых p_i мала, и меньший вес – при которых p_i, велика. Очевидно, этого удается достичь, если взять с обратно пропорциональными вероятностям p_i. Взяв в качестве весов $c_i = \frac{n}{p_i}$, можно доказать, что при $n \to \infty$ статистика

$$U = \chi^2 = \sum_{i=1}^{m} \frac{n}{p_i} (w_i - p_i)^2$$

или

$$U = \chi^2 = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i}$$

имеет χ^2-распределение с $k = m - r - 1$ степенями свободы, где m – число интервалов эмпирического распределения (вариационного ряда); r – число параметров теоретического распределения, вычисленных по экспериментальным данным.

Числа $n_i = nw_i$ и np_i называются соответственно ***эмпирическими*** и ***теоретическими частотами.***

Схема применения критерия χ^2 сводится к следующему:

1. Определяется мера расхождения эмпирических и теоретических частот.
2. Для выбранного уровня значимости, а по таблице χ^2-распределения находят критическое значение $\chi^2_{\alpha;k}$ при числе степеней свободы $k = m - r - 1$.
3. Если фактически наблюдаемое значение χ^2 больше критического, т.е. $\chi^2 > \chi^2_{\alpha;k}$, то гипотеза H_0 отвергается; если $\chi^2 \le \chi^2_{\alpha;k}$, то гипотеза H_0 не противоречит опытным данным.

Замечание. *Как уже отмечено, статистика*

$$\chi^2 = \sum_{i=1}^{m} (n_i - np_i)^2 / np_i$$

имеет χ^2-распределение лишь при $n \to \infty$, поэтому необходимо, чтобы в каждом интервале было достаточное количество наблюдений, по крайней мере 5 наблюдений. Если в каком-нибудь интервале число наблюдений $n_i < 5$, имеет смысл объединить соседние интервалы, чтобы в объединенных интервалах n_i было не меньше 5.

Пример. Для эмпирического распределения рабочих цеха по выработке по данным первых двух граф табл. 2 подобрать соответствующее

теоретическое распределение и на уровне значимости $\alpha = 0{,}05$ проверить гипотезу о согласованности двух распределений с помощью критерия χ^2.

Решение. По виду гистограммы распределения рабочих по выработке (рис. 2) можно предположить нормальный закон распределения признака. Параметры нормального закона a и σ^2, являющиеся соответственно математическим ожиданием и дисперсией случайной величины X неизвестны, поэтому заменяем их «наилучшими» оценками по выборке – несмещёнными и состоятельными оценками соответственно выборочной средней $\bar{x}$ и «исправленной» выборочной дисперсией $\hat{s}^2$. Так как число наблюдений $n = 100$ достаточно велико, то вместо «исправленной» $\hat{s}^2$ можно взять «обычную» выборочную дисперсию s^2. Ранее были вычислены $\bar{x} = 119{,}2(\%), s^2 = 87{,}48, s = 9{,}35(\%).$

Для расчета вероятностей p_i попадания случайной величины X в интервал $[x_i, x_{i+1}]$ используем функцию Лапласа в соответствии со свойством нормального распределения:

$$p_i(x_i \le X \le x_{i+1}) = \left[\Phi\left(\frac{x_{i+1}-a}{\sigma}\right) - \Phi\left(\frac{x_i - a}{\sigma}\right)\right] \approx \left[\Phi\left(\frac{x_{i+1}-119{,}2}{9{,}35}\right) - \Phi\left(\frac{x_i - 119{,}2}{9{,}35}\right)\right].$$

Например, $p_1(-\infty \le X \le 100) = \left[\Phi\left(\frac{100-119{,}2}{9{,}345}\right) - \Phi(-\infty)\right] = 0{,}0202$ и соответствующая первому интервалу теоретическая частота $np_i = 100 \cdot 0{,}0166 \approx 1{,}7$ и т.д.

Для определения статистики χ^2 удобно составить таблицу:

Таблица 8

i	Интервал $[x_i, x_{i+1}]$	Эмпирические частоты n_i	Вероятности p_i	Теоретические частоты np_i	$(n_i - np_i)^2$	$\frac{(n_i - np_i)^2}{np_i}$
1	$-\infty$-100	3 } 10	0,0202	2,02 } 7,92	4,326	0,546
2	100-106	7 }	0,059	5,9 }		
3	106-112	11	0,141	14,1	9,61	0,682
4	112-118	20	0,228	22,8	7,84	0,344
5	118-124	28	0,247	24,7	10,89	0,441
6	124-130	19	0,182	18,2	0,64	0,035
7	130-136	10 } 12	0,087	8,7 } 12,3	0,09	0,0073
8	136-$+\infty$	2 }	0,0359	3,6 }		
$\sum$		100	1	100	-	$\chi^2 = 2{,}055$

Учитывая, что в рассматриваемом эмпирическом распределении частоты первого и последнего интервалов $(n_1 = 3, n_8 = 2)$ меньше 5, при использовании критерия χ^2-Пирсона целесообразно объединить указанные интервалы с соседними (см. табл. 8).

Итак, фактически наблюдаемое значение статистики $\chi^2 = 2{,}055$.

Так как новое число интервалов (с учетом объединения крайних) $m=6$, а нормальный закон распределения определяется $r=2$ параметрами, то число степеней свободы $k=m-r-1=6-2-1=3$. Соответствующее критическое значение статистики χ^2 (см. приложение 3) $\chi^2_{0,05;3}=3{,}18$. Так как $\chi^2<\chi^2_{0,05}$, то гипотеза о выбранном теоретическом нормальном законе с параметрами $N(119{,}2;87{,}48)$ согласуется с опытными данными.

Замечание. *Для графического изображения эмпирического и выравнивающего его теоретического нормального распределений необходимо использовать одинаковый для двух распределений масштаб по оси ординат.*

Так, если при построении гистограммы эмпирического распределения по оси ординат откладывать плотность частости $\frac{n_i}{n\Delta x}$ (где n_i – частота $i-го$ интервала $(i=1,2,...,m)$, Δx – величина интервала, m – число интервалов, n – число наблюдений, объем выборки), то выравнивать такую гистограмму будет теоретическая нормальная кривая с плотностью $\varphi_N(x)=\frac{1}{\sqrt{2\pi\sigma}}e^{-(x-\alpha)^2/2\sigma^2}$, где в качестве параметров a и σ^2 используются их состоятельные выборочные оценки: соответственно средняя $\bar{x}$ и дисперсия $\hat{s}^2$ (либо $s^2\approx\hat{s}^2$ при больших n).

В заключение отметим, что при проверке ряда гипотез, например, гипотез о законе распределения на заданном уровне значимости, контролируется лишь ошибка первого рода, но нельзя сделать вывод о степени риска, связанного с принятием неверной гипотезы, т.е. с возможностью совершения ошибки второго рода.

Применение критерия Колмогорова. На практике кроме критерия χ^2 часто используется критерий Колмогорова, в котором в качестве меры расхождения между теоретическим и эмпирическим распределениями рассматривают максимальное значение абсолютной величины разности между эмпирической функцией распределения $F_n(x)$ и соответствующей теоретической функцией распределения

$$D=\max\left|F_n(x)-F(x)\right|,$$

называемое ***статистикой критерия Колмогорова***.

Доказано, что какова бы ни была функция распределения $F(x)$ непрерывной случайной величины X, при неограниченном увеличении числа наблюдений $(n\to\infty)$ вероятность неравенства $P(D\sqrt{n}\ge\lambda)$ стремится к пределу

$$P(\lambda)=1-\sum_{k=-\infty}^{+\infty}(-1)^k e^{-2k^2\lambda^2}.$$

Задавая уровень значимости α, из соотношения

$$P(\lambda_\alpha)=\alpha$$

можно найти соответствующее критическое значение λ_α. В табл. 9 приводятся критические значения λ_α критерия Колмогорова для некоторых α.

Таблица 9. Критические значения λ_α критерия Колмогорова

Уровень значимости α	0,40	0,30	0,20	0,10	0,05	0,025	0,01	0,005	0,001	0,0005
Критическое значение λ_α	0,89	0,97	1,07	1,22	1,36	1,48	1,63	1,73	1,95	2,03

Схема применения критерия Колмогорова следующая:

1. Строятся эмпирическая функция распределения $F_n(x)$ и предполагаемая теоретическая функция распределения $F(x)$.

2. Определяется мера расхождения D между теоретическим и эмпирическим распределением и вычисляется величина $\lambda = D\sqrt{n}$.

3. Если вычисленное значение λ окажется больше критического λ_α, определенного на уровне значимости α, то нулевая гипотеза H_0 о том, что случайная величина X имеет заданный закон распределения, отвергается. Если $\lambda \le \lambda_\alpha$, то считают, что гипотеза H_0 не противоречит опытным данным.

Пример. По данным табл. 3 с помощью критерия Колмогорова на уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0 о том, что случайная величина X – выработка рабочих предприятия – имеет нормальный закон распределения с параметрами $a = 119{,}2; \sigma^2 = 87{,}48$, т.е. $N(119{,}2;87{,}48)$. Значения эмпирической функции распределения $F_n(x)$, или накопленной частости, вычислены выше в табл. 3, а ее график приведен на рис. 2б. Для построения теоретической функции распределения для нормального закона воспользуемся выражением через функцию Лапласа:

$$F(x) = \frac{1}{2} + \frac{1}{2}\Phi\left(\frac{x - 119{,}2}{9{,}35}\right).$$

Например, $F(94) = \frac{1}{2} + \frac{1}{2}\Phi(\frac{94 - 119{,}2}{9{,}35}) = \frac{1}{2} + \frac{1}{2}\Phi(-2{,}69) = 0{,}5 - 0{,}5 \cdot 0{,}9928 = 0{,}0036 \approx 0{,}004$

и т.д. Результаты вычислений сведем в табл. 10, а график $F(x)$ представим на рис. 8.

Таблица 10

x	94	100	106	112	118	124	130	136	142
$F_n(x)$	0,010	0,030	0,100	0,210	0,410	0,690	0,880	0,980	1,000
$F(x)$	0,004	0,021	0,080	0,221	0,449	0,695	0,878	0,964	0,993

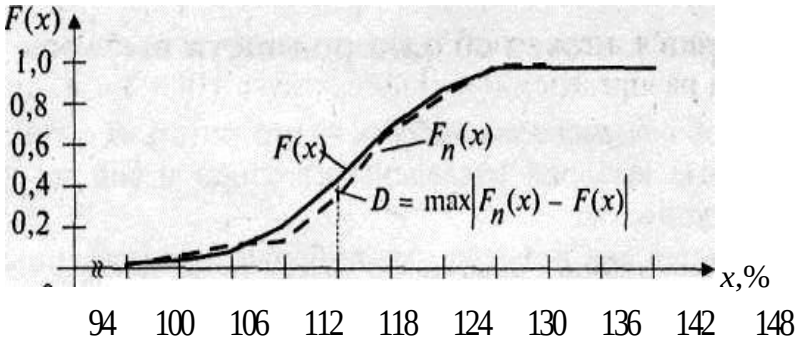


Рис. 8.

Из рис. 8 следует, что

$$D = |F_n(118 - F(118))| = |0{,}410 - 0{,}449| = 0{,}039.$$

Тогда величина $\lambda = D\sqrt{n} = 0{,}039\sqrt{100} = 0{,}39$.

Критическое значение критерия Колмогорова по табл. 9 равно $\lambda_{0,05} = 1{,}36$. Так как $\lambda < \lambda_{0,05}(0{,}39 < 1{,}36)$, то гипотеза H_0 согласуется с опытными данными.

