Министерство образования республики Беларусь
Учреждение образования
МОГИЛЕВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ПРОДОВОЛЬСТВИЯ

Кафедра высшей математики

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебно-методическое пособие
для студентов – технологов дневной и заочной форм обучения

Могилев 2011

УДК

Рассмотрены и утверждены
на заседании кафедры
высшей математики
Протокол № 1 от 21.09.2010г.

Составитель О.А. Шендрикова

Рецензенты кандидат физико-математических наук, доцент, зав. кафедрой высшей математики УО МГУП
В. Э. Гарист
кандидат физико-математических наук, профессор,
УО МГУ им. А. А. Кулешова
Б.Д. Чеботаревский

Содержание

Введение

В своем знаменитом романе «Двенадцать стульев» Ильф и Петров утверждали: «Статистика знает все. Известно, сколько и какой пищи съедает в год средний гражданин республики… Известно, сколько в стране охотников, балерин, … станков, велосипедов, памятников, маяков и швейных машинок… Как много жизни, полной пыла, страстей и мысли, глядит на нас со статистических таблиц!...» Это ироническое описание дает довольно точное представление о статистике (от лат. *status* – состояние) – науке, изучающей, обрабатывающей и анализирующей количественные данные о самых разнообразных массовых явлениях в жизни.

Сегодня уже трудно себе представить область деятельности человека, в которой не использовалась бы математическая статистика. Так *экономическая статистика* изучает изменение цен, спроса и предложения на товары, прогнозирует рост и падение производства и потребления. *Медицинская статистика* изучает эффективность различных лекарств и методов лечения, вероятность возникновения заболевания в зависимости от возраста, пола, наследственности, условий жизни, вредных привычек, прогнозирует распространение эпидемий. *Демографическая статистика* изучает рождаемость, численность населения, его состав (возрастной, национальный, профессиональный). Существует еще статистика финансовая, налоговая, биологическая, метеорологическая и др.

Статистика имеет многовековую историю. Уже в Древнем Мире вели статистический учет населения. Впервые идеи статистических методов высказал английский математик Томас Симпсон (1710-1761). Однако в ту пору его идеи не нашли никакого использования. Лишь в XX веке произошло становление математической статистики как науки, основанной на законах теории вероятностей. Соединение накопленных к этому времени практических методов обработки данных с математическим аппаратом теории вероятностей превратило эти две отрасли человеческого знания в мощный инструмент для исследования законов природы и общества.

Чтобы конкретизировать сказанное, рассмотрим лишь один из немногих вопросов, возникающих в обществе, решению которого способствует математическая статистика. Это задача проверки качества изготовляемой продукции. От повышения качества продукции зависит прогресс производства, наиболее полное удовлетворение потребностей общества в производимом товаре. Оборудование высокого качества работает дольше и успешнее без частых профилактических осмотров и восстановлений, без дорогостоящих замен узлов и деталей, без неоправданно частых остановок на ремонт и на приведение в порядок. В результате каждая денежная единица, вложенная в процессе производства для целей повышения качества продукции, оборачивается многократной экономией для предприятия в процессе эксплуатации. Повышение качества продукции влияет и на конкурентоспособность предприятия на рынке, экономию материалов и запасных частей, как в процессе изготовления, так и в процессе использования изготовляемых изделий. Однако проверка качества продукции - не такой уж простой процесс, как может показаться.

Во-первых, часто проверка качества связана с частичной или полной порчей изделия (при проверке качества произведенных продуктов питания никто не будет проверять каждую банку консерв в отдельности). Во-вторых, современные объемы производства так велики, что проверка качества изготовленной продукции потребует такой затраты труда и средств, что зачастую они сравнимы с затратами на ее производство. Во многих случаях нужно искать достаточно надежные методы проверки качества больших партий изделий, когда приходится проверять качество не каждого изделия данной партии, а только сравнительно небольшой ее части. Такие методы получили наименование *статистических методов выборочного контроля.*

Но это только один из вопросов, которыми занимается математическая статистика. Важно не только правильно организовывать отбор материала, проводить анализ полученных данных, но и уметь на основе этих данных сделать необходимые выводы. Известный математик Б.В. Гнеденко описывает случай из своей практики, наглядно демонстрирующий важность рассмотренной проблемы так: «Как-то на открытии Дома техники ко мне обратились молодые инженеры, физики и математики, работавшие на одном современном заводе. Они показали тщательно собранные и обработанные статистические данные относительно качества одного из весьма важных элементов электронной техники. Выход дефектных изделий в ту пору был очень высок. Ни к тому, как были собраны исходные статистические материалы, ни к их обработке у меня не было никаких претензий. Но они упустили из вида самое важное: из полученных результатов не сделали необходимых выводов и не попытались проанализировать причины преждевременных массовых отказов. А именно в этом и заключается смысл сбора статистических данных и их обработки. Сделав правильные выводы, поняв причину такого высокого процента дефектных изделий, в течение каких-нибудь трех лет, эта группа молодых людей зарегистрировала несколько изобретений, которые позволили автоматизировать технологический процесс, в том числе и контроль качества исходных материалов, что в конечном итоге привело к снижению доли дефектных изделий до десятых долей процента».

В данном пособии мы рассмотрим на простейших примерах основные вопросы раздела высшей математики «Математическая статистика». Решение некоторых задач будет приводиться с использование электронных таблиц Excel. Такие задачи обозначены ▌.

Предмет и задачи математической статистики

Рассмотрим серию опытов:

1) На точных аналитических весах взвешивается одно и тоже тело. Условия проведения опыта остаются неизменными, однако, результаты взвешиваний каждый раз немного отличаются друг от друга. Отчего так происходит?

2) Из орудия стреляют по мишени. Вид орудия, снаряда, угол стрельбы остается неизменным. Теоретически, траектория движения снарядов должна быть одинаковой. Но на практике она несколько разная, причем точки попадания снаряда по мишени разбросаны. Что послужило помехой?

Очевидно, что во всех этих опытах, как, впрочем, и во многих явлениях, которые встречаются в природе, присутствуют в той или иной мере элементы случайности.

В первом опыте на результат могло повлиять положение тела и разновесов на чаше весов, вибрация аппаратуры, смещение головы наблюдателя и т.д. Во втором случае – дефекты при изготовлении снарядов, метеоусловия, которые менялись незначительно от выстрела к выстрелу и т.д. Поэтому, как бы точно мы не старались соблюдать все условия проведения опыта, результаты будут несколько отличаться друг от друга.

Вернемся к опыту со стрельбой из орудия. Допустим, перед нами стоит задача изготовления прицельного орудия. Для этого нам важно знать, в том числе и траекторию движения снаряда, изменение скорости ветра. Небольшие дефекты при изготовлении снарядов не окажут существенного влияния на решение поставленной задачи. Поэтому мы их в данном вопросе можем не учитывать. Для построения математической модели из бесчисленного множества факторов, мы выбрали самые, решающие, значительные; влиянием остальных просто пренебрегли. Это – классическая схема многих точных наук, которая называется *детерминистской*.

А что, если мы производим стрельбу по цели, которая меньше зоны рассеивания снарядов. Тогда, например, важно, какой процент выпущенных снарядов попадет в цель, сколько нужно потратить снарядов, чтобы с достаточной степенью надежности попасть в цель, и т.д. Чтобы ответить на эти вопросы уже нельзя пренебрегать случайными явлениями, надо рассмотреть рассеивание снарядов со стороны закономерностей, присущих ему как случайному явлению.

Так, рассматривая в совокупности однородные случайные явления, мы часто обнаруживаем закономерности. В первом опыте, при небольшом количестве взвешиваний, результаты представляют собой набор хаотичных чисел. Но как только число взвешиваний увеличивается, нетрудно заметить, что результаты взвешиваний группируются около некоторого среднего значения. Тоже происходит и с выстрелами. При большом количестве выстрелов видно, что в зоне рассеивания снарядов попадания гуще в центральной области, чем по краям. Даже, казалось бы, на первый взгляд, непредсказуемый уличный травматизм, обнаруживает в массе своей определенные закономерности, которые учитываются при работе служб скорой помощи.

Выявленные в подобного рода задачах закономерности называются *статистическими*. Они исследуются методами специальных математических дисциплин – теории вероятностей и математической статистики.

Математическая статистика – раздел математики, в котором рассматриваются приближенные методы нахождения законов числовых характеристик СВ по результатам экспериментов или наблюдений.

Предметом математической статистики является изучение случайных величин (СВ) по результатам наблюдений.

Статистический прогноз не отвечает на вопрос, что произойдет при таких-то условиях, он указывает только границы, в которых, с достаточно высокой степе-

нью достоверности, будут заключены интересующие нас параметры. Чем обширнее изучаемый массив случайных явлений, тем уже эти границы, тем точнее и определеннее становится вероятностный прогноз.

Задачи математической статики могут быть классифицированы следующим образом.

1) Упорядочивание статистического материала, полученного в результате наблюдений (опыта, эксперимента), представление в удобном для обозрения и анализа виде, например, в виде таблиц и графиков.

2) Оценка, хотя бы приблизительно, интересующих нас характеристик наблюдаемой СВ, а также определение точности, с которой оцениваются эти характеристики при данном количестве опытов. Зачастую тип распределения случайной величины известен, а неизвестны только параметры распределения. Например, рассматривая распределение числа успехов в последовательности испытаний Бернулли, достоверно известен тип распределения – биномиальное. Однако вероятность успеха в одном испытании p может быть неизвестна и ее следует определить, исходя из наблюдаемого числа успехов.

3) Проверка статистических гипотез. Относящиеся сюда задачи имеют несколько разновидностей. Одной из важнейших является задача проверки гипотезы о законе распределения случайной величины. Например, по имеющимся опытным данным, относящимся к одной или нескольким случайным величинам, мы предполагаем, что СВ имеет определенный тип распределения, однако у нас нет уверенности. Поэтому, на основании наблюдаемых значений СВ необходимо оценить параметры распределения и проверить, насколько хорошо наблюдаемые значения соответствуют гипотезе об этом законе распределения. Аналогичным образом формулируются задачи и в других случаях. Так, например, можно интересоваться гипотезой о том, что две СВ независимы, что с течением времени вероятность некоторого случайного события (например, получения бракованного изделия) не увеличивается и т.д.

4) Разработка методов, позволяющих по результатам исследования некоторой части совокупности объектов, делать обоснованные выводы о распределении признака изучаемых объектов по всей совокупности.

Генеральная совокупность и выборка. Повторная и бесповторная выборки. Репрезентативная выборка

Исходным материалом для статистического исследования является совокупность результатов наблюдений (опытов, испытаний или экспериментов) относительно некоторого признака. Например, рассматривая партию пирожных типа «Эклер», можно исследовать: вкусовые характеристики пирожных (качественный признак), размер и вес (количественный признак). Каждый такой признак образует СВ, наблюдения над которой мы производим.

Генеральной совокупностью называется совокупность всех возможных значений, или реализаций, исследуемых объектов, подлежащих изучению относительно качественного или количественного признака.

В дальнейшем можно говорить либо признак, либо случайная величина (СВ).

Но на практике проводить изучение над всеми объектами, так называемое *сплошное обследование*, дорого (перепись населения), экономически нецелесообразно (например – при проверке качества продукции, никто не проверяет каждую единицу товара). В таких случаях случайным образом отбирают из всей совокупности ограниченное число объектов и подвергают их исследованию т.е. пользуются т.н. *выборочным обследованием.*

Выборочной совокупностью или просто **выборкой** называется совокупность объектов, отобранных случайным образом из генеральной совокупности.

Объемом совокупности (выборочной или генеральной) называют число объектов этой совокупности.

Конкретные значения выборки, полученные в результате наблюдений (испытаний), называют **реализацией выборки** или **вариантами выборки** и обозначают $x_1, x_2, ..., x_n$.

Пример 1

1) Предположим, что имеется партия рыбных консерв в 10000 банок. Потому как исследовать все банки консерв не представляется возможным, чтобы можно было судить хотя бы приблизительно об относительной доле брака, отбирают и контролируют 100 банок. В этом примере генеральной совокупностью является исходная партия консервных банок. Объем генеральной совокупности $N = 10000$. Выборкой является множество банок, взятых из генеральной совокупности для контроля. Объем выборки $n = 100$.

2) Всю партию полученного полимера нет возможности проверять на качество. Поэтому довольствуются случайным образом отобранным числом готового полимера, которое и является выборкой. А вся готовая продукция в данном случае представляет собой генеральную совокупность.

Выбор элементов генеральной совокупности можно организовать двояким способом: *выбор без повторений* (возвращений), при котором отобранный объект в генеральную совокупность не возвращается, и *выбор с повторениями* (возвращениями) – отобранный объект перед отбором следующего возвращается в генеральную совокупность. На практике обычно пользуются бесповторным случайным отбором.

Для получения хороших оценок характеристик генеральной совокупности необходимо, чтобы выборка была *репрезентативной*, т.е. достаточно правильно представляла изучаемые признаки генеральной совокупности. В силу закона больших чисел можно утверждать, что выборка будет репрезентативной, если ее осуществить случайно. Выборка называется *случайной,* если любой объект генеральной совокупности с одинаковой вероятностью может попасть в эту выборку.

Пример 2

Генеральная совокупность –100 ящиков с оборудованием, содержимое которых нужно проверить на соответствие документам. Проще всего отобрать ближайшие 10 ящиков и проверить их, но поставщик мог позаботиться об укомплек-

товании какого то числа первых и какого то числа последних ящиков, поэтому такая выборка не будет репрезентативной.

