Министерство образования и науки Российской Федерации

Магнитогорский государственный технический
университет им. Г. И. Носова

Н.И. Барышникова
А.Р. Ишбирдин
М.М. Ишмуратова

МЕТОДЫ СБОРА, ОБРАБОТКИ ДАННЫХ И ПРЕДСТАВЛЕНИЯ РЕЗУЛЬТАТОВ В НАУЧНЫХ ИССЛЕДОВАНИЯХ В ПИЩЕВОЙ ПРОМЫШЛЕННОСТИ

Утверждено Редакционно-издательским советом университета в качестве учебного пособия

Магнитогорск
2013

УДК 658.362

Рецензенты:

Заведующий кафедрой пищевой биотехнологии
Уральского государственного экономического университета,
доктор технических наук, профессор
Г.Б. Пищиков

Заведующая кафедрой экологии
Башкирского государственного университета
Е.И. Новоселова

Барышникова, Н.И.
Методы сбора, обработки данных и представления результатов в научных исследованиях в пищевой промышленности: учеб. пособие / Н.И. Барышникова, А.Р. Ишбирдин, М.М. Ишмуратова. Изд-во Магнитогорск. гос. техн. ун-та им. Г.И. Носова, 2013. 55 с.

Состоит из разделов, содержащих описание методов сбора первичного материала, основных методов статистической обработки материала и специфики представления результатов научной работы.

Пособие, как дополнительное методическое обеспечение, предназначено для студентов, обучающихся по дисциплинам «Методы и средства научных исследований» и «Статистические методы контроля и управления качеством» специальностей 200503 «Стандартизация и сертификация», 260501 «Технология продуктов общественного питания», 260303 «Технология молока и молочных продуктов», 260301 «Технология мяса и мясных продуктов», и направлений 221700.62 «Стандартизация и метрология», 260200.62 «Продукты питания животного происхождения», 260100.62 «Продукты питания из растительного сырья».

УДК 658.362

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Научно-исследовательская работа студентов (НИРС) является неотъемлемой частью подготовки специалиста с высшим образованием. Выполняется НИРС в процессе специализации и оформляется в виде квалификационной работы (дипломной работы специалиста, магистерской диссертации), а также в виде конкурсных научных работ студента и публикаций в научной печати.

Научной работа может считаться только тогда, когда ее результаты и выводы обеспечены сбором репрезентативных данных и их статистической обработкой, подтверждающей (или опровергающей) выдвинутые исследователем гипотезы.

Статистика – это наука о том, как обрабатывать данные. Статистические методы активно применяются в технических исследованиях, экономике, социологии, медицине, биологии, геологии, истории и т.д. С обработкой результатов наблюдений, измерений, испытаний, опытов, анализов имеют дело специалисты во всех отраслях практической деятельности, почти во всех областях научных исследований. При этом в каждой из этих отраслей для решения специальных задач сложились свои правила и особенности применения статистического аппарата.

Статистический анализ является заключительной стадией статистического исследования. Он включает в себя обработку полученных в исследованиях или эксперименте статистических данных, интерпретацию полученных результатов с целью получения объективных выводов о состоянии изучаемого явления и закономерностях его развития. В процессе статистического анализа изучаются структура, динамика и взаимосвязь явлений и процессов.

Краткая история математической статистики. Статистика существовала еще в глубокой древности, однако, как наука она возникла лишь в XVII веке. Термин «статистика» произошел от латинского слова «статус», т.е. состояние, определенное положение вещей. В науку термин введен немецким ученым Готфридом Ахенвалем, который в 1746 г. начал читать в Марбукском, а затем в Геттенгенском университетах новую дисциплину, названную им «статистика».

Математическая статистика как наука начинается с работ знаменитого немецкого математика Карла Фридриха Гаусса (1777-1855 гг.), который на основе теории вероятностей исследовал и обосновал метод наименьших квадратов, созданный им в 1795 г. и примененный для обработки астрономических данных (с целью уточнения орбиты малой планеты Церера). Его именем часто называют одно из наиболее популярных распределений вероятностей – нормальное, а в теории случайных процессов основной объект изучения – гауссовские процессы.

В конце XIX – начале XX вв. крупный вклад в математическую статистику внесли английские исследователи, прежде всего К. Пирсон (1857-1936 гг.) и Р.А. Фишер (1890-1962 гг.). В частности, Пирсон разработал критерий «хи-квадрат» проверки статистических гипотез, а Фишер – дисперсионный анализ, теорию планирования эксперимента, метод максимального правдоподобия оценки параметров.

В 30-е годы XX в. поляк Ежи Нейман (1894-1977 гг.) и англичанин Э. Пирсон развили общую теорию проверки статистических гипотез, а советские математики академик А.Н. Колмогоров (1903-1987 гг.) и член-корреспондент АН СССР Н.В. Смирнов (1900-1966 гг.) заложили основы непараметрической статистики. В сороковые годы XX в. румын А. Вальд (1902-1950 гг.) построил теорию последовательного статистического анализа.

Математическая статистика бурно развивается и в настоящее время. Так, за последние 40 лет можно выделить четыре принципиально новых направления исследований:

– разработка и внедрение математических методов планирования экспериментов;

– развитие статистики объектов нечисловой природы как самостоятельного направления в прикладной математической статистике;

– развитие статистических методов, устойчивых по отношению к малым отклонениям, от используемой вероятностной модели;

– широкое развертывание работ по созданию компьютерных пакетов программ, предназначенных для проведения статистического анализа данных.

Внутренняя структура статистики как науки была выявлена и обоснована при создании в 1990 г. Всесоюзной статистической ассоциации.

Математическая статистика исходит из сформулированных в 1930-1950 гг. постановок математических задач, происхождение которых связано с анализом статистических данных. Начиная с 70-х годов XX в. исследования по математической статистике посвящены обобщению и дальнейшему математическому изучению этих задач.

В настоящее время статистическая обработка данных проводится, как правило, с помощью соответствующих программных продуктов (например, широко применяемых в практике и доступных MS *Excell* и *Statistica*).

1. ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ И ВЫБОРКА

Генеральная совокупность – множество однородных объектов, из которого можно выделить некоторое подмножество, называемое **выборочной совокупностью** или **выборкой**. Более строгое с математической точки зрения определение выборки звучит так: *выборкой называют последовательность независимых одинаково распределенных случайных величин*. **Объемом совокупности** (генеральной или выборочной) называется число ее объектов.

Выборочным методом называется метод исследования общих характеристик генеральной совокупности на основе изучения свойств выборки. Статистическое исследование в центр внимания всегда ставит выборку. Основная особенность выборки как множества значений случайной величины – это отличие отдельных вариант друг от друга, явление *изменчивости*. Причиной изменчивости является воздействие на случайную величину систематического (доминирующего) фактора и отклоняющих ее случайных факторов.

Например, генеральную совокупность могут составлять качественные показатели всей продукции, произведенной по единой технологии на аналогичном оборудовании разных производителей. Выборками могут служить показатели продукции (например, массовая доля отдельных ингредиентов), произведенной за смену на одной единице оборудования.

2. ПОСТРОЕНИЕ ВАРИАЦИОННОГО РЯДА

Первым этапом статистического изучения явления служит построение вариационного ряда – упорядоченного распределения единиц совокупности по возрастающим (чаще) или по убывающим (реже) значениям признака и подсчет числа единиц с тем или иным значением признака.

Промежуток $x_{набл} = [x_{(1)} – x_{(n)}] = [x_{min\_набл} – x_{max\_набл}]$ между крайними членами вариационного ряда называется **интервалом варьирования**, его длина $W_n = x_{(n)} – x_{(1)} = x_{max\_набл} – x_{min\_набл}$ называется **размахом выборки**.

Крайние члены вариационного ряда $x_{min\_набл} = x_{(1)} = \min\{x_k\}$ для k=1...n и $x_{max\_набл} = x_{(n)} = \max\{x_k\}$ для k=1...n называются **экстремальными значениями**.

Характеристиками вариационного ряда являются *средние величины*. В исследованиях обычно применяются различные виды средних величин: средняя арифметическая, средняя геометрическая, медиана, мода и другие. Наиболее распространенными являются средняя арифметическая, медиана и мода.

Средняя арифметическая применяется в тех случаях, когда между определяющим свойством и данным признаком имеется прямо пропорциональная зависимость.

Средняя арифметическая представляет собой частное от деления суммы величин на их число и вычисляется по формуле

$$\overline{x} = \frac{\sum x_i}{n},$$

где $\overline{x}$ – средняя арифметическая; x_i – результаты отдельных наблюдений; Σ – сумма результатов всех наблюдений (приемов, действий); n – количество наблюдений (приемов, действий).

Характеристиками вариационного ряда являются также медиана и мода.

Медианой (Me) называется мера среднего положения, характеризующая значение признака на упорядоченной (построенной по признаку возрастания или убывания) шкале, которое соответствует середине исследуемой совокупности. Медиана может быть определена для порядковых и количественных признаков. Место расположения этого значения определяется по формуле

$Me_{набл} = x_{(m)}$, где m=(n+1)/2 при нечетном n, $Me_{набл} = (x_{(m)}+x_{(m+1)})/2$, где m = n/2 при четном n.

Пример. Для вариационного ряда 3-3-3-4-4-4-4-4-5-5-5-5-6-6-6-6 m=16/2=8, восьмое и девятое значения в вариационном ряду – 4 и 5. $Me_{набл}$= (4+5)/2=4,5.

Мода (Mo) – наиболее часто встречающееся типичное значение признака среди других значений. Она соответствует классу с максимальной частотой. Этот класс называется модальным значением.

Пример.

Если на вопрос анкеты: «Какой шоколад Вы предпочитаете?», ответы распределились:

1 – горький	– 25.
2 – темный	– 54.
3 – молочный	– 253.
4 – с наполнителем	– 173.
5 – белый	– 28.

Очевидно, что наиболее типичным значением здесь является – «молочный», которое и будет модальным.

Существуют три формы вариационного ряда: ранжированный ряд, дискретный ряд, интервальный ряд.

Ранжированный ряд – это перечень членов вариационного ряда в порядке возрастания (убывания) изучаемого признака. Если признак принимает небольшое число целых значений, строится дискретный вариационный ряд.

Любое статистическое исследование должно начинаться с установления характера распределения изучаемых признаков. Распределение – это соотношение между значениями случайной величины и частотой их встречаемости. Статистическая теория началась с идеи подсчитать, как часто случается то или иное событие. Бóльшая повторяемость одних значений по сравнению с другими заставляет задумываться о причинах, о закономерностях наблюдаемых процессов. В качестве первичного описания любого явления может выступить частотное распределение. Если значения признака откладывать по оси абсцисс, а частоты их встречаемости по оси ординат, то можно построить *гистограмму*, частотную диаграмму, удобную для целей иллюстрации и исследования.

Основой для построения гистограммы служит вариационный ряд – представленный в виде таблицы ряд значений изучаемого признака (первый столбец), расположенных в порядке возрастания с соответствующими им частотами их встречаемости в выборке (второй столбец).

Пример. Изучение качества макаронных изделий дало следующие результаты (показатель – число деформированных макарон в 1 пачке): 5565564445646646455853655555636464625653763468635565438475431653456744656465.

Для дискретного признака (число деформированных макарон) строится вариационный ряд с учетом встречаемости конкретных значений.

Число деформаций, x	1	2	3	4	5	6	7	8
Частота, a	1	1	8	16	23	21	3	3

Гистограмма, построенная по данным о частоте деформаций (рис. 1), сразу же обнаруживает характерное поведение случайной величины – высокие частоты встречаемости значений в центре распределения и низкие по периферии.

Если численность дискретных единиц совокупности достаточно велика, ранжированный ряд становится громоздким. В таких случаях вариационный ряд строится с помощью группировки единиц совокупности по значениям изучаемого признака.

Если же изучаемый признак непрерывен (таковы размерно-весовые характеристики), то для построения вариационного ряда сначала преобразуют его в интервальный ряд. Весь диапазон изменчивости признака разбивается на серию равных интервалов (классов вариант), затем подсчитывают, сколько вариант попало в каждый интервал.

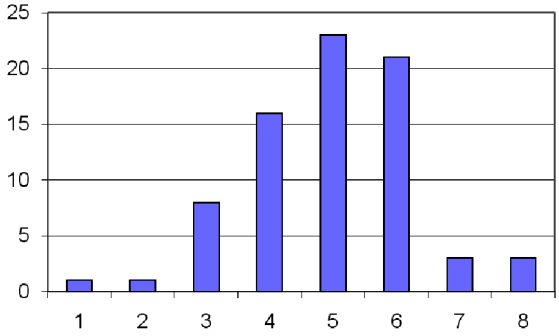


Рис. 1. Распределение числа деформаций макарон

Определение числа групп. Число групп в дискретном вариационном ряду определяется числом реально существующих значений варьирующего признака (как в примере с распределением частоты деформаций). Если же признак может принимать хотя и дискретные значения, но их число очень велико (например, поголовье скота на 1 января года в разных сельхозпредприятиях может составлять от нуля до десятков тысяч голов) или вариационный ряд состоит из непрерывных размерно-весовых характеристик, тогда строится интервальный вариационный ряд.

При построении интервального вариационного ряда необходимо выбрать оптимальное число групп (интервалов признака) и установить длину интервала. Поскольку при анализе вариационного ряда сравнивают частоты в разных интервалах, необходимо, чтобы величина интервала была постоянной. Оптимальное число групп выбирается так, чтобы в достаточной мере отразилось разнообразие значений признака в совокупности и в то же время закономерность распределения, его форма не искажалась случайными колебаниями частот. Если групп будет слишком мало, не проявится закономерность вариации; если групп будет чрезмерно много, случайные скачки частот исказят форму распределения.

Чаще всего число групп в вариационном ряду устанавливают, придерживаясь формулы, рекомендованной американским статистиком Стерджессом (Sturgess).

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

где k – число групп; n – объем выборки.

Эта формула показывает, что число групп – функция объема данных.

Предположим, необходимо построить вариационный ряд распределения веса куриных яиц (без сортового разбора).

Номер	Вес, г	Номер	Вес, г	Номер	Вес, г
1	52,1	19	64,8	37	69,1
2	54,4	20	65,1	38	69,5
3	56,3	21	65,4	39	69,6
4	57,4	22	65,6	40	71,2
5	58,4	23	65,8	41	71,4
6	59,1	24	65,9	42	71,5
7	60	25	66,2	43	71,6
8	60,2	26	66,6	44	71,7
9	60,5	27	66,8	45	72,4
10	60,7	28	67,2	46	73,7
11	60,8	29	67,3	47	73,8
12	60,8	30	67,4	48	74,1
13	61	31	67,6	49	74,9
14	61,2	32	67,7	50	75,2
15	61,3	33	68	51	75,4
16	62,3	34	68,1	52	75,9
17	63,7	35	68,3	53	77,1
18	64	36	68,8	54	81,6

Выполнить эту процедуру можно в среде программы Excel в следующей последовательности: ввести вариационный ряд (B1:BC1); определить объем выборки, минимальное и максимальное значения ряда, интервал (разница между максимальным и минимальным значениями). Ввести в ячейку C6 формулу расчета числа классов вариационного ряда и определить интервал класса. Определить правые границы классов, последовательно суммируя, начиная с минимального значения, правую границу и интервал класса.

Для подсчета частот на листе Excel следует вызвать программу (макрос) построения вариационного ряда командой меню Данные (Сервис) – Анализ данных – Гистограмма и заполнить окно. Каждое действие выполняется в два приема. Сначала нужно установить курсор в нужное окошко, щелкнув туда мышкой, затем мышкой же выделять соответствующие диапазоны ячеек листа Excel, нажимая левую кнопку над первой ячейкой диапазона и отпуская над последней.