Критерий Колмогорова достаточно часто применяется на практике благодаря своей простоте. Однако в принципе его применение возможно лишь тогда, когда теоретическая функция распределения $F(x)$ задана полностью. Но такой случай на практике встречается весьма редко. Обычно из теоретических соображений известен лишь вид функции распределения, а ее параметры определяются по эмпирическим данным. При применении критерия χ^2 это обстоятельство учитывается соответствующим уменьшением числа степеней свободы. Такого рода поправок в критерии Колмогорова не предусмотрено. Поэтому, если при неизвестных значениях параметров применить критерий Колмогорова, взяв за значения параметров их оценки, то получим завышенное значение вероятности $P(\lambda)$, а значит, большее критическое значение λ_α. В результате есть риск в ряде случаев принять нулевую гипотезу H_0 о законе распределения случайной величины как правдоподобную, в то время как на самом деле она противоречит опытным данным.

Примеры решения задач

1. Сравниваются средние доходы двух фирм *X* и *Y* в двух однотипных отраслях, имеющих нормальное распределение с дисперсиями 1 млн. долларов и 4 млн. долларов соответственно. В первой отрасли по выборке из 20 фирм получен средний доход 1 млн. долларов, а во второй по выборке из 25 фирм получен средний доход 0,9 млн. долларов. Определить, есть ли основание отклонить гипотезу H_0: $MX=MY$ с уровнем значимости $a=0{,}05$.

Решение. В качестве критерия проверки H_0 примем СВ $U_{набл} = \dfrac{\bar{x} - \bar{y}}{\sqrt{\dfrac{\sigma_x^{\,2}}{n} + \dfrac{\sigma_y^{\,2}}{k}}}$.

По условию $\bar{x} = 1; \bar{y} = 0{,}9; \sigma_x^{\,2} = 1; \sigma_y^{\,2} = 4; n = 20; k = 25$, тогда

$$U_{набл} = \frac{1 - 0{,}9}{\sqrt{\dfrac{1}{20} + \dfrac{4}{25}}} = \frac{7}{\sqrt{21}} \approx 0{,}22.$$

При конкурирующей гипотезе H_1: $MX \neq MY$ определяем две критические точки $U_{1-\frac{\alpha}{2}}$ и $U_{\frac{\alpha}{2}}$ из условий $\Phi\left(U_{\frac{\alpha}{2}}\right)=\frac{1-\alpha}{2}=\frac{1-0{,}05}{2}=0{,}475; U_{1-\frac{\alpha}{2}}=U_{\frac{\alpha}{2}}$. Из приложения 2 находим $U_{\frac{\alpha}{2}}=U_{1-\frac{\alpha}{2}}=1{,}95$. Так как $U_{набл.} \approx 0{,}22 < 1{,}95$, то нет основания для отклонения гипотезы H_0: $MX=MY$.

2. Используя критерий Пирсона, при уровне значимости 0,05 проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X с эмпирическим распределением выборки объема $n=200$:

x_i	5	7	9	11	13	15	17	19	21
n_i	15	26	25	30	26	21	24	20	13

Решение. Сначала найдем выборочное среднее $\bar{x}_в$ и выборочное среднеквадратическое отклонение $\sigma_в$. Перейдем к условным вариантам по формуле $v_i=\frac{x_i-x_0}{h}$, где $h=2$, $x_0=13$, тогда $v_i=\frac{x_i-13}{2}$. Все необходимые вычисления приведем в таблице:

x_i	n_i	v_i	$n_i \cdot v_i$	$n_i \cdot v_i^2$
5	15	-4	-60	240
7	26	-3	-78	234
9	25	-2	-50	100
11	30	-1	-30	30
13	26	0	0	0
15	21	1	21	21
17	24	2	48	96
19	20	3	60	180
21	13	4	52	208
Сумма	200		-37	1109

Следовательно, $\bar{v}=\frac{\sum_i m_i \cdot v_i}{n}=\frac{-37}{200}=-0{,}185$, тогда

$\bar{x}=\bar{v}_i \cdot 2+13=-0{,}185 \cdot 2+13=12{,}63$; $\sigma_v^{\,2}=\frac{\sum_i m_i \cdot v_i^{\,2}}{n}-\left(\frac{\sum_i m_i \cdot v_i}{n}\right)^2=\frac{1109}{200}-\left(\frac{-37}{200}\right)^2=5{,}511$;

$\sigma_в^2=\sigma_v^{\,2} \cdot h^2=5{,}511 \cdot 2^2=22{,}043$, тогда $\sigma_в=\sqrt{22{,}044}=4{,}695$. Далее вычислим теоретические частоты по формуле $m_i'=\frac{n \cdot h}{\sigma_в} \cdot \varphi(u_i)=\frac{200 \cdot 2}{4{,}695} \cdot \varphi(u_i)=85{,}2 \cdot \varphi(u_i)$, где $u_i=\frac{(x_i-\bar{x})}{\sigma_в}$. Значение функции $\varphi(u)$ найдем из приложения 1.

Все необходимые вычисления приведены в таблице:

x_i	$u_i = \frac{(x_i - \bar{x})}{\sigma_в}$	$\varphi(u)$	$n_i' = 85,2 \cdot \phi(u_i)$
5	-1,6251	0,1074	9,0056
7	-1,1991	0,1942	16,5458
9	-0,7731	0,2966	25,2703
11	-0,3471	0,3752	31,9670
13	0,0788	0,3977	33,8840
15	0,5047	0,3503	29,9989
17	0,9307	0,2589	22,0582
19	1,3567	0,1582	13,4786
21	1,7827	0,0818	6,9693

Так как $\chi^2_{набл} = \sum \frac{\left(n_i - n_i'\right)^2}{n_i'}$, то производя соответствующие вычисления в таблице

n_i	n_i'	$n_i - n_i'$	$\left(n_i - n_i'\right)^2$	$\frac{\left(n_i - n_i'\right)^2}{n_i'}$
15	9,0056	5,9944	35,9328	3,9900
26	16,5458	9,4542	89,3819	5,4020
25	25,2703	-0,2703	0,07306	0,0028
30	31,9670	-1,967	3,8690	0,1210
26	33,8840	-7,884	62,1574	1,8344
21	29,9989	-8,9989	80,9802	2,6994
24	22,0582	1,9418	3,7705	0,1709
20	13,4786	6,5214	42,5286	3,1552
13	6,9693	6,0307	36,3693	5,2185
Сумма				22,59

получаем $\chi^2_{набл} = 22,59$. Найдем $\chi^2_{кр}$ по уровню значимости α =0,05 и числу степеней свободы $k=s-r-1=9-2-1=6$ из приложения 4 $\chi^2_{кр}(0,05;6) = 12,6$. Так как $\chi^2_{набл} > \chi^2_{кр}$, то гипотезу о нормальном распределении генеральной совокупности отвергаем, т.е. эмпирические и теоретические частоты различаются значимо.

3. Распределение СВ X задается следующим вариационным рядом:

Интервалы СВ X	m_i
5 - 10	2
10 - 15	14
15 - 20	11
20 - 25	9
25 - 30	4

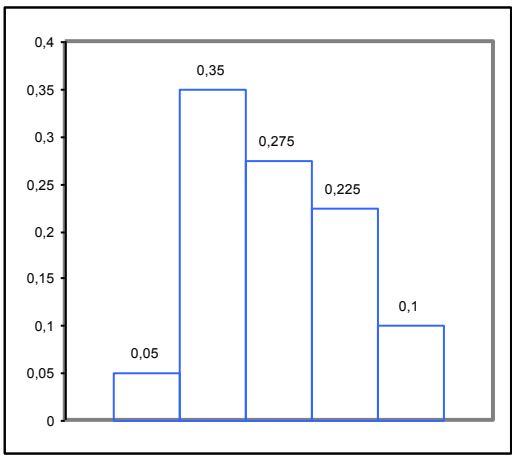
Выдвинуть гипотезу о виде распределения этой случайной величины. Используя критерий Пирсона, при уровне значимости 0,05 проверить, согласуется ли выдвинутая гипотеза с эмпирическим распределением выборки.

Решение. Найдем объем выборки *n*: $\sum_{i=1}^{l} m_i = n \quad \Rightarrow \quad n = 2+14+11+9+4 = 40.$

Составим следующую таблицу:

x	m_i	$p_i = m_i / n$	Середины интервалов
5-10	2	0,05	7,50
10-15	14	0,35	12,5
15-20	11	0,275	17,5
20-25	9	0,225	22,5
25-30	4	0,1	27,5
	40		

Построим полигон полученного распределения:



Используя таблицу, найдем выборочное среднее значение дисперсию и среднее квадратическое отклонение интервального вариационного ряда.

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{5} x_i m_i = \frac{695}{40} = 17{,}375;$$

$$\overline{X^2} = \frac{1}{n}\sum_{i=1}^{5} x_i^2 m_i = \frac{13250}{40} = 331{,}25$$

$$\sigma^2 = \overline{X^2} - \left(\overline{X}\right)^2 = 331{,}25 - 17{,}375^2 = 29{,}36;$$

$$\sigma = \sqrt{29{,}36} = 5{,}42$$

По виду полигона частот выдвинем гипотезу, что случайная величина распределена по нормальному закону. Проверим эту гипотезу с помощью критерия χ^2 Пирсона при уровне значимости $\alpha = 0{,}05$. Будем считать, что случайная величина X распределена по нормальному закону распределения с функцией распределения $F(x) = \frac{1}{2}\left(1 + \Phi\left(\frac{x - 17{,}375}{5{,}42}\right)\right)$. По формуле $P(x_1 < X < x_2) = \frac{1}{2}\left(\Phi\left(\frac{x_2 - a}{\sigma}\right) - \Phi\left(\frac{x_1 - a}{\sigma}\right)\right)$ вычислим вероятность $p_i \quad \left(i = \overline{1,5}\right)$ того, что случайная величина содержится в интервале (x_i; x_{i+1}), а затем по формуле

$m_i' \approx n \cdot p_i$ найдем соответствующую теоретическую частоту случайной величины X в этом интервале.

	$P(x_1 < X < x_2)$	$m_i' \approx n \cdot p_i$	p_i'
P(5<X<10)=	0,0755952	$m_1' \approx 3$	3,0238073
P(10<X<15)=	0,2438195	$m_2' \approx 10$	9,7527793
P(15<X<20)=	0,355296	$m_3' \approx 14$	14,211842
P(20<X<25)=	0,2343409	$m_4' \approx 9$	9,373635
P(25<X<30)=	0,0698188	$m_5' \approx 3$	2,7927537

Проверим степень согласия теоретического и эмпирического распределений с помощью критерия Пирсона. Для этого заполним таблицу:

m_i	m_i'	$m_i - m_i'$	$\left(m_i - m_i'\right)^2$	$\frac{\left(m_i - m_i'\right)^2}{m_i'}$
2	3	1	1	0,3333333
14	10	-4	16	1,6
11	14	3	9	0,6428571
9	9	0	0	0
4	3	-1	1	0,3333333
			СУММА	2,9095238

$$\chi_{набл}^2 = \sum \frac{\left(m_i - m_i'\right)^2}{m_i'} = 2{,}9095238$$

Итак, $\chi_{набл}^2 = 2{,}91$. Найдем число степеней свободы: $5-2=3$. Находим области при уровне значимости 0,05 $\chi^2_{прак} = 7{,}82$. Так как $\chi_{набл}^2 < \chi_{кр}^2$, то нет оснований отвергать гипотезу о нормальном распределении СВ X.

Задачи для самостоятельного решения

1. По двум независимым выборкам, объемы которых n=40 и m=50, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние: $\bar{x} = 130$ и $\bar{y} = 140$. Генеральные дисперсии известны: DX=80, DY=100. Требуется при уровне значимости 0,01 проверить нулевую гипотезу H_0: $MX=MY$ при конкурирующей гипотезе H_1: $MX \neq MY$.