Лучше всего по накладной случайно отобрать ящики для проверки. Случайность отбора гарантирует, что поставщики не смогут предугадать, какие ящики будут отобраны для проверки.

Если объем генеральной совокупности конечен, то для обеспечения равной возможности попадания объектов в выборку применяют различные приемы, в частности, используют *генераторы случайных величин* (т.е. таблицы случайных чисел). Для того чтобы отобрать, например, 50 объектов из пронумерованной генеральной совокупности, открывают любую страницу таблицы случайных чисел и выписывают подряд 50 чисел; в выборку попадают те объекты, номера которых совпадают с выписанными случайными числами. Если окажется, что случайное число таблицы превышает число N, то такое случайное число следует пропустить. При осуществлении бесповторной выборки случайные числа таблицы, уже встречавшиеся ранее, следует также пропустить.

Другой способ получения случайной выборки можно осуществить с помощью Excel. Идея заключается в том, чтобы перемешать элементы генеральной совокупности случайным образом, а затем отобрать необходимое количество элементов.

Пример 3

Построим с помощью программы Excel случайную выборку объемом n=5 из генеральной совокупности объемом N=20. В одном столбце расположим числа от 1 до 20. Во втором столбце введем в ячейку B1 формулу (=СЛЧИС ()), а далее протянем вдоль всего столбца получаем столбец случайных чисел.

Затем выделяем оба столбца, выполняем команду (Данные→Сортировка→Значения столбца случайных чисел (2 столбец)→по возрастанию). В результате числа в первом столбце будут упорядочены случайным образом. Для получения искомой выборки нам необходимо взять первые 5 чисел, в нашем случае это 10, 14, 1, 19, 11, (рисунок 1). Полученная таким образом случайная выборка обладает тем же свойством репрезентативности, что и выборка, построенная использованием таблицы случайных чисел.

	A	B
1	10	0.63501
2	14	0.0431
3	1	0.3598
4	19	0.2584
5	11	0.40625
6	8	0.26734
7	16	0.22728
8	15	0.79281
9	12	0.08502
10	13	0.40559
11	2	0.81167
12	9	0.24934
13	20	0.12837
14	6	0.20007
15	7	0.01053
16	18	0.9759
17	3	0.79312
18	5	0.46302
19	17	0.19601

Рисунок 1 – Случайная выборка

Помимо использования таблиц случайных чисел применяют различные способы отбора: *типический*, при котором генеральную совокупность делят на «типические» части и отбор осуществляется из каждой части (например, при проведении опроса о вкусовых пристрастиях разделить опрашиваемых по возрасту, полу и т.д.). Так типический способ отбора применяется, если обследуемый признак заметно колеблется в различных типических частях генеральной совокупности (так мнение людей о производимых продуктах питания может быть разным у людей разного возраста); *механический*, при котором отбор производится через определенный интервал (например, отбирают каждый 30 произведенный товар); *серийный,* при котором объекты из генеральной совокупности выбираются не по одному, а «сериями» (например, если изделия изготавливаются большой группой

станков-автоматов, то подвергают сплошному обследованию продукцию только нескольких станков).

На практике часто применяют комбинированный отбор.

Вопросы для самопроверки

1) Что называется генеральной совокупностью, а что – выборкой?

2) Почему приходится прибегать к формированию выборки?

3) Результаты переписи населения, проводимой в стране, являются генеральной совокупностью или выборкой?

4) На заводе по производству деталей для машин сельского хозяйства возникли некоторые проблемы с качеством изготовления поршней. Для проведения анализа принято решение собрать информацию о продукции, выпущенной в определенный день. Для каждого из указанных ниже методов извлечения выборки определите, является ли она репрезентативной:

а) первые 10 произведенных поршней;

б) образованная в конце дня случайная выборка;

в) все явно бракованные поршни вместе со случайной выборкой из очевидно стандартных поршней.

5) В ходе опроса предстоит выяснить отношение жителей региона к постройке торгового центра. Из каких категорий, на ваш взгляд, следует выбирать людей при построении выборки?

Вариационный ряд. Эмпирическая функция распределения и ее свойства

Рассмотрим следующий пример.

Пример 4:

Хлебозавод специализируется на выпуске городских булок (4 линии) и подмосковных батонов (1 линия). Проверяется соответствие тестовых заготовок нормам. Для обеспечения точности выборки с учетом влияния сменности в работе предприятия на каждой из пяти линий сделали в разные смены по несколько выборок.

Ниже приведены данные распределения веса тестовых заготовок для выпечки городских булок (0,2 кг).

197, 198, 200, 201, 203, 202, 199, 200, 200, 198, 200, 201, 202, 200, 199, 201, 199, 200, 200, 203.

Проанализируем полученные результаты.

Понятно, что чем больше данных, тем труднее производить анализ. В данном случае объем выборки равен 20, но в статистике зачастую данных бывает гораздо больше. Чтобы не погрязнуть в море цифр их представляют в удобном для человека виде – в виде таблиц. Первое, что мы должны сделать – это упорядочить полученные данные.

╟ Операция расположения значений СВ. по неубыванию называется **ранжированием** статистических данных.

╟ Последовательность значений СВ. X, полученная после ранжирования, $x_1, x_2, \ldots, x_n$ называется **вариационным рядом.**

В нашем случае, вариационный ряд имеет вид

197, 198, 198, 199, 199, 199, 200, 200, 200, 200, 200, 200, 200, 201, 201, 201, 202, 202, 203, 203.

▌ Если объем данных велик, то ранжирование статистических данных будет трудоемким. Поэтому в Excel, воспользовавшись меню Данные→Сортировка, можно легко и быстро упорядочить данные.

╟ Числа n_i, показывающие, сколько раз встречаются варианты x_i (i = 1, 2, …, n) в ряде наблюдений, называются **частотами**.

╟ Отношение частот к объему выборки называется **относительными частотами**, которые находятся по формуле (1)

$$w_i = \frac{n_i}{n} \qquad (1)$$

╟ Перечень вариант и соответствующих им частот или относительных частот называется **статистическим рядом**.

Объем n данной в примере 4 выборки равен 20. Подсчитаем частоту и относительную частоту вариант. Масса 197 г., например, встречается 1 раз - это частота, а относительная частота для этой варианты равна $w_i = \frac{1}{20} = 0{,}05$ и т.д. по всем данным. Полученные результаты занесем в таблицу. В первой строке запишем все значения выборки (варианты), во вторую строку – соответствующие им значения частот или относительных частот. Получим статистическое распределение выборки, представленное в таблице 1.

Таблица 1 – Статистический ряд распределения

Значения вариант x_i	197	198	199	200	201	202	203
Значения частот n_i	1	2	3	7	3	2	2
Значения относит. частот w_i	0,05	0,1	0,15	0,35	0,15	0,1	0,1

Такие систематизированные данные легче анализировать. В этом плане относительная частота является довольно содержательной характеристикой. Например, считается, что норма тестовой заготовки 200г. Получается, что только 35% заготовок соответствуют норме. Обратите внимание, что сумма частот будет равна объему выборки (в данном примере – количеству отобранных тестовых заготовок), а сумма относительных частот равна единице.

Статистический ряд распределения в математической статистике является аналогом ряда распределения СВ X в теории вероятностей.

В случае, когда n велико или СВ X является непрерывной, составляют *интервальный статистический ряд.* Для этого в первую строку таблицы статистического распределения вписывают частичные промежутки: $[x_0, x_1), [x_1, x_2), ..., [x_{k-1}, x_k]$, которые как правило берутся одинаковыми по длине – шаг разбиения $h = x_0 - x_1 = x_2 - x_1 = ...$

Длину промежутка можно найти по формуле

$$h = \frac{x_{\max} - x_{\min}}{m}, \qquad (2)$$

где m находится по формуле Стерджесса

$$m = 1 + 3{,}322 \lg n; \qquad (3)$$

$x_{\max} - x_{\min}$ – разность между наибольшим и наименьшим значениями СВ.

За начало первого интервала рекомендуется брать $x_{нач} = x_{\min} - \frac{h}{2}$. Во вторую строчку таблицы записывают количество результатов наблюдений n_i, $(i = \overrightarrow{1, k})$, попавших в каждый интервал.

Пример 5

Пусть объём выборки тестовых заготовок, рассмотренной в примере 4, увеличился до 40.

197, 198, 200, 201, 203, 202, 199, 200, 200, 198, 200, 201, 202, 200, 199, 201, 199, 200, 200, 203, 197, 200, 200, 198, 200, 200, 199, 200, 200, 202, 200, 200, 197, 200, 199, 200, 200, 200, 198, 200.

Найдем прежде всего длину интервала. В нашем случае $n = 40$. По формуле (3) $m = 1 + 3{,}322 \lg n = 1 + 3{,}322 \lg 40 = 6$, тогда длина интервала по формуле (2) равна $h = \frac{203 - 197}{6} \approx 1$. За начало первого интервала возьмем $x_{нач} = x_{\min} - \frac{h}{2} = 197 - \frac{1}{2} \approx 196{,}5$.

Итак, статистический ряд распределения будет иметь вид, представленный в таблице 2.

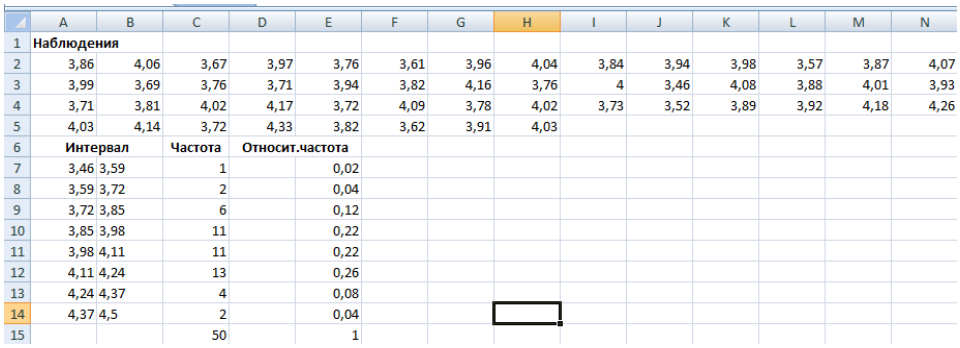
Таблица 2 – Интервальный ряд распределения тестовых заготовок

$[x_{k-1}; x_k]$	[196,5; 197,5)	[197,5; 198,5)	[198,5; 199,5)	[199,5; 200,5)
n_i	3	4	5	20
$[x_{k-1}; x_k]$	[200,5; 201,5)	[201,5; 202,5)	[202,5; 203,5)	
n_i	3	3	2	

В случае, когда данных достаточно много и систематизировать их вручную достаточно трудоемко, процедуру подсчета частот можно выполнить с помощью электронных таблиц Excel.

Пример 6.

Наблюдения за жирностью молока дали следующие результаты, представленные на рисунке 2.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Наблюдения													
2	3,86	4,06	3,67	3,97	3,76	3,61	3,96	4,04	3,84	3,94	3,98	3,57	3,87	4,07
3	3,99	3,69	3,76	3,71	3,94	3,82	4,16	3,76	4	3,46	4,08	3,88	4,01	3,93
4	3,71	3,81	4,02	4,17	3,72	4,09	3,78	4,02	3,73	3,52	3,89	3,92	4,18	4,26
5	4,03	4,14	3,72	4,33	3,82	3,62	3,91	4,03						
6	Интервал		Частота	Относит.частота										
7	3,46	3,59	1		0,02									
8	3,59	3,72	2		0,04									
9	3,72	3,85	6		0,12									
10	3,85	3,98	11		0,22									
11	3,98	4,11	11		0,22									
12	4,11	4,24	13		0,26									
13	4,24	4,37	4		0,08									
14	4,37	4,5	2		0,04									
15			50		1									

Рисунок 2 – Нахождение частот и относительных частот

Построим интервальный статистический ряд распределения частот и относительных частот. Вначале, определим длину интервала по формулам 1 и 2, а также начальное значение первого интервала. Имеем, *h = 0,13*, тогда за начало интервала возьмем 3,46. Значение частот и относительных частот найдем с помощью Excel.

В нашем случае:

1) Заполняем столбец интервал. В ячейку А7 вводим начальное значение3,46, согласно нашему подсчету. В следующую ячейку А8 вставляем формулу (=СУММ(А7; 0,13)) (т.к. найденная ранее длина интервала равна 0,13) нажимаем Enter и протягиваем по всей длине столбца, так, чтобы последнее значение этого столбца не превышало максимального значения выборки.

2) Заполняем столбец «Частота» с помощью встроенной функции ЧАСТОТА. Для этого выделяем блок ячеек С7:С14. С помощью меню выбираем: f(x)→Статистические→ЧАСТОТА. В *массив данных* вводим диапазон наблюдений, т.е. А2:N5 указателем мыши. В рабочее поле *двоичный массив* вводим диапазон интервалов А7:А14 аналогичным образом. После нажатия клавиш Ctrl+Shift+Enter в столбце «Частота» С7:С14 появится массив частот. Для проверки правильности нахождения частот достаточно найти сумму значений этого столбца, выделив столбец С7:С14 и нажав $\sum$, ниже мы увидим значение50, что соответствует объему выборки.

3) Для нахождения относительных частот в ячейку Е7 вводим формулу (=С7/С$15), нажимаем Enter, далее, протягивая левой кнопкой мыши, копируем введенную формулу в диапазон Е8:Е14, рисунок 2. Если мы просуммируем все значения из этого диапазона, то убедимся, что получится 1.