В качестве «Входного интервала» задать массив ячеек, содержащих исходные значения вариант (B1:BC1). «Интервал карманов» – это блок значений правых границ классовых интервалов (C9:C15). Для «Выходного интервала» достаточно указать мышью одну ячейку (B17), это будет верхняя левая ячейка для блока результатов подсчета частот. После этого

нажать ОК. Если все сделано правильно, появятся представленные ниже результаты. Однако необходимо помнить, что на листе Excel значения частот ставятся в соответствие не центрам классовых интервалов, но их правым (большим) границам.

C6 | fx | =1+3,32*LOG10(64)

	A	B	C	D	E	F	G	H	I	J
1	ВАРИАЦИОННЫЙ РЯД	52,1	54,4	56,3	57,4	58,4	59,1	60	60,2	60,5
2	ОБЪЕМ ВЫБОРКИ		64							
3	МИНИМУМ		52,1							
4	МАКСИМУМ		81,6							
5	ИНТЕРВАЛ		29,5							
6	ЧИСЛО КЛАССОВ ВАРИАЦИОННОГО РЯДА (ОКРУГЛ)		7							
7	ИНТЕРВАЛ КЛАССА		4,2							
8										
9	ПРАВЫЕ ГРАНИЦЫ КЛАССОВ	1	56,3							
10		2	60,5							
11		3	64,7							
12		4	69,0							
13		5	73,2							
14		6	77,4							
15		7	81,6							
16										
17		*Карман*	*Частота*							
18		56,3	3							
19		60,5	6							
20		64,7	9							
21		69,0	18							
22		73,2	9							
23		77,4	8							
24		81,6	1							
25		Еще	0							
26										

3. НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

Представленные в таблице и на гистограммах предыдущего раздела данные распределения частот показывают характерную для многих признаков форму распределения: чаще встречаются значения средних интервалов признака, реже – крайние; малые и большие значения признака. Форма этого распределения близка к рассматриваемому в курсе математической статистики закону нормального (или Гауссова) распределения.

Закон нормального распределения выступает центральной моделью для строгого описания действительности, хотя статистическая теория предлагает множество других математических моделей.

Когда говорят, что данный признак имеет нормальное распределение, подразумевается, что «стохастическое поведение» этой случайной величины очень хорошо описывается формулой

$$p = \frac{1}{\sqrt{2\pi}} e^{-t^2/2},$$

аппроксимирующей специфическое соотношение между значениями (плотностью) непрерывной случайной величины (t) и частотой (вероятностью) встречаемости ее значений (p).

Форма нормального распределения – характерная «колоколообразная кривая» (рис. 2).

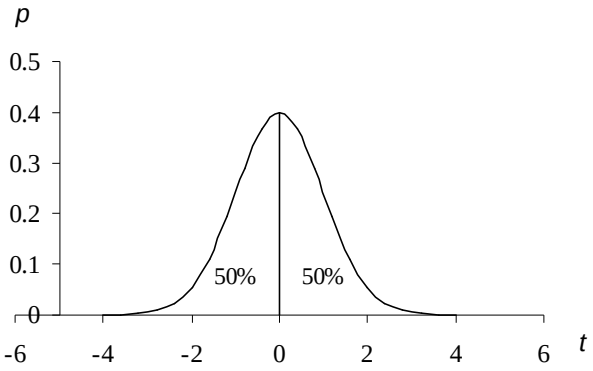


Рис. 2. Нормальное распределение

Практика показывает, что эта формула подходит к очень большому числу количественных характеристик. Модель нормального распределения чаще других используют для описания случайных событий. Ее применение (предположение о «нормальности» изучаемых признаков) дает в руки исследователя множество полезных и удобных инструментов решения задач. Это и интервальная оценка для прогноза ожидаемых значений случайной величины, и метод расчета наиболее теоретически обоснованных общих характеристик выборки (средних, дисперсий) и показателей сопряженной изменчивости разных признаков (корреляции), и пр. На идее нормального распределения базируются конструкции всевозможных статистических критериев для сравнения параметров разных выборок и проверки статистических гипотез. Кроме нормального закона статистической наукой обнаружены другие виды поведения случайных величин, которые основаны либо на том или ином допущении о нарушении условий формирования нормального закона, либо на специфическом преобразовании случайной величины, исходно распределенной нормально.

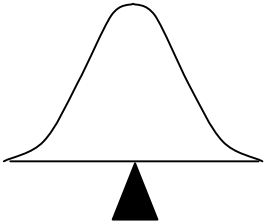
В большинстве случаев, связанных с измерениями в естественных дисциплинах, можно говорить о нормальном законе.

4. СУЩНОСТЬ И ЗНАЧЕНИЕ СРЕДНИХ ВЕЛИЧИН В СТАТИСТИКЕ

Точная форма нормального распределения определяется только двумя параметрами: **средним арифметическим и средним квадратичным** (стандартным) **отклонениями**.

Из нескольких видов средних (средняя арифметическая – простая и взвешенная, средняя гармоническая, средняя квадратичная) в практике естественнонаучных исследований наибольшее значение имеет средняя арифметическая величина, вокруг которой «концентрируются» варианты.

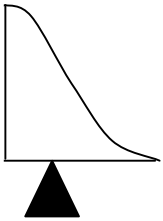
Физической аналогией может послужить такой образ средней арифметической для признака с нормальным распределением: средняя – это та точка *вырезанного из картонки распределения*, опираясь на которую левая и правая симметричные половинки уравновешивают друг друга.



При полном соответствии распределения нормальному значения средней арифметической, медианы и моды совпадают. Поэтому сравнение этих величин может послужить простейшим тестом соответствия изучаемого распределения нормальному.

Дисперсия и среднее квадратичное отклонение. Среднее квадратичное отклонение (или **стандартное отклонение**, σ) – вторая по значению константа вариационного ряда. Она является мерой разнообразия входящих в группу объектов и показывает, на сколько *в среднем* отклоняются варианты от средней арифметической изучаемой совокупности.

Продолжим рассмотрение физической аналогии, предложенной для средней. Разрежем вырезанное из картонки нормальное распределение по вертикальной линии строго пополам, начиная с точки средней арифметической. Стандартное отклонение для признака с нормальным распределением – это та точка *половинки*, вырезанной из картонки фигуры распределения, опираясь на которую левая и правая *несимметричные* части уравновешивают друг друга.



Дисперсия равна среднему квадрату отклонений значения варианты от среднего значения. Она выступает как одна из характеристик индивидуальных результатов разброса значений исследуемой переменной вокруг среднего значения. Дисперсией часто пользуются в расчетах, но более удобна характеристика **среднее квадратичное отклонение,** т.к. оно

имеет ту же размерность, что и исходные величины. Расчет непосредственно среднего отклонения от средней арифметической невозможен по той причине, что равновероятные отклонения вправо и влево от средней дают в сумме нуль.

Вычисление дисперсии осуществляется путем определения: отклонения от среднего значения; квадрата указанного отклонения; суммы квадратов отклонения и среднего значения квадрата отклонения (табл. 1).

Таблица 1

Пример вычисления дисперсии

№ п/п	Значение показателя	Отклонение от среднего	Квадрат отклонения
1	1	2 – 1 = +1	1
2	3	2 – 3 = –1	1
3	3	2 – 3 = – 1	1
4	0	2 – 0 = +2	4
5	4	2 – 4 = –2	4
6	1	2 – 1 = +1	1
$\sum_{i=1}^{n} x_i = 12$		$\sum_{i=1}^{n} (\overline{x} - x_i)^2 = 12$	
$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = 2$		$\sigma^2 = \frac{\sum_{i=1}^{n} (\overline{x} - x_i)^2}{n-1} = 2,4$	

Значение дисперсии используется в различных статистических расчетах, но не имеет непосредственного наблюдаемого характера. Величиной, непосредственно связанной с содержанием наблюдаемой переменной, является среднее квадратичное отклонение.

Среднее квадратичное отклонение подтверждает типичность и показательность средней арифметической, отражает меру колебания численных значений признаков, из которых выводится средняя величина. Оно равно корню квадратному из дисперсии и определяется по формуле

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}},$$

где σ – средняя квадратичная.

При малом числе наблюдения (действий) – менее 100 – в значении формулы следует ставить не «n», а «n – 1».

Оценивая результаты исследования важно определить рассеивание случайной величины около среднего значения. Это рассеивание описывается с помощью закона Гаусса (закона нормального распределения веро-

ятности случайной величины). Суть закона заключается в том, что при измерении некоторого признака в данной совокупности элементов всегда имеют место отклонения в обе стороны от нормы вследствие множества неконтролируемых причин, при этом, чем больше отклонения, тем реже они встречаются.

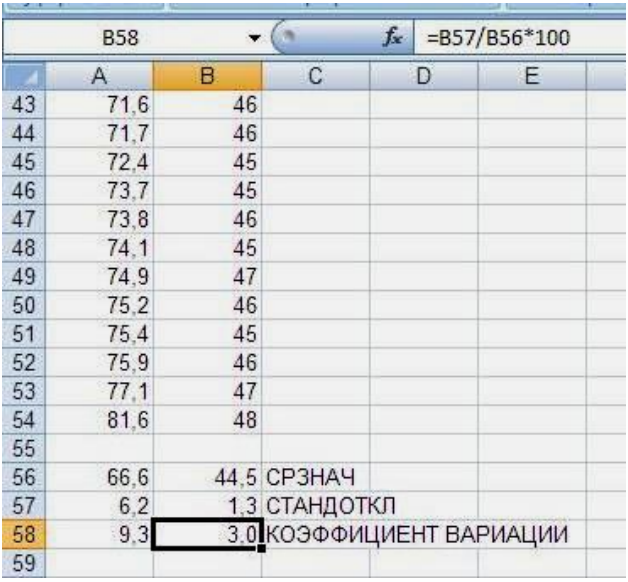
Стандартное отклонение – величина именованная, поэтому с ее помощью можно сравнивать характер варьирования лишь одних и тех же признаков. Чтобы сопоставить изменчивость разнородных признаков, выраженных в различных единицах измерения, а также нивелировать влияние масштаба измерений, используют так называемый **коэффициент вариации** (CV), безразмерную величину, отношение выборочной оценки стандартного отклонения к собственной средней.

$$CV = \frac{\sigma}{\overline{x}} \cdot 100\%.$$

Пример. В таблице приведены размерно-весовые характеристики куриных яиц. Вычислить средние величины распределений и уровень изменчивости признаков.

Номер	Вес, г	Ширина, мм	Номер	Вес, г	Ширина, мм	Номер	Вес, г	Ширина, мм
1	52,1	41	19	64,8	43	37	69,1	44
2	54,4	42	20	65,1	44	38	69,5	45
3	56,3	44	21	65,4	44	39	69,6	44
4	57,4	43	22	65,6	44	40	71,2	44
5	58,4	43	23	65,8	45	41	71,4	44
6	59,1	44	24	65,9	45	42	71,5	45
7	60	42	25	66,2	45	43	71,6	46
8	60,2	43	26	66,6	45	44	71,7	46
9	60,5	42	27	66,8	45	45	72,4	45
10	60,7	45	28	67,2	45	46	73,7	45
11	60,8	43	29	67,3	44	47	73,8	46
12	60,8	43	30	67,4	46	48	74,1	45
13	61	43	31	67,6	45	49	74,9	47
14	61,2	44	32	67,7	45	50	75,2	46
15	61,3	45	33	68	46	51	75,4	45
16	62,3	44	34	68,1	44	52	75,9	46
17	63,7	45	35	68,3	44	53	77,1	47
18	64	45	36	68,8	45	54	81,6	48

Расчеты можно провести в среде Excel. Для этого создадим по столбцам массивы данных веса и ширины яиц (A1:A54) и (B1:B54). В свободные ячейки (A56 и A57) через опцию «вставка функции» (*fx*) из статистических категорий вставим соответственно СРЗНАЧ и СТАНДОТКЛОН для блока данных. Введя в ячейку A58 формулу, рассчитаем коэффициент вариации признака. Выделив ячейки (A56:A58) переведем расчеты в блок B56:B58.



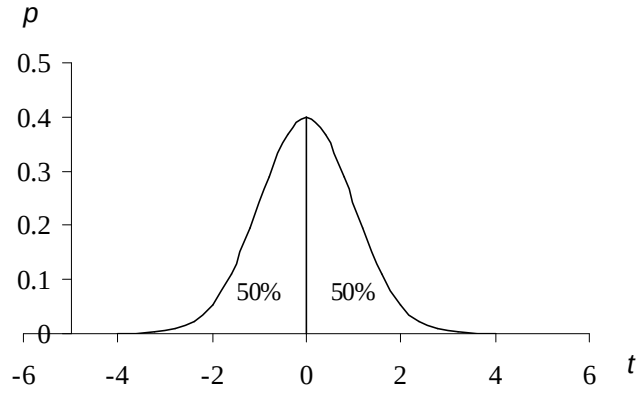
Из результатов видно, что вес яиц более изменчивый признак, чем ширина.

5. СВОЙСТВА НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

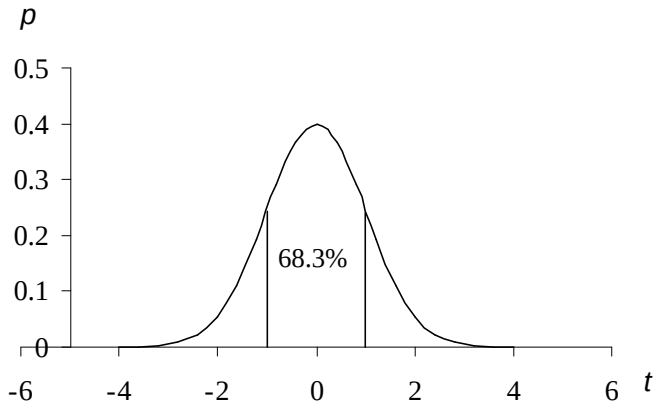
Уравнение нормального распределения определяет ход кривой линии, имеющей характерную колоколообразную форму, т.е. позволяет вычислить *ординаты нормальной кривой*, или «плотность вероятности» (*p*). Вероятность – численная мера возможного, определяется как отношение числа вариант (исходов испытаний) определенного вида к общему числу вариант (опытов). Поскольку нормальное распределение характерно для непрерывных случайных величин, говорят не о вероятности какого-то определенного значения варианты, но о «плотности вероятности», отражая тем самым плавность изменения вероятности значений для разных значений *t*, чем ближе к центру распределения, тем плотность вероятности выше. С помощью уравнения плотности вероятности можно рассчи-

тать (интегрируя) вероятность появления нового значения случайной величины в том или ином интервале значений *t*. Итак, формула количественно выражает вполне определенные свойства поведения случайной величины, из которых можно назвать следующие практически важные следствия:

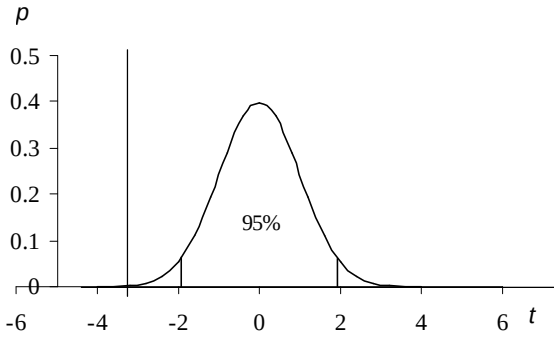
1. Все варианты лежат в интервале плюс-минус бесконечность. Иными словами, с вероятностью $P = 1$ ($P = 100\%$) мы вправе ожидать появление новой варианты в пределах от $-\infty$ до $+\infty$. Слева и справа от средней арифметической лежат по 50% вариант, т.е. с вероятностью $P = 0{,}5$ (50%) можно предсказать появление новой варианты в интервалах $\overline{x} - \infty$ и $\overline{x} + \infty$.