2. По выборке объема n=100, извлеченной из двумерной нормальной генеральной совокупности, найдем выборочный коэффициент корреляции $r_в$=0,2. Требуется при уровне значимости 0,05 проверить нулевую гипотезу о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе H_1: $r_г \neq 0$.

3. Используя критерий Пирсона, при уровне значимости 0,05 установить, случайно или значимо расхождение между эмпирическими частотами m_i и теоретическими частотами m_i', которые вычислены исходя из гипотезы о нормальном распределении генеральной совокупности:

m_i	5	10	20	8	7
m_i'	6	14	18	7	5

4. По выборке объема n=30 найден средний вес $\bar{x}$=130г изделий, изготовленных на первом станке; по выборке объема m=40 найден средний вес $\bar{y}$=125г изделий, изготовленных на втором станке. Генеральные дисперсии известны: DX=60г^2, DY=80г^2. Требуется, при уровне значимости 0,05 проверить нулевую гипотезу H_0: $MX=MY$ при конкурирующей гипотезе $MX \neq MY$. Предполагается, что случайные величины X и Y распределены нормально, а выборки независимы.

5. По выборке объема n=62, извлеченной из нормальной двухмерной генеральной совокупности, найден выборочный коэффициент корреляции $r_в$=0,3. Требуется при уровне значимости 0,05 проверить нулевую гипотезу о равенстве нулю генерального коэффициента корреляции при конкурирующей гипотезе H_1: $r_г \neq 0$.

6. Используя критерий Пирсона, при уровне значимости 0,05 проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X с эмпирическим распределением:

$x_i - x_i'$	(-20)–(-10)	(-10)–0	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
Частота m_i	20	47	80	89	40	16	8

Вопросы для самоконтроля

1. Что такое статистическая гипотеза?
2. Какова цель проверки гипотез?
3. Что такое нулевая и альтернативная гипотезы?
4. Приведите общую схему проверки гипотез.
5. Что такое ошибки первого и второго рода?
6. Что такое уровень значимости?

4. ЭЛЕМЕНТЫ РЕГРЕССИОННОГО И КОРРЕЛЯЦИОННОГО АНАЛИЗА

В естественных науках часто речь идет о функциональной зависимости, когда каждому значению одной переменной соответствует вполне определенное значение другой (например, скорость падения тела в вакууме в зависимости от времени и т.п.).

В социологии в большинстве случаев между переменными величинами существуют зависимости, когда каждому значению одной переменной соответствует не какое-то определенное, а множество возможных значений другой переменной. Иначе говоря, каждому значению одной переменной соответствует определенное (условное) распределение другой переменной. Такая зависимость получила название ***статистической (стохастической)***.

Возникновение понятия статистической связи обуславливается тем, что зависимая переменная подвержена влиянию ряда неконтролируемых или неучтенных факторов, а также тем, что измерение значений переменных неизбежно сопровождается некоторыми случайными ошибками. Примером статистической связи является зависимость урожайности от количества внесенных удобрений, производительности труда на предприятии от его энерговооруженности и т.п.

В силу неоднозначности статистической зависимости между X и Y для исследователя, в частности, представляет интерес усредненная по x схема зависимости, т.е. закономерность в изменении условного математического ожидания $M_{X=x}(Y)$ (математического ожидания случайной переменной Y, вычисленного в предположении, что переменная X приняла значение x) в зависимости от x.

4.1. Выборочные уравнения регрессии

Корреляционной зависимостью между двумя переменными величинами X и Y называется функциональная зависимость между значениями одной из них и условным математическим ожиданием другой.

Корреляционная зависимость может быть представлена в виде:

$M_{X=x}(Y)=\varphi(x)$ или $M_{Y=y}(X)=\psi(y)$

где $\varphi(x)\neq const$, $\psi(y)\neq const$.

Эти уравнения называются ***модельными уравнениями регрессии*** (или просто ***уравнениями регрессии***) соответственно Y по X и X по Y, функции $\varphi(x)$ и $\psi(y)$ – ***модельными функциями регрессии*** (или ***функциями регрессии***), а их графики – ***модельными линиями регрессии*** (или ***линиями регрессии***).

Для отыскания модельных уравнений регрессии, вообще говоря, необходимо знать закон распределения двумерной случайной величины (X, Y). На практике исследователь, как правило, располагает лишь выборкой пар значений (x_i, y_i) ограниченного объема. В этом случае речь может идти об оценке (приближенном выражении) по выборке функции регрессии. Такой наилучшей (в смысле метода наименьших квадратов) оценкой является выборочная линия (кривая) регрессии Y по X: $y=\varphi(x,b_0,b_1,...b_p)$, где x –

фиксированное значение случайной величины *X*, *y* – соответствующее значение переменной *У*; $b_0, b_1, ...b_p$ – параметры кривой; аналогично определяется выборочная линия (кривая) регрессии *X* по *У*: $x = \psi(y, c_0, c_1, ...c_p)$. Параметры должны быть выбраны так, чтобы данные кривые проходили вблизи соответствующих кривых регрессии.

Эти уравнения называют также ***выборочными уравнениями регрессии*** соответственно *У* по *X* и *X* по *У*.

Основное уравнение линейной регрессии выглядит следующим образом:

$$y = ax + b,$$

где *Y* – изучаемый признак, переменная, которая испытывает на себе влияние другой переменной;

X – переменная, оказывающая влияние на переменную *Y*;

a – коэффициент регрессии, определяющий наклон линии регрессии по отношению к осям *X* и *Y*;

b – константа, определяющая высоту линии регрессии над осью *X*.

По сути, это уравнение прямой в декартовой системе координат и решение уравнения регрессии сводится к нахождению коэффициента регрессии и свободного члена в уравнении регрессии, которые определяются по формулам:

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2};$$

$$b = \overline{y} - a\overline{x}.$$

Аналогичным образом можно построить уравнение зависимости не только *Y* от *X*, но и *X* от *Y*, что требует замены переменных местами. Однако, статистика не может заменить собой логики, и поэтому математически такая перестановка осуществима, но логически может быть совершенно не оправдана. Можно, например, изучать зависимость успеваемости от интеллекта у школьников первого класса, но вряд ли целесообразно будет изучение зависимости интеллекта от успеваемости у тех же школьников, если интеллект у них формировался задолго до того, как появилась успеваемость. Вопрос о причинно-следственных связях в таких случаях всегда остается на совести исследователя.

Статистические связи между переменными можно изучать методами корреляционного и регрессионного анализа.

Регрессионный анализ – это статистический метод изучения изменения значений одной переменной от изменения значений другой переменной на единицу измерения. В широком смысле регрессионный анализ изучает связь между переменными, зависимость одной переменной от другой и влияние одной переменной на другую. Регрессионный анализ дает ответ на один очень важный вопрос: как изменится значение одной переменной, если значение

другой переменной изменилось на некоторое количество единиц ее измерения. Такого рода задача может возникнуть в том случае, если необходимо знать какой тестовый балл окажется у испытуемого по тесту А, если нам известен его тестовый балл по тесту Б и насколько возрастет (уменьшится) тестовый балл данного испытуемого по одному тесту, если изменится тестовый балл по другому. В качестве еще одного примера применения регрессионного анализа можно привести следующие **примеры**:

1. Как изменится мотивация персонала фирмы, если зарплата будет увеличена на определенное количество денежных единиц.
2. Насколько изменится спрос на товар, если общее время показа рекламы по телевидению увеличится на определенное количество минут.
3. Насколько точно можно оценить успеваемость по интеллекту.
4. Как изменится самооценка подростка, если его социометрический статус возрастет.
5. Как зависит оценка студента на экзамене от успеваемости в течение семестра.

Правда, нужно отметить, что для решения такого рода задач необходима предварительная статистика, т.е. исследователь должен располагать данными измерений двух случайных величин, зависимость одной из которых от другой он исследует.

Замечание. *Основной задачей регрессионного анализа является установление формы и изучение зависимости между переменными.*

Корреляционный анализ (correlation analysis) – статистический метод изучения взаимосвязи между двумя и более случайными величинами. В качестве случайных величин в эмпирических исследованиях выступают значения переменных, измеряемые свойства исследуемых объектов наблюдения. Суть корреляционного анализа заключается в расчете коэффициентов корреляции. Коэффициенты корреляции могут принимать, как правило, положительные и отрицательные значения. Знак коэффициента корреляции позволяет интерпретировать направление связи, а абсолютное значение – силу связи.

Способ расчета коэффициентов корреляции зависит от шкал измерения переменных, между которыми исследуется взаимосвязь. Для переменных, измеряемых в количественной шкале (интервальной шкале или шкале отношений), рассчитывают ковариацию или корреляционный момент, а на его основе линейный коэффициент корреляции (коэффициент корреляции Пирсона).

Для оценки силы и направления связи между переменными, измеренными в порядковой шкале, используются непараметрические ранговые коэффициенты корреляции: коэффициент ранговой корреляции Кендалла и коэффициент корреляции Спирмена. Также часто используют коэффициент корреляции знаков Фехнера, коэффициент множественной ранговой корреляции (коэффициент Конкордации). Существуют меры оценки связи и между дихотомическими переменными.

Корреляционный анализ используется в экономике, социологии и психологии, медицине, управления качеством, биометрии и других сферах. Популярность корреляционного анализа объясняется тем, что коэффициенты корреляции относительно просты в расчете, и их применение не требует специальной математической подготовки. С другой стороны – коэффициенты корреляции легко интерпретировать.

Однако корреляционный анализ имеет свою специфику и методику. Очень важно использование этого метода только при соблюдении предпосылок расчета того, или иного, коэффициента корреляции. Методика корреляционного анализа предполагает, не просто расчет коэффициентов корреляции, но и обязательную проверку их значимости, в основе которой лежит принцип проверки статистических гипотез, построение интервальных оценок коэффициентов корреляции.

Нередки случаи возникновения так называемых «ложных корреляций», приводящим к ложным выводам. В этом случае при анализе взаимосвязи между количественными переменными рассчитывают и анализируют частные коэффициенты корреляции.

Замечание. *Основной задачей корреляционного анализа – выявление связи между случайными переменными и оценка ее тесноты. Корреляционный анализ не позволяет определить форму связи между переменными и предсказывать значения одной зависимой переменной по одной или нескольким независимым. Для этого, например, для количественных переменных применяется, как говорилось ранее, регрессионный анализ.*

Данные о статистической зависимости удобно задавать в виде корреляционной таблицы, в клетках которой записываются частоты всевозможных сочетаний значений признаков *X* и *У*.

4.2. Коэффициент линейной корреляции и его свойства

Особенности коэффициента корреляции. Коэффициент корреляции показывает сразу два параметра статистической связи – ее направление и тесноту. Направление связи может быть положительным, когда большему значению одной переменной соответствует большее значение другой переменной и отрицательным, когда большему одной переменной соответствует меньшее значение другой переменной. Коэффициент корреляции всегда находится в пределах от – 1 до +1. При этом, если он оказывается положительным, то говорят о положительной корреляции между двумя переменными, а если отрицательным – то, соответственно об отрицательной. Абсолютное значение коэффициента корреляции показывает тесноту или степень выраженности такой связи. При коэффициенте корреляции равном нулю признается отсутствие связи, но даже тогда, когда он оказывается больше нуля, еще не следует делать вывод о наличии корреляционной связи. О связи между двумя переменными можно говорить лишь в том случае, если значение коэффициента корреляции оказывается выше критического для соответствующего числа наблюдений, если речь идет о положительной связи, и ниже критического, если об отрицательной.