Одним из способов распределения вариационного ряда является построение функции распределения. В связи с тем, что эта функция находится опытным (эмпирическим) путем, ее называют *эмпирической функцией распределения*.

‖ **Эмпирической функцией** распределения называется функция $F^*(x)$, определяющая для каждого значения x относительную частоту события $X < x$.

$$F^*(x) = \frac{n_x}{n} \qquad (4)$$

где n_x – число вариант (наблюдений), меньшее x;

n – общее число наблюдений (объем выборки).

Свойства эмпирической функции распределения $F^*(x)$:

- значения эмпирической функции принадлежат отрезку [0;1];
- $F^*(x)$ – неубывающая функция;
- если x_1 – наименьшая варианта, то $F^*(x) = 0$, при $x \le x_1$;
- если x_k – наибольшая варианта, то $F^*(x) = 1$, при $x > x_k$;

Пример 7

Построим эмпирическую функцию распределения по данным о потреблении мяса и мясопродуктов в Республике Беларусь за период с 1990 по 2010 (на душу населении в год, кг):

58	60	60	62	59	59	57	58	58	57
59	59	62	60	60	58	60	59	62	57

Упорядочим этот вариационный ряд:

57	57	57	58	58	58	58	59	59	59
59	59	60	60	60	60	60	62	62	62

Наименьшая варианта (масса указанной продукции, потребляемой в год на душу населения, кг.) равна 57, следовательно $F^*(x)=0$ при $x\le 57$ (см. свойства эмпирической функции распределения); значение при $57\le x<58$, а именно $x_1=57$ наблюдалось 3 раза и $F^*(x)=\frac{3}{20}=0{,}15$ (по формуле (4)); при $58\le x<59$, а именно $x_1=57,\ x_2=58,$ значение наблюдалось 3+4 = 7 раза и $F^*(x)=\frac{7}{20}=0{,}35$, и т.д.

Таким образом, искомая $F^*(x)=\begin{cases}0 & при\ x\le 57,\\ 0{,}15 & при\ 57<x\le 58,\\ 0{,}35 & при\ 58<x\le 59,\\ 0{,}6 & при\ 59<x\le 60,\\ 0{,}85 & при\ 60<x\le 62,\\ 1 & при\ \ x>62.\end{cases}$

Построим график эмпирической функции распределения.

График этой функции (рисунок 3) уже дает некоторое общее представление о характере распределения.

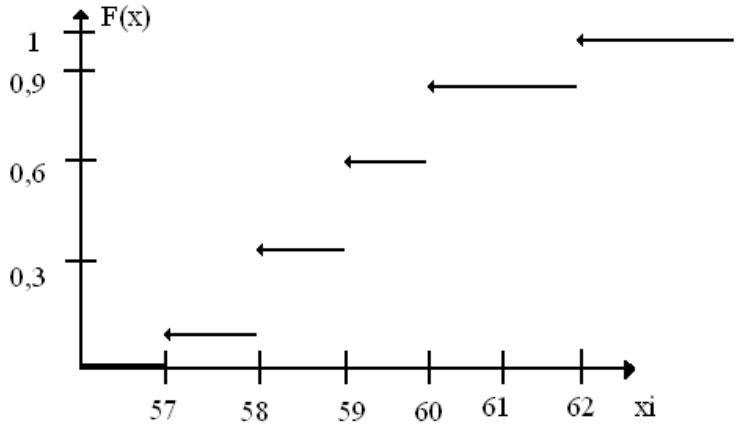


Рисунок 3 – График эмпирической функции распределения

При неограниченном увеличении числа n скачки кривой $F^*(x)$ станут более мелкими; кривая станет плавне, будет приближаться (сходится по вероятности) к функции распределения *F (x)* случайной величины *X*

На практике, как правило, приходится обрабатывать выборку большого объема. В связи с этим построение эмпирической функции распределения для минимизации временных затрат удобнее проводить с помощью Excel.

Пример 8

Построим эмпирическую функцию распределения по данным о потреблении мяса и мясопродуктов в Республике Беларусь только за последние 50 лет. Прежде чем находить эмпирическую функцию распределения, найдем значения частот и относительных частот аналогичным образом, рассмотренном в примере 6.

Для нахождения значений эмпирической функции распределения в ячейку M9 (рисунок 4) перепишем значение ячейки H9, а затем в ячейку M10 введем формулу (=M9+H10). Нажимаем Enter. Протягиванием (за правый нижний угол при нажатой левой кнопки мыши) скопируем введенную формулу в диапазон M10:M14. Далее строим диаграмму эмпирической функции распределения с помощью мастера диаграмм. Для этого выделяем область M9:M14, далее выбираем функцию Мастер диаграмм, выбираем вид графика и соответствующий макет

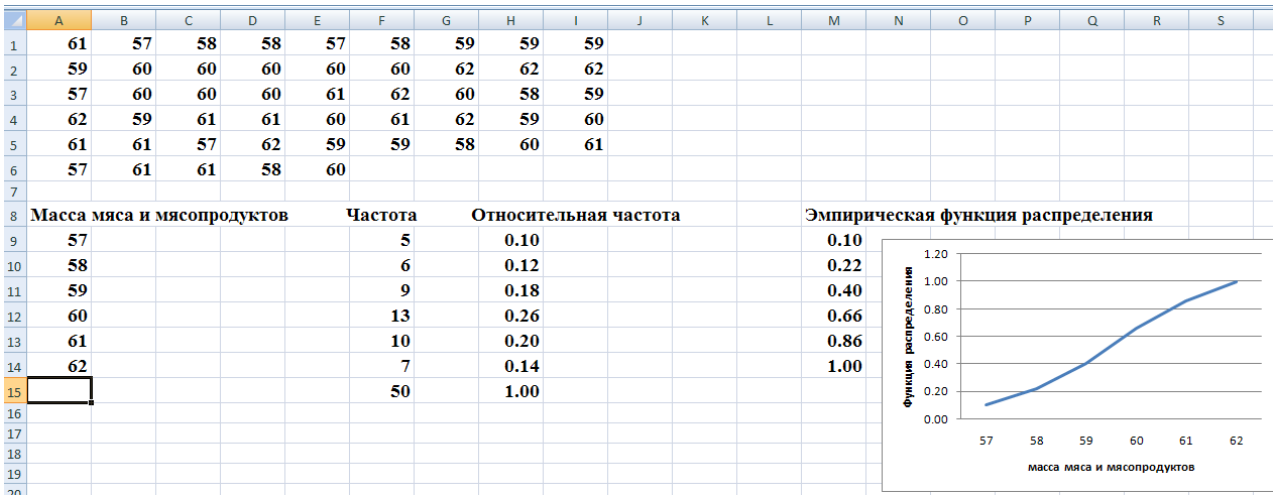


Рисунок 4 – Построение эмпирической функции распределения в Excel

Вопросы для самопроверки

1) Что называется вариационным рядом?

2) Что называется частотой, а что – относительной частотой?

3) Чему равна сумма частот, относительных частот вариационного ряда?

4) Что называется эмпирической функцией распределения?

Графическое изображение вариационных рядов: полигон, гистограмма

В целях наглядности строят различные графики статистического распределения. По графику можно предположить какая существует зависимость между изучаемыми величинами, в частности, общее представление о законе распределения СВ дают полигон и гистограмма.

Полигоном частот называют ломаную линию, отрезки которой соединяют точки с координатами $(x_1, n_1), (x_2, n_2), \ldots, (x_k, n_k)$, а **полигоном относительных частот** – ломаную линию, отрезки которой соединяют точки с координатами $(x_1, w_1), (x_2, w_2), \ldots, (x_k, w_k)$.

Для построения полигона частот или относительных частот, необходимо на оси абсцисс отложить значение вариант x_i, а на оси ординат – соответствующие им значения частот n_i или значения относительных частот w_i. Масштаб выбирают

таким образом, чтобы рисунок имел желательный размер и обеспечивал необходимую наглядность. Полученные точки соединяют отрезками прямых. Отсюда и название: *полигон* в переводе с греческого означает *многоугольник*.

Пример 9

▌ В результате серии экспериментов по подбрасыванию 10 монет фиксировалось количество монет, выпавших «орлом». Результаты представлены следующим числовым рядом: 5, 4, 5, 6, 2, 6, 8, 6, 3, 4, 5, 8, 5, 2, 5, 2, 5, 7, 3, 3, 5, 4, 5, 5, 6, 5, 7, 6, 3, 5, 5, 5, 5, 6, 5, 5, 5, 4, 7, 4, 5, 4, 5, 7, 7, 7, 6, 6, 4, 4. Построим полигон частот.

Все вычисления частот и относительных частот, а также построение полигона частот проводим в Excel. Заполняем ячейки A2:N5 результатами наблюдений. В ячейки C8:C17 заполняем все значения, которые может принять комбинация выпавших монет. В ячейки B8:B17 заполняем значение частот, аналогично тому, как мы находили частоты в примере о продолжительности работы ламп. Затем на основе полученных данных строим соответствующий график, изображенный на рисунке 5. Действительно, чаще всего орел выпадет на пяти из десяти монет, реже на четырех и шести и т.д.

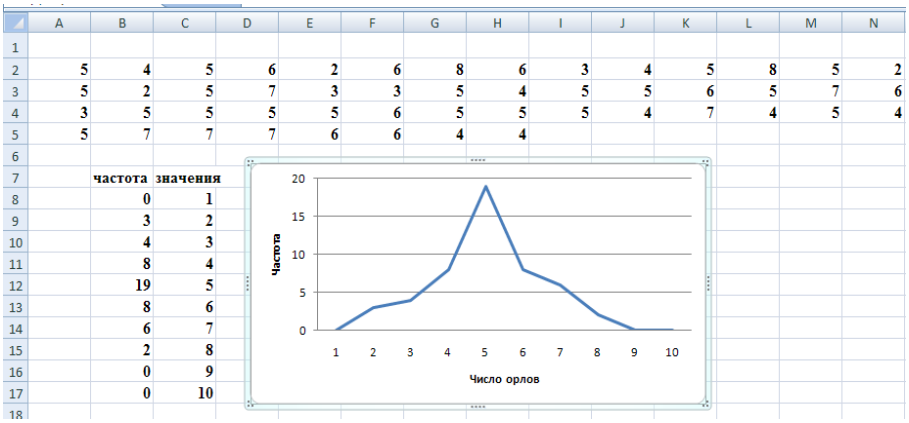


Рисунок 5 – Полигон частот

Кроме дискретных вариационных рядов рассматриваются интервальные вариационные ряды, в которых значения признака могут меняться непрерывно (например, урожай какой-либо зерновой культуры).

В случае, когда задан интервальный статистический ряд, то строят гистограмму.

▐ **Гистограммой частот** называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{n_i}{h}$.

Отметим, что площадь i-го частичного прямоугольника равна $h_i \cdot \frac{n_i}{h_i} = n_i$ – сумме частот вариант i-го интервала. А площадь гистограммы частот равна сумме всех частот, т.е. объему выборки.

Действительно, если S_i – площадь i-го прямоугольника, то

$S_i = h_i \cdot \frac{n_i}{h_i} = n_i$, тогда $S = \sum_{i=1}^{k} S_i = \sum_{i=1}^{k} n_i = n$.

Гистограммой относительных частот – фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{w_i}{h}$.

Для того чтобы построить гистограмму частот, на оси абсцисс откладываем значения интервалов, а на оси ординат – значения высот. Далее строят прямоугольники, основаниями которых служат длины интервалов, а высоты – равны расстоянию $\frac{n_i}{h}$.

Отметим, что площадь i-го частичного прямоугольника равна $h_i \cdot \frac{w_i}{h_i} = w_i$ – сумме относительных частот вариант, половин i-го интервала. Площадь S гистограммы относительных частот равна сумме относительных частот, т.е. 1.

Действительно, если S_i – площадь i-го прямоугольника, то

$$S_i = h_i \cdot \frac{w_i}{h_i} = w_i, \text{ тогда } S = \sum_{i=1}^{k} S_i = \sum_{i=1}^{k} w_i = \sum_{i=1}^{k} \frac{n_i}{n} = 1$$

Пример 10

В таблице 3 приведено распределение числа взрослых рабочих-женщин цеха по росту. Постройте по этим данным гистограмму частот.

Таблица 3 – Распределение числа рабочих-женщин по росту

Рост, см	143-146	146-149	149-152	152-155	155-158	158-161	161-164	164-167
Число женщин	1	2	8	26	65	120	181	201
Рост, см	167-170	170-173	173-176	176-179	179-182	182-185	185-188	Итого
Число женщин	170	120	64	28	10	3	1	1000

При построении полигона частот аналогичным образом заносим все данные по столбцам. Затем, выделив эти данные, выбираем меню Гистограмма.

Из предложенного перечня разнообразных макетов гистограмм конструируем наиболее подходящий график, представленный на рис. 6.