2. В интервале от $\overline{x} - 1S$ до $\overline{x} + 1S$ лежат 68,3% всех вариант; с вероятностью $P = 0{,}683$ ($P = 68{,}3\%$) можно прогнозировать появление новой варианты на расстоянии $\pm 1S$ от средней, или в диапазоне $\overline{x} \pm S$.



3. Между $\overline{x}$ –1,96σ до $\overline{x}$ +1,96σ лежат 95% вариант. Это позволяет с 95%-й вероятностью предполагать, что новая варианта окажется в интервале $\overline{x}$ ±1,96σ (округленно $\overline{x}$ ±2σ – так называемое правило двух стандартных отклонений).



4. С вероятностью P = 0,99 значение новой варианты будет заключено в пределах $\overline{x}$ ±2,58σ и с вероятностью P = 0,999 – в интервале $\overline{x}$ ±3,3σ.

Исходя из сказанного, можно оценить вероятность появления новых значений признака. В отношении непрерывных случайных величин (метрических признаков) эта процедура сводится к так называемой интервальной оценке. Для полученных ранее характеристик, массы куриных яиц, средней $\overline{x}$ = 66,6 и стандартного отклонения σ = 6,2 (г), находим доверительные интервалы: $\overline{x}$ ±1σ = 66,6±6.2, $\overline{x}$ ±1,96σ = 66,6±12,1. Новое значение признака с вероятностью P = 0,68 ожидается в пределах 60,4-72,8 г, а с вероятностью P = 0,95 – между 54,5 и 78,7 г.

Важнейшее значение для практического применения имеет «соглашение о 95%». В соответствии с ним совокупности, состоящей из 95% объектов, мы доверяем так же, как и 100%-й. Термин «*доверительная вероятность* P = 0,95» означает, что, согласно принятому допущению, 95% вариант достаточно полно характеризуют изучаемое явление (в данном случае изменчивость веса куриных яиц), что позволяет ограничиться рассмотрением вариант в области $\overline{x}$ ±1,96σ, охватывающей эту 95%-ю совокупность.

При этом в биометрии обычно довольствуются доверительной вероятностью P = 0.95 (уровень значимости α = 0,05), хотя в наиболее ответственных исследованиях принимают и более строгие уровни – P = 0,99 и P = 0,999. Однако это имеет смысл лишь при очень больших выборках исходных данных, точно описывающих закономерности изменчивости признаков. Обычно же выборки не очень велики, что позволяет ограничиться меньшей степенью доверительной вероятности P = 0,95.

«Уровень значимости» – понятие, альтернативное доверительной вероятности и, соответственно, составляет разность между единицей и доверительной вероятностью ($\alpha = 1–P$). Для доверительной вероятности 0,95 уровень значимости составляет 0,05, а для 0,99 и 0,999 – соответственно 0,01 и 0,001. Уровень значимости, равный 0,05 (5%), можно интерпретировать так: имеется всего 5% шансов, что полученная величина не будет соответствовать изучаемой совокупности. Уровень значимости – это тот теоретический процент вариант нормального распределения, который можно отбросить, не учитывать, дабы с меньшими усилиями получить основную информацию об изучаемом явлении. Можно потратить очень много времени на поиски яйца весом в 100 г, но так и не собрать выборку, достаточную по объему, чтобы это реализовать (рекордный вес куриного яйца зарегистрирован на Кубе – 148 г). Поэтому использование доверительной вероятности и уровня значимости можно назвать средством (теоретической базой) разумного ограничения материала (времени и масштабов исследования), позволяющего получить достоверную общую информацию за счет исключения ничтожной доли частной (излишне конкретной). В итоге такой прием дает возможность найти границы нормальной изменчивости изучаемых признаков и отбросить ошибочные, наведенные и артефактные значения.

5.1. Ошибка репрезентативности выборочных параметров

По части (выборке) никогда не удается полностью охарактеризовать целое, всегда остается вероятность того, что оценка генеральной совокупности на основе выборочных данных недостаточно точна, имеет некоторую большую или меньшую ошибку. Такие ошибки, представляющие собой ошибки обобщения, экстраполяции, связанные с перенесением результатов, полученных при изучении выборки, на всю генеральную совокупность, называются ошибками репрезентативности (репрезентативность – степень соответствия выборочных показателей генеральным параметрам). Отличия значений выборочных параметров от генеральных называются ошибкой репрезентативности данного параметра, или просто (статистической) ошибкой.

Если для каждой из нескольких выборок одной генеральной совокупности рассчитать средние значения, получим разные величины. Эти отличия и есть ошибки репрезентативности, связанные с неточностью оценок по небольшим выборкам. Если теперь мы найдем среднеквадратичное отклонение этих отдельных средних от общей, оно будет характеризовать средний диапазон отклонения выборочных оценок от генеральных значений. Эта величина называется ошибкой средней арифметической (или стандартной ошибкой) и является по существу *средним квадратичным отклонением множества выборочных средних от генеральной средней*. На практике обычно нет возможности делать несколько выборок

и вычислять несколько выборочных средних, чтобы по ним проводить расчеты. Статистическая теория показывает, что ошибка средней в $\sqrt{n}$ раз меньше, чем стандартное отклонение. Значит, ошибку можно рассчитать для единичной отдельной выборки по формуле

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

Статистические ошибки служат мерой тех пределов, в которых выборочные частные оценки могут отклоняться от параметров генеральной совокупности. Как следует из конструкции расчетной формулы, величина ошибки тем больше, чем больше варьирование признака (σ) и чем меньше выборка (n). При увеличении объема выборки ошибки репрезентативности стремятся к нулю (следствие закона больших чисел).

Ошибку репрезентативности имеют все статистические параметры, рассчитанные по выборке: средняя, стандартное отклонение, коэффициент вариации, показатели асимметрии и эксцесса. Для разных типов распределений расчетные формулы могут немного изменяться. Для нормального распределения они имеют следующий вид:

ошибка средней $m_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$;

ошибка стандартного отклонения $m_{\sigma} = \frac{\sigma}{\sqrt{2 \cdot n}}$;

ошибка коэффициента вариации $m_{CV} = \frac{CV}{\sqrt{2 \cdot n}}$.

Вычисленные значения ошибок подставляют к соответствующим параметрам со знаками плюс-минус (параметр ± ошибка) и в такой форме представляют в научных отчетах и публикациях.

Вернемся к примеру с весом куриных яиц и определим соответствующие ошибки:

средней арифметической $m_{\bar{x}} = \frac{6,2}{\sqrt{54}} = 0,84$, x = 66,6±0.8 г;

стандартного отклонения $m_{\sigma} = \frac{6.2}{\sqrt{2 \cdot 54}} = 0,59$, σ = 6,2±0,6 г;

коэффициента вариации $m_{CV} = \frac{9.6}{\sqrt{2 \cdot 54}} = 0,92$, CV = 9,6±0,9%.

Не следует путать статистическую ошибку с методическими ошибками и ошибками точности (точности измерений, анализов, подсчетов и т.д.), хотя методические погрешности и увеличивают ошибку репрезента-

тивности, но другим путем – методические огрехи увеличивают изменчивость признака, стандартное отклонение. Чем лучше взята выборка, чем больше ее размеры, т.е. чем вернее отражает она генеральную совокупность (все явление, весь процесс в полном объеме), тем меньше статистическая ошибка и расхождение между значениями признаков в выборочной и генеральной совокупностях. При всей неизбежности статистической ошибки она может быть сведена к минимуму отбором достаточного числа особей (вариант). С ростом объема выборки оценки параметров стабилизируются, а их ошибки репрезентативности уменьшаются.

5.2. Доверительный интервал

При конкретных наблюдениях параметры генеральной совокупности остаются неизвестными, о них судят по выборочным оценкам, используя для этого величину ошибок репрезентативности. Границы, в которых с той или иной вероятностью находится параметр генеральной совокупности, называются доверительными, а интервал, заключенный между этими границами, – доверительным интервалом. Теоретические исследования поведения выборочных средних (как случайных величин) показали, что они подчиняются нормальному закону, большинство из них (95%) находится поблизости от генеральной средней – в диапазоне $\overline{x}_{ген.} \pm 1{,}96 \cdot m$. Это обстоятельство позволяет делать обратное заключение – генеральная средняя находится в диапазоне $\overline{x}_{выбор.} \pm 1{,}96 \cdot m$, т.е. предсказывать ширину интервала, в котором находится генеральный параметр, давать *интервальную оценку* генеральному параметру. В соответствии с законом нормального распределения можно ожидать, что генеральный параметр (истинное значение) окажется в интервале

$$\text{от } \overline{x} - T \cdot m \text{ до } \overline{x} + T \cdot m,$$

где m – ошибка средней арифметической; T – квантиль распределения Стьюдента (см. приложение) при данном числе степеней свободы (df) и уровне значимости (обычно $\alpha = 0{,}05$).

Сказанное можно перефразировать так: с вероятностью $P = 0{,}95$ можно ожидать, что генеральная средняя находится в доверительном интервале $\overline{x} \pm T \cdot m$, построенном вокруг выборочной средней арифметической M.

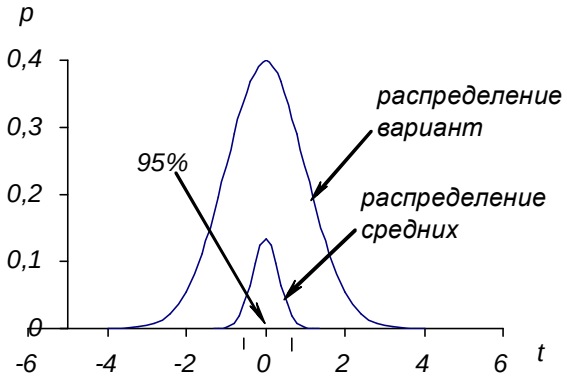
Возвращаясь к примеру о весе куриных яиц, мы теперь можем записать доверительные интервалы при разных уровнях вероятности (граничные значения T взяты для случая $n = \infty$):

Для $P = 0{,}95$ $\overline{x} \pm T \cdot m = 66{,}6 \pm 1{,}96 \cdot 0{,}84 = 66{,}6 \pm 1{,}6$ г.

Для $P = 0{,}99$ $\overline{x} \pm T \cdot m = 66{,}6 \pm 2{,}58 \cdot 0{,}84 = 66{,}6 \pm 2{,}2$ г.

Для $P = 0{,}999$ $\overline{x} \pm T \cdot m = 66{,}6 \pm 3{,}30 \cdot 0{,}84 = 66{,}.6 \pm 2{,}8$ г.

Таким образом, искомая генеральная средняя величина веса яиц с вероятностью $P = 95\%$ находится в пределах 65,0-68,2 г, с вероятностью $P = 99\%$ – 64,4-68,8, для $P = 99{,}9\%$ – 63,8-69,4 г.



Если объем выборки, для которой были получены параметры и вычислялась ошибка репрезентативности m, был невелик (n<500), то необходимо вводить поправки на объем выборки, расширяя область возможного пребывания генерального параметра. Это понятно, поскольку при дефиците информации любые заключения не могут быть очень точными. Рассчитаем доверительный интервал для тех же данных, но с объемом n = 54 экз. Ошибка средней арифметической составит

$$m_{\bar{x}} = \frac{6,2}{\sqrt{54}} = 0,8 \text{ г}, \ \bar{x} = 66,6 \pm 0,8 \text{ г}.$$

При уровне значимости $\alpha = 0,05$ и числе степеней свободы $df = n–1 = 54–1 = 53$ табличная величина статистики Стьюдента равна T = 2,008, тогда доверительный интервал составит

$$\bar{x} \pm T \cdot m = 66,6 \pm 2,008 \cdot 0,8 = 66,6 \pm 1,7 \text{ г} – \text{от } 64,9\text{-}68,3 \text{ г}.$$

Аналогичным образом можно построить доверительный интервал для стандартного отклонения, коэффициента вариации, а также других статистических параметров (коэффициентов асимметрии, эксцесса, регрессии, корреляции), рассмотренных в следующих разделах.

5.3. Определение точности опыта

Статистическая ошибка позволяет судить о надежности полученных результатов, т.е. о том, достаточное ли количество случаев при данной величине изменчивости было получено, чтобы по части характеризовать целое. В практике биометрического анализа используется относительная ошибка измерений – «показатель точности опыта» или «показатель точности оценки параметров» – отношение ошибки средней к самой средней арифметической, выраженное в процентах

$$\varepsilon = \frac{m}{\bar{X}} \cdot 100\%.$$

Чем точнее определена средняя, тем меньше будет ε, и наоборот. Точность считается хорошей, если ε меньше 3%, и удовлетворительной при $3\% < \varepsilon < 5\%$. Если относительная ошибка превышает 5%, полученные данные следует уточнить (повторить опыт, собрать дополнительный материал и т. д.).

В разобранном выше примере определения характеристик массы куриного яйца показатель точности составил $\varepsilon = (0{,}8/66{,}6) \cdot 100 = 1{,}3\%$, что говорит о достаточной надежности выборочной оценки.

5.4. Оптимальный объем выборки

В исследованиях, при планировании экспериментов, для определения величины контрольных групп часто заранее требуется установить число наблюдений, достаточное для получения правильного представления о явлении в целом (получить репрезентативные оценки генеральной совокупности). Можно говорить о двух разных подходах для непрерывных и дискретных признаков.

Идея первого метода состоит в том, чтобы, используя известные соотношения (все формулы были представлены выше) между средней, стандартным отклонением, ошибкой средней, плотностью вероятности распределения Стьюдента (на их основе вычисляются коэффициент вариации, показатель точности оценок, доверительный интервал), найти число степеней свободы, соответствующее доверительному интервалу для средней при уровне значимости $\alpha = 0{,}05$. Иными словами, решается задача, прямо противоположная рассмотренной в предыдущем разделе.

Объем выборки, достаточной для получения результата заданной точности, находят по формуле

$$n = \left(\frac{T \cdot CV}{\varepsilon} \right)^2,$$

где n – объем выборки; T – граничное значение из таблицы распределения Стьюдента (см. приложение), соответствующее принятому уровню значимости при планируемом объеме выборки (в крайнем случае можно взять значение $T = 1{,}96$ для $df = \infty$); CV – приблизительное значение коэффициента вариации (%); ε – планируемая точность оценки (погрешности) (%).

Рассчитаем необходимый объем условной выборки, обеспечивающий хорошую точность $\varepsilon = 3\%$, для уровня значимости $\alpha = 0{,}05$ ($T = 1{,}98$, для $df \approx 100$) и для коэффициента вариации $CV = 10\%$ (такова относительная изменчивость многих размерно-весовых признаков).

$$n = \left(\frac{1.98 \cdot 10}{3} \right)^2 = 43{,}56 \approx 44 \text{ шт.}$$

5.5. Асимметрия и эксцесс

В практике исследований нередки случаи, когда числовые значения признаков дают распределения, в той или иной мере отличающиеся от нормального. Иногда обнаруживается асимметричное, в других сериях – эксцессивное распределение (рис. 3, 4). Для асимметричных вариационных кривых характерно появление «хвоста» – сдвиг частот от средних значений вправо или влево. В распределении эксцессивных признаков наблюдается чрезмерное накапливание или, наоборот, снижение частот в центральных классах вариационного ряда, вследствие этого вершина кривой распределения либо сильно поднимается и заостряется (положительный эксцесс, островершинность), либо, напротив, опускается, приобретая вид широкого плато (отрицательный эксцесс, туповершинность) или даже седловины между двумя боковыми вершинами. Для нормального распределения коэффициенты асимметрии и эксцесса равны нулю.