Необходимо подчеркнуть, что коэффициент корреляции предназначен лишь для измерения линейных связей между переменными. По этой причине в реальных условиях почти невозможно получить коэффициент корреляции равный единице. Например, если рассчитать коэффициент корреляции между расстоянием планет Солнечной системы от Солнца и их периодом обращения, то коэффициент корреляции окажется равным 0.998, несмотря на то, что связь здесь прямая: чем дальше планета удалена от Солнца, тем больше ее период обращения. Причина этого заключается в том, что связь между расстоянием от Солнца и периодом обращения для планет Солнечной системы на графике отображается не прямой, а слегка изогнутой линией, следуя известным законам небесной механики И. Кеплера.

Что касается социологических измерений, то здесь коэффициент корреляции равный 0,8–0,9 признается достаточно высоким, а связь статистически значимой (достоверной) даже для небольшого числа наблюдений. Например, если при первичном и повторном тестировании большая часть испытуемых показала один и тот же результат по тесту *X*, и коэффициент корреляции оказался в указанных пределах, то тест может быть признан надежным, несмотря на то, что у части испытуемых результат повторного тестирования отличался от первичного. В реальных экспериментальных условиях наличие небольшого разброса данных может свидетельствовать не об отсутствии связи, а о некоторой ошибке измерения, или влиянии неучтеного фактора на исход эксперимента.

Виды коэффициентов корреляции. Наиболее известным и часто применяемым в социологических исследованиях является коэффициент корреляции К. Пирсона для двух переменных, измеренных в шкалах интервалов или отношений:

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

Вторым часто используемым в социологии коэффициентом корреляции является коэффициент ранговой корреляции Ч. Спирмена, который обозначается греческой буквой "ρ" (ро):

$$\rho = 1 - \frac{6\sum (R_x - R_y)^2}{n^3 - n},$$

где $(R_x - R_y)^2$ – квадрат разности между соответствующими парами рангов.

Он предназначен для определения связи между двумя переменными, измеренными в шкалах порядка. Достоинством ρ-Спирмена является то, что он нетруден в вычислениях и применим для первичной оценки связи, так как множество переменных легко поддается ранжированию. Однако, оценка такой связи будет более грубая, чем при применении коэффициента корреляции Пирсона, так как при переходе от шкалы более высокого порядка к шкале более низкого порядка информативность данных снижается.

Особенности интерпретации коэффициента корреляции. В отношении коэффициента корреляции рядом авторов часто употребляется понятие

зависимости между переменными. В действительности, говоря о корреляции можно говорить лишь о статистической связи. Например, если обнаруживается положительная корреляция между успехами учеников по математике и английскому языку, то из этого не следует, что оценки по одному предмету зависят от оценок по другому, так как они выставляются независимо друг от друга. Скорее всего, за всеми этими оценками стоят факторы интеллекта и мотивации, проявлениями которых и являются успехи по учебным предметам. Также неправомерно, в таких случаях, говорить о причинной связи между двумя переменными, если коэффициент корреляции оказывается высоким. Связь между уровнем дохода в семье и величиной IQ у детей вполне может оказаться достоверной, так как дети из обеспеченных семей имеют больше шансов на получение хорошего образования, но из этого не следует, что количество денег положительно влияет на умственные способности. Статистические методы не могут заменить собой логику и здравый смысл, и констатация причинной связи или зависимости на основе вычислений коэффициента корреляции лежит исключительно на совести исследователя.

При интерпретации нулевого значения коэффициента корреляции необходимо учитывать, что ноль не всегда означает отсутствие связи. Если связь между переменными носит нелинейный характер и на графике отображается кривой, то коэффициент корреляции получится близким или равным нулю несмотря на очевидный характер связи. Действительное отсутствие связи на графике будет отображаться множеством рассеянных точек.

Сфера применения коэффициента корреляции. Использование коэффициентов корреляции в социально-гуманитарных исследованиях насчитывает уже почти столетнюю историю, и в основном они применяются в следующих случаях:

1. Для проверки гипотезы о связи различных явлений и переменных: социальных и социально-психологических, социально-психологических и психологических, психических и психофизиологических, психофизиологических и физиологических. Результаты таких исследований помогают составить системную картину психических явлений и явлений окружающего мира.
2. В психодиагностике для определения надежности и валидности теста, при создании и адаптации психологических методик.
3. В методе репертуарных решеток Келли для определения связей между конструктами индивидуального сознания.
4. В факторном анализе – методе исследования латентной структуры сложных психологических явлений и переменных, таких как интеллект, личность и т.д.

Как говорилось ранее, основное уравнение линейной регрессии имеет вид

$$y = ax + b,$$

коэффициенты *a* и *b* которого удобнее высчитывать по формулам:

$$a = r_{xy}\frac{\sigma_y}{\sigma_x},\ \ b = \bar{y} - a\bar{x}.$$

Проверка значимости коэффициента линейной корреляции. Поскольку коэффициент корреляции r_{xy} вычисляется по значениям переменных, случайно попавшим в выборку из генеральной совокупности, то возникает вопрос, объясняется ли это действительно существенной линейной корреляционной связью между переменными X и Y в генеральной совокупности или является следствием случайности отбора между переменными в выборку.

В практических исследованиях обычно проверяется гипотеза H_0 об отсутствии линейной корреляционной связи между переменными в генеральной совокупности, т.е. H_0: $r_{xy}=0$. При справедливости этой гипотезы статистика

$$T_{набл} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^{\ 2}}}$$

имеет t-распределение Стьюдента с $k=n-2$ степенями свободы. Поэтому гипотеза H_0 отвергается, т.е. коэффициент корреляции значимо (существенно) отличается от нуля, если

$$T_{набл} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^{\ 2}}} > t_{1-\alpha;k},$$

где $t_{1-\alpha;k}$ – табличное значение t-критерия Стьюдента, определенное на уровне значимости α при числе степеней свободы $k=n-2$.

Пример. Определяется наличие линейной зависимости между уровнем инфляции X и безработицы Y в стране за последние 11 лет. По статистическим данным найден выборочный коэффициент корреляции $r_в=0{,}34$. Существует ли значимая линейная связь между указанными показателями в стране на интервале в 11 лет при уровне значимости $\alpha=0{,}05$.

Решение. Требуется при уровне значимости $\alpha=0{,}05$ проверить нулевую гипотезу о равенстве нулю коэффициента корреляции при конкурирующей гипотезе H_1: $r_г \neq 0$

Найдем наблюдаемое значение критерия:

$$T_{набл} = \frac{r_в\sqrt{n-2}}{\sqrt{1-r_в^{\ 2}}} = \frac{0{,}34\sqrt{11-2}}{\sqrt{1-(0{,}34)^2}} = 3{,}254.$$

По таблице критических точек Стьюдента (приложение 3), по уровню значимости $\alpha=0{,}05$ и числу степеней свободы $k=n-2=11-2=9$ находим критическую точку $t_{кр}(0{,}05;\ 9)=2{,}26$.

Так как $T_{набл.}>t_{кр}$, то приходится отвергнуть нулевую гипотезу о равенстве нулю коэффициента корреляции, т.е. коэффициент корреляции $r_в$ значимо отличается от нуля и между уровнями инфляции X и безработицы Y существует линейная зависимость.

4.3. Лабораторная работа 3. Корреляционный и регрессионный анализ

Цель: *научиться проводить корреляционный и регрессионный анализ в Excel.*

Задание 1. В восьми районах собраны сведения о числе обращений граждан в поликлинику с сердечно-сосудистыми заболеваниями за год, численности населения (тыс. чел.) и размере среднего ежемесячного дохода на душу населения (у. е.):

Число заболеваний	300	133	100	200	120	270	120	260
Доход	180	200	330	180	310	180	310	200

Требуется определить, имеется ли связь между этими величинами.

Решение.

1. На новом листе создайте следующую таблицу, начиная с ячейки *A1*.

	A	B	C	D	E	F	G	H	I
1	Число заболеваний	300	133	100	200	120	270	120	260
2	Доход	180	200	330	180	310	180	310	200

2. Вычислите значение коэффициента корреляции между величинами. Для этого курсор установите на свободную ячейку *A3*. На панели инструментов необходимо нажать кнопку *Вставка функции*. В появившемся диалоговом окне *Мастер функций* выберите категорию *Статистические* и функцию КОРРЕЛ, после чего нажмите ОК. Указателем мыши введите диапазон данных *Число заболеваний* в поле *Массив 1* (B1:I1). В поле *Массив 2* введите диапазон данных *Доход* (B2:I2). Нажмите ОК. В ячейке A3 появится значение коэффициента корреляции.

3. Проверьте гипотезу о значимости коэффициента корреляции (α=0,05). Для этого в ячейку A4 введите формулу: =(ABS(A3)*КОРЕНЬ(8-2))/(КОРЕНЬ(1-(A3*A3))) (используйте кнопку *Вставка функции*). В ячейку B4 вставьте функцию СТЬЮДРАСПОБР (используйте кнопку *Вставка функции*). В поле *Вероятность* введите 0,05, а в поле *Степени свободы* введите значение равное $k=n-2$, где n – объем выборки, т.е. 6. В ячейке B4 вы получите критическое значение t-критерия Стьюдента — 2,446913641.

4. Сравните значение ячейки A4 и критическое значение и сделайте вывод.

5. Сохраните документ на диске D (Имя папки – номер группы) под именем *Корреляция*.

Задание 2. Имеются ежемесячные данные наблюдений за состоянием погоды и посещаемостью музеев и парков.

Число ясных дней	Количество посетителей музея	Количество посетителей парка
8	495	132
14	503	348
20	380	643
25	305	765
20	348	743
15	465	541

Необходимо определить, существует ли взаимосвязь между состоянием погоды и посещаемостью музеев и парков.

Решение.

1. Перейдите на второй лист рабочей книги.
2. Для выполнения корреляционного анализа введите в диапазон A1:G3 исходные данные.

	A	B	C	D	E	F	G
1	Ясные дни	8	14	20	25	20	15
2	Посещаемость музея	495	503	380	305	348	465
3	Посещаемость парка	132	348	643	765	743	541

3. В меню *Сервис*, выберите команду *Анализ данных* и далее в появившемся списке *Инструменты анализа* выберите строку *Корреляция*. В появившемся диалоговом окне укажите *Входной интервал* B1:G3. Укажите, что данные рассматриваются по строкам. Укажите выходной диапазон. Для этого поставьте флажок в левое поле *Выходной интервал* и введите A4. Нажмите ОК. В результате вы получите корреляционную матрицу.
4. *Интерпретация результатов*. Корреляция между состоянием погоды и посещаемостью музея равна –0,92, а между состоянием погоды и посещаемостью парка – 0,95, между посещаемостью парка и музея $r = -0.89$.

Таким образом, в результате выявлены зависимости: сильная степень обратной линейной взаимосвязи между посещаемостью музея и количеством солнечных дней ($r = -0.92$) и практически линейная (очень сильная прямая) связь между посещаемостью парка и состоянием погоды ($r = 0.95$). Между посещаемостью музея и парка также имеется сильная обратная взаимосвязь ($r = -0.89$).

5. Сохраните документ.

Частная корреляция

Задание 3. В условиях предыдущей задачи психолога опять интересуют три вопроса: в какой степени тактичность (переменная *X*) связана с требовательностью (переменная *Y*), при условии того, что критичность (переменная *Z*) при этом остается неизменной; в какой степени тактичность связана с критичностью при условии того, что требовательность остается неизменной; в какой степени требовательность связана с критичностью, при условии того, что тактичность остается неизменной?

Решение.