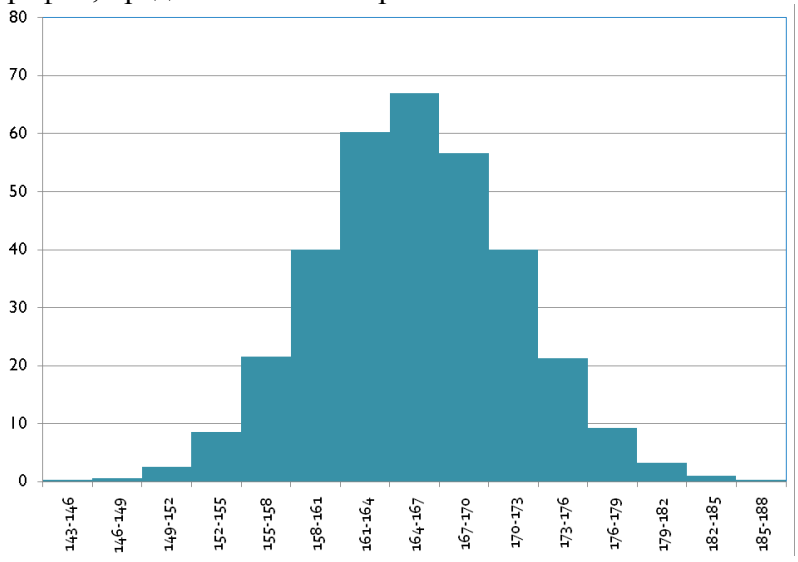


Рисунок 6 – Гистограмма частот

Вопросы для самопроверки

1) Проанализируйте построенный график по данным примера о распределении числа взрослых рабочих-женщин цеха по росту.

Числовые характеристики вариационного ряда: выборочное среднее, мода и медиана, выборочная дисперсия, среднее квадратическое отклонение

В результате исследований, связанных с массовыми явлениями, собирается много числовых данных. До сих пор мы решали задачи, связанные с обработкой статистических данных, представлением их в более наглядном и компактном виде таблицы, диаграммы. Необходимо «заменить» всю совокупность числовых данных выборки одним-двумя числовыми параметрами, которые описывают СВ суммарно. Зачастую при решении задач, связанных с конечным результатом опыта, нет надобности в знании закона распределения отдельных случайных величин, достаточно знать их числовые характеристики. Возникает проблема – найти такие числовые характеристики, которые довольно полно характеризовали бы полученный числовой материал.

Итак, числа, характеризующие наиболее существенные черты распределения, называются *числовыми характеристиками СВ.* Важнейшие среди этих характеристик являются *характеристики положения*, фиксирующие некоторое среднее значение СВ, около которого группируются все возможные значения. Так, если вспомнить выражения «средний балл», «средняя зарплата», «средний доход», то интуитивно мы подразумеваем определенную числовую характеристику, описывающую ее *положение* на числовой оси. Среди этих характеристик важную роль в статистике играют: *выборочная средняя, мода, медиана.*

Выборочной средней называется среднее арифметическое всех значений признака выборочной совокупности.

Если все значения $x_1, x_2, \ldots, x_n$ признака выборки объема n различны, то

$$\overline{x_B} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i \quad \text{или} \quad \overline{x_B} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot w_i \qquad (5)$$

где $w_i = \frac{n_i}{n}$.

Именно эта величина будет, скорее всего, ориентиром для преподавателя при выставлении итоговой оценки в баллах студенту за работу в течение семестра. При этом среднее значение ряда вполне может не совпадать ни с одной из его оценок.

Рассмотрим примеры, иллюстрирующие основные свойства выборочной средней.

1) Выборочная средняя постоянной величины равно этой постоянной. Это значит, что если при исследовании признака x он n раз принимал одно и тоже значение c, то и выборочная средняя $\overline{x}_B = c$.

2) Если каждое значение признака Z равно сумме (разности) выборочных средних значений признаков X и Y, то выборочная средняя признака Z равна сумме (разности) средних арифметических признаков X и Y. Так, средняя сти-

пендия студента за два месяца равна сумме средних стипендий студента за каждый из этих двух месяцев.

3) Если ко всем вариантам выборки прибавить одно и тоже число, то и выборочная средняя увеличится на это число. Например, если стипендию студента увеличить на одну и ту же сумму, то и средняя стипендия студента увеличится на ту же сумму.

4) Если все варианты выборки умножить (разделить) на одно и то же число, то выборочная средняя умножится (разделится) на это же число. Если значение стипендии каждого студента группы удвоить, то и значение средней стипендии студентов этой группы удвоится.

5) Если все частоты умножить (разделить) на одно и то же число, выборочная средняя не изменится. Например, при увеличении вдвое численности рабочих одного отдела, получающих одинаковую зарплату, значение средней зарплаты работников этого отдела останется прежним.

6) Выборочная средняя, вычисленная по данным всех элементов совокупности, равна взвешенной средней для так называемых частичных средних, т.е. средних, найденных для отдельных частей совокупности, причем частота для каждой частичной средней равна количеству элементов в соответствующей части совокупности.

Пусть выборка состоит из следующих элементов: $x_1, x_2, \ldots, x_k$, y_1, $y_2, \ldots, y_l, z_1, z_2, \ldots, z_m$, причем $k+l+m=n$. Тогда выборочная средняя этой выборки равна $k\overline{x_B}+l\overline{y_B}+m\overline{z_B}$.

Это свойство дает возможность упростить вычисление средней зарплаты, например работников трех цехов. Для этого достаточно вычислить выборочную среднюю зарплату работников по каждому цеху (групповое среднее), а затем вычислить среднюю этих частичных сумм (общее среднее), приняв в качестве их частот количество рабочих в соответствующих цехах.

Так выборочная средняя наиболее полезно в качестве обобщающего показателя при отсутствии резко выделяющихся наблюдений, или как их называют выбросов, т.е. когда набор данных представляет собой более-менее однородную группу. Эти свойства особенно полезны, когда есть необходимость планировать общую сумму для большой группы. Сначала вычисляют выборочную среднюю для меньшей выборки данных, хорошо представляющей большую группу. Затем полученное значение умножают на количество элементов в большой группе. С помощью выборочной средней решается еще одна из задач математической статистики – оценка неизвестного параметра распределения совокупности.

Пример 11

Число бракованных изделий при производстве мороженного в течение 20 дней равнялось 24, 28, 5, 10, 22, 5, 19, 0, 3, 18, 20, 4, 0, 1, 21, 17, 10, 20, 4, 7

1) Найдите выборочную среднюю дневного выпуска бракованных изделий

2) Если бы регистрация числа бракованных изделий продолжалась еще 23 дня и если бы характеристика числа бракованных изделий была такой же, как и в

остальные дни, то какое число бракованных изделий можно было бы ожидать за 23 дня?

Решение

1) С помощью рассмотренных выше свойств выборочного среднего, найдем выборочную среднюю дневного выпуска бракованных изделий (по формуле (5)). Имеем:

$$\overline{x_B} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i = \frac{24+28+5\cdot 2+19+3+18+20\cdot 2+4\cdot 2+1+21+17+10+7}{20} = 11.9$$

2) За 23 дня соответственно можно было бы ожидать $23 \cdot 11.9 \approx 274$ бракованных изделий.

Для того, чтобы вычислить выборочное среднее по данным *интервального вариационного ряда*, необходимо за значение признака для всех элементов в данном интервале взять середину интервала. При этом допускается определенная неточность, но обычно в различных интервалах погрешности будут разных знаков, а потому при большом количестве наблюдений они в значительной мере «гасят» друг друга.

Например, студент получил в течение семестра по высшей математике следующие оценки: 7, 6, 7, 7, 4, 6, 7, 7. Многие из вас, наверняка бы поставили студенту оценку 7, аргументируя это тем, что эта оценка чаще всего была получена студентом. Эта величина в статистике называется *модой* M_o.

Модой называется число, которое наиболее часто встречается в вариационном ряду, или варианта, имеющая наибольшую частоту.

В нашем примере это число 7. Если выборочное среднее можно вычислить для каждого вариационного ряда, то моды у вариационного ряда может вообще и не быть (Приведите пример).

Значение моды для интервального ряда можно получить по формуле (6)

$$M_o = x_0 + h\frac{n_2 - n_1}{(n_2 - n_1) + (n_2 - n_3)}, \qquad (6)$$

где x_0 – начальное значение модального интервала, т.е. интервала, который содержит моду;

h – длина модального интервала;

n_1 – частота интервала, предшествующего модальному,

n_2 – частота модального интервала,

n_3 – частота интервала, следующего за модальным.

Особенностью моды является то, что ее можно использовать не только в случае дискретного ряда распределения. Например, проводится опрос в ходе которого выясняется какой предмет нравиться студентам больше всего. Модой этого опроса окажется предмет, который чаще всего называют студенты. Это одна из причин по которой мода широко используется при изучении спроса и проведении других социологических исследований. Например, при решении вопросов: в пачки какого веса фасовать масло, какие открывать авиарейсы и т.д. предварительно изучается спрос и выявляется мода – наиболее часто встречающийся заказ.

При рассмотрении, например, производственных вопросов, когда набор данных представляет собой описание причин выхода из строя сложного устройст-

ва с соответствующими частотами, мода помогает сосредоточить внимание на самой важной категории. Если набор данных представляет собой описание последовательных этапов производства сложного устройства с указанием количества блоков, находящихся на разных стадиях производства, то мода указывает на стадию производства, на которой находится наибольшее количество блоков, т.е. на «узкое» место в производстве.

Тем не менее, следует осторожно относиться к использованию моды. Например, пусть в группе, к которой 22 студента, выполняется тест, состоящий из 25 заданий. На тестирование явилось 20 человек, двое не явились; все тестировавшиеся показали различные результаты. Модой является результат «не явился», т.к. это наиболее часто повторяющийся вариант. Конечно, этот результат является плохой характеристикой результатов тестирования.

Еще одной важной средней характеристикой вариационного ряда является его *медиана* M_e.

Медианой называют варианту, которая делит вариационный ряд на две части, равные по числу вариант. Если число вариант нечетное $n = 2k + 1$, то

$$M_e^* = x_{(k+1)}. \qquad (7)$$

Если число вариант четное, т.е. $n = 2k$, то необходимо взять два средних числа и найти их полусумму, т.е.

$$M_e^* = \frac{x_k + x_{k+1}}{2}. \qquad (8)$$

Пример 12

На студенческой спортивной олимпиаде проводится несколько квалификационных забегов на 100 м, из которых в финал проходят ровно половина от числа всех спортсменов. Пред вами результаты всех спортсменов. Какой результат позволяет пройти в финал?

15,5; 16,8; 21,8; 18,4; 16,2; 32,3; 19,9; 15,5; 14,7; 19,8; 20,5; 15,4.

Упорядочим данный ряд: 14,7; 15,4; 15,5; 15,5; 16,2; 16,8; 18,4; 19,8; 19,9; 20,5; 21,8; 32,3.

Выборочная средняя равна $\overline{x_B} = 18{,}9$, мода $M_o = 15{,}5$. Медиану найдем по формуле (8) $M_e = \frac{16{,}8 + 18{,}4}{2} = 17{,}6$. Для этого примера лучшей характеристикой является медиана, так как данное значение выборочной средней не позволяет пройти в финал половине спортсменов, есть спортсмен с результатом 18,4. А мода дает слишком завышенную оценку для среднего результата.

Преимуществом медианы перед выборочным средним является т.н. «устойчивость к ошибкам». Например, если бы в задании о спортивной студенческой олимпиаде вместо числа 21,8 записали бы с ошибкой число 218, то значение выборочной средней изменилось бы существенно, а вот значение медианы осталось бы прежним.

Пример 13

Вычислим медиану по данным о распределении работников цеха по тарифным разрядам, представленным в таблице 4,

Таблица 4 – Распределение работников цеха по тарифным разрядам

Тарифные разряды	1	2	3	4	5	6	Всего
Число работников	4	6	12	16	44	18	100

Очень часто встречается ошибочное вычисление медианы в такого рода задачах. Не учитывая ни частоты вариант, ни общего количества элементов в качестве медианы берут полусумму средних вариант, т.е. в нашем случае $M_e = \frac{3+4}{2} = 3{,}5$, что не верно. Для получения правильного результата составим таблицу накопленных частот (таблица 5) в скобках показано, как находятся накопленные частоты.

Таблица 5 – Таблица накопленных частот

Тарифные разряды	1	2	3	4	5	6
Накопл. частоты	4	10 (4+6)	22 (4+6+12)	38 (4+6+12+16)	82	100

Определяем 1-ую накопленную частоту, большую половины общего количества элементов. В данном случае это 82. Итак элементам с номерами 50 и 51 отвечает значение тарифного разряда равное 5, поэтому $M_e=5$. Это значение медиана означает, что примерно половина всех работников цеха имеет разряд 5 и меньше, а половина – 5 и больше.

Для интервального упорядоченного вариационного ряда медиана вычисляется по формуле (9)

$$M_e = x_n + \frac{\frac{n}{2} - S_{M_{e-1}}}{n_{M_e}} \cdot h, \qquad (9)$$

где x_n – начало медианного интервала,

h – ширина медианного интервала,

n_{M_e} – частота медианного интервала,

$S_{M_{e-1}}$ – сумма частот интервалов, предшествующих медианному,

n – объем выборки.

Однако не всегда эти характеристики дают полное представление о поведении изучаемой величины. Например, на планете Меркурий средняя температура составляет $+15^0$С, получается вполне пригодный для жизни климат. На самом же деле температура на Меркурии колеблется от -150^0С до $+350^0$С. Эти показатели температуры делают Меркурий не пригодным для жизни.

Поэтому в некоторых случаях мало значений одних средних характеристик (характеристик положения). Необходимо знать еще и *характеристики разброса или рассеяния*. К таким характеристикам относятся: выборочная дисперсия, среднее квадратическое отклонение. Дадим определения этим характеристикам в таблице 6.

