МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

КУРГАНСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Кафедра «Автомобильный транспорт и автосервис»

ОБРАБОТКА РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТА С ПОМОЩЬЮ MICROSOFT EXCEL

ЧАСТЬ 1

Лабораторный практикум
по дисциплине «Компьютерные технологии в науке и образовании» для студентов квалификации (степени) Магистр направления 190500 (552100)

Курган 2011

Кафедра: «Автомобильный транспорт и автосервис»
Дисциплина: «Компьютерные технологии в науке и образовании» специализированной подготовки направления 190600 (552100) «Эксплуатация транспортных средств» (квалификация выпускника (степень) Магистр)
Составил: канд. техн. наук, доцент Д.И. Дик

Утвержден на заседании кафедры «6» мая 2011 г.

Рекомендован методическим советом университета «2» июня 2011 г.

СОДЕРЖАНИЕ

1 ВВЕДЕНИЕ В СТАТИСТИЧЕСКИЕ МЕТОДЫ

Математическая статистика занимается как статистическим описанием результатов опытов и наблюдений, так и построением и проверкой подходящих математических моделей, содержащих понятие вероятности.

Статистическое описание и вероятностные модели применяются к физическим процессам, обладающим тем свойством, что хотя результат отдельного измерения физической величины x не может быть предсказан с достаточной точностью, значение некоторой подходящей функции $y = y\,(x_1, x_2, \ldots, x_n)$ от множества результатов $x_1, x_2, \ldots, x_n$ повторных измерений может быть предсказано с существенно лучшей точностью. Такая функция называется **статистикой** (наиболее известная статистика — выборочное среднее), а указанное свойство физического процесса — его статистической устойчивостью. Статистическая устойчивость в каждом конкретном случае есть физический закон, который может быть проверен только опытом.

В классической вероятностной модели наблюдаемая физическая величина x рассматривается как одномерная случайная величина с подлежащей определению или оценке плотности вероятности $\varphi(x)$. Каждая выборка $(x_1, x_2, \ldots, x_n)$ значений x рассматривается как результат n независимых повторных измерений. При этом $x_1, x_2, \ldots, x_n$ представляют собой взаимно независимые случайные величины с одинаковой плотностью вероятности $\varphi(x)$. Такая выборка называется **случайной выборкой** объема n. Поскольку возможны различные реализации выборки она сама представляет собой n-мерную случайную величину $(x_1, x_2, \ldots, x_n)$. Плотность ее распределения называется **функцией правдоподобия**:

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

Исходя из того, что выборка является случайной величиной, каждая статистика, определяемая как некоторая функция

$$y = y\,(x_1, x_2, \ldots, x_n)$$

выборочных значений $x_1, x_2, \ldots, x_n$ также представляет собой случайную величину[1], распределение которой (так называемое **выборочное распределение** статистики y) однозначно определяется функцией правдоподобия, а, следовательно, и распределением величины x. Каждое выборочное распределение зависит, как правило, от объема выборки[2].

---

[1] Например, имеется несколько выборок одного случайного процесса. Очевидно, что статистка этого случайного процесса, например, выборочное среднее, будет различаться для каждой из выборок. Таким образом, сама статистика также является случайной величиной со своим законом распределения.

[2] Возьмем для примера выборочное среднее. Если размер выборки велик, то по закону больших чисел значение статистики для выборки будет близко к среднему значению теоретического распределения физической величины, таким образом, рассеивание выборочного среднего между различными выборками будет незначительным. Очевидно, что с уменьшением размера выборки из фактора случайности точность оценки среднего значения теоретического распределения будет ухудшаться, и рассеивание будет увеличиваться.

Когда возрастает объем выборки, многие выборочные статистики сходятся по вероятности к соответствующим параметрам теоретического распределения величины *x*. Поэтому каждую выборку рассматривают как выборку из теоретически бесконечной *генеральной совокупности*, распределение признака в которой совпадает с теоретическим распределением вероятностей величины *x*. Последнее называется **распределением генеральной совокупности**, а его параметры – параметрами генеральной совокупности. Во многих приложениях теоретическая генеральная совокупность есть идеализация действительной совокупности, из которой получена выборка.

2 ОПИСАТЕЛЬНАЯ СТАТИСТИКА

2.1 Подготовка данных