На эти вопросы может ответить вычисление коэффициентов частной корреляции по формулам:

$$r_{xy(z)} = \frac{r_{xy} - r_{xz} \cdot r_{yz}}{\sqrt{\left(1 - r_{xz}^{\,2}\right) \cdot \left(1 - r_{yz}^{\,2}\right)}},\ r_{xz(y)} = \frac{r_{xz} - r_{xy} \cdot r_{yz}}{\sqrt{\left(1 - r_{xy}^{\,2}\right) \cdot \left(1 - r_{yz}^{\,2}\right)}},$$

$$r_{yz(x)} = \frac{r_{yz} - r_{xy} \cdot r_{xz}}{\sqrt{\left(1 - r_{xy}^{\,2}\right) \cdot \left(1 - r_{xz}^{\,2}\right)}}.$$

1. Перейдите на четвертый лист рабочей книги.

2. Выделите данные в диапазоне *A1:J12* на *листе 3* и скопируйте их на *лист 4*.

	A	B	C	D	E	F	G	H	I	J
1	№ испытуемого	Тактичность X	Требовательность Y	Критичность Z	X*X	Y*Y	Z*Z	X*Y	Y*Z	X*Z
2	1	70	18	36	4900	324	1296	1260	648	2520
3	2	60	17	29	3600	289	841	1020	493	1740
4	3	70	22	40	4900	484	1600	1540	880	2800
5	4	46	10	12	2116	100	144	460	120	552
6	5	58	16	31	3364	256	961	928	496	1798
7	6	69	18	32	4761	324	1024	1242	576	2208
8	7	32	9	13	1024	81	169	288	117	416
9	8	62	18	35	3844	324	1225	1116	630	2170
10	9	46	15	30	2116	225	900	690	450	1380
11	10	62	22	36	3844	484	1296	1364	792	2232
12	Сумма	575	165	294	34469	2891	9456	9908	5202	17816

3. Вычислите r_{xy}, r_{xz} и r_{yz}. Для этого, в отличие от предыдущей задачи, воспользуемся пакетом *Анализ данных*. В меню *Сервис*, выберите команду *Анализ данных* и далее в появившемся списке *Инструменты анализа* выберите строку *Корреляция*. В появившемся диалоговом окне укажите *Входной интервал* B2:D11. Укажите, что данные рассматриваются по столбцам. Укажите выходной диапазон. Для этого поставьте флажок в левое поле *Выходной интервал* и введите A14. Нажмите ОК. В результате вы получите корреляционную матрицу.

	Столбец 1	*Столбец 2*	*Столбец 3*
Столбец 1	1		
Столбец 2	0,863766147	1	
Столбец 3	0,852243213	0,948685377	1

Таким образом, r_{xy} =0,863766147, r_{xz} =0,852243213, а r_{yz} =0,948685377.

4. Для ответа на первый вопрос задачи рассчитайте частный коэффициент корреляции $r_{xy(z)}$, т.е. в ячейку A19 введите формулу =(B16-B17*C17)/КОРЕНЬ((1-B17*B17)*(1-C17*C17)) (используйте кнопку *Вставка функции*). Для ответа на второй вопрос вычислите $r_{xz(y)}$, т.е. в ячейку A20 введите формулу

=(B17-B16*C17)/КОРЕНЬ((1-B16*B16)*(1-C17*C17)).

Аналогично, для ответа на третий вопрос в ячейку A21 введите =(C17-B16*B17)/КОРЕНЬ((1-B16*B16)*(1-B17*B17)).

5. Проверьте значимость частных коэффициентов корреляции. Сначала проверим на значимость первый коэффициент. Для этого в ячейку B19 введите формулу =A19*КОРЕНЬ((10-2)/(1-A19*A19)) и скопируйте ее (растяните) в диапазон B20:B21. Далее определите критические значения, используя значения t-критерия Стьюдента. Для этого установите курсор в ячейку C20 и введите (используйте кнопку *Вставка функции*) формулу

=СТЬЮДРАСПОБР(0,05;8), а в ячейку C21 введите формулу

=СТЬЮДРАСПОБР(0,01;8).

6. Проанализируйте результаты.
7. Сохраните документ.

Регрессионный анализ

Задание 4. В условиях предыдущей задачи психолога интересует вопрос: при увеличении величины экспертных баллов на 1 при оценке тактичности, на

какую величину экспертных баллов увеличится или уменьшится экспертная оценка требовательности и критичности? Иными словами, требуется построить уравнение множественной регрессии вида: $X = a_0 + a_1 Y + a_2 Z$.

Решение.

1. Перейдите на пятый лист рабочей книги.
2. Выделите данные в диапазоне *A1:D12* на *листе 3* и скопируйте их на *лист 5*.
3. В меню *Сервис*, выберите команду *Анализ данных* и далее в появившемся списке *Инструменты анализа* выберите строку *Регрессия*. Нажмите ОК.
4. В появившемся диалоговом окне задайте *Входной интервал Y* – B2:B11. Так же укажите *Входной интервал X* — C2:D11.
5. Установите флажок в поле *График подбора*.
6. Далее укажите выходной интервал. Для этого поставьте переключатель в положение *Выходной интервал,* затем наведите указатель мыши на правое поле ввода *Выходной интервал* и, установите курсор на ячейку A14. Нажмите ОК.
7. В выходном диапазоне появятся результаты регрессионного анализа и графики предсказанных точек.
8. *Интерпретация результатов*. В таблице *Дисперсионный анализ* оценивается достоверность полученной модели по уровню значимости критерия Фишера (столбец *Значимость F* = 0,007088452), т.е. это значение меньше 0,05 и модель значима. Степень описания моделью процесса – *R-квадрат* равен 0,756851372, что говорит о средней точности аппроксимации (модель довольно хорошо описывает зависимость тактичности от требовательности и критичности).

Далее необходимо определить значения коэффициентов модели. Они определяются из таблицы. В столбце *Коэффициенты* – в строке *Y*-пересечение приводится свободный член $a_0 = 18{,}46892259$; в строках соответствующих переменных приводятся значения коэффициентов при этих переменных $a_1 = 1{,}596475848$ и $a_2 = 0{,}431606324$.

В столбце *P-Значение* приводится достоверность отличия соответствующих коэффициентов от нуля. В данном случае все коэффициенты не значимы (т.е. *P-Значение* >0,05).

Таким образом, уравнение множественной регрессии будет иметь вид: $X = 18.47 + 1.59 \cdot Y + 0.43 \cdot Z$.Сохраните документ.

Примеры решения задач

1. Туристическая компания предлагает места в гостиницах приморского курорта. Менеджера компании интересует, насколько возрастает привлекательность компании в зависимости от ее расстояния до пляжа. С этой целью по 10 гостиницам города была выяснена среднегодовая наполняемость номеров и расстояние в километрах до пляжа:

Расстояние (*X*)	0,1	0,1	0,2	0,2	0,3	0,4	0,5	0,6	0,7	0,9
Наполняемость в % (*У*)	91	98	95	89	82	80	79	75	73	72

Найти выборочный коэффициент корреляции и напишите уравнения прямых линейной регрессии *У* на *Х* и *Х* на *У*. Проанализируйте полученные результаты.

Решение. Вычислим числовые характеристики выборки:

$$\bar{x} = \frac{0,1+0,1+0,2+0,2+0,3+0,4+0,5+0,6+0,7+0,9}{10} = 0,4;$$

$$\overline{x^2} = \frac{0,1^2+0,1^2+0,2^2+0,2^2+0,3^2+0,4^2+0,5^2+0,6^2+0,7^2+0,9^2}{10} = 0,226;$$

$$\sigma_x{}^2 = \overline{x^2} - \left(\bar{x}\right)^2 = 0,226 - 0,4^2 = 0,066;$$

$$\sigma_x = \sqrt{0,066} \approx 0,257;$$

$$\bar{y} = \frac{91+98+95+89+82+80+79+75+73+72}{10} = 83,4;$$

$$\overline{y^2} = \frac{91^2+98^2+95^2+89^2+82^2+80^2+79^2+75^2+73^2+72^2}{10} = 7033,4;$$

$$\sigma_y{}^2 = \overline{y^2} - \left(\bar{y}\right)^2 = 7033,4 - 83,4^2 = 77,84;$$

$$\sigma_y = \sqrt{77,84} \approx 8,823;$$

$$\overline{xy} = \frac{0,1\cdot 91+0,1\cdot 98+0,2\cdot 95+0,2\cdot 89+0,3\cdot 82+0,4\cdot 80+}{10}$$

$$\frac{+0,5\cdot 79+0,6\cdot 75+0,7\cdot 73+0,9\cdot 72}{} = 31,27.$$

Находим выборочный коэффициент корреляции:

$$r_{xy} = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\sigma_x \cdot \sigma_y} = \frac{31,27 - 0,4\cdot 83,4}{0,257\cdot 8,823} \approx -0,92.$$

Формула прямой регрессии *У* на *Х* имеет вид:

$$y - 83,4 = -0,92 \cdot \frac{8,823}{0,257}(x - 0,4);$$

$$y = -31,6x + 96,03.$$

Формула прямой регрессии *Х* на *У* имеет вид:

$$x - 0,4 = -0,92 \cdot \frac{0,257}{8,823}(y - 83,4);$$

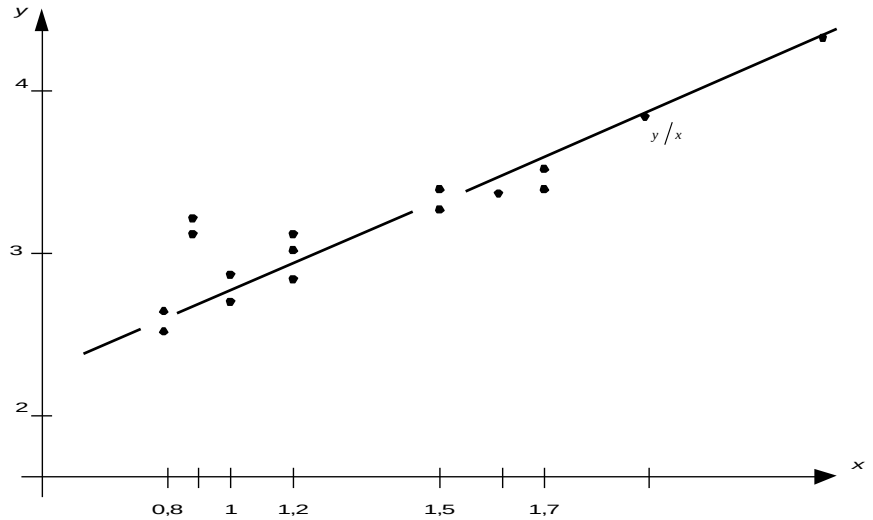
$$x - 0,4 = -0,0268y + 2,635.$$

2. Туристическая компания предлагает места в гостиницах приморского курорта. Менеджера компании интересует, насколько возрастает привлекательность компании в зависимости от ее расстояния до пляжа. С этой целью по 14 гостиницам города была выяснена среднегодовая стоимость одного номера (*У*, тыс. руб.) и расстояние в километрах до пляжа (*Х*, км):

X	0,8	1,5	0,9	0,8	1,2	1,5	1,0	1,6	1,2	1,0	0,9	1,2	1,7	1,6
Y	2,7	3,2	2,7	2,6	3,1	3,3	2,8	3,4	2,9	2,9	2,6	3	3,4	3,3

Представить полученные результаты графически. Найти коэффициент корреляции. Что можно сказать о зависимости между этими двумя характеристиками? Проверить коэффициент линейной корреляции на значимость. Написать уравнение линейной регрессии.