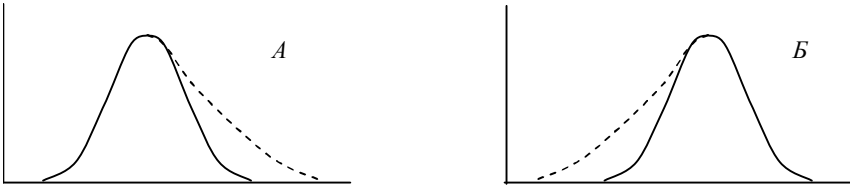


Рис. 3. Асимметрия распределения (обозначена пунктиром относительно нормальной кривой): А – положительная (правосторонняя), Б – отрицательная (левосторонняя)

При расчете коэффициентов асимметрии А и эксцесса Е используются следующие базовые формулы:

$$A = \frac{1}{n} \cdot \frac{\sum (x - \overline{x})^3}{\sigma^3},$$

$$E = \frac{1}{n} \cdot \frac{\sum (x - \overline{x})^4}{\sigma^4} - 3.$$

Более сложные формулы, дающие несмещенные оценки, реализованы в среде Excel в виде функций: для оценки асимметрии =СКОС (диапазон) и для оценки эксцесса =ЭКСЦЕСС (диапазон). Диапазон ячеек должен содержать все значения изучаемой выборки.

Рис. 4. Эксцесс распределения (обозначен пунктиром относительно нормальной кривой): А – положительный (островершинность), Б – отрицательный (туповершинность)

Показатели асимметрии и эксцесса используются в качестве тестов для проверки соответствия эмпирического распределения нормальному. Статистическая значимость этих показателей говорит о нарушении нормальной формы кривой распределения. Критерии Стьюдента для $df = \infty$ проверяют нулевую гипотезу Но: «коэффициент асимметрии (эксцесса) существенно от нуля не отличается, следовательно, асимметрия (эксцесс) достоверно не выражена»:

$$T = \frac{A - 0}{m_A};$$

$$T = \frac{E - 0}{m_E},$$

где m – статистическая ошибка соответствующего коэффициента.

Точная и приближенная формулы для расчета статистической ошибки показателя асимметрии и эксцесса составляют:

$$m_A = \sqrt{\frac{6 \cdot n \cdot (n-1)}{(n-2) \cdot (n+1) \cdot (n+3)}} \approx \sqrt{\frac{6}{n}};$$

$$m_E = \sqrt{\frac{24 \cdot n \cdot (n-1)^2}{(n-3) \cdot (n-2) \cdot (n+3) \cdot (n+5)}} \approx \sqrt{\frac{24}{n+5}} \approx 2 \cdot \sqrt{\frac{6}{n}}.$$

Проведем вычисление коэффициентов асимметрии и эксцесса для данных по куриным яйцам:

A =СКОС (A1:A54) = -0,03819;

m_A = КОРЕНЬ(6·54·(54-1)/((54-2)·(54+1)·(54+3))) = 0,324556,

T_A = 0,11767;

E =ЭКСЦЕСС(A1:A54) = -0,22731;

m_E = КОРЕНЬ(24·54·(54-1)^2/((54-3)·(54-2)·(54+3)·(54+5))) = 0,638893;

T_E = 0,35579.

Табличное значение критерия Стьюдента составляет $T_{(0.05,\infty)} = 1{,}96$. Поскольку полученное значение $T_A = 0{,}12$ меньше табличного (1,96), коэффициент асимметрии значимо не отличается от нуля. Поскольку полученное значение $T_E = 0{,}36$ меньше табличного (1,96), коэффициент эксцесса значимо от нуля не отличается. Распределение в целом соответствует нормальному.

Вычислить все рассмотренные параметры вариационного ряда можно в среде Excel с помощью макроса, который вызывается командой меню Данные (Сервис)/ Анализ данных/ Описательная статистика.

Например, обработка данных по массе куриных яиц дает следующие результаты.

Среднее	66,60185
Стандартная ошибка	0,8404
Медиана	67
Мода	60,8
Стандартное отклонение	6,175652
Дисперсия выборки	38,13868
Эксцесс	-0,22731
Асимметричность	-0,03819
Интервал	29,5
Минимум	52,1
Максимум	81,6
Сумма	3596,5
Счет	54

6. ОБЩАЯ ХАРАКТЕРИСТИКА МЕТОДОВ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ДАННЫХ

Анализ взаимосвязи между большим количеством переменных осуществляется путем использования методов статистической обработки.

Цель применения подобных методов – сделать наглядными скрытые закономерности, выделить наиболее существенные взаимосвязи между переменными.

При всем кажущемся многообразии вариантов проявления различного рода закономерностей, можно выделить всего 4 класса статистических задач вида «доказать отличия»:

1. Доказать чужеродность варианты в выборке
(или «классифицировать объекты»).

2. Доказать отличие двух выборок.

3. Доказать отличие нескольких выборок

(или «доказать влияние фактора на признак»).
4. Найти зависимость между признаками
(или «доказать сопряженность варьирования признаков»).
Выбрать методы решения статистических задач позволит табл. 2.

Таблица 2

Определитель статистического метода

Что доказать?	Что изучено?	Метод
Чужеродность варианты в выборке	Величина	Сравнение средней и варианты
Достоверность отличия двух выборок	Величина	Сравнение средних арифметических
	Изменчивость	Сравнение дисперсий
	Распределение частот	Сравнение эмпирического и теоретического распределений
		Сравнение двух эмпирических распределений
	В целом	Сравнение двух наборов значений
Достоверность отличия нескольких выборок	Величина	Дисперсионный анализ
	Изменчивость	Сравнение серии дисперсий
	Распределение частот	Сравнение нескольких эмпирических распределений
	В целом	Непараметрический дисперсионный анализ
Достоверность влияния фактора на признак	Величина	Дисперсионный анализ
Достоверность влияния признака на признак	Величина	Регрессионный анализ
Достоверность сопряженности варьирования двух признаков	Величина	Корреляционный анализ

6.1. Доказать чужеродность варианты в выборке

Часто встречается ситуация, когда одна из полученных вариант сильно отличается от остальных. Эти отклонения могли возникнуть в результате неточности измерений, ошибок внимания, методических погрешностей и т. д.

Можно ли такие резко выделяющиеся значения использовать при дальнейших расчетах?

Общепринятой безразмерной характеристикой отклонения отдельной варианты от средней арифметической служит *нормированное отклонение*, оно показывает, на сколько стандартных отклонений отклоняется та или иная варианта от среднего уровня варьирующего признака и выражается формулой

$$t = \frac{x - \overline{x}}{\sigma} \sim t_{табл},$$

где t – критерий выпада (исключения); x – выделяющееся значение признака; M – средняя величина для группы вариант; $t_{табл}$ – стандартные значения критерия выпадов, определяемые свойствами нормального распределения.

Если значение критерия больше табличного, то это означает, что данное значение не относится к анализируемой совокупности, а есть проявление каких-то особых закономерностей, ошибок и пр. и должно быть поэтому исключено из рассмотрения (отброшено). При этом иногда рекомендуют значения параметров (M, S) рассчитывать без учета «подозрительной» варианты. После такой «чистки» параметры выборки должны быть рассчитаны заново.

Пример. В выборке куриных яиц одно яйцо имеет сильно превышающий остальные вес – 81,6 г. Значение признака за пределами доверительного интервала $\overline{x} \pm 1{,}96\sigma = 66{,}6 \pm 12{,}1$ для $P = 0{,}95$ (между 54,5 и 78,7 г). Критерий исключения $t = (81{,}6\text{-}66{,}6)/6{,}2 = 2{,}428594$. Табличное значение критерия t для n=50 и α=0,05 составляет 2,03. Критерий выпада (2,43) превышает табличное значение (2,03), в итоге этот сомнительный результат должен быть отброшен. Однако для n=50 и α=0,01 табличное значение критерия t равно 2,71. Следовательно, при более жестких требованиях к результатам анализа это значение укладывается в пределы одной генеральной совокупности с другими значениями выборки.

6.2. Доказать отличие двух выборок

Сравнение двух выборок по величине признака. Задача сравнения выборочных средних – это вопрос о том, действовал ли в одной из выборок новый систематический фактор по сравнению с другой выборкой? В терминах статистики отличия между средними могут иметь два противоположных источника:

1. Обе выборки взяты из одной генеральной совокупности, но средние отличаются в силу ошибки репрезентативности.

2. Выборки взяты из разных генеральных совокупностей, отличие средних вызвано, в основном, действием разных доминирующих факторов (а также и случайно).

Статистическая задача состоит в том, чтобы сделать обоснованный выбор. Исходно предполагается (нулевая гипотеза – Но): «достоверных отличий между средними нет».

Для исключения чужеродных («выскакивающих») вариант мы применяли закон нормального распределения: в диапазоне четырех стандартных отклонений, $\overline{x} \pm 1{,}96 \cdot \sigma$, отклонение вариант от средней происходит по случайным причинам; за границами этого диапазона лежат чужеродные для данной выборки значения. Поскольку выборочные средние имеют нормальное распределение (см. разд. «Ошибка репрезентативности выборочных параметров»), критерий отличия двух выборочных средних также базируется на свойствах *нормального распределения*: в границах $\overline{x}_{общ.} \pm 1{,}96 \cdot m$ (или приблизительно $\overline{x}_{общ.} \pm 2 \cdot m$) выборочные средние арифметические отличаются от общей (генеральной) средней по случайным причинам. Критерий отличия средних формируется по типу критерия «исключения», если одну из выборочных средних ($\overline{x}_1$) принять в качестве генеральной средней, другую взять как «подозрительную» варианту ($\overline{x}_2$), а роль характеристики варьирования играет обобщенная ошибка репрезентативности (m_d).

$$t = \frac{x - \overline{x}}{\sigma} \Rightarrow t = \frac{\overline{x}_1 - \overline{x}_2}{m_d}.$$

Обобщенная ошибка получена объединением двух ошибок, рассчитанных по сравниваемым выборкам (для случая, когда выборочные дисперсии отличаются несильно)

$$m_d = \sqrt{m_1^2 + m_2^2},$$

которые, в свою очередь, определены рассмотренным выше соотношением

$$m = \frac{\sigma}{\sqrt{n}}.$$

Тогда рабочая формула для T критерия отличия средних будет

$$T = \frac{\left|\overline{x}_1 - \overline{x}_2\right|}{\sqrt{m_1^2 + m_2^2}} \sim T_{(\alpha,\, df)}.$$

Следует помнить, что разность средних нужно брать по модулю, т.е. без учета знака. Полученное этим способом значение критерия T Стьюдента сравнивают с табличным при выбранном уровне значимости (обычно для $\alpha = 0{,}05$) и числе степеней свободы (объемы выборок без числа ограничений, $df = n_1 + n_2 - 2$). Результатом такого сравнения должен стать один из двух вариантов следующего статистического вывода. Если полученное значение (величина) критерия больше табличного, значит, различия между параметрами при заданном уровне значимости и установленном числе степеней свободы статистически значимы. Если же по-

лученная величина критерия меньше табличной, то при данном уровне значимости и числе степеней свободы различия между параметрами статистически не значимы. Последнее говорит о том, что различия случайны, никакого определенного вывода сделать нельзя, нулевая гипотеза остается не опровергнутой.

Табличные значения критерия следует брать из таблицы Стьюдента (см. приложение). Обычно эта статистика соответствует нормальному распределению, но в случае небольших выборок дает необходимую поправку на объем выборки, предупреждает возможность сделать слишком жесткий вывод по недостаточным данным. По этой причине критерий различия средних арифметических носит название критерия Стьюдента*. Одно из необходимых требований к применению этого критерия – это уверенность в том, что изучаемые признаки имеют распределение, в целом соответствующее нормальному.

Пример. Исследовали спагетти марок «Макфа» и «Смак»: измеряли длину спагетти до (1) и после (2) варки в одинаковом режиме. Вопрос – можно ли отнести изделия двух производителей к одной генеральной совокупности?

Расчеты можно провести в среде Excel с помощью макроса, который вызывается командой меню Данные (Сервис)/ Анализ данных/ Описательная статистика. Входной интервал задать по всему массиву данных (B3:E12). В параметрах вывода выставить флажок на Итоговая статистика и вывести результат на Новый рабочий лист.

Затем, удалив лишние столбцы с наименованиями параметров описательной статистики, скопировать и выставить результаты в соответствующие ячейки рабочего листа с массивом первичных данных. Ввести в ячейки (B29 и D29) формулу расчета критерия T соответственно для пар средних значений длины спагетти до и после варки.

Из результатов видно, что расчетный критерий сравнения продукции разных производителей до варки (0,827) меньше табличного значения критерия для $\alpha = 0{,}05$ и $df = n_1+n_2-2=18$ (2,101). Следовательно, продукцию разных производителей можно отнести к одной генеральной совокупности – произведенной по единой технологии на аналогичном оборудовании. Однако оценка значимости различия средних значений длины спагетти после варки показывает, что расчетное значение критерия (6,091) превышает табличное значение. Таким образом, вареные спагетти можно отнести к разным генеральным совокупностям. Причиной этому может быть то, что продукция двух наименований произведена из муки разного качества.

---

* **Критерий Стьюдента** был разработан английским химиком У. Госсетом, когда он работал на пивоваренном заводе Гиннеса и по условиям контракта не имел права открытой публикации своих исследований. Поэтому публикации своих статей по t-критерию У. Госсет сделал в 1908 г. в журнале «Биометрика» под псевдонимом «Student», что в переводе означает «Студент». В отечественной же литературе принято писать «Стьюдент»

6.3. Доказать отличие нескольких выборок (или «доказать влияние фактора на признак»)

Дисперсионный анализ (от латинского dispersio – рассеивание; разработан биологом Р. Фишером в 1925 г.) – статистический метод, используемый для изучения одной или нескольких одновременно действующих и независимых переменных на изменчивость наблюдаемого признака. Его особенность состоит в том, что наблюдаемый признак может быть только количественным, в то же время объясняющие признаки могут быть как количественными, так и качественными.

Целью дисперсионного анализа является проверка значимости различия между средними с помощью сравнения дисперсий. Дисперсию измеряемого признака разлагают на независимые слагаемые, каждое из которых характеризует влияние того или иного фактора или их взаимодействия. Последующее сравнение таких слагаемых позволяет оценить значимость каждого изучаемого фактора, а также их комбинации.

Дисперсионный анализ позволяет решать следующие задачи: измерение силы влияний; определение достоверности влияний; оценка генеральных параметров влияния в форме доверительных границ; оценка разности частных средних; функциональный (регрессионный) анализ ряда частных средних и т.д.