Таблица 6 – Определение дисперсии и среднего квадратического отклонения

Величина, обозначение.	Определение	Формула
D_B Выборочная дисперсия	Среднее арифметическое квадратов отклонений значений выборки от выборочной средней	$D_B = \frac{1}{n}\sum_{i=1}^{k}(x_i - \bar{x})^2 \cdot n_i$ или $D_B = \sum_{i=1}^{k}(x_i - \bar{x}) \cdot w_i$, где $w_i = \frac{n_i}{n}$.
σ_B Среднее квадратическое отклонение	Квадратный корень из дисперсии	$\sigma_B = \sqrt{D_B}$

Дисперсия имеет размерность, равную квадрату размерности элементов выборочной совокупности, т.е. если значение ряда измеряется допустим в рублях, то у дисперсии будут «квадратные рубли». Чтобы иметь показатели той же размерности, что и размерность элементов данной совокупности, рассматривают среднее квадратическое отклонение. Этот показатель отражает картину отклонения отдельных значений от среднего значения совокупности. Схематически это можно изобразить следующим образом (рисунок 7)

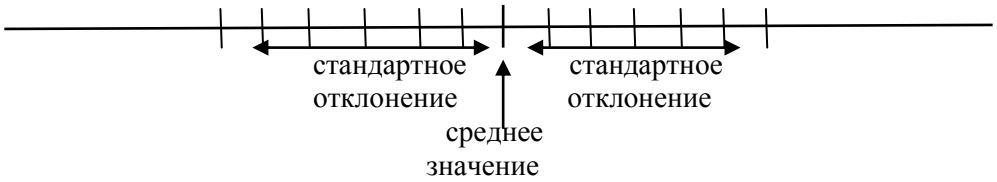


Рисунок 7 – Отклонения значений от среднего значения совокупности

Пример 14

Пусть в группе проведено тестирование студентов по высшей математике. Результаты тестирования следующие: 9, 17, 19, 18, 23, 25, 19, 13, 10, 19, 18, 19, 25, 10, 15, 16, 15, 20, 12. Балл у студента N оказался равным 17.

Требуется определить, типичен ли результат учащегося N для всего класса.

Для этого найдем средний бал класса и среднее квадратическое отклонение (по формуле (5)).

$$\overline{x_B} = \frac{9+17+19+18+23+\ldots+12}{20} = 17{,}35$$

$$D_B = \frac{(9-17{,}35)^2 + (17-17{,}35)^2 + \ldots + (12-17{,}35)^2}{20} = 24{,}45,\ \sigma_B = \sqrt{24{,}45} = 4{,}94.$$

Теперь найдем разность между баллом учащегося N и средним баллом класса: $17{,}35 - 17 = 0{,}35$. Полученное значение гораздо меньше среднего квадратического отклонения, поэтому несмотря на то, что результат учащегося меньше среднего, он является типичным для класса.

Пример 15

По данным выборки характеристической вязкости полимера (Дл/г) найти значение основных числовых характеристик.

0,76; 0,76; 0,76; 0,77; 0,76; 0,76; 0,76; 0,76; 0,77; 0,77; 0,77; 0,77;0,78; 0,75; 0,76; 0,76; 0,76; 0,76;0,77; 0,76; 0,77; 0,77; 0,77; 0,77; 0,76; 0,77; 0,77; 0,77; 0,78; 0,78; 0,78; 0,77; 0,77; 0,77; 0,78; 0,78; 0,76; 0,76; 0,77; 0,77; 0,75; 0,77; 0,77; 0,76; 0,77; 0,77; 0,77;

Для вычисления выборочного среднего с помощью электронных таблиц Excel можно использовать функцию (СРЗНАЧ). Выбирают ячейку, куда помещают среднее (в нашем примере B2). В главном меню выбирают (Вставка→Функция), затем в качестве категории функции выбирают (Статистические) и в качестве названия функции – (СРЗНАЧ). Появится диалоговое окно. Перетаскивая курсор мыши, выделяют необходимый список чисел (A1:A46), а затем нажимают клавишу Enter для завершения процесса. Аналогичным образом вычисляется мода, в Excel можно использовать функцию (МОДА). Медиана вычисляется с помощью функции (МЕДИАНА). Выборочная дисперсия вычисляется с помощью функции (ДИСП). Среднее квадратическое отклонение или стандартное отклонение вычисляется с помощью функции (СТАНДОТКЛОН).

Для нахождения всех характеристик одновременно можно воспользоваться встроенным Пакетом анализа. Для его установки необходимо в меню Сервис выбрать команду Надстройки, в появившемся списке установить флажок Пакет анализа. Для определения характеристик выборки используется процедура Описательная статистика, для выполнения которой необходимо выполнить команду Сервис→Анализ данных, в появившемся списке Инструменты анализа выбрать строку Описательная статистика и нажать кнопку ОК (см. пример) Далее, в появившемся диалоговом окне указать входной диапазон, в нашем случае – столбец данных. В качестве выходного диапазона выбрать любую пустую, удобную ячейку документа (C1). В разделе Группировка установить в положение по столбцам. Установить флажок в поле Итоговая статистика. Нажать кнопку ОК (в результате появится Столбец 1 в ячейке C1) (рисунок 8). Помимо рассмотренных ранее числовых характеристик в данном столбце содержатся такие характеристики, как

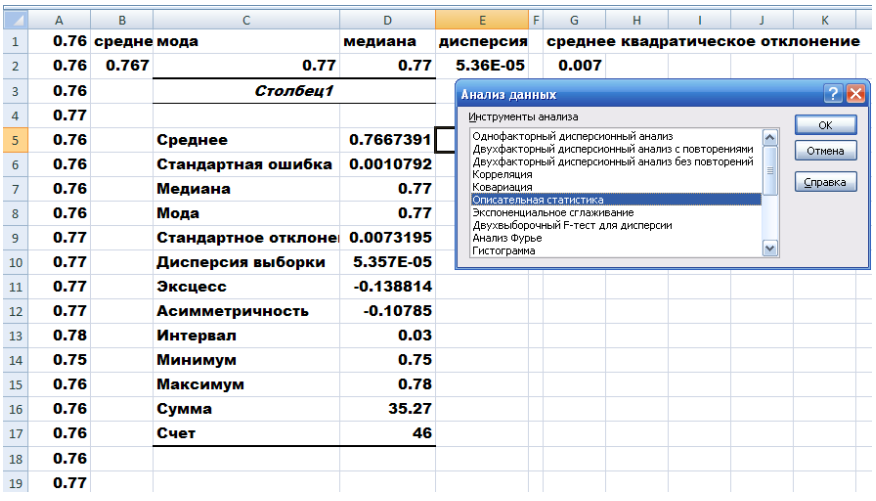
○ Эксцесс – это степень выраженности «хвостов» распределения, т.е. частоты появления удаленных от среднего значения;

○ Асимметрия – величина, характеризующая несимметричность распределения элементов выборки относительно среднего значения. Принимает значения от -1 до 1. В случае симметричного распределения асимметрия равна 0;

○ Интервал (амплитуда, вариационный размах) – это разница между максимальным и минимальным значениями элементов выборки.

○ Минимум, максимум – значение минимального и максимального элемента выборки соответственно;

○ Счет – количество элементов в выборке, т.е. её объем;



	A	B	C	D	E	F	G	H	I	J	K
1	0.76	средне	мода		медиана	дисперсия		среднее квадратическое отклонение			
2	0.76	0.767		0.77	0.77	5.36E-05	0.007				
3	0.76		*Столбец1*								
4	0.77										
5	0.76		Среднее		0.7667391						
6	0.76		Стандартная ошибка		0.0010792						
7	0.76		Медиана		0.77						
8	0.76		Мода		0.77						
9	0.77		Стандартное отклоне		0.0073195						
10	0.77		Дисперсия выборки		5.357E-05						
11	0.77		Эксцесс		-0.138814						
12	0.77		Асимметричность		-0.10785						
13	0.78		Интервал		0.03						
14	0.75		Минимум		0.75						
15	0.76		Максимум		0.78						
16	0.76		Сумма		35.27						
17	0.76		Счет		46						
18	0.76										
19	0.77										

Рисунок 8 – Нахождение основных числовых характеристик

○ Минимум, максимум – значение минимального и максимального элемента выборки соответственно;

○ Счет – количество элементов в выборке, т.е. её объем;

Вопросы для самопроверки

1) Достаточно ли знать среднее количество бракованных изделий, выпускаемых рабочим в каждом из трех цехов предприятия, чтобы найти среднее число

бракованных изделий, выпускаемых рабочим на всем предприятии, если на этом предприятии три цеха?

2) Как, имея данные о количестве заказов, поступающих в две мастерские ежедневно в течение месяца, сравнить загруженность этих мастерских работой?

3) Требуется выяснить потребность населения некоторого города в определенном товаре. Как может помочь выборочное среднее в решении этой проблемы?

4) Есть данные о спаде производства на некотором предприятии на протяжении 5 лет: 20, 30, 25, 20, 15%. Какая из статистических характеристик наилучшим образом описывает средний ежегодный спад производства?

5) Как изменится мода совокупности, если ко всем ее значениям прибавить одно и то же число, все ее значения умножить на одно и тоже число?

6) Как изменится медиана совокупности, если все ее значения умножить на одно и то же число?

Точечные оценки параметров распределения СВ по данным выборки

Следующей задачей математической статистики является оценивание неизвестных параметров генеральной совокупности. Например, следует оценить число изделий, которое необходимо дополнительно выпустить предприятию для замены вышедших из строя; денежные затраты населения на определенный вид продукции и т.д.

Параметр генеральной совокупности или просто параметр – это показатель, вычисленный для всей генеральной совокупности (например, математическое ожидание, дисперсия и т.д. генеральной совокупности) Параметр является фиксированным, но зачастую неизвестным числом, т.к. по многим причинам исследовать всю генеральную совокупность не представляется возможным.

Пусть изучается количественный признак X генеральной совокупности с законом распределения, зависящим от одного или нескольких параметров $\theta_1, \theta_2, \ldots, \theta_n$. Часть этих параметров может быть неизвестна. *Задача математической статистики* – получить оценки неизвестных параметров распределения. По выборке $x_1, x_2, \ldots, x_n$, полученной в результате наблюдений, необходимо оценить неизвестный параметр θ.

В результате приходится анализировать данные выборки.

▌ Функцию результатов наблюдений (т.е. выборочную функцию) называют **статистикой,** (например, выборочная средняя, медиана, выборочная дисперсия и т.д.).

▌ **Статистической оценкой** $\tilde{\theta}$ параметра θ теоретического распределения (генеральной совокупности) называют его приближенное значение, зависящее от наблюдаемых значений признака.

Оценка $\tilde{\theta}$ есть значение некоторой функции результатов наблюдений $\tilde{\theta} = \tilde{\theta}(x_1, x_2, \ldots, x_n)$.

▌ Статистика, которая используется в качестве приближенного значения неизвестного параметра генеральной совокупности называется **точечной оценкой.**

К оценке любого параметра предъявляют ряд требований, которым она должна удовлетворять, чтобы быть приближенной, «близкой» к истинному значению параметра и иметь практическую ценность.

‖ Оценочную функцию некоторого параметра генеральной совокупности называют **несмещенной**, если ее математическое ожидание равно этому параметру, т.е. $M(\tilde{\theta}) = \theta$, в противном случае ее называют **смещенной**, т.е. $M(\tilde{\theta}) \neq \theta$.

Требование несмещенности особенно важно при малом числе наблюдений. Несмещенность оценки означает, что использование такой оценки не приведет к систематической ошибке.

Среди несмещенных оценок одного и того же параметра выделяют эффективные оценки.

‖ Несмещенная оценка $\tilde{\theta}_n$ параметра θ называется **эффективной**, если она имеет (при заданном объеме выборки) наименьшую возможную дисперсию. при $n \to \infty$.

При рассмотрении выборок большого объема (n велико) к статистическим оценкам предъявляют требование состоятельности.

‖ Если оценка параметра при $n \to \infty$ сходится по вероятности к оцениваемому параметру, то ее называют **состоятельной**, т.е $\forall \varepsilon > 0$ выполняется

$$\lim_{n\to\infty} p\,(\tilde{\theta}_n) \xrightarrow{\text{по вероят}} = \theta.$$

‖ Несмещенная оценка $\tilde{\theta}_n$ параметра θ называется **эффективной**, если она имеет наименьшую дисперсию.

Пусть $x_1, x_2, x_3, \ldots, x_n$ - выборка, полученная в результате проведения n независимых наблюдений за СВ X. Будем рассматривать эти значения как независимые, одинаково распределенные СВ $X_1, X_2, X_3, \ldots, X_n$, которые имеют одинаковые числовые характеристики, т.е. $M(X_1) = M(X_2) = \ldots = M(X_n) = a$, $D(X_1) = D(X_2) = \ldots = D(X)$, тогда:

1) *Выборочное среднее* $\overline{x_B} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i$ – несмещенная и состоятельная оценка математического ожидания $M(X)$.

2) *Исправленная выборочная дисперсия* $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2 = \frac{n}{n-1}D_B$ – несмещенная и состоятельная оценка дисперсии.

Рассмотрим биномиальное распределение. Пусть p - вероятность «успеха» в каждом испытании Бернулли. Проводят n испытаний Бернулли, пусть m - число «успехов» в них. Тогда в качестве оценочной функции параметра p генеральной совокупности можно принять частоту «успеха» $\bar{p} = \frac{m}{n}$. Таким образом, оценочной функцией для вероятности p является случайная величина $\bar{p} = \frac{m}{n}$, а оценкой неизвестной вероятности – ее значение при некотором значении m. Например, при производстве 20 деталей, бракованных оказалось 2, поэтому вероят-

ность производства бракованных деталей оказалась 2/20=0,1. А оценочной функцией является частота. В приведенном примере она равна 0,1, в другой раз число бракованных изделий может составить 1, тогда оценкой неизвестной вероятности служит значение 0,05. А оценочная функция одна и та же – это частота выпуска бракованных изделий.