Решение. Представим результаты в графическом виде:



Коэффициент линейной корреляции Пирсона находим по формуле:

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}.$$

Сначала найдем средние значения $\bar{x}$ и $\bar{y}$. Всего в таблице 14 значений, значит, в нашем случае n=14.

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

$$= \frac{0,8+1,5+0,9+0,8+1,2+1,5+1,0+1,6+1,2+1,0+0,9+1,2+1,7+1,6}{14} =$$

$$= \frac{16,9}{14} = 1,207.$$

Аналогично $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{41,9}{14} = 2,993.$

Вычисляем $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 1,552$; $\overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2 = 9,036$.

Находим выборочные дисперсии:

$\sigma_x^2 = \overline{x^2} - \left(\bar{x}\right)^2 = 1,552 - \left(1,207\right)^2 = 0,095$, $\sigma_x = \sqrt{0,095} \approx 0,308$,

$\sigma_y^2 = \overline{y^2} - \left(\overline{y}\right)^2 = 9{,}036 - \left(2{,}993\right)^2 = 0{,}078$, $\sigma_y = \sqrt{0{,}078} \approx 0{,}279$.

Далее

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i = \frac{0{,}8 \cdot 2{,}7 + 1{,}5 \cdot 3{,}2 + 0{,}9 \cdot 2{,}7 + 0{,}8 \cdot 2{,}6 + 1{,}2 \cdot 3{,}1 +}{14}$$

$$\frac{+1{,}5 \cdot 3{,}3 + 1{,}0 \cdot 2{,}8 + 1{,}6 \cdot 3{,}4 + 1{,}2 \cdot 2{,}9 + 1{,}0 \cdot 2{,}9 + 1{,}6 \cdot 3{,}3 + 0{,}9 \cdot 2{,}6 +}{14}$$

$$\frac{+1{,}2 \cdot 3{,}0 + 1{,}7 \cdot 3{,}4}{14} = \frac{51{,}76}{14} = 3{,}697.$$

И, наконец, вычисляем коэффициент линейной корреляции Пирсона:

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \cdot \sigma_y} = \frac{3{,}697 - 1{,}207 \cdot 2{,}993}{0{,}308 \cdot 0{,}279} = \frac{0{,}084}{0{,}086} \approx 0{,}98.$$

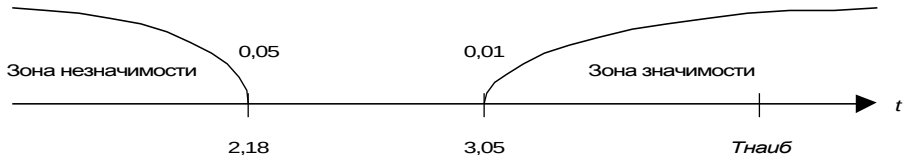
Полученное значение коэффициента корреляции (близкое к 1) свидетельствует о линейной корреляционной зависимости между признаками *X* и *Y*, близкой к функциональной. Подтвердим это проверкой коэффициента линейной корреляции на значимость.

Находим статистику: $T = r \cdot \sqrt{\frac{n-2}{1-r^2}} = 0{,}98 \cdot \sqrt{\frac{14-2}{1-\left(0{,}98\right)^2}} = 17{,}06$.

По числу степеней свободы $\nu = n - 2 = 14 - 2 = 12$ и по уровням значимости 0,05 и 0,01 с помощью критерия Стьюдента из приложения 3 находим критические значения:

$$t_{\text{кр.}} = \begin{cases} 2{,}18 & \text{для } P \le 0{,}05 \\ 3{,}05 & \text{для } P \le 0{,}01 \end{cases}.$$

Строим ось значимости.

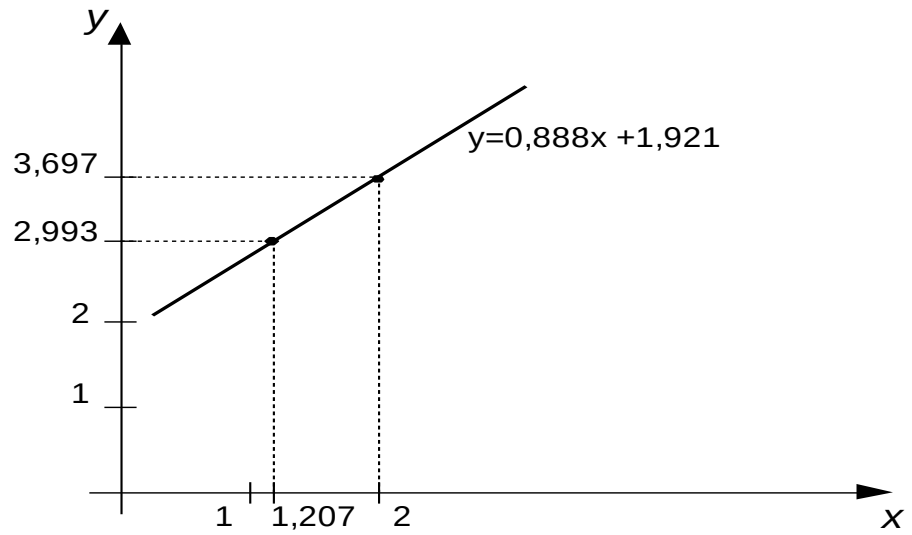


Значение *T* попало в зону значимости, следовательно, коэффициент линейной корреляции значимо отличается от нуля. Напишем уравнение линейной регрессии $y = ax + b$, где

$$a = r\frac{\sigma_y}{\sigma_x} = 0{,}98 \cdot \frac{0{,}279}{0{,}308} = 0{,}888;\ b = \overline{y} - r\frac{\sigma_y}{\sigma_x}\overline{x} = 1{,}921.$$

Следовательно, искомое уравнение регрессии имеет вид: $y = 0{,}888x + 1{,}921$.

Эта прямая проходит через точки $\left(\overline{X};\overline{Y}\right) = \left(1{,}207;\ 2{,}993\right)$ и $\left(2;\ 3{,}697\right)$. Построим эту прямую.



3. Из интервальной совокупности произведена выборка. Результаты измерения признаков X и Y у членов выборки приводятся в следующей корреляционной таблице:

Y \ X	15-25	25-35	35-45	45-55	55-65	65-75
200-300	2	3				1
300-400	1	5	4	3	1	
400-500		1	10	2		
500-600	1		18	8	2	3
600-700		1	3	7	10	5
700-800	1				3	5

Найти выборочный коэффициент линейной корреляции, коэффициент детерминации. Написать уравнения прямых регрессий Y на X и Y на X и построим их.

Решение. Сначала интервалы значений признаков X и Y заменим их серединами. Далее перейдем к условным вариантам по формулам:

$$u_i = (x_i - c_1)/h_1;\ v_j = (y_j - c_2)/h_2,$$

где в качестве ложных нулей возьмем c_1=40; c_2=550; шаги: h_1=10; h_2=100. Перейдем к новой корреляционной таблице

	V \ U	20	30	40	50	60	70	
		-2	-1	0	1	2	3	n_v
250	-3	2	3				1	6
350	-2	1	5	4	3	1		14
450	-1		1	10	2			13
550	0	1		18	8	2	3	32
650	1		1	3	7	10	5	26
750	2	1				3	5	9
	n_u	5	10	35	20	16	14	**100**

Все необходимые вычисления приведены в следующей таблице:

V \ U	-2	-1	0	1	2	3	n_v	$n_v \cdot v$	$n_v \cdot v^2$	$\sum n_{uv} \cdot u$	$v \cdot \sum n_{uv} u$
-3	2	3				1	6	-18	54	-4	12
-2	1	5	4	3	1		14	-28	56	-2	4
-1		1	10	2			13	-13	13	1	-1
0	1		18	8	2	3	32	0	0	19	0
1		1	3	7	10	5	26	26	26	41	41
2	1				3	5	9	18	36	19	38
n_u	5	10	35	20	16	14	**100**	**-15**	**185**		**94**
$n_u \cdot u$	-10	-10	0	20	32	42	**74**				
$n_u \cdot u^2$	20	10	0	20	64	126	**240**				
$u \cdot \sum n_{uv} v$	12	19	0	-1	28	36	**94**				

Контроль

Найдем $\bar{u}$ и $\bar{v}$: $\bar{u} = \dfrac{\sum n_u \cdot u}{n} = \dfrac{74}{100} = 0{,}74$; $\bar{v} = \dfrac{\sum n_v \cdot v}{n} = \dfrac{-15}{100} = -0{,}15$.

Найдем величины $\bar{u}^2$ и $\bar{v}^2$ $\bar{u}^2 = \dfrac{\sum n_u \cdot u^2}{n} = \dfrac{240}{100} = 2{,}4$; $\bar{v}^2 = \dfrac{\sum n_v \cdot v^2}{n} = \dfrac{185}{100} = -1{,}85$.

Найдем σ_u и σ_v $\sigma_u = \sqrt{\bar{u}^2 - (\bar{u})^2} = \sqrt{2{,}4 - (0{,}74)^2} = \sqrt{2{,}4 - 0{,}5476} = 1{,}361$;

$$\sigma_v = \sqrt{\bar{v}^2 - (\bar{v})^2} = \sqrt{1{,}85 - (-0{,}15)^2} = \sqrt{1{,}85 - 0{,}0225} = 1{,}352.$$

Найдем $\sum\sum n_{uv} uv$, для чего записываем последние два столбца таблицы. Поясним, как вычислялись эти столбцы. Находим произведение частот n_{uv} на соответствующие варианты u и суммируем их. Например, для первой строки $-2 \cdot 2 + (-1) \cdot 3 + 3 \cdot 1 = -4$, для второй $-2 \cdot 1 + (-1) \cdot 5 + 0 \cdot 4 + 1 \cdot 3 + 2 \cdot 1 = -2$ и так далее. Результаты запишем в столбец $\sum n_{uv} u$.

Наконец, умножаем варианту v на соответствующие значения $\sum n_{uv} u$ и помещаем в последнем столбце таблицы. Например, в первой строке таблицы $v \cdot \sum n_{uv} u = -3 \cdot (-4) = 12$. Сложив все числа последнего столбца, получим сумму $\sum\sum n_{uv} uv = \sum v \sum n_{uv} u = 94$. Для контроля аналогичные вычисления производят по столбцам и находят $\sum u \sum n_{uv} v = 94 = \sum v \sum n_{uv} u$. В итоге $\overline{uv} = \dfrac{\sum\sum n_{uv} uv}{n} = \dfrac{94}{100} = 0{,}94$.

Найдем искомый коэффициент корреляции

$$r_{uv} = \frac{\overline{uv} - \bar{u} \cdot \bar{v}}{\sigma_u \cdot \sigma_v} = \frac{0{,}94 - 0{,}74 \cdot (-0{,}15)}{1{,}361 \cdot 1{,}352} = \frac{1{,}051}{1{,}84} = 0{,}571.$$

Так как коэффициент линейной корреляции не изменяется при линейной замене, то $r_{uv} = r_{xy} = 0{,}571$.

Найдем коэффициент детерминации

$$D = r^2 = (0{,}571)^2 = 0{,}326.$$

Найдем $\bar{x}$ и $\bar{y}$. Учитывая, что $u = \dfrac{x - c_1}{h_1}$; $v = \dfrac{y - c_2}{h_2}$, получаем

$$\bar{x} = \bar{u} \cdot h_1 + c_1 = 0{,}74 \cdot 10 + 40 = 47{,}4,$$
$$\bar{y} = \bar{v} \cdot h_2 + c_2 = -0{,}15 \cdot 100 + 550 = 535.$$

Найдем σ_x и σ_y: $\sigma_x = h_1 \cdot \sigma_u = 10 \cdot 1{,}361 = 13{,}61$; $\sigma_y = h_2 \cdot \sigma_v = 100 \cdot 1{,}352 = 135{,}2$.