B29 =(ABS(B14-D14)/КОРЕНЬ(B15*B15+D15*D15))

	A	B	C	D	E	F	G
1		"Макфа"		"Смак"			
2		1	2	1	2		
3		24,8	29,5	24,9	25,5		
4		24,8	28,6	24,6	25,9		
5		25,2	30	25,3	25,9		
6		24,9	29,5	25,1	25,6		
7		24,4	28,5	24,8	25,7		
8		24,9	28,3	25,2	26		
9		25,2	28	25	25,8		
10		23,8	27,5	25,2	25,5		
11		25	25,9	25	27,5		
12		25,3	29,6	24,6	25,9		
13							
14	Среднее	24,83	28,54	24,97	25,93		
15	Стандартная ошибка	0,14067	0,38736	0,07753	0,18321		
16	Медиана	24,9	28,55	25	25,85		
17	Мода	24,8	29,5	24,6	25,9		
18	Стандартное отклонение	0,44485	1,22493	0,24518	0,57937		
19	Дисперсия выборки	0,19789	1,50044	0,06011	0,33567		
20	Эксцесс	2,60996	1,21935	-0,95159	7,59129		
21	Асимметричность	-1,51709	-1,02947	-0,41503	2,61448		
22	Интервал	1,5	4,1	0,7	2		
23	Минимум	23,8	25,9	24,6	25,5		
24	Максимум	25,3	30	25,3	27,5		
25	Сумма	248,3	285,4	249,7	259,3		
26	Счет	10	10	10	10		
27	CV	1,8	4,3	1,0	2,2		
28							
29	T-критерий отличия средних	0,872		6,091			
30							
31	Критерий Стьюдента (0,05; 18)	2,101					
32							

В зависимости от количества факторов, определяющих вариацию результативного признака, дисперсионный анализ подразделяют на *однофакторный* и *многофакторный*.

При проведении дисперсионного анализа изучаются три основных вида статистических влияний: факториальное, случайное и общее.

***Факториальное влияние* – это простое или комбинированное статистическое влияние изучаемых факторов.**

***Случайное влияние* – это действие тех многих факторов, которые не организованы в изучаемом дисперсионном комплексе и составляют общий фон, на котором действуют организованные факторы.**

***Общее влияние* – это влияние всех организованных и неорганизованных факторов, определивших такое развитие признака, которое наблюдалось в дисперсионном комплексе. Общее влияние служит базой для определения доли влияний – факториальных и случайных.**

Дисперсионный анализ позволяет оценить достоверность отличия нескольких выборочных средних одновременной, т.е. изучить влияние одного контролируемого фактора на результативный признак путем оценки его относительной роли в общей изменчивости этого признака, вызванной влиянием всех факторов.

Задача дисперсионного анализа состоит в том, чтобы охарактеризовать силу и достоверность влияния фактора на признак, причем только на *величину* (средний уровень) признака, но не на его изменчивость. Дисперсионный анализ есть метод сравнения нескольких средних арифметических. В этом смысле он подобен методу сравнения двух средних арифметических с помощью критерия Стьюдента.

В дисперсионном анализе использован такой же показатель достоверности влияния фактора, но адаптированный к случаю сравнения нескольких выборок (критерий Фишера).

$$F = \sigma^2_{факт.} / \sigma^2_{случ.}.$$

В качестве обобщенной *меры отличия нескольких выборочных средних* выступает дисперсия, рассеяние выборочных средних ($\overline{x}_j$) вокруг общей средней ($\overline{x}_{общ.}$)

$$\sigma^2_{факт.} = \sum^{k} (\overline{x}_j - \overline{x}_{общ.})^2 / df_{факт.},$$

где $df_{факт.} = k–1$; $j = 1, 2, \ldots k$; k – число сравниваемых средних.

В качестве обобщенной *меры случайного варьирования* служит дисперсия вариант (x_i) вокруг средней в каждой градации ($\overline{x}_j$):

$$\sigma^2_{случ.} = \sum^{k} \sum^{n_j} (x_{ij} - \overline{x}_j)^2 / df_{случ.},$$

где $df_{случ.} = n–1$; $i = 1, 2, \ldots n$, n – число вариант всех выборок.

Нулевая гипотеза гласит: «влияние фактора на признак отсутствует». Проверяют гипотезу по критерию Фишера

$$F = \sigma^2_{факт.} / \sigma S^2_{случ.} \geq F_{(\alpha,\, df1,\, df2)},$$

где $df_1 = k–1$, $df_2 = n–k$; k – число градаций результативного признака; n – общий объем всех выборок по всем градациям.

Влияние считается достоверным, если величина расчетного критерия равна или превышает свое табличное значение с принятым уровнем значимости (обычно $\alpha = 0{,}05$).

Не вдаваясь в подробности техники расчетов, приведем пример выполнения однофакторного дисперсионного анализа в среде Excel.

Пример. Для выявления значимости влияния технологического режима производства (например, температуры при выпечке хлеба) сняты показатели качества производимой продукции (толщина корки) при четырех градациях температурного режима.

Введем подписанные метками (*A*1, *A*2…) данные в четыре столбца, отдельно для каждой градации.

	A	B	C	D
1	A1	A2	A3	A4
2	1	2	2	2
3	1	2	2	4
4	2	1	3	3
5	1	3	2	2
6	1	2	3	4

Вызовем программу обработки командой Данные (Сервис) / Анализ данных…/ Однофакторный дисперсионный анализ, ОК. Заполним окно макроса, ОК. На новом листе появятся результаты расчетов.

	A	B	C	D	E	F	G
1	Однофакторный дисперсионный анализ						
2							
3	ИТОГИ						
4	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
5	Столбец 1	5	6	1,2	0,2		
6	Столбец 2	5	10	2	0,5		
7	Столбец 3	5	12	2,4	0,3		
8	Столбец 4	5	15	3	1		
9							
10							
11	Дисперсионный анализ						
12	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
13	Между группами	8,55	3	2,85	5,7	0,00750894	3,238871522
14	Внутри групп	8	16	0,5			
15							
16	Итого	16,55	19				
17							

Чтобы все надписи были видны, нужно изменить ширину столбцов. Это можно сделать, нажав на серый квадрат слева вверху листа (над 1, левее А), перевести курсор на границу между любыми столбцами (курсор примет форму креста со стрелками, направленными в стороны) и дважды кликнуть левой кнопкой мыши. Ширина каждого столбца будет автоматически определена по максимально длинному содержимому какой-либо ячейки этого столбца.

Поскольку полученное значение критерия (F = 5,7) больше табличного ($F_{(0.05,3,19)}$ = 3,1), отличие факториальной и случайной дисперсий достоверно, влияние фактора значимо.

6.4. Найти зависимость между признаками (или «доказать сопряженность варьирования признаков»)

Корреляционный анализ. Корреляция (термин ввел Ф. Гальтон в 1888 г.; в переводе означает «соотношение» или «взаимосвязь») – есть наличие взаимной согласованности в изменчивости двух или нескольких признаков (явлений). Она оценивается с помощью значения коэффициента корреляции, который является мерой степени и величины этой связи.

Задача корреляционного анализа сводится к установлению направления и формы связи между признаками, измерению ее тесноты и к оценке достоверности выборочных показателей корреляции.

Корреляционная связь не является точной зависимостью одного признака от другого, поэтому она может иметь различную степень – от полной независимости до очень сильной связи. Кроме того, характер связи между разными признаками может быть различен. Поэтому возникла необходимость определять форму, направление и степень корреляционных связей.

По форме корреляционная связь между признаками может быть *линейной* и *криволинейной (нелинейной)*, а по направлению – *положительной* и *отрицательной*.

Положительная корреляция отражает однотипность в изменении признаков: с увеличением значений первого признака увеличиваются значения и другого, или с уменьшением первого уменьшается второй.

Отрицательная корреляция указывает на увеличение первого признака при уменьшении второго или уменьшение первого признака при увеличении второго.

Когда исследуется корреляция между количественными признаками, значения которых можно точно измерить в единицах метрических шкал, то очень часто принимается модель двумерной нормально распределенной генеральной совокупности. Такая модель отображает зависимость между переменными величинами x_i и y_i графически в виде геометрического места точек в системе прямоугольных координат. Эту графическую зависимость называют также *диаграммой рассеивания* или *корреляционным полем* (рис. 5).

Данная модель двумерного нормального распределения (корреляционное поле) позволяет дать наглядную графическую интерпретацию коэффициента корреляции.

Таким образом, визуальный анализ корреляционного поля помогает выявить не только наличия статистической зависимости (линейную или нелинейную) между исследуемыми признаками, но и ее тесноту и форму. Это имеет существенное значение для следующего шага в анализе: выбора и вычисления соответствующего коэффициента корреляции.

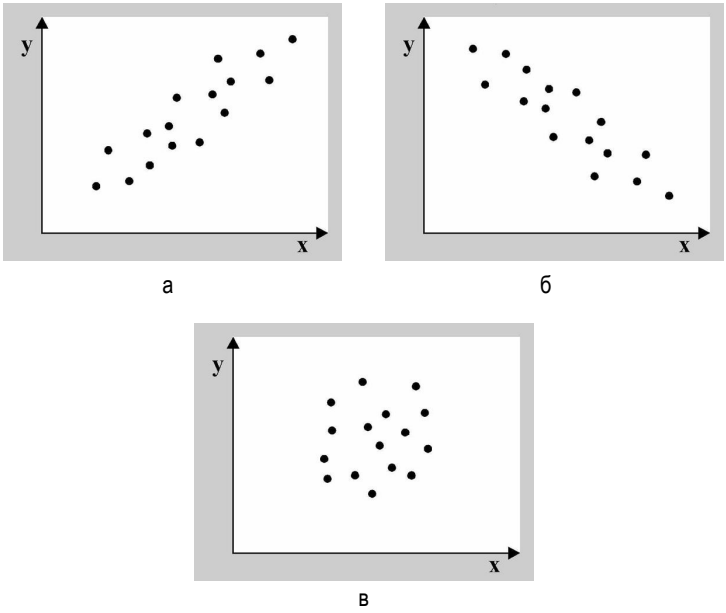


Рис. 5. Графическая интерпретация взаимосвязи между показателями: а – прямая корреляция; б – обратная корреляция; в – отсутствие корреляции

Коэффициент корреляции (***r***) – удобный показатель связи, получивший широкое применение в практике. К их основным свойствам необходимо отнести следующие:

1. Коэффициенты корреляции способны характеризовать только линейные связи, т.е. такие, которые выражаются уравнением линейной функции. При наличии нелинейной зависимости между варьирующими признаками следует использовать другие показатели связи.

2. Значения коэффициентов корреляции – это отвлеченные числа, лежащие в пределах от -1 до +1, т.е. $-1 < r < 1$.

3. При независимом варьировании признаков, когда связь между ними отсутствует, $r = 0$.

4. При положительной, или прямой, коэффициент корреляции приобретает положительный (+) знак и находится в пределах от 0 до +1, т.е. $0 < r < +1$.

5. При отрицательной, или обратной, коэффициент корреляции сопровождается отрицательным (–) знаком и находится в пределах от 0 до –1, т.е. $-1 < r < 0$.

6. Чем сильнее связь между признаками, тем ближе величина коэффициента корреляции к |1|. Если $r = \pm 1$, то корреляционная связь переходит в функциональную, т.е. каждому значению признака *x* будет соответствовать одно строго определенное значение признака *y*.

7. Только по величине коэффициентов корреляции нельзя судить о достоверности корреляционной связи между признаками. Этот параметр зависит от числа степеней свободы $k = n–2$, где *n* – число коррелируемых пар показателей *x* и *y*. Чем больше *n*, тем выше достоверность связи при одном и том же значении коэффициента корреляции.

В практической деятельности, когда число коррелируемых пар признаков *x* и *y* невелико ($n \leq 30$), то при оценке зависимости между показателями используют следующую градацию (Плохинский Н.А., 1970):

1) *высокая степень взаимосвязи* – значения коэффициента корреляции находятся в пределах от 0,76 до 1,00;

2) *средняя степень взаимосвязи* – значения коэффициента корреляции находятся в пределах от 0,51 до 0,75;

3) *слабая степень взаимосвязи* – значения коэффициента корреляции находятся от 0,25 до 0,50.

В качестве оценки генерального коэффициента корреляции используется *коэффициент корреляции (r) Браве–Пирсона.* Для его определения принимается предположение о двумерном нормальном распределении генеральной совокупности, из которой получены экспериментальные данные. Это предположение может быть проверено с помощью соответствующих критериев значимости.

Коэффициент корреляции Браве-Пирсона ($\mathbf{r^P_{xy}}$) относится к параметрическим коэффициентам и для практических расчетов вычисляется по формуле

$$r_{xy}^{P} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \cdot \sum(y_i - \bar{y})^2}}.$$

Вычислив это значение $\mathbf{r^P_{xy}}$, необходимо определить достоверность найденного коэффициента корреляции, сравнив его фактическое значение с табличным для $k = n - 2$ по таблице «Коэффициент Стьюдента» (см. приложение). Если $\mathbf{r_\varphi \geq r_{st}}$, то можно говорить о том, что между признака-

ми наблюдается статистически значимая взаимосвязь. Если $\mathbf{r_{\varphi} \leq r_{st}}$, то между признаками корреляционная взаимосвязь статистически не значима.

Пример. Произведены измерения веса (г), длины (мм) и ширины (мм) куриных яиц. Вычислить попарные коэффициенты корреляции и оценить статистическую значимость взаимосвязей признаков.

Расчеты можно провести в среде Excel. Вызовем программу обработки командой Данные (Сервис) / Анализ данных…/ Корреляция, ОК. Заполним окно макроса, выделив блок данных с метками и поставив галочку в поле «Метки в первой строке» и выходной интервал (E2), ОК. На листе появится результат.

	A	B	C	D	E	F	G	H
1	вес, г	длина, мм	ширина, мм			*вес, г*	*длина, мм*	*ширина, мм*
2	52,1	53	41		вес, г	1		
3	54,4	54	42		длина, мм	0,805760392	1	
4	56,3	54	44		ширина, мм	0,798316774	0,394548111	1
5	57,4	57	43					
6	58,4	57	43					
7	59,1	58	44					
8	60	60	42					
9	60,2	58	43					
10	60,5	59	42					
11	60,7	55	45					
12	60,8	59	43					
13	60,8	58	43					
14	61	58	43					
15	61,2	56	44					
16	61,3	58	45					
17	62,3	58	44					
18	63,7	57	45					
19	64	57	45					
20	64,8	61	43					
21	65,1	58	44					
22	65,4	59	44					
23	65,6	59	44					
24	65,8	59	45					
25	65,9	58	45					
26	66,2	59	45					
27	66,6	58	45					
28	66,8	59	45					
29	67,2	57	45					
30	67,3	62	44					

По таблице «Минимальные значения коэффициента корреляции *r*, достоверно отличные от нуля» (см. приложение) для $df = 52$ и $\alpha=0,05$ находим пороговое значение коэффициента корреляции (0,273). Все попарные связи исследованных параметров оказались статистически значимы.

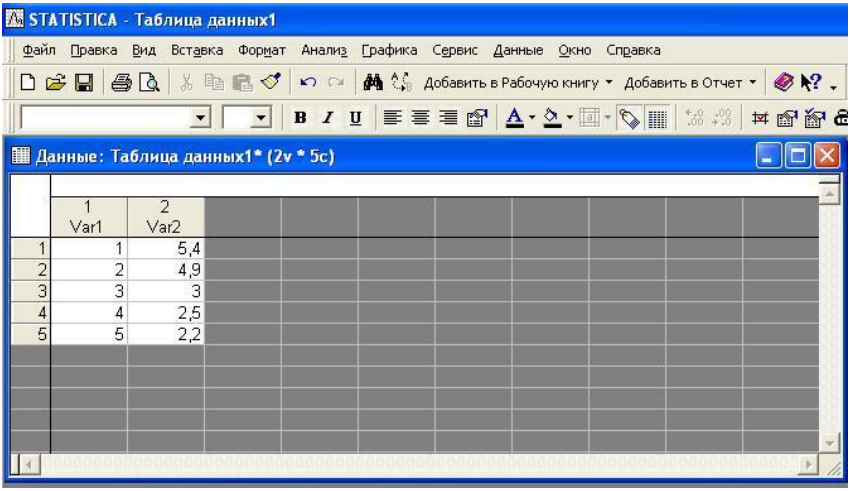
Регрессионный анализ. Во многих естественных науках часто приходится статистически анализировать влияние одного фактора на другой. Подобные задачи возникают тогда, когда такие факторы не являются независимыми, но их функциональная зависимость неизвестна (или ее невозможно найти аналитически). Примерами могут служить зависимость

между надоями молока и ее жирностью, содержанием сахара в являющихся сырьем для вина ягодах и крепостью произведенного вина, сроком хранения молока и ее кислотностью и т.п.