3) *Частота* $\bar{p} = \frac{m}{n}$ появления события A в n независимых испытаниях является несмещенной, состоятельной и эффективной оценкой неизвестной вероятности p этого события.

Вопросы для самопроверки

1) В чем суть задачи оценивания неизвестных параметров генеральной совокупности?

2) Стандартное отклонение выборки равно 8,5. Является ли это число оценочной функцией или оценкой среднего квадратическое отклонения генеральной совокупности?

3) Какие дополнительные соображения, кроме несмещенности оценки, влияют на выбор метода оценивания?

Доверительная вероятность. Доверительные интервалы

Оценочная функция есть случайная величина, имеющая некоторый разброс около истинного значения параметра, а поэтому, принимая истинное значение параметра равным числовому значению оценочной функции или оценке, мы допускаем определенную ошибку. Другими словами, построить оценку неизвестного параметра по результатам наблюдений – значит, найти «хорошее» приближенное значение этого неизвестного параметра. С приближенными вычислениями и понятиями абсолютной погрешности вы уже неоднократно сталкивались. Поэтому говоря о приближениях или пользуясь приближенными значениями, надо ясно себе представлять и точность приближения и границы абсолютной погрешности. Например, 1 м может считаться приближенным значением и для длины 910 мм, и для 1007 мм, и для 993,3 мм. Границы абсолютной погрешности составят соответственно 100 мм, 10 мм, 1 мм. Без указания, с какой точностью взяты приближенные значения, сами по себе они практически не имеют смысла.

Эта общая идея находит свое применение и в статистике. Точечные оценки параметров распределения, рассмотренные выше при выборке малого объема могут привести к грубым (значительным) ошибкам. По этой причине при небольшом объеме выборки пользуются *интервальными оценками*.

Задачу интервального оценивания можно сформулировать так: по данным выборки построить числовой интервал $(\tilde{\theta}_1, \tilde{\theta}_2)$ относительно которого с заранее выбранной вероятностью γ можно сказать, что этот интервал накрывает точное значение оцениваемого параметра.

‖ Интервал $(\tilde{\theta}_1, \tilde{\theta}_2)$ накрывающий с вероятностью γ истинное значение параметра θ, называется **доверительным интервалом**, а вероятность γ - **надежностью оценки** или **доверительной вероятностью** (рисунок 9).

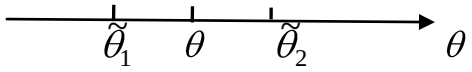


Рисунок 9 – Доверительный интервал

Интервал вида $(\tilde{\theta}-\varepsilon,\ \tilde{\theta}+\varepsilon)$, такой, что $p\{\theta \in (\tilde{\theta}-\varepsilon,\ \tilde{\theta}+\varepsilon)\} = p\{|\theta-\tilde{\theta}|<\varepsilon\} = \gamma$, где число $\varepsilon>0$ характеризует точность оценки: чем меньше разность $|\theta-\tilde{\theta}|$, тем точнее оценка. Интервал $(\tilde{\theta}-\varepsilon,\ \tilde{\theta}+\varepsilon)$ имеет случайные концы, их называют доверительными границами. Доверительные границы будут различными для различных выборок.

Обычно величина γ задается заранее и зависит от конкретно решаемой задачи. Так, степень доверия авиапассажира к надежности самолета, очевидно, должна быть выше степени доверия покупателя к надежности телевизора, лампочки, игрушки и т.д. Надежность принято выбирать равной 0,9; 0,95; 0,99 или 0,999. Тогда практически достоверно что доверительный интервал $(\tilde{\theta}-\varepsilon,\ \tilde{\theta}+\varepsilon)$ накроет параметр θ .

Вопросы для самопроверки

1) Какую дополнительную информацию о генеральной совокупности дает доверительный интервал по сравнению с точечной оценкой параметра?

2) Представьте, что вы строите доверительный интервал для неизвестной вероятности выпуска бракованных парашютов. Удовлетворяет ли вас доверительная вероятность, равная 0,99?

Построение доверительных интервалов для оценки математического ожидания нормального распределения при известном и неизвестном среднем квадратическом отклонении.

Рассмотрим построение доверительного интервала для оценки математического ожидания нормального распределения при *известном среднем квадратическом отклонении.*

Пусть количественный признак X генеральной совокупности имеет нормальное распределение с известным σ и неизвестным математическим ожиданием a. Так как выборочная средняя $\overline{x_B}$ меняется от выборки к выборке, его можно рассматривать как СВ $\overline{X_B}$. Выборочные значения признака $x_1,\ x_2,\ x_3,\ldots,x_n$ также меняются от выборки к выборке. Будем рассматривать их, как одинаково распределенные СВ $X_1,\ X_2,\ X_3,\ldots,X_n$ ($M(X_1)=M(X_2)=\ldots=M(X_n)=a$, $D(X_1)=D(X_2)=\ldots=D(X)=\sigma$).

Пусть для СВ X σ – известно , доверительная вероятность γ (надежность) задана. Тогда доверительный интервал для $a=M(X)$ находят по формуле

$$\left(\overline{X}-t\cdot\frac{\sigma}{\sqrt{n}},\ \overline{X}+t\cdot\frac{\sigma}{\sqrt{n}}\right), \qquad (10)$$

где t определяется из уравнения $\Phi_0(t)=\frac{\gamma}{2}$ или $\Phi_0(t)=\frac{1+\gamma}{2}$, (при заданной γ по таблице функции Лапласа находим аргумент t).

Пример 16

Произведено 5 независимых наблюдений над СВ X. Результаты наблюдений таковы: $x_1=-25,\ x_2=34,\ x_3=-20,\ x_4=10,\ x_5=21,\ \sigma=20$. Найти точечную оценку для $a=M(X)$, а также построить для него доверительный интервал с надежностью 0,95 (т.е. 95%-й доверительный интервал).

Найдем $\bar{x}_B$: $\bar{x}_B=\frac{-25+34-20+10+21}{5}=4$, т.е. $\bar{x}_B=4$. Учитывая, что по условию $\gamma=0{,}95$ и $\Phi_0(t)=\frac{\gamma}{2}$, получаем $\Phi_0(t)=0{,}475$. По таблице значений функций $\Phi_0(t)$ выясняем, что $t=1{,}96$. Так как σ – известно, то для нахождения доверительного интервала воспользуемся формулой (10).

Тогда $t\cdot\frac{\sigma}{\sqrt{n}}=\frac{1{,}96\cdot 20}{\sqrt{5}}\approx 17{,}5$. Значит доверительный интервал для $a=M(X)$ таков: $(4-17{,}5;\ 4+17{,}5)$, т.е. $(-13{,}5;\ 21{,}5)$

Теперь рассмотрим как строится доверительный интервал для оценки математического ожидания нормального распределения при *неизвестном среднем квадратическом отклонении.*

Пусть σ – неизвестна, доверительная вероятность γ (надежность) задана. Тогда доверительный интервал для $a=M(X)$ находят по формуле

$$\left(\overline{X}-t_\gamma\cdot\frac{s}{\sqrt{n}},\ \overline{X}+t_\gamma\cdot\frac{s}{\sqrt{n}}\right), \qquad (11)$$

где t_γ определяется по формуле $2\cdot\int_0^{t_\gamma} f_T(t,n-1)\,dt=\gamma$,

s – исправленное среднее квадратическое отклонение СВ X $s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}$.

На практике величину t_γ находим пользуясь таблицей распределения Стьюдента в зависимости от заданной доверительной вероятности γ и числа степеней свободы $n-1$ (t_γ – квантиль уровня $1-\gamma$);

Пример 17

Рассмотрим предыдущий пример, считая, что σ – неизвестно. Построим 95%-й доверительный интервал для $a=M(X)$, воспользовавшись формулой (11).

Оценку $\bar{x}_B=4$ мы уже знаем. Теперь найдем $s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}=\sqrt{\frac{(-25-4)^2+(34-4)^2+(-20-4)^2+(10-4)^2+(21-4)^2}{4}}\approx 25{,}7.$

По таблице значений $t_\gamma = t(\gamma, n)$ по заданным $\gamma = 0{,}95$ и $n-1=4$ находим $t_\gamma = 2{,}78$. Следовательно, величина $t_\gamma \cdot \frac{s}{\sqrt{n}} = 2{,}78\frac{25{,}4}{2{,}24} \approx 31{,}9$. Итак, с надежностью $\gamma = 0{,}95$ параметр a заключен в доверительный интервал: $(-27{,}9;\ 35{,}9)$.

В Excel для более точного вычисления границ доверительного интервала и при числе элементов в выборке меньше 30 можно воспользоваться функцией (ДОВЕРИТ) или процедурой Описательная статистика. Функция (ДОВЕРИТ) (*альфа; станд_отклон; размер*) определяет полуширину доверительного интервала и содержит следующие параметры: *альфа* – уровень значимости, используемый для вычисления доверительной вероятности, скажем при 95%-ом доверительном интервале *альфа* равно 0,5 т.е. доверительная вероятность равна 100*(1-*альфа*)%; *станд_отклон* – стандартное отклонение для интервала данных, предполагается известным; *размер* – это объем выборки.

Пример 18

Найти границы 95%-го доверительного интервала для среднего значения, если у 25 аккумуляторов среднее время разряда составило 140 часов, а стандартное отклонение – 2,5 часа.

Для определения доверительно интервала воспользуемся рассмотренной нами ранее функцией ДОВЕРИТ. Для этого на панели инструментов выбираем последовательно Стандартная→Вставка функции→Мастер функций→Статистические→ДОВЕРИТ и нажимаем ОК (рисунок 9). В рабочее поле появившегося диалогового окна ДОВЕРИТ с клавиатуры вводим условия задачи: Альфа – 0,05; Станд_отклон – 2,5; Размер – 25. В установленной курсором ячейке появилось значение - 0,979981. Другими словами, с 95%-ым уровнем надежности можно утверждать, что средняя продолжительность разряда аккумулятора составляет (140+0,979981; 140-0,979981) или от 139,02 и до 140,98 часа.

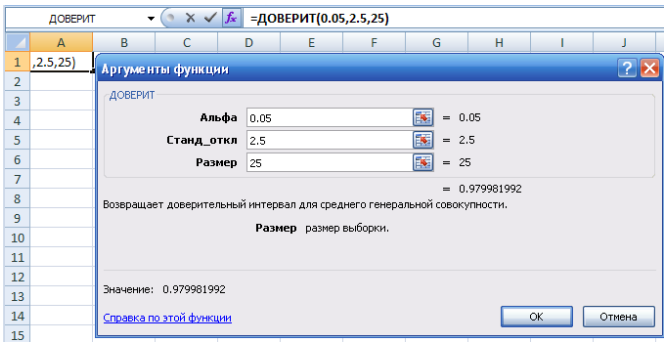


Рисунок 9 – Построение доверительного интервала

Построение доверительного интервала для среднего квадратического отклонения нормального распределения

Пусть σ – неизвестна, доверительная вероятность γ (надежность) задана. Если $a = M(X)$ – *известно*, то доверительный интервал для среднего квадратического отклонения σ имеет вид:

$$\left(\frac{\sqrt{n}\cdot S_0}{\chi_2}, \frac{\sqrt{n}\cdot S_0}{\chi_1}\right), \qquad (12)$$

где n – объем выборки,

$$S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2,$$

$\chi_2^2 = \chi^2_{\frac{1-\gamma}{2};n}$, $\chi_1^2 = \chi^2_{\frac{1+\gamma}{2};n}$ – определяются по таблице квантилей $\chi^2_{\alpha,n}$ распределения χ^2_n.

Если $a = M(X)$ – *неизвестно*, то доверительный интервал для σ имеет вид

$$\left(\frac{\sqrt{n-1} \cdot s}{\chi_2}, \frac{\sqrt{n-1} \cdot s}{\chi_1} \right), \qquad (13)$$

где s – исправленное среднее квадратическое отклонение СВ X,

χ_2, χ_1 – определяются по таблице $\chi^2_{\alpha,n}$ при $k = n-1$ и $\alpha = \frac{1+\gamma}{2}$ и $\alpha = \frac{1-\gamma}{2}$ соответственно.

Пример 19.

Чтобы оценить среднее квадратическое отклонение нормально распределенной СВ *X*, была сделана выборка объема в 30 единиц. Найдено значение $s = 1{,}5$. Необходимо найти доверительный интервал, покрывающий σ с вероятностью $\gamma = 0{,}90$. В нашем случае величина $a = M(X)$ неизвестна. Построим соответствующий доверительный интервал.

Для этого, по заданному $n = 30$ и $\gamma = 0{,}90$, по таблице $\chi^2_{\alpha,n}$, находим $\chi_1^2 = \chi^2_{\frac{1+0.9}{2};30-1} = \chi^2(0{,}95;29) = 17{,}7$. $\chi_2^2 = \chi^2_{\frac{1-0.9}{2};30-1} = \chi^2(0{,}05;29) = 42{,}6$. Тогда доверительный интервал имеет вид: $\left(\frac{\sqrt{30-1} \cdot 1{,}5}{\sqrt{42{,}6}}, \frac{\sqrt{30-1} \cdot 1{,}5}{\sqrt{17{,}7}} \right)$, согласно формуле (13), или $1{,}238 < \sigma < 1{,}920$.