Подставив найденные величины в соотношения

$$y_x - \bar{y} = r_в \frac{\sigma_y}{\sigma_x}(x - \bar{y}); \quad x_y - \bar{x} = r_в \frac{\sigma_x}{\sigma_y}(y - \bar{y}),$$

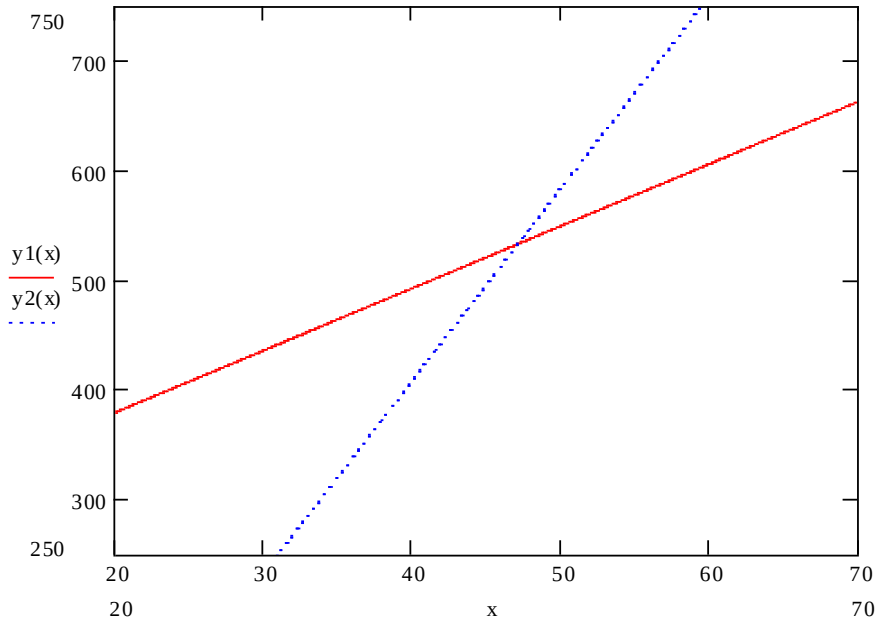
получим уравнение регрессии Y на X

$y_x - 535 = 0{,}571 \cdot \frac{135{,}2}{13{,}61}(x - 47{,}4)$, или окончательно: $y_x = 5{,}672x + 266{,}3$.

Уравнение регрессии X на Y

$x_y - 47{,}4 = 0{,}571 \cdot \frac{13{,}61}{135{,}2}(y - 535)$, или окончательно: $x_y = 16{,}65 + 0{,}057y$.

Построим эти линии. Заметим, что линии регрессии пересекаются в точке с координатами ($\bar{X}; \bar{Y}$), где $\bar{X} = 47{,}5$, а $\bar{Y} = 530$. На графике $y_1(x) = y_x = 5{,}672x + 266{,}3$, а $y_2(x) = \frac{x_y - 16{,}65}{0{,}057}$.



Задачи для самостоятельного решения

1. В ходе социологического эксперимента измерялись характеристики X и Y. В декартовой системе координат дать графическое представление полученных данных. Найти коэффициент корреляции. Что можно сказать о зависимости между этими двумя характеристиками? Проверить коэффициент линейной корреляции на значимость. Написать уравнение линейной регрессии.

x	0,8	1,0	2,7	3,4	0,3	-0,3	-0,6	-0,5	1,0	-2,0
y	-1,0	0,4	-0,1	-0,4	-1,6	0,3	0,8	0,9	0,0	-0,5

2. В ходе социологического эксперимента измерялись характеристики *X* (стаж работы, года) и *Y* (выработка одного рабочего за смену, шт.). Данные сведены в таблицу:

Стаж работы (*X*, года)	1	3	5	9	12	15	18	20	22
Выработка одного рабочего за смену (*У*, шт.)	18	28	32	34	35	36	40	42	43

Найти выборочный коэффициент корреляции и напишите уравнения прямых линейной регрессии *У* на *X* и *X* на *У*. Проанализируйте полученные результаты.

3. Имеется случайная выборка из 10 семей для изучения связи между числом членов семью (*X*) и числом автомобилей (*У*).

X	1	1	1	2	2	3	3	4	5	6
У	1	0	2	1	2	1	3	2	3	3

Найти выборочный коэффициент корреляции и напишите уравнения прямых линейной регрессии *У* на *X* и *X* на *У*. Проанализируйте полученные результаты. Проверить коэффициент линейной корреляции на значимость.

4. Имеется случайная выборка из 10 семей для изучения связи между числом членов семью (*X*) и числом телевизоров (*У*) в домохозяйстве.

X	1	1	1	2	2	3	3	4	5	6
У	1	1	2	1	1	1	2	2	3	3

Найти выборочный коэффициент корреляции и напишите уравнения прямых линейной регрессии *У* на *X* и *X* на *У*. Проанализируйте полученные результаты.

5. Опрос 10 студентов факультета философии и социальных наук БГУ позволяет выявить зависимость меду средним баллом по результатам предыдущей сессии (*X*) и числом часов в неделю, затраченных студентом на самостоятельную подготовку (*У*):

X	4	5	5	6	7	7	8	9	9	10
У	1	5	10	15	13	20	22	25	26	30

Найти выборочный коэффициент корреляции и напишите уравнения прямых линейной регрессии *У* на *X* и *X* на *У*. Проанализируйте полученные результаты.

6. Эмпирическая зависимость между пробегом автомобилей (*X*, лет) и стоимостью ежемесячного технического обслуживания (*У*, руб.) приведена в таблице:

X	10	14	15	16	18	20	21	22	24	25
У	150	125	180	220	225	230	280	255	290	310

Найти выборочный коэффициент корреляции и напишите уравнения прямых линейной регрессии *У* на *X* и *X* на *У*. Проанализируйте полученные результаты.

7. В ходе социологического эксперимента 50 рабочих измерялись характеристики *X* (стаж работы, года) и *Y* (выработка одного рабочего за смену, шт.). Данные сведены в таблицу:

Y \ X	5	10	15	20	25	30	35	40	n_y
100	2	1							3
120	3	4	3						10
140			5	10	8				23
160				1		6	1	1	9
180							4	1	5
n_x	5	5	8	11	8	6	5	2	n=50

Найти выборочный коэффициент линейной корреляции между измеряемыми характеристиками (стажем работы и выработкой одного рабочего за смену) и написать уравнения регрессии.

8. В ходе социологического эксперимента 50 автомобилей выявлена эмпирическая зависимость между пробегом автомобилей (X, лет) и стоимостью ежемесячного технического обслуживания (Y, руб.):

Y \ X	18	23	28	33	38	43	48	n_y
125		1						1
150	1	2	5					8
175		3	2	12				17
200			1	8	7			16
225					3	3		6
250						1	1	2
n_x	1	6	8	20	10	4	1	n=50

Найти выборочный коэффициент линейной корреляции между пробегом автомобилей и стоимостью ежемесячного технического обслуживания. Найти выборочное уравнение прямой линии регрессии Y на X.

Вопросы для самоконтроля

1. Формула коэффициента линейной корреляции.
2. Свойства коэффициента линейной корреляции.
3. Применение коэффициента линейной корреляции. Коэффициент детерминации.
4. Уравнения прямых регрессий Y на X и X на Y.
5. Проверка значимости коэффициента линейной корреляции.

ЛИТЕРАТУРА

1. Абчук В.А. Экономико-математические методы. – СПб.: Союз, 1999.
2. Ахтямов, А.М. Математика для социологов и экономистов: учеб.пособие / А.М. Ахтямов. – М.: Физматлит, 2004. – 464 с.
3. Белько, И. В. Теория вероятностей, математическая статистика, математическое программирование : учеб. пособие для студ. учреждений высш. образования по экон. спец. / И. В. Белько, И. М. Морозова, Е. А. Криштапович. – Минск : Новое знание, 2016 ; Москва : ИНФРА-М. – 298 с. : ил.
4. Велько, О.А. Основы высшей математики для социологов: Учебно-методическое пособие / О.А. Велько, М.В. Мартон, Н.А. Моисеева. – Минск: БГУ, 2020. – 303 с.
5. Велько, О. А. Основы высшей математики и теории вероятностей: учебная программа УВО по учебной дисциплине по специальности 1-23 01 15 Социальные коммуникации [Электронный ресурс] / О.А. Велько // Белорусский государственный университет. – Минск, 2021. – Режим доступа: https://elib.bsu.by/handle/123456789/269619. – Дата доступа: 2.07.2021.
6. Велько, О.А. Основы высшей математики. Учебная программа УВО для специальности 1-23 01 05 Социология [Электронный ресурс] / О.А. Велько, Н.А. Моисеева // Белорусский государственный университет. – Минск, 2019. – Режим доступа:http://elib.bsu.by/handle/123456789/233274. – Дата доступа: 12.07.2019.
7. Велько, О. А. Основы высшей математики : электронный учебно-методический комплекс для специальности 1-23 01 05 «Социология» / О. А. Велько, Н. А. Моисеева; БГУ, Механико-математический фак., Каф. общей математики и информатики. – Минск: БГУ, 2020. – 257 с.: ил. – Библиогр.: с. 255–257. [Электронный ресурс]. – 2020. – Режим доступа: http://elib.bsu.by/handle/123456789/241078. Дата доступа: 06.03.2020.
8. Велько О.А. Теория вероятностей и математическая статистика: сб. задач / О.А. Велько, Е.В. Воронкова, Г.К. Игнатьева, Л.В. Корчёмкина, И.П. Мацкевич, С.А. Мызгаева; под общ. ред. И. П. Мацкевича. – Минск: МИУ, 2003. – 56 с.
9. Велько, О. А. Теория вероятностей и математическая статистика: учебная программа учреждения высшего образования по учебной дисциплине для специальности 1-23 01 05 Социология [Электронный ресурс] / О.А. Велько // Белорусский государственный университет. – Минск, 2021. – Режим доступа: https://elib.bsu.by/handle/123456789/269618. – Дата доступа: 2.07.2021.

10. Гайшун, Л.Н. Теория вероятностей: Учебное пособие для студентов экономических специальностей / Л.Н. Гайшун, Г.К. Игнатьева, О.А. Велько. – Минск: МИУ, 2002. – 167 с.

11. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 1997.

12. Гмурман, В.Е. Руководство к решению задач по теории вероятностей и математической статистике: учебное пособие / В.Е. Гмурман. – 7-е., изд., доп. – М.: Высш.шк., 2003. – 405 с.

13. Гурский Е.И. Сборник задач по теории вероятностей и математической статистике. – Мн.: Выш. шк., 1984.

14. Кузьмин, К.Г. Теория вероятностей и математическая статистика: учебно-методическое пособие для студентов специальности 1-25 01 03 «Мировая экономика» / К.Г. Кузьмин, Н.И. Широканова. – Минск: БГУ, 2009. – 89 с.

15. Мацкевич И.П., Свирид Г.П., Булдык Г.М. Сборник задач и упражнений по высшей математике. Теория вероятностей и математическая статистика. – Мн.: Выш. шк., 1996.

16. Мацкевич, И.П. Математические методы в психологии / И.П. Мацкевич, О.А. Велько, Е.В. Воронкова, С.Л. Гуринович. – 3-е изд. – Минск: МИУ, 2009. – 188 с.

17. Мацкевич, И.П. Статистические методы в психологии: Учебно-методический комплекс / И.П. Мацкевич, О.А. Велько, Е.В. Воронкова, С.Л. Гуринович. – 2-е изд. – Минск: МИУ, 2012. – 194 с.