Вероятностный подход к решению подобных задач исходит из предположения, что система рассматриваемых величин обладает определенным *совместным распределением вероятностей*.

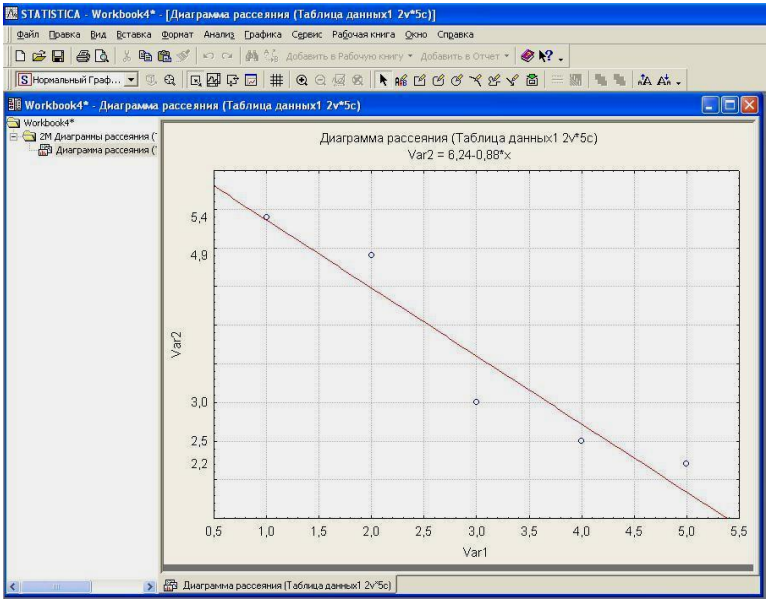
Не вдаваясь в подробности содержания алгоритмов расчета линий регрессии отметим, что выполнить этот вид анализа можно, например, в среде программы Statistica. Предварительно заполнив таблицу данных (или экспортировав ее из программы Excel), вызовем программу обработки данных командой Графика / Диаграммы рассеяния / Переменные. Определим переменные. Как правило, по оси абсцисс откладывают независимые переменные (например, градации технологического режима), а по оси ординат – зависимые переменные (например, показатели качества продукции). Далее – ОК / Дополнительно. Выбираем тип графика (Простой) и Подгонка (для начала Линейная). Далее – ОК. Для сложных форм зависимости переменных можно произвести Подгонку, щелкнув по линии регрессии правой кнопкой мыши и выбрав подходящую опцию для отображения криволинейной регрессии.

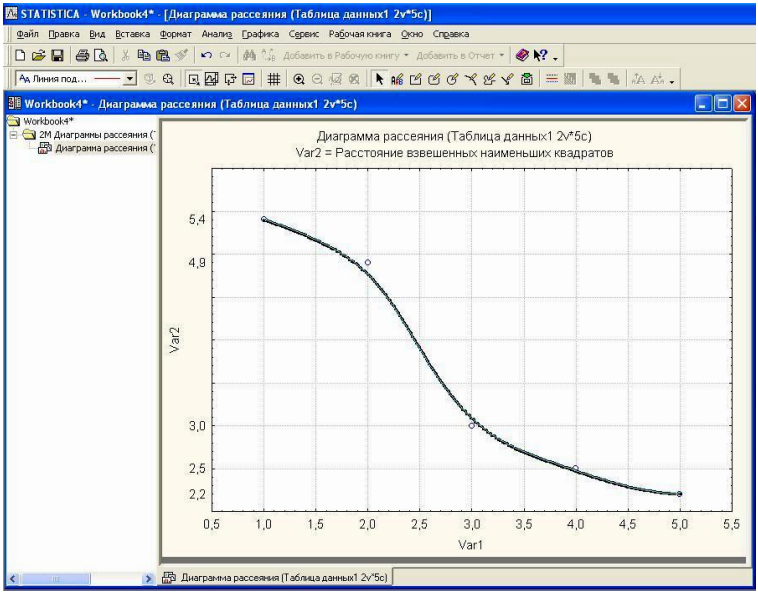
Пример. При разных градациях технологического режима (Var1) получены показатели качества продукции (Var2). Определить наличие и характер зависимости качества продукции от технологического режима ее производства.



	1 Var1	2 Var2
1	1	5,4
2	2	4,9
3	3	3
4	4	2,5
5	5	2,2

Регрессионный анализ, выполненный в программе Statistica по описанному выше алгоритму, дал следующие результаты.





Вторая линия получена с помощью метода наименьших квадратов. Суть метода в следующем: сумма квадратов отклонений экспериментальных точек от сглаживающей кривой должна быть минимальной.

7. КЛАССИФИКАЦИЯ ОБЪЕКТОВ

Методы многомерного анализа. Теоретической основой для методов многомерной статистики служит понятие гиперпространства, или многомерного пространства. В отличие от привычного физического трехмерного пространства, имеющего три ортогональных (взаимно перпендикулярных) оси, многомерное пространство имеет множество осей координат, в качестве которых выступают признаки (переменные) изучаемых объектов. Отдельный объект, охарактеризованный по нескольким признакам, рассматривается как отдельная точка, а множество объектов – как облако точек. Если объекты отличаются друг от друга по разным признакам, то они будут занимать разное положение в многомерном пространстве; объекты оказываются рассеянными в нем.

Главной характеристикой объектов становится расстояние между ними в этом гиперпространстве, а главной особенностью всей выборки – форма облака рассеяния со своими пустотами и сгущениями объектов. Методы многомерной статистики изучают информацию, «закодированную» в порядке расположения объектов друг относительно друга.

Кластерный анализ позволяет выделить ведущий признак и иерархию взаимосвязей признаков. Слово кластер (cluster)с английского переводится как сгусток, пучок, группа (родственные понятия – класс, таксон, сгущение).

Кластерный анализ – это совокупность методов, позволяющих классифицировать многомерные наблюдения, каждое из которых описывается набором исходных переменных X_1, X_2,..., X_m.

Целью кластерного анализа является образование групп схожих между собой объектов, которые принято называть кластерами.

Например, если вы должны кластеризовать типы еды в кафе, то можете принять во внимание количество содержащихся в ней калорий, цену, субъективную оценку вкуса и т.д.

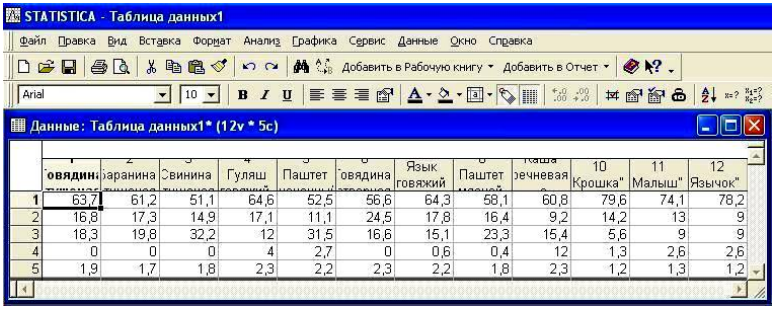
Пример. Классифицировать мясные консервы по показателям химического состава, приведенным в следующей таблице:

Консервы	Массовая доля, %				
	Вода	Белки	Жиры	Углеводы	Зола
Говядина тушеная	63,7	16,8	18,3	0	1,9
Баранина тушеная	61,2	17,3	19,8	0	1,7
Свинина тушеная	51,1	14,9	32,2	0	1,8
Гуляш говяжий	64,6	17,1	12,0	4,0	2,3
Паштет печеночный	52,5	11,1	31,5	2,7	2,2
Говядина отварная	56,6	24,5	16,6	0	2,3
Язык говяжий в желе	64,3	17,8	15,1	0,6	2,2

Окончание таблицы

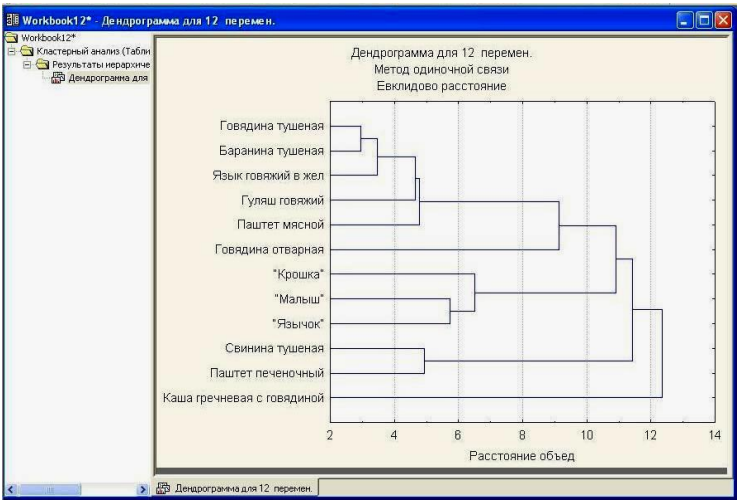
Консервы	Массовая доля, %				
	Вода	Белки	Жиры	Углеводы	Зола
Паштет мясной	58,1	16,4	23,3	0,4	1,8
Каша гречневая с говядиной	60,8	9,2	15,4	12,0	2,3
«Крошка»	79,6	14,2	5,6	1,3	1,2
«Малыш»	74,1	13,0	9,0	2,6	1,3
«Язычок»	78,2	9,0	9,0	2,6	1,2

Последовательность действий при выполнении анализа в среде Statistica следующая: экспортировать матрицу данных в программу Statistica, предварительно транспонировав ее через программу Excel.



	1	2	3	4	5	6	7	8	9	10	11	12
	овядин	аранина	Свинина	Гуляш	Паштет	овядина	Язык говяжий	Паштет	ечневая	Крошка"	Малыш"	Язычок"
1	63,7	61,2	51,1	64,6	52,5	56,6	64,3	58,1	60,8	79,6	74,1	78,2
2	16,8	17,3	14,9	17,1	11,1	24,5	17,8	16,4	9,2	14,2	13	9
3	18,3	19,8	32,2	12	31,5	16,6	15,1	23,3	15,4	5,6	9	9
4	0	0	0	4	2,7	0	0,6	0,4	12	1,3	2,6	2,6
5	1,9	1,7	1,8	2,3	2,2	2,3	2,2	1,8	2,3	1,2	1,3	1,2

Далее команды Анализ / Многомерный разведочный анализ / Кластерный анализ / Иерархическая классификация, ОК / Выбрать все, ОК / ОК.



Из рисунка видно, что консервы с близким составом образовали отдельные кластеры: консервы со сбалансированным содержанием белков и жиров и отсутствием (или низким содержанием) углеводов (говядина тушеная, баранина тушеная, язык говяжий в желе, гуляш говяжий и паштет мясной) – к ним примыкает отличающаяся высоким содержанием белка говядина отварная; углеводсодержащие консервы для детского питания («Крошка», «Малыш», «Язычок») с низким содержанием жира; консервы с высоким содержанием жира (свинина тушеная и паштет печеночный и, наконец, отдельную группу составляет каша гречневая с говядиной с высоким содержанием углеводов.

Факторный анализ – многомерный статистический метод, применяемый для изучения взаимосвязей между значениями переменных. Факторный анализ позволяет решить две важные проблемы исследователя: описать объект измерения *всесторонне* и в то же время *компактно*. С помощью факторного анализа возможно выявление скрытых переменных факторов, отвечающих за наличие линейных статистических связей корреляций между наблюдаемыми переменными.

Например, анализируя оценки, полученные по нескольким шкалам, исследователь замечает, что они сходны между собой и имеют высокий коэффициент корреляции, он может предположить, что существует некоторая латентная переменная, с помощью которой можно объяснить наблюдаемое сходство полученных оценок. Такую латентную переменную называют *фактором*. Данный фактор влияет на многочисленные показатели других переменных, что приводит нас к возможности и необходимости выделить его как наиболее общий, более высокого порядка.

Для выявления наиболее значимых факторов и, как следствие, факторной структуры, наиболее оправданно применять **метод главных компонентов** (англ. *Principal component analysis, PCA*). Суть данного метода состоит в замене коррелированных компонентов некоррелированными факторами. Другой важной характеристикой метода является возможность ограничиться наиболее информативными главными компонентами и исключить остальные из анализа, что упрощает интерпретацию результатов. Достоинство метода главных компонентов также в том, что он – единственный математически обоснованный метод факторного анализа.

Практическое выполнение факторного анализа начинается с проверки его условий. Обязательные условия факторного анализа:

1. Все признаки должны быть количественными.
2. Число признаков должно быть в два раза больше числа переменных.
3. Выборка должна быть однородна.
4. Исходные переменные должны быть распределены симметрично.
5. Факторный анализ осуществляется по коррелирующим переменным.

При анализе в один фактор объединяются сильно коррелирующие между собой переменные, как следствие происходит перераспределение дисперсии между компонентами и получается максимально простая и наглядная структура факторов. После объединения коррелированность компонент внутри каждого фактора между собой будет выше, чем их коррелированность с компонентами из других факторов. Эта процедура также позволяет выделить латентные переменные, что бывает особенно важно при анализе социальных представлений и ценностей.

Виды медоносных растений	Лес	Лес	Лес	Лес	Лес	Лес	Лесная поляна	Влажный луг	Влажный луг	Влажный луг
	1	2	3	4	5	6	7	8	9	10
Клевер горный	0,0	1,0	0,2	0,0	0,1	0,0	0,0	0,1	0,0	2,8
Клевер ползучий	2,3	0,2	0,4	0,0	0,0	0,0	23,6	0,0	0,2	0,0
Дягиль	3,8	2,8	0,0	2,3	2,4	10,6	2,1	2,8	4,9	15,1
Дудник лесной	0,5	0,0	0,0	0,0	0,0	2,7	0,0	0,0	0,0	0,0
Василек ложнофригийский	0,0	0,0	0,0	0,0	0,0	0,6	17,0	5,9	9,4	0,0
Бутень прескотта	3,3	0,2	0,0	0,0	0,3	3,2	0,4	0,1	0,0	0,1
Чистотел большой	1,8	7,9	7,2	0,0	2,1	5,6	0,2	0,0	0,2	0,1
Таволга вязолистная	17,3	9,8	4,7	28,6	0,3	6,5	1,0	45,6	41,5	55,9
Таволга обыкновенная	3,8	1,8	0,0	1,3	0,1	0,3	0,2	0,5	0,0	0,0
Борщевик сибирский	0,5	0,0	0,6	1,0	39,2	1,8	0,2	1,0	0,0	3,3
Зверобой	0,0	0,0	0,0	1,6	0,6	5,3	0,2	0,0	0,0	0,0
Котовник венгерский	0,0	0,2	0,3	0,0	0,1	0,0	4,8	0,0	0,0	2,1
Душица	2,0	2,4	0,4	0,3	1,7	0,9	2,3	0,0	1,5	0,5
Кровохлебка	0,3	0,0	0,4	0,0	0,0	0,0	21,4	27,3	12,2	0,0
Морковник	4,6	0,4	0,0	0,6	0,7	2,1	0,6	0,4	0,0	0,0
Липа	42,4	41,7	39,0	54,7	42,0	27,7	10,1	4,8	13,5	5,7
Вероника широколистная	0,8	8,3	22,9	0,3	0,0	0,0	0,2	0,5	0,0	2,5

8. ОБЩИЕ ПОНЯТИЯ О СТАТИСТИЧЕСКИХ ТАБЛИЦАХ

Таблицы относятся к наиболее простому способу представления данных. Они состоят из колонок со значениями двух или более связанных переменных. С помощью этого метода трудно получить прямое и ясное указание на связь между переменными, но он часто является первым этапом регистрации информации и служит основой для выбора последующей формы графического представления данных.