Статистическая проверка статистических гипотез

Следующей задачей математической статистики является задача статистической проверки статистических гипотез.

Человеку часто приходится принимать то или иное решение. В большинстве принимаемых решений содержится элемент риска. Однако, во многих случаях степень риска может быть снижена. Например, принятию решения о переходе на новую технологию производства какого-либо изделия должна предшествовать экспериментальная проверка этой технологии, сбор необходимой информации, ее обработка, проверка того, говорят ли собранные данные в пользу новой технологии.

Статистической называют гипотезу о виде неизвестного распределения или о параметрах известных распределений. Например, группа исследователей в области животноводства разработала новый вид кормов, полагая, что эти корма способны повысить жирность молока. Однако может быть и так, что они ошибаются. Как быть? В статистике поступают так. Выдвигают *нулевую гипотезу* H_0, которая состоит в том, что при использовании данного корма средняя жирность молока остается неизменной. Наряду с этим рассматривают гипотезу, состоящую в том,

что данный вид кормов эффективен (*альтернативная гипотеза* H_1). Предполагается, что закон распределения СВ X – жирность молока известен.

Процедура проверки статистических гипотез в определенной степени напоминает проведение обоснований математических утверждений: чтобы опровергнуть утверждение, достаточно привести пример, подтверждающий, что оно не имеет места. Тем не менее, в статистике не говорят, что нулевая гипотеза верна или неверна, а говорят осторожнее: нулевая гипотеза не опровергается, т.к. полученные данные не противоречат ей.

Следует понимать, что есть определенный риск и в том случае, когда нулевая гипотеза отвергается (она может оказаться верной) и в том случае, когда нулевая гипотеза не отвергается (она может оказаться неверной).

Если отвергается нулевая гипотеза в случае, когда она верна, то говорят, что совершается ошибка I рода, ее вероятность обозначается α и называется *уровнем значимости*. Если же нулевая гипотеза не отвергается в то время, когда она неверна, то совершается ошибка II рода, её вероятность обозначается β. В нашем случае: ошибка первого рода – на самом деле жирность молока не изменилась, но жирность молока, полученная по выборке, значительно отличается от прежней; ошибка второго рода – на самом деле жирность молока изменилась, но жирность молока, полученная по выборке, незначительно отличается от прежней.

В общем случае, в таблице 7 представлены верные решения и типы ошибок при проверке статистических гипотез:

Таблица 7 – Типы ошибок при проверке гипотез

Статистическое решение	**Фактическая оценка нулевой гипотезы**	
	Верна	**Неверна**
Не отвергать нулевую гипотезу	Правильное решение, его вероятность равна $1-\alpha$.	Ошибка второго рода, её вероятность равна β.
Отвергнуть нулевую гипотезу	Ошибка первого рода, её вероятность равна α – уровень значимости.	Правильное решение, его вероятность равна $1-\beta$ – мощность критерия.

Так, для нашего случая данная таблица будет иметь вид, представленный в таблице 8

Таблица 8 – Проверка гипотезы о качестве новых кормов

Принятое решение	**Истинное положение**	
	Жирность молока не изменилась	**Жирность молока изменилась**
Жирность молока не изменилась	*Правильное решение*	*Ошибка второго рода* Принято решение о неэффективности нового рациона – возможен возврат к прежнему рациону
Жирность молока изменилась	*Ошибка первого рода* Принято решение об эффективности нового рациона кормления – будет повсеместно введен новый рацион,	*Правильное решение*

	но он не оправдает надежд	

Последствия ошибок первого и второго рода могут быть совершенно разными. В частности, в книгах по контролю качества продукции, считается, что α – риск «производителя», т.е. забраковка всей партии изделий, удовлетворяющих стандарту. А β – риск «потребителя», т.е. прием по выборке всей партии товаров, не удовлетворяющих стандарту.

Принципы проверки статистических гипотез

Обобщая, изложенное выше, рассмотрим одну из схем последовательности рассуждений, которыми пользуются в статистике:

1) *Формулируем нулевую гипотезу*. Нулевую гипотезу обозначают H_0. Она представляет собой такое утверждение, которое принимается тогда, когда нет убедительных аргументов для его отклонения. Альтернативную гипотезу обозначают H_1. Ей отдают предпочтение только тогда, когда есть убедительное статистическое доказательство, которое отвергает приемлемость нулевой гипотезы.

2) *Получают статистические (результаты наблюдений) данные о событиях, относительно которых была сформулирована нулевая гипотеза.*

3) *Определяют вероятность того, что полученный результат мог быть получен при условии, что нулевая гипотеза верна.* Результаты наблюдений зависят от случая. Поэтому статистические гипотезы носят не категорический характер, а характер правдоподобного утверждения, которое имеет вполне определенную вероятность.

4) *Если вероятность получения данного результата при условии, что нулевая гипотеза верна, мала, нулевую гипотезу отвергают на уровне значимости, равном этой вероятности.* Если вероятность получения данного результата мала, то он практически невозможен, т.е. в однократном эксперименте практически не может быть получен. Другими словами, соответствующее событие практически не должно произойти, если верна нулевая гипотеза. Если же в конкретном эксперименте событие произошло, то нулевая гипотеза неверна, и ее отвергают на определенном уровне значимости. В этом и состоит сущность ***принципа практической уверенности.***

5) *Признают, что и в том случае, когда отвергают и когда не отвергают нулевую гипотезу, возможен определенный риск.*

Иногда используют другой порядок проверки статистических гипотез. А именно:

1) Формулируют нулевую и альтернативную гипотезы H_0 и H_1.

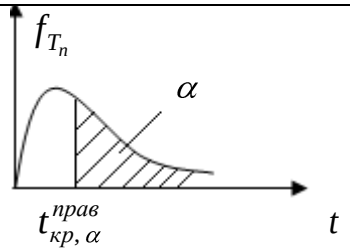
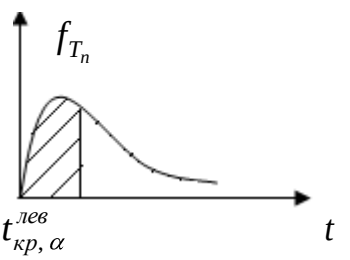
2) Проверку гипотезы осуществляют на основании результатов выборки $X_1, X_2,\ldots,X_n$, из которых формируют функцию выборки $T_n = T(X_1, X_2,\ldots,X_n)$, называемую *статистикой критерия*. Статистическим критерием (или просто критерием) называют случайную величину, которая служит для проверки нулевой гипотезы. В каждом конкретном случае она, как правило, может быть выбрана из следующих: υ – нормальное распределение, χ^2 – хиквадрат распределение (Пирсона), t – распределение Стьюдента и т.д.

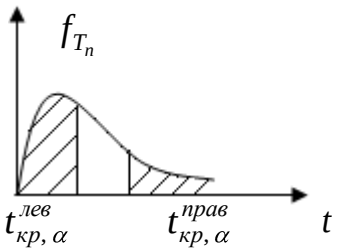
3) Назначают *уровень значимости*, т.е. число α, равное вероятности наступления события S при условии, что верна нулевая гипотеза. Это вероятность ошибки первого рода. Это число должно быть малым, если мы не хотим отвергнуть нулевую гипотезу безосновательно. Значение α выбирают исходя из того риска, на который согласен пойти исследователь, отвергнув верную гипотезу. В одних случаях считается возможным пренебречь событиями, вероятность которых меньше 0,05, т.е. в среднем в 5 из 100 случаев испытаний верная гипотеза будет отвергнута, в других случаях, когда речь идет, например, о гибели судна, разрушении сооружений, нельзя пренебречь обстоятельствами, которые могут появиться с вероятностью, равной 0,001. Обычно для α используют стандартные значения: α = 0,05; 0,01; 0,005; 0,001.

Вероятность ошибки второго рода обозначается через β. Величину $1-\beta$, т.е. вероятность недопущения ошибки второго рода называют *мощностью критерия*. Чем больше мощность критерия, тем вероятность ошибки второго рода меньше, что весьма желательно наряду с уменьшением α. При заданном объеме выборки невозможно одновременно уменьшить α и β: если уменьшать α, то растет β. Например, если принять $\alpha = 0$, то будут приниматься все гипотезы, в том числе и неверные, т.е. возрастет β. Одновременное уменьшение ошибок первого и второго рода возможно лишь при увеличении объема выборки. На практике при заданном уровне значимости α подбирают критерий с наибольшей мощностью.

4) По статистике критерия T_n и уровню значимости α определяют критическую область S, т.е. область отклонения гипотезы H_0, и $\overline{S}$ – область принятия этой гипотезы. Для ее отыскания достаточно найти критическую точку $t_{кр}$, т.е. границу, отделяющую области S и $\overline{S}$. Границы областей определяются, соответственно по таблице 9:

Таблица 9 – Критические области

Критическая область	Формула	Графическая интерпретация
Правосторонняя	$P(T_n > t_{кр}) = \alpha$	
Левосторонняя	$P(T_n < t_{кр}) = \alpha$	

Двусторонняя	$P(T_n < t^{л}_{кр}) = P(T_n > t^{п}_{кр}) = \frac{\alpha}{2}$	

Вопросы для самопроверки

1) В чем заключается цель проверки статистических гипотез?

2) В каких случаях может быть принято правильное решение?

3) Может ли вероятность ошибки первого рода равняться нулю?

4) Что показывает уровень значимости?

5) Применение новой технологии производства некоторого продукта в течение нескольких дней привело к увеличению объема выпускаемой продукции. Является ли это достаточным основанием для внедрения новой технологии?

Критерии согласия Пирсона, Колмогорова

Критерием согласия называют статистический критерий проверки гипотезы о предполагаемом законе неизвестного распределения. Он используется для проверки согласия предполагаемого вида распределения с опытными данными на основании выборки.

Имеется несколько критериев согласия: χ^2 – «хи квадрат» (Пирсона), Колмогорова, Фишера, Смирнова и др.

Критерий согласия Пирсона – наиболее часто употребляемый критерий для проверки гипотезы о нормальном законе распределения генеральной совокупности. Изложим описание этого критерия сначала для случая дискретной СВ X.

Допустим, что проведено n независимых опытов, в каждом из которых СВ X приняла значение x_i, $i = \overline{1, k}$. Результаты эксперимента помещены в таблицу 10

Таблица 10 – Полученный ряд распределения

x_i	x_1	x_2	…	x_k
w_i	w_1	w_2	…	w_k

Здесь $w_i = \frac{n_i}{n}$ – частота события $X = x_i$,

n_i – число опытов, в которых появилось это событие.

Мы выдвигаем гипотезу о том, что СВ X имеет ряд распределения представленный в таблице 11:

Таблица 11 – Предполагаемый ряд распределения

x_i	x_1	x_2	…	x_k
p_i	p_1	p_2	…	p_k

А отклонения частот w_i от вероятностей p_i объясняются случайными причинами. Чтобы проверить правдоподобность этой гипотезы, надо выбрать какую-то меру расхождения статистического распределения теоретическим. Этой мерой расхождения является величина

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}, \qquad (14)$$

которая, как доказал Пирсон, при достаточно большом объеме выборки ($n \to \infty$) имеет закон распределения, приближающийся к χ^2, независимо от того, какому закону подчинена генеральная совокупность.

Проверка гипотезы о нормальном распределении сводится к следующему алгоритму:

1) по формуле (14) вычисляем $\chi^2_{набл}$ - выборочное значение статистики критерия;

2) находим число степеней свободы $\kappa = l - 3$ (l – число различных значений вариант);

3) выбрав уровень значимости α критерия по таблице χ^2 – распределения, находим критическую точку $\chi^2_{\alpha,k}$.

4) если $\chi^2_{набл} \le \chi^2_{\alpha,k}$, то гипотеза H_0 не противоречит опытным данным; если $\chi^2_{набл} > \chi^2_{\alpha,k}$, то гипотеза отвергается.

Пусть *вариационный ряд непрерывен*, тогда проверка гипотезы о нормальном распределении сводится к следующему алгоритму:

1) вычисляем выборочную среднюю x_B и выборочное среднее квадратичное отклонение σ_B, причем вместо вариант x_i берем среднее арифметическое концов интервала $x_i^* = \frac{x_i + x_{i+1}}{2}$;

2) нормируем СВ X, т.е переходим к новой СВ Y: $y = \frac{x_i - x_B}{\sigma_B}$;

3) вычисляем теоретические вероятности попадания в интервалы $(y_i; y_{i+1})$

$$P(y_i \le y \le y_{i+1}) = \Phi(y_{i+1}) - \Phi(y_i),$$

где $\Phi(y)$ – функция Лапласа;

4) по формуле (14) вычисляем $\chi^2_{набл}$ – выборочное значение статистики критерия;

5) находим число степеней свободы $\kappa = l - 3$ (l – число интервалов выборки. Далее, согласно п.3 п.4 предыдущего алгоритма необходимо проверить гипотезу H_0.

Пример 20

Измерены 100 преформ для изготовления пластиковых бутылок; отклонения от заданного размера приведены в таблице 12:

Таблица 12 – Отклонение преформ от заданного размера

$[x_i;\ x_{i+1})$	[-3;-2)	[-2;-1)	[-1;0)	[0;1)	[1;2)	[2;3)	[3;4)	[4;5)
n_i	3	10	15	24	25	13	7	3

Необходимо при заданном уровне значимости $\alpha = 0{,}01$ проверить гипотезу о том, что отклонение проектного размера подчиняется нормальному закону распределения. Заметим, что интервалы должны содержать не менее 5-8 вариант. Интервалы с меньшим количеством вариант следует объединить.

Число наблюдений в крайних интервалах меньше 5. Поэтому объединим их с соседними. Получим ряд распределения, представленный в таблице 13

Таблица 13 – Ряд распределения

$[x_i;\ x_{i+1})$	[-3;-1)	[-1;0)	[0;1)	[1;2)	[2;3)	[3;5)
n_i	13	15	24	25	13	10

Пусть СВ X – отклонение от заданного размера. Для вычисления вероятностей p_i необходимо вычислить параметры, определяющие нормальный закон распределения (a и σ). Их точечные оценки вычислим по выборке:

$$\bar{x}_B = \frac{(-2\cdot 13 + (-0{,}5)\cdot 15 + \ldots + 4\cdot 10)}{100} \approx 0{,}885 \approx 0{,}9,$$

$$D_B = \frac{((-2)^2\cdot 13 + ((-0{,}5)^2\cdot 15 + \ldots + 4^2\cdot 10)}{100} - (0{,}885)^2 \approx 2{,}809,\ \sigma = \sqrt{2{,}809} \approx 1{,}676 \approx 1{,}7$$

Так как СВ X определена на $(-\infty;\ +\infty)$, то крайние интервалы заменяем на $(-\infty;\ -1)$ и $(3;\ +\infty)$. Тогда

$$p_1 = p\{-\infty < X < -1\} = \Phi_0\left(\frac{-1-0{,}9}{1{,}7}\right) - \Phi_0(-\infty) = \frac{1}{2} - \Phi_0(1{,}12) = 0{,}1314\ldots$$

$$p_6 = p\{3 < X < +\infty\} = \Phi_0(+\infty) - \Phi_0\left(\frac{3-0{,}9}{1{,}7}\right) = \frac{1}{2} - \Phi_0(1{,}24) = 0{,}1075.$$

Полученные данные представим в таблице 14.

Таблица 14 – Результаты вычислений

$[x_i;\ x_{i+1})$	[-∞; -1)	[-1; 0)	[0; 1)	[1; 2)	[2; 3)	[3; +∞)
n_i	13	15	24	25	13	10
p_i	0,13	0,17	0,23	0,22	0,15	0,11
$n\cdot p_i$	13,14	16,67	22,58	21,83	15,03	10,75

Далее, вычислим по формуле (14)

$\chi^2_{набл}$:

$$\chi^2_{набл} = \sum_{i=1}^{k}\frac{(n_i - np_i)^2}{np_i} = \frac{(13-13{,}14)^2}{13{,}14} + \frac{(15-16{,}67)^2}{16{,}67} + \ldots + \frac{(10-10{,}75)^2}{10{,}75} = 1{,}045.$$

Находим число степеней свободы, $\kappa = l - 3 = 6 - 3 = 3$. По условию $\alpha = 0{,}01$ и $\kappa = 3$, по таблице χ^2 – распределения находим $\chi^2_{\alpha,k} = 11{,}3$. Получили, что $\chi^2_{набл} < \chi^2_{\alpha,k}$. Следовательно, у нас нет оснований отвергнуть проверяемую гипотезу и данные не противоречат нормальному распределению.

В Excel критерий Пирсона реализован в функции ХИ2ТЕСТ, которая вычисляет вероятность совпадения наблюдаемых (фактических) значений и теоретических (гипотетических) значений. Если вычисленная вероятность ниже уровня значимости (0,05), то нулевая гипотеза отвергается и утверждается, что наблюдаемые значения не соответствуют нормальному закону распределения. Если вычисленная вероятность близка к 1, то можно говорить о высокой степени соответствия экспериментальных данных нормальному закону распределения. Функция имеет следующие параметры: ХИ2ТЕСТ (*фактический _интервал; ожидаемый_интервал*). Здесь *фактический_интервал* – это интервал данных, которые содержат наблюдения, подлежащие сравнению с ожидаемыми значениями; *ожидаемый_интервал* – это интервал данных, который содержит теоретические (ожидаемые) значения для соответствующих наблюдаемых.

Пример 21

Получены данные о длине семян подсолнечника (рисунок 10). Проверить соответствие данных нормальному закону распределения.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	10	11,45	10,5	11,51	11,06	11,14	11,73	11,06	10,46	11,19		Среднее	Станд.отклонен.
2	11,7	11,95	12,27	11,43	10,41	12,75	11,7	11,86	10,03	11,99		11,3064	0,87362197
3	10,72	11,35	12,02	10,62	11,11	10,62	12,39	11,22	11,21	12,2			
4	12,28	11,45	12,7	11,96	12,6	11,14	9,88	10,67	9,94	11,28			
5	11,6	7,99	11,97	10,75	11	11,78	11,1	12,46	11,3	11,82			
6													
7	Длина семян		Частота	Отн.частота		Накопленная частота			Теоретические частотсти			Теоретические частоты	
8	7,64		0	0		0			0,00			0,00	
9	8,34		1	0,02		0,02			0,00			0,07	
10	9,04		0	0		0,02			0,02			0,79	
11	9,74		0	0		0,02			0,09			4,58	
12	10,44		5	0,1		0,12			0,28			13,96	
13	11,14		12	0,24		0,36			0,45			22,42	
14	11,84		18	0,36		0,72			0,38			18,95	
15	12,54		11	0,22		0,94			0,17			8,43	
16	13,24		3	0,06		1			0,04			1,97	
17													
18			50	1									

Рисунок 10 - Данные о длине семян

Частоты, относительные частоты найдем аналогично тому, как мы находили в примере 6 и 7. Теперь найдем теоретические частоты и теоретические частости.

Для этого дополнительно в ячейке L2 найдем выборочное среднее с помощью встроенной функции СРЗНАЧ, и в ячейке M2 стандартное отклонение с помощью функции СТАНДОТКЛОН для данных из диапазона A1:J5. Затем с помощью функции НОРМРАСП находим теоретические частости, заполнив поля: *x* – A8; *среднее* - L2; *стандартное_отклонение* - M2, *интегральный* – 0. Нажимаем ENTER, в результате в ячейке I8 получаем первое значение. Протягивая, скопируем содержимое этой ячейки в диапазон ячеек I9:I16.

Теперь найдем значение теоретических частот. Установим курсор в ячейку L8 и введем формулу C$18*I8. Нажимаем ENTER/ Протягивая, скопируем содержимое ячейки L8 в диапазон ячеек L9:L16.

Теперь с помощью функции ХИ2ТЕСТ определяем соответствие данных нормальному распределению. В рабочие поля этой функции вводим значения: *фактический* – C8:C16, *ожидаемый* – L8:L16 (рисунок 11).

Поскольку полученная вероятность 0, 0002, меньше, чем уровень значимости $\alpha = 0{,}05$, то можно утверждать, что данные не соответствуют нормальному закону распределения.

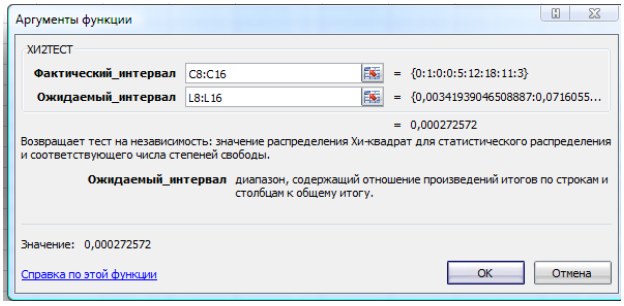


Рисунок 11 – Функция ХИ2ТЕСТ

Критерий Колмогорова применяется только для случая непрерывных случайных величин. Он связывает эмпирическую функцию распределения $F_n^*(x)$ с функцией распределения $F(x)$ непрерывной СВ X.

Пусть $x_1, x_2,\ldots,x_n$ – конкретная выборка из распределения с неизвестной непрерывной функцией распределения $F(x)$ и $F_n^*(x)$ – эмпирическая функция распределения. Выдвигается нулевая гипотеза $H_0: F(x) = F(x_0)$, альтернативная: $H_1: F(x) \neq F(x_0)$. Колмогоров доказал, что функция распределения величины

$$D = \max\left|F^*(x) - F(x)\right| \qquad (15)$$

В случае непрерывной функции *F(x)* при $n \to \infty$, имеет предел

$$K(\lambda) = \sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2\lambda^2} \qquad (16)$$

Функция $K(\lambda)$ получила название функции Колмогорова.

Проверку гипотезы с помощью критерия Колмогорова проводят в ниже приведенном порядке:

1) располагаем результаты наблюдений по возрастанию их значений в виде интервального вариационного ряда;

2) находим эмпирическую функцию распределения $F^*(x)$;

3) вычисляем, пользуясь предполагаемой функцией $F(x)$, значения теоретической функции распределения, соответствующие наблюдаемым значениям СВ X;

4) находим для каждого x_i модуль разности между эмпирической и теоретической функциями распределении;.

5) определяем $\lambda = D\sqrt{n} = \max\left|F^*(x) - F(x)\right|\sqrt{n}$;

6) находим критические значения λ_α в зависимости от уровня значимости по таблице15

Таблица 15 – Критические значения λ_α

α	0,50	0,40	0,30	0,20	0,10	0,05	0,02	0,01	0,001
λ_α	0,828	0,895	0,974	1,073	1,224	1,358	1,510	1,627	1,950

Если опытное значение $\lambda \geq \lambda_\alpha$, то гипотеза о соответствии теоретического закона распределения с данными выборки отклоняется. Если $\lambda \leq \lambda_\alpha$, то гипотеза принимается.

Пример 22

Пусть СВ X – результат измерения длин печенья. Результаты измерения длин (мм) 1000 штук изготавливаемого печенья, помещены в таблицу 16, предварительно их упорядочив.

Таблица 16 – Результаты измерений длины печенья

x_i	98,0	98,5	99,0	99,5	100,0	100,5	101,0	101,5	102,0	102,5
n_i	21	47	87	158	181	201	142	97	41	25

С помощью критерия согласия Колмогорова проверить гипотезу о нормальном законе распределения случайной величины X с математическим ожиданием $M(X)=a=100{,}25$ мм и средним квадратическим отклонением $\sigma=1$ мм при уровне значимости $\alpha=0{,}05$.

Поскольку результаты измерений печенья, приведенные в таблице, помещены в порядке возрастания СВ X и задано математическое ожидание, то вычислим эмпирическую функцию распределения $F^*(x_i)=\sum_{i=1}^{10}\frac{n_i}{1000}$; разность x_i-a; теоретическую функцию $F(x)$ по формуле $F(x)=\frac{1}{2}+\Phi(x-a)$, ($\Phi(x)$ – функция Лапласа); разности $F^*(x_i)-F(x_i)$. Результаты расчетов приведены в таблице17

Таблица 17 – Результаты расчетов

x_i	x_i-a	$\Phi(x_i-a)$	$F(x_i)$	$F^*(x_i)$	$\lvert F^*(x_i)-F(x_i)\rvert$
98,0	-2,25	-0,4877	0,0123	0	0,0123
98,3	-1,75	-0,4599	0,0401	0,021	0,0191
99,0	-1,25	-0,3944	0,1056	0,068	0,0376
99,3	-0,75	-0,2734	0,2266	0,155	0,0716
100,0	-0,25	-0,0987	0,4013	0,313	0,0883
100,3	0,25	0,0987	0,5987	0,494	0,1057
101,0	0,75	0,2734	0,7734	0,655	0,0784
101,3	1,25	0,3944	0,8944	0,837	0,0574
102,0	1,75	0,4599	0,9599	0,934	0,0259
102,3	2,25	0,4877	0,9877	0,975	0,0127

Из этой таблицы выбираем наибольшую из разностей $\left|F^*(x_i)-F(x_i)\right|=D=0{,}1057$. Находим $\lambda=D\sqrt{n=0{,}1057}\cdot 31{,}622=3{,}342$. По значению $\alpha=0{,}05$ из таблицы 1 находим значение $\lambda_\alpha=1{,}358$. Видим, что $\lambda\geq\lambda_\alpha$, значит гипотеза о подчинении случайной величины нормальному закону распределения с параметрами $M(X)=a=100{,}25$ мм и $\sigma=1$ мм отклоняется.

Литература:

1. Гмурман В.Е. Теория вероятностей и математическая статистика. – М: Высш. Шк., 1977. – 480 с.

2. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высш. Шк., 1979. – 400 с.

3. Гусак А. А. Высшая математика: В 2 т. Т. 2. Учеб. Для студентов вузов. – 5-е изд. – Мн.: ТетраСистемс, 2004. – 448 с.

4. Гнеденко Б.В. Курс теории вероятностей. – М., 1965.

5. Письменный Д.Т. Конспект лекций по теории вероятностей и математической статистике. – М.: Айрис-прес, 2004. – 256 с.

6. Гельман В.Я. Решение математических задач средствами Excel: Практикум. – СПб.: Питер, 2003. – 240 с.

Щендрикова О.А.

Математическая статистика: теория, примеры решения типовых задач с использованием MS Excel для студентов – технологов дневной и заочной форм обучения. – Могилев: УО МГУП, 2010. –

В учебно-методическом пособии дано краткое изложение теоретической части раздела «Математическая статистика» учебной дисциплины «Высшая математика». Подробно разобраны примеры решения типовых задач, в том числе с применением MS Excel. Включены вопросы для самопроверки.