18. Моисеева, Н. А. Основы высшей математики : электронный учебно-методический комплекс для специальности 1-23 01 15 «Социальные коммуникации» / Н. А. Моисеева, О. А. Велько; БГУ, Механико-математический фак., Каф. общей математики и информатики. – Минск: БГУ, 2020. – 193 с.: ил. – Библиогр.: с. 191–193. [Электронный ресурс]. – 2020. – Режим доступа: http://elib.bsu.by/handle/123456789/241081. Дата доступа: 06.03.2020.

19. Моисеева, Н. А. Основы высшей математики и теории вероятностей : электронный учебно-методический комплекс для специальности: 1-23 01 15 «Социальные коммуникации» [Электронный ресурс] / Н. А. Моисеева, О. А. Велько ; БГУ, Механико-математический фак., Каф. общей математики и информатики. – Минск : БГУ, 2021. – 239 с. : ил., табл. – Библиогр.: с. 238–239. – Режим доступа: https://elib.bsu.by/handle/123456789/274772: 26.01.2022.

20. Ниворожкина Л.И, Морозова З.А. Основы статистики с элементами теории вероятностей для экономистов. – Ростов на Дону : Феникс, 1996.

21. Петров, В.А. Теория вероятностей и математическая статистика: Учебно-методический комплекс / В.А. Петров, Г.К. Игнатьева, О.А. Велько. – Минск: МИУ, 2007. – 268 с.

22. Путькина, Л. В. Информатика и математика для гуманитарных вузов : учеб.пособие / Л. В. Путькина, Т. Г. Пискунова, Т. Б. Антипова. – Санкт-Петербург : СПбГУП, 2014. – 236 с. : ил.

23. Сборник индивидуальных заданий по теории вероятностей и математической статистике / Под ред. Рябушко А.П. – Мн.: Выш. шк., 1992.

ПРИЛОЖЕНИЕ 1

Таблица значений функции $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

	0	1	2	3	4	5	6	7	8	9
0,0	0,398	3989	3989	3988	3986	3984	398	398	397	397
0,1	3970	3965	3961	3956	3951	3945	393	393	392	391
0,2	3910	3902	3894	3885	3876	3867	385	384	383	382
0,3	3814	3802	3790	3778	3765	3752	373	372	371	369
0,4	3683	3668	3652	5637	3621	3605	358	357	355	353
0,5	3521	3503	3485	3467	3448	3429	341	339	337	335
0,6	3332	3312	3292	3271	3261	3230	320	318	316	314
0,7	3123	3101	3079	3056	3064	3011	298	296	294	292
0,8	2897	2874	2850	2827	2803	2780	275	273	270	268
0,9	2661	2637	2613	2589	2565	2541	251	249	246	244
1,0	2420	2396	2371	2347	2323	2299	227	225	222	220
1,1	2179	2155	2131	2107	2083	2059	203	201	198	196
1,2	1942	1919	1895	1872	1849	1826	180	178	175	173
1,3	1714	1691	1669	1647	1626	1604	158	156	153	151
1,4	1497	1476	1456	1435	1415	1394	137	135	133	131
1,5	1295	1276	1257	1238	1219	1200	118	116	114	112
1,6	1109	1092	1074	1057	1040	1023	100	098	097	095
1,7	0940	0925	0909	0893	0878	0863	084	083	081	080
1,8	0790	0775	0761	0748	0734	0721	070	069	068	066
1,9	0656	0644	0632	0620	0608	0596	058	057	056	055
2,0	0,054	0529	0519	0508	0498	0488	047	046	045	044
2,1	0440	0431	0422	0413	0404	0396	038	037	037	036
2,2	0355	0347	0339	0332	0325	0317	031	030	029	029
2,3	0283	0277	0270	0264	0258	0252	024	024	023	022
2,4	0224	0219	0213	0208	0203	0198	019	018	018	018
2,5	0175	0171	0167	0163	0158	0154	015	014	014	013
2,6	0136	0132	0129	0126	0122	0119	011	011	011	010
2,7	0104	0101	0099	0096	0093	0091	008	008	008	008
2,8	0079	0077	0075	0073	0071	0069	006	006	006	006
2,9	0060	0058	0056	0055	0053	0051	005	004	004	004
3,0	0,004	0043	0042	0040	0039	0038	003	003	003	003
3,1	0033	0032	0031	0030	0029	0028	002	002	002	002
3,2	0024	0023	0022	0022	0021	0020	002	001	001	001
3,3	0017	0017	0016	0016	0015	0015	001	001	001	001
3,4	0012	0012	0012	0011	0011	0010	001	001	000	000
3,5	0009	0008	0008	0008	0008	0007	000	000	000	000
3,6	0006	0006	0006	0005	0005	0005	000	000	000	000

ПРИЛОЖЕНИЕ 2

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

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,000	0,30	0,117	0,60	0,225	0,90	0,315
0,01	0,004	0,31	0,121	0,61	0,229	0,91	0,318
0,02	0,008	0,32	0,125	0,62	0,232	0,92	0,321
0,03	0,012	0,33	0,129	0,63	0,235	0,93	0,323
0,04	0,016	0,34	0,133	0,64	0,238	0,94	0,326
0,05	0,019	0,35	0,136	0,65	0,242	0,95	0,328
0,06	0,023	0,36	0,140	0,66	0,245	0,96	0,331
0,07	0,027	0,37	0,144	0,67	0,248	0,97	0,334
0,08	0,031	0,38	0,148	0,68	0,251	0,98	0,336
0,09	0,035	0,39	0,151	0,69	0,254	0,99	0,338
0,10	0,039	0,40	0,155	0,70	0,258	1,00	0,341
0,11	0,043	0,41	0,159	0,71	0,261	1,01	0,343
0,12	0,047	0,42	0,162	0,72	0,264	1,02	0,346
0,13	0,051	0,43	0,166	0,73	0,267	1,03	0,348
0,14	0,055	0,44	0,170	0,74	0,270	1,04	0,350
0,15	0,059	0,45	0,173	0,75	0,273	1,05	0,353
0,16	0,063	0,46	0,177	0,76	0,276	1,06	0,355
0,17	0,067	0,47	0,180	0,77	0,279	1,07	0,357
0,18	0,071	0,48	0,184	0,78	0,282	1,08	0,359
0,19	0,075	0,49	0,187	0,79	0,285	1,09	0,362
0,20	0,079	0,50	0,191	0,80	0,288	1,10	0,364
0,21	0,083	0,51	0,195	0,81	0,291	1,11	0,366
0,22	0,087	0,52	0,198	0,82	0,293	1,12	0,368
0,23	0,091	0,53	0,201	0,83	0,296	1,13	0,370
0,24	0,094	0,54	0,205	0,84	0,299	1,14	0,372
0,25	0,098	0,55	0,208	0,85	0,302	1,15	0,374
0,26	0,102	0,56	0,212	0,86	0,305	1,16	0,377
0,27	0,106	0,57	0,215	0,87	0,307	1,17	0,379
0,28	0,110	0,58	0,219	0,88	0,310	1,18	0,381
0,29	0,114	0,59	0,222	0,89	0,313	1,19	0,383

1,20	0,3849	1,53	0,4370	1,86	0,4686	2,38	0,4913
1,21	0,3869	1,54	0,4382	1,87	0,4693	2,40	0,4918
1,22	0,3883	1,55	0,4394	1,88	0,4699	2,42	0,4922
1,23	0,3907	1,56	0,4406	1,89	0,4706	2,44	0,4927
1,24	0,3925	1,57	0,4418	1,90	0,4713	2,46	0,4931
1,25	0,3944	1,58	0,4429	1,91	0,4719	2,48	0,4934
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	0,4793	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	0,4986
1,52	0,4357	1,85	0,4678	2,36	0,4909	3,20	0,4993
3,40	0,4996	3,80	0.4999	4,50	0,4999		
3,60	0,4998	4,00	0,4999	5,00	0,4999		

ПРИЛОЖЕНИЕ 3

Таблица значений критических точек $t_{кр}$ распределения Стьюдента, соответствующих вероятности p и степени свободы v

v \ p	0,1	0,05	0,01	0,005
1	6,31	12,71	63,66	127,32
2	2,92	4,30	9,92	14,09
3	2,35	3,18	5,84	7,45
4	2,13	2,78	4,60	5,60
5	2,01	2,57	4,03	4,77
6	1,94	2,45	3,71	4,32
7	1,89	2,36	3,50	4,03
8	1,86	2,31	3,36	3,83
9	1,83	2,26	3,25	3,69
10	1,81	2,23	3,17	3,58
11	1,80	2,20	3,11	3,50
12	1,78	2,18	3,05	3,43
13	1,77	2,16	3,01	3,37
14	1,76	2,14	2,98	3,33
15	1,75	2,13	2,95	3,29
16	1,75	2,12	2,92	3,25
17	1,74	2,11	2,90	3,22
18	1,73	2,10	2,88	3,20
19	1,73	2,09	2,86	3,17
20	1,72	2,09	2,85	3,15
25	1,71	2,06	2,79	3,08
30	1,70	2,04	2,75	3,03
40	1,68	2,02	2,70	2,97
60	1,67	2,00	2,66	2,91
120	1,66	1,98	2,62	2,86
∞	1,64	1,96	2,58	2,81

ПРИЛОЖЕНИЕ 4

Таблица вероятностей P для критерия χ^2 Пирсона

Квантили χ^2 распределения с m степенями свободы.

m	Уровень значимости α					
	0.01	0.025	0.05	0.95	0.975	0.99
1	6.6	5.0	3.8	0.0	0.0	0.0
2	9.2	7.4	6.0	0.1	0.1	0.0
3	11.3	9.3	7.8	0.4	0.2	0.1
4	13.3	11.1	9.5	0.7	0.5	0.3
5	15.1	12.8	11.1	1.1	0.8	0.6
6	16.8	14.4	12.6	1.6	1.2	0.9
7	18.5	16.0	14.1	2.2	1.7	1.2
8	20.1	17.5	15.5	2.7	2.2	1.6
9	21.7	19.0	16.9	3.3	2.7	2.1
10	23.2	20.5	18.3	3.9	3.2	2.6
11	24.7	21.9	19.7	4.6	3.8	3.1
12	26.2	23.3	21.0	5.2	4.4	3.6
13	27.7	24.7	22.4	5.9	5.0	4.1
14	29.1	26.1	23.7	6.6	5.6	4.7
15	30.6	27.5	25.0	7.3	6.3	5.2
16	32.0	28.8	26.3	8.0	6.9	5.8
17	33.4	30.2	27.6	8.7	7.6	6.4
18	34.8	31.5	28.9	9.4	8.2	7.0
19	36.2	32.9	30.1	10.1	8.9	7.6
20	37.6	34.2	31.4	10.9	9.6	8.3
21	38.9	35.5	32.7	11.6	10.3	8.9
22	40.3	36.8	33.9	12.3	11.0	9.5
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
50	76.2	71.4	67.5	34.8	32.4	29.7
100	135.8	129.6	124.3	77.9	74.2	70.1
1000	1107.0	1089.5	1074.7	927.6	914.3	898.9

ПРИЛОЖЕНИЕ 5

Таблица значений критических точек
распределения Фишера-Снедекора
при уровне значимости α = 0,05
Значения $F_{\text{кр}}(0{,}05; \nu_1, \nu_2)$ корней уравнения $P(F(\nu_1, \nu_2) > F_p) = 0{,}05$.

ν_1 \ ν_2	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,71	224,58	230,16	233,99	238,88	243,91	249,05	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,02	1,83	1,61	1,25
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00