Практикой выработаны следующие основные правила составления и оформления статистических таблиц:

1. Таблица должна быть по возможности небольшой по размерам (облегчается анализ данных). Целесообразно построить несколько небольших взаимосвязанных таблиц, чем одну большую.

2. Таблица должна иметь кратко, ясно и точно сформулированное название, заголовки строк подлежащего и граф сказуемого. В названии необходимо отразить объект изучения, территорию и период времени, к которым относятся приводимые данные.

3. Строки подлежащего и графы сказуемого обычно размещаются по принципу от частного к общему. Если приводятся не все слагаемые, то сначала показывают общие итоги, а затем выделяют наиболее важные их составные части («в том числе», «из них»).

4. Таблица должна обязательно содержать необходимые итоги (групповые, общие, проверочные), их отсутствие затрудняет анализ и даже обесценивает таблицу.

5. Строки в подлежащем и графы в сказуемом часто нумеруют порядковыми номерами. При этом в сказуемом нумеруются только графы, в которые вписываются цифры. Графы для обозначений подлежащего и единиц его измерения обычно обозначаются буквами («а», «б»… или «А», «Б»…).

6. При заполнении таблицы необходимо строго соблюдать следующие условные обозначения: если данное явление (событие) отсутствует, ставить знак « – » (тире), если отсутствуют сведения, ставится знак « … » (многоточие) или пишут «нет сведений», если сведения имеются, но числовое значение меньше принятой в таблице точности, то ставится « 0,0 ».

7. Округлённые числа приводятся в таблице с одинаковой степенью точности (до 0,1; до 0,01 и т.д.) для всей графы однородных показателей. Не следует округлять проценты до целых чисел (округление значений, близких к 100, исказит картину). Когда показатели в процентах выражены большими числами, целесообразно заменить их выражением «во столько-то раз больше или меньше».

8. Если приводятся не только зафиксированные при наблюдении (первичные) данные, но и данные, полученные в результате расчетов, целесообразно об этом сделать оговорку в таблице или в примечании к ней.

9. Таблица может сопровождаться примечаниями, в которых указываются источники данных, более подробное содержание показателей и другие необходимые пояснения (например, методика расчёта).

9. ГРАФИЧЕСКИЕ ИЗОБРАЖЕНИЯ В СТАТИСТИКЕ

Полученный в результате статистического исследования материал нередко изображается с помощью точек, геометрических линий и фигур или географических картосхем, т.е. ***графиков***.

Графики придают изложению статистических данных бóльшую наглядность, чем таблицы, выразительность, облегчают их восприятие и анализ. Статистический график позволяет зрительно оценить характер изучаемого явления, присущие ему закономерности, тенденции развития, взаимосвязи с другими показателями, географическое разрешение изучаемых явлений. Еще в древности китайцы говорили, что одно изображение заменяет тысячу слов. Графики делают статистический материал более понятным, доступным и неспециалистам, привлекают внимание широкой аудитории к статистическим данным, популяризируют статистику и статистическую информацию.

При любой возможности анализ статистических данных рекомендуется всегда начинать с их графического изображения. График позволяет сразу получить общее представление обо всей совокупности статистических показателей. Графический метод анализа выступает как логическое продолжение табличного метода и служит целям получения обобщающих статистических характеристик процессов, свойственных массовым явлениям.

При помощи графического изображения статистических данных решаются многие задачи статистического исследования:

1) наглядное представление величины показателей (явлений) в сравнении друг с другом;

2) характеристика структуры какого-либо явления;

3) изменение явления во времени;

4) ход выполнения плана;

5) зависимость изменения одного явления от изменения другого;

6) распространенность или размещение каких-либо величин по территории.

Другими словами, в статистических исследованиях применяются самые разнообразные графики.

В каждом графике выделяют (различают) следующие основные элементы:

А) Пространственные ориентиры задаются в виде системы координатных сеток. В статистических графиках чаще всего применяется система прямоугольных координат. Иногда используется принцип полярных (угловых) координат (круговые графики).

На осях системы координат в определенном порядке располагаются характеристики статистических признаков изображаемых явлений или процессов. По вертикальной оси (оси ***y***) откладывают значения, называемые ординатами, которые показывают величину зависимой переменной, т.е. функции. Это «неизвестное количество», иными словами переменная, значения которой не выбираются экспериментатором. Горизонтальная ось ***x*** несет значения, называемые абсциссами, которые показывают величину независимой переменной. Это «известное количество», т.е. переменная, значения которой выбираются экспериментатором.

Признаки, располагаемые на осях координат, могут быть качественными или количественными.

Б) Графический образ статистических данных представляет собой совокупность линий, фигур, точек, образующих геометрические фигуры разной формы (окружность, квадраты, прямоугольники и т.п.) с различной штриховкой, окраской, густотой нанесения точек.

Любое явление, изучаемое статистикой, можно представить в графической форме. Для этого требуется найти правильное графическое решение, определить тот графический образ, который лучше всего соответствует данному явлению, нагляднее изображает статистические данные. Графический образ должен соответствовать цели графика. Поэтому перед построением графика необходимо уяснить сущность явления и цель, которая ставится перед графическим изображением. Выбранная форма графика должна соответствовать внутреннему содержанию и характеру статистического показателя. Например, сравнение на графике производится по таким измерениям, как площадь, длина одной из сторон фигур, местонахождением точек, их густотой и т.д.

Так, для изображения изменений явления во времени наиболее естественным типом графика является линия. Для рядов распределения – полигон или гистограмма.

В) Поле графика – это пространство, в котором располагаются графические образы (геометрические тела, образующие графики). Поле графика характеризуется по размерам и пропорциям. Размер поля зависит от назначения графика. Пропорции и размер графика (формат графика) должны соответствовать также сущности изображаемых явлений. Для статистических исследований часто используются графики с неравными сторонами, н*о иногда удобна квадратная форма графиков.*

Г) Масштабные ориентиры, обеспечивающие геометрическому образу количественную определенность, – это использованная в графике система масштабных шкал. Масштабом графика называется условная мера перевода статистической числовой величины в графическую.

Масштабная шкала – это линия, отдельные точки которой могут быть в соответствии с принятым масштабом прочитаны как определенное значение статистического показателя. Масштаб выбирается с таким рас-

четом, чтобы на графике могли поместиться самая большая и самая маленькая из изображаемых величин при максимальном использовании поля графика.

Каждая ось должна начинаться с 0, но если все значения одной переменной расположены близко друг к другу, например, между 6,12 и 6,68 лежат десять точек, то, чтобы разместить эти точки, потребуется крупный масштаб. В этом случае ось также начинается с 0, но сразу после 0 на оси делается отметка о разрыве в виде знака -//-/.

Масштабные шкалы бывают равномерными и неравномерными, прямолинейными (обычно располагаются по осям координат) и криволинейными (круговые в секторных диаграммах).

Д) Экспликация графика – это словесное пояснение его содержания (название графика и соответствующие пояснения отдельных его частей).

Название графика должно точно и кратко раскрывать его содержание. Пояснительные тексты могут располагаться в пределах графического образа, рядом с ним или выноситься за его пределы, вдоль масштабных шкал. Они помогают мысленно перейти от геометрических образов к явлениям и процессам, изображенным на графике.

Особенность графических изображений в их выразительности, доходчивости и обозримости. Однако графические изображения не только иллюстративны, они носят и аналитический характер.

Графические способы изображения могут быть сгруппированы по различным признакам: по форме графического образа, по типу шкалы, поля, задачам изображения и т.д.

По *виду поля графика* различают диаграммы и статистические карты.

По *форме графического образа* различают линейные, плоскостные, объемные, точечные, фоновые, изобразительные диаграммы и карты.

По *типу шкалы*: **линейные равномерные (арифметические), линейные неравномерные (функциональные, логарифмические), криволинейные** и др.

По *задачам изображения* можно выделить: графики статистического и динамического сравнения; графики структуры и структурных сдвигов или структурно-динамические; графики динамики или динамические; графики контроля выполнения плана; графики пространственного (территориального) размещения и пространственной распространенности; графики вариационных рядов; графики зависимости варьирующих признаков и взаимосвязи и др.

Каждый из основных видов графических изображений в статистической практике строится с учетом определенных правил. В статистических исследованиях для выяснения характерных черт и особенностей массовых явлений, познания типичного в этих явлениях и решения других задач широко используется сравнение одних абсолютных, средних и относительных статистических величин с другими.

Анализ – это, прежде всего, сравнение и сопоставление статистических данных. Нередко возникает необходимость сопоставления результатов статистического исследования конкретного явления с величинами типичного (идеального) явления аналогичной природы. Поэтому наглядное представление (графическое изображение) сравнения статистических показателей относится к наиболее распространенным графикам в статистике. Для этих целей применяются *диаграммы*.

Диаграмма – это графическое изображение, наглядно показывающее соотношение между сравниваемыми величинами. Диаграмма представляет собой чертеж, на котором статистические данные условно изображаются геометрическими линиями, фигурами и телами различных размеров.

Различают следующие основные виды графиков (диаграмм) сравнения: круговые, столбиковые, полосовые, квадратные, фигурные.

Круговые диаграммы (разновидность структурной диаграммы) – при их помощи достигается наглядное представление структуры совокупности или изображения состава целого, разделенного на части (рис. 6).

Используются для отображения соотношения составляющих единство частей, например, структуру успеваемости в группе (доля студентов, сдавших экзамен на отлично, хорошо и удовлетворительно) или содержания веществ в продукте.

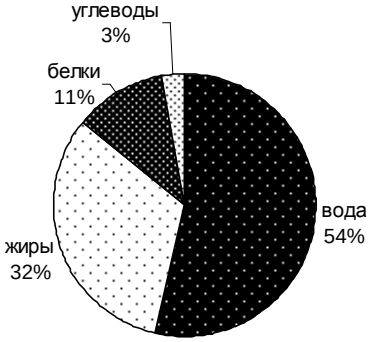


Рис. 6. Массовая доля (%) веществ в мясных консервах «Паштет печеночный»

Столбиковые диаграммы – являются наиболее простым и наглядным графиком для сравнения величин статистических показателей (рис. 7, 8).

При построении столбиковых диаграмм необходимо начертить систему прямоугольных координат. Основания столбиков одинакового размера размещаются на оси абсцисс, а вершина столбика будет соответствовать величине показателя, нанесенного в соответствующем масштабе на ось ординат. Каждый отдельный столбик соответствует отдельному объекту (показателю). Общее число столбиков равно числу сравнивае-

мых величин. Расстояние между столбиками берется одинаковое, а иногда столбики располагаются вплотную друг к другу. Вертикальная шкала всегда начинается с нуля и охватывает весь диапазон изображаемых данных. Для целей наглядности допускается разрыв по шкале данных (обычно начальных). С помощью столбиковых диаграмм легко изобразить также структуру или процесс развития явления во времени.

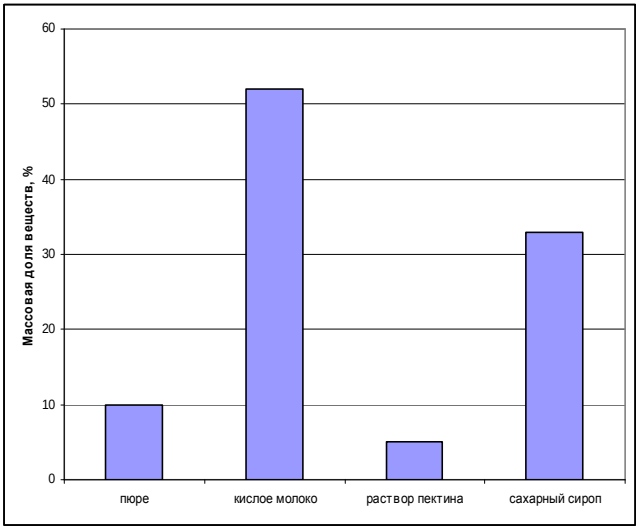


Рис. 7. Массовая доля веществ в молочном напитке «Черносмородиново-молочный»

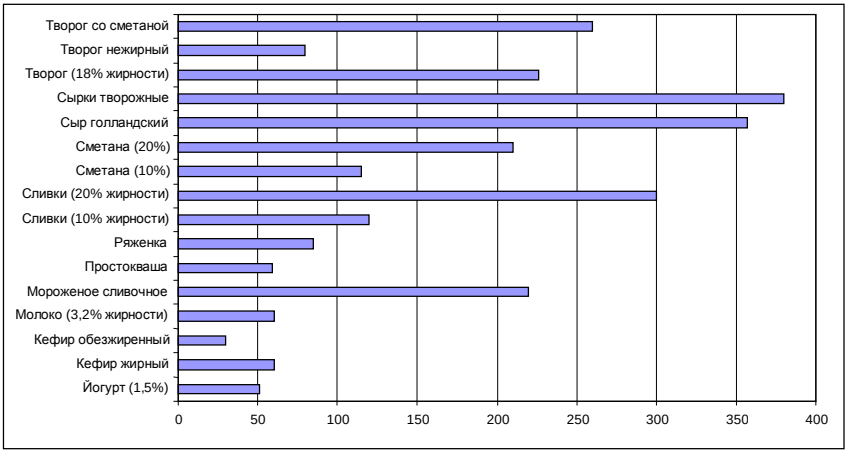


Рис. 8. Содержание энергии в 100 г молочных продуктов, кал.

Полосовые диаграммы – обычно используются для отображения отношения между непрерывной зависимой переменной (например, содержанием энергии) и нечисловой независимой переменной (например, различными видами пищи). В этом случае столбики размещаются не по вертикали, а по горизонтали, т.е. основание полос (объекты, данные) располагаются на оси ординат, а масштаб – на оси абсцисс.

Ширина полос также (как столбцов в столбиковой диаграмме) должна быть одинаковой. Расстояние между ними берется одинаковым или полосы строятся вплотную. Шкала горизонтальной полосовой диаграммы должна начинаться также с нуля, ее разрыв обычно не допускается.

ВОПРОСЫ ДЛЯ САМОСТОЯТЕЛЬНОГО ИЗУЧЕНИЯ

1. Какие ученые работали над развитием математической статистики?
2. Что считают генеральной совокупностью?
3. Как строится корреляционный ряд?
4. Какие средние величины применяют в обработке данных, полученных при исследованиях? Дайте им характеристику.
5. Охарактеризуйте закон нормального распределения.
6. Каковы свойства нормального распределения?
7. Как выбирается оптимальное число групп?
8. В чем заключается сущность средней арифметической, среднего квадратичного отклонения и дисперсии?
9. Что такое доверительный интервал?
10. Что называют ошибкой репрезентативности параметра?
11. Как определяют точность опыта?
12. Что такое асимметрия и эксцесс?
13. Каковы четыре класса статистических задач? Дайте им характеристику.
14. Каковы свойства коэффициента корреляции?
15. В чем особенность кластерного и факторного анализов?
16. Каковы основные правила составления и оформления статистических таблиц?
17. Каковы основные правила составления и оформления графиков?

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Волкова П.А., Шипунов А.Б. Статистическая обработка данных в учебно-исследовательских работах. М.: Экопресс, 2008. 60 с.
2. Грин Н., Стаут У., Тейлор Д. Биология. Т.1. М.: Мир, 1990. 368 с.
3. Практикум по статистике / А.П. Зинченко, А.Е. Шибалкин, О.Б. Тарасова и др. М.: КолосС, 2007. 413 с.
4. Ивантер Э.В., Коросов А.В. Введение в количественную биологию: учеб. пособие. Петрозаводск: ПетрГУ. 2003. 304 с.
5. Лакин Г.Ф. Биометрия. М.: Высш. шк., 1990. 352 с.
6. Орлов А.И. Математика случая: Вероятность и статистика – основные факты: учеб. пособие. М.: МЗ-Пресс, 2004. 110 с.
7. Плохинский Н.А. Биометрия. М.: Изд-во Моск. ун-та, 1970. 367 с.
8. Ростова Н.С. Корреляции: структура и изменчивость. СПб.: Изд-во СПб. ун-та, 2002. 308 с.
9. Шмидт В.М. Математические методы в ботанике: учеб. пособие. Л.: Изд-во Ленингр. ун-та, 1984. 288 с.

ПРИЛОЖЕНИЕ

Значение критерия *t* для отбраковки «выскакивающих» вариант

n	*α*			*n*	*α*		
	0.05	0.01	0.001		0.05	0.01	0.001
5	3.04	5.04	9.43	20	2.15	2.93	3.98
6	2.78	4.36	7.41	25	2.11	2.85	3.82
7	2.62	3.96	6.37	30	2.08	2.80	3.72
8	2.51	3.71	5.73	35	2.06	2.77	3.65
9	2.43	3.54	5.31	40	2.05	2.74	3.60
10	2.37	3.41	5.01	45	2.04	2.72	3.57
11	2.33	3.31	4.79	50	2.03	2.71	3.53
12	2.29	3.23	4.62	60	2.02	2.68	3.49
13	2.26	3.17	4.48	70	2.01	2.67	3.46
14	2.24	3.12	4.37	80	2.00	2.66	3.44
15	2.22	3.08	4.28	90	2.00	2.65	3.42
16	2.20	3.04	4.20	100	1.99	2.64	3.41
17	2.18	3.01	4.13	0	1.96	2.58	3.29
18	2.17	2.98	4.07				

Минимальные значения коэффициента корреляции *r*,
достоверно отличные от нуля (*df* = *n*–2)

	α			*α*			*α*	
df	0.05	0.01	*df*	0.05	0.01	*df*	0.05	0.01
1	0.997	1	**16**	0.468	0.59	**40**	0.304	0.393
2	0.95	0.99	**17**	0.456	0.575	**45**	0.288	0.372
3	0.878	0.959	**18**	0.444	0.561	**50**	0.273	0.354
4	0.811	0.917	**19**	0.433	0.549	**60**	0.25	0.325
5	0.754	0.874	**20**	0.423	0.537	**70**	0.232	0.302
6	0.707	0.834	**21**	0.413	0.526	**80**	0.217	0.283
7	0.666	0.798	**22**	0.404	0.515	**90**	0.205	0.267
8	0.632	0.765	**23**	0.396	0.505	**100**	0.195	0.254
9	0.602	0.735	**24**	0.388	0.496	**125**	0.174	0.228
10	0.576	0.708	**25**	0.381	0.487	**150**	0.159	0.208
11	0.553	0.684	**26**	0.374	0.478	**200**	0.138	0.181
12	0.532	0.661	**27**	0.367	0.47	**300**	0.113	0.148
13	0.514	0.641	**28**	0.361	0.463	**400**	0.098	0.128
14	0.497	0.623	**30**	0.349	0.449	**500**	0.088	0.115
15	0.482	0.606	**35**	0.325	0.418	**1000**	0.062	0.081

Значения критерия Стьюдента

Число степеней свободы, *df*	Доверительная вероятность (*P*) Уровень значимости (α)		
	P = 0.095 α = 0.05	*P* = 0.099 α = 0.01	*P* = 0.0999 α = 0.001
2	4.303	9.925	31.598
3	3.182	5.841	12.941
4	2.776	4.604	8.610
5	2.571	4.032	6.859
6	2.447	3.707	5.959
7	2.365	3.499	5.405
8	2.306	3.355	5.041
9	2.262	3.250	4.781
10	2.228	3.169	4.587
11	2.201	3.106	4.437
12	2.179	3.055	4.318
13	2.160	3.012	4.221
14	2.145	2.977	4.140
15	2.131	2.947	4.073
16	2.120	2.921	4.015
17	2.110	2.898	3.965
18	2.101	2.878	3.922
19	2.093	2.861	3.883
20	2.086	2.845	3.850
22	2.074	2.819	3.792
25	2.060	2.787	3.725
30	2.042	2.750	3.646
35	2.030	2.724	3.591
40	2.021	2.704	3.551
45	2.014	2.690	3.520
50	2.008	2.678	3.496
55	2.004	2.669	3.476
60	2.000	2.660	3.460
70	1.994	2.648	3.435
80	1.989	2.638	3.416
90	1.986	2.631	3.402
100	1.982	2.625	3.390
120	1.980	2.617	3.373
>120	1.960	2.5758	3.2905

Значения критерия Фишера F при уровне значимости α=0,05
(число степеней свободы указано для дисперсии знаменателя – в строке, для дисперсии числителя – в столбце)

df_2 \ df_1	1	2	3	4	5	6	7	8	9	10	12	15	20	30	∞
1	161.0	200.0	216.0	225.0	230.0	234.0	237.0	239.0	241.0	242.0	244.0	246.0	248.0	250.0	254.0
2	18.5	19.0	19.2	19.3	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.4
3	10.1	9.6	9.3	9.1	9.0	8.9	8.9	8.9	8.8	8.8	8.7	8.7	8.7	8.6	8.5
4	7.7	6.9	6.6	6.4	6.3	6.2	6.1	6.0	6.0	5.9	5.9	5.9	5.8	5.8	5.6
5	6.6	5.8	5.4	5.2	5.1	5.0	4.9	4.8	4.8	4.7	4.7	4.6	4.6	4.5	4.4
6	6.0	5.1	4.7	4.5	4.4	4.3	4.2	4.2	4.1	4.1	4.0	4.0	3.9	3.8	3.7
7	5.6	4.7	4.4	4.1	4.0	3.9	3.8	3.7	3.7	3.6	3.6	3.5	3.4	3.4	3.2
8	5.3	4.5	4.1	3.8	3.7	3.6	3.5	3.4	3.4	3.3	3.3	3.2	3.2	3.1	3.0
9	5.1	4.3	3.9	3.6	3.5	3.4	3.3	3.2	3.2	3.1	3.1	3.0	2.9	2.9	2.7
10	5.0	4.1	3.7	3.5	3.3	3.2	3.1	3.1	3.0	3.0	2.9	2.9	2.8	2.7	2.5
11	4.8	4.0	3.6	3.4	3.2	3.1	3.0	3.0	2.9	2.9	2.8	2.7	2.7	2.6	2.4
12	4.7	3.9	3.5	3.3	3.1	3.0	2.9	2.9	2.8	2.8	2.7	2.6	2.5	2.5	2.3
13	4.7	3.8	3.4	3.2	3.0	2.9	2.8	2.8	2.7	2.7	2.6	2.5	2.5	2.4	2.2
14	4.6	3.7	3.3	3.1	3.0	2.9	2.8	2.7	2.7	2.6	2.5	2.5	2.4	2.3	2.1
15	4.5	3.7	3.3	3.1	2.9	2.8	2.7	2.6	2.6	2.5	2.5	2.4	2.3	2.2	2.1
16	4.5	3.6	3.2	3.0	2.8	2.7	2.7	2.6	2.5	2.5	2.4	2.3	2.3	2.2	2.0
17	4.4	3.6	3.2	2.9	2.8	2.7	2.6	2.5	2.5	2.4	2.4	2.3	2.2	2.1	2.0
18	4.4	3.5	3.2	2.9	2.8	2.7	2.6	2.5	2.5	2.4	2.3	2.3	2.2	2.1	1.9
19	4.4	3.5	3.1	2.9	2.7	2.6	2.5	2.5	2.4	2.4	2.3	2.2	2.2	2.1	1.9
20	4.3	3.5	3.1	2.9	2.7	2.6	2.5	2.4	2.4	2.3	2.3	2.2	2.1	2.0	1.8
21	4.3	3.5	3.1	2.8	2.7	2.6	2.5	2.4	2.4	2.3	2.2	2.2	2.1	2.0	1.8
22	4.3	3.4	3.0	2.8	2.7	2.5	2.5	2.4	2.3	2.3	2.2	2.1	2.1	2.0	1.8
23	4.3	3.4	3.0	2.8	2.6	2.5	2.4	2.4	2.3	2.3	2.2	2.1	2.0	1.9	1.8
24	4.3	3.4	3.0	2.8	2.6	2.5	2.4	2.4	2.3	2.2	2.2	2.1	2.0	1.9	1.7
26	4.2	3.4	3.0	2.7	2.6	2.5	2.4	2.3	2.3	2.2	2.1	2.1	2.0	1.9	1.7
28	4.2	3.3	2.9	2.7	2.6	2.4	2.4	2.3	2.2	2.2	2.1	2.0	2.0	1.9	1.6
30	4.2	3.3	2.9	2.7	2.5	2.4	2.3	2.3	2.2	2.2	2.1	2.0	1.9	1.8	1.6
40	4.1	3.2	2.8	2.6	2.4	2.3	2.2	2.2	2.1	2.1	2.0	1.9	1.8	1.7	1.5
60	4.0	3.1	2.8	2.5	2.4	2.2	2.2	2.1	2.0	2.0	1.9	1.8	1.7	1.6	1.4
120	3.9	3.1	2.7	2.4	2.3	2.2	2.1	2.0	2.0	1.9	1.8	1.7	1.7	1.6	1.2
∞	3.8	3.0	2.6	2.4	2.2	2.1	2.0	1.9	1.9	1.8	1.7	1.7	1.6	1.5	1.0

Значения критерия Фишера F при уровне значимости $\alpha = 0.01$
(число степеней свободы указано для дисперсии знаменателя – в строке,
для дисперсии числителя – в столбце)

df_2 \ df_1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**	**30**	∞
1	4052	4999	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6261	6366
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5
3	31.4	30.8	29.5	28.7	28.4	27.9	27.7	27.5	27.3	27.2	27.0	26.9	26.7	26.5	26.1
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.8	13.5
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.0	10.0	9.7	9.5	9.4	9.0
6	13.7	10.9	9.8	9.1	8.7	8.5	8.3	8.1	8.0	7.9	7.7	7.6	7.4	7.2	6.9
7	12.3	9.5	8.5	7.8	7.5	7.2	7.0	6.8	6.7	6.6	6.5	6.3	6.2	6.0	5.6
8	11.3	8.7	7.6	7.0	6.6	6.4	6.2	6.0	5.9	5.8	5.7	5.5	5.4	*5.2*	4.9
9	10.6	8.0	7.0	6.4	6.1	5.8	5.6	5.5	5.3	5.3	5.1	5.0	4.8	4.6	4.3
10	10.0	7.6	6.5	6.0	5.6	5.4	5.2	5.1	4.9	4.8	4.7	4.6	4.4	4.2	3.9
11	9.7	7.2	6.2	5.7	5.3	5.1	4.9	4.7	4.6	4.5	4.4	4.2	4.1	3.9	3.6
12	9.3	6.9	5.9	5.4	5.1	4.8	4.6	4.5	4.4	4.3	4.2	4.0	3.9	3.7	3.4
13	9.1	6.7	5.7	5.2	4.9	4.6	4.4	4.3	4.2	4.1	4.0	3.8	3.7	3.5	3.2
14	8.9	6.5	5.6	5.0	4.7	4.5	4.3	4.1	4.0	3.9	3.8	3.7	3.5	3.3	3.0
15	8.7	6.4	5.4	4.9	4.6	4.3	4.1	4.0	3.9	3.8	3.7	3.5	3.4	3.2	2.9
16	8.5	6.2	5.3	4.8	4.4	4.2	4.0	3.9	3.8	3.7	3.5	3.4	3.3	3.1	2.7
17	8.4	6.1	5.2	4.7	4.3	4.1	3.9	3.8	3.7	3.6	3.5	3.3	3.2	3.0	2.6
18	8.3	6.0	5.1	4.6	4.2	4.0	3.8	3.7	3.6	3.5	3.4	3.2	3.1	2.9	2.6
19	8.2	5.9	5.0	4.5	4.2	3.9	3.8	3.6	3.5	3.4	3.3	3.1	3.0	2.8	2.5
20	8.1	5.8	4.9	4.4	4.1	3.9	3.7	3.6	3.5	3.4	3.2	3.1	2.9	2.8	2.4
21	8.0	5.8	4.9	4.4	4.0	3.8	3.6	3.5	3.4	3.3	3.2	3.0	2.9	2.7	2.4
22	7.9	5.7	4.8	4.3	4.0	3.8	3.6	3.4	3.3	3.3	3.1	3.0	2.8	2.7	2.3
23	7.9	5.7	4.8	4.3	3.9	3.7	3.5	3.4	3.3	3.2	3.1	2.9	2.7	2.6	2.3
24	7.8	5.6	4.7	4.2	3.9	3.7	3.5	3.4	3.3	3.2	3.0	2.9	2.7	2.6	2.2
26	7.7	5.5	4.6	4.1	3.8	3.6	3.4	3.3	3.2	3.1	3.0	2.8	2.7	2.5	2.1
28	7.6	5.4	4.6	4.1	3.7	3.5	3.4	3.2	3.1	3.0	2.9	2.7	2.6	2.4	2.1
30	7.6	5.4	4.5	4.0	3.7	3.5	3.3	3.2	3.1	3.0	2.8	2.7	2.5	2.4	2.0
40	7.3	5.2	4.3	3.8	3.5	3.3	3.1	3.0	2.9	2.8	2.7	2.5	2.4	2.2	1.8
60	7.1	5.0	4.1	3.6	3.3	3.1	2.9	2.8	2.7	2.6	2.5	2.3	2.2	2.0	1.6
120	6.8	4.8	3.9	3.5	3.2	3.0	2.8	2.7	2.6	2.5	2.3	2.2	2.0	1.9	1.4
∞	6.6	4.6	3.8	3.3	3.0	2.8	2.6	2.5	2.4	2.3	2.2	2.5	1.9	1.7	1.0

Учебное издание

БАРЫШНИКОВА Надежда Ивановна
ИШБИРДИН Айрат Римович
ИШМУРАТОВА Майя Мунировна

МЕТОДЫ СБОРА, ОБРАБОТКИ ДАННЫХ И ПРЕДСТАВЛЕНИЯ РЕЗУЛЬТАТОВ В НАУЧНЫХ ИССЛЕДОВАНИЯХ В ПИЩЕВОЙ ПРОМЫШЛЕННОСТИ

Учебное пособие

Редактор Т.А. Колесникова
Компьютерная верстка Г.Н. Лапиной

Подписано в печать 11.11.2013. Рег. № 16-13. Формат 60х84/16. Бумага тип. № 1.
Плоская печать. Усл.печ.л. 3,50. Тираж 100 экз. Заказ

Издательский центр ФГБОУ ВПО «МГТУ»
455000, Магнитогорск, пр. Ленина, 38
Полиграфический участок ФГБОУ ВПО «МГТУ»

Н.И. Барышникова
А.Р. Ишбирдин
М.М. Ишмуратова

МЕТОДЫ СБОРА, ОБРАБОТКИ ДАННЫХ И ПРЕДСТАВЛЕНИЯ РЕЗУЛЬТАТОВ В НАУЧНЫХ ИССЛЕДОВАНИЯХ В ПИЩЕВОЙ ПРОМЫШЛЕННОСТИ

Магнитогорск
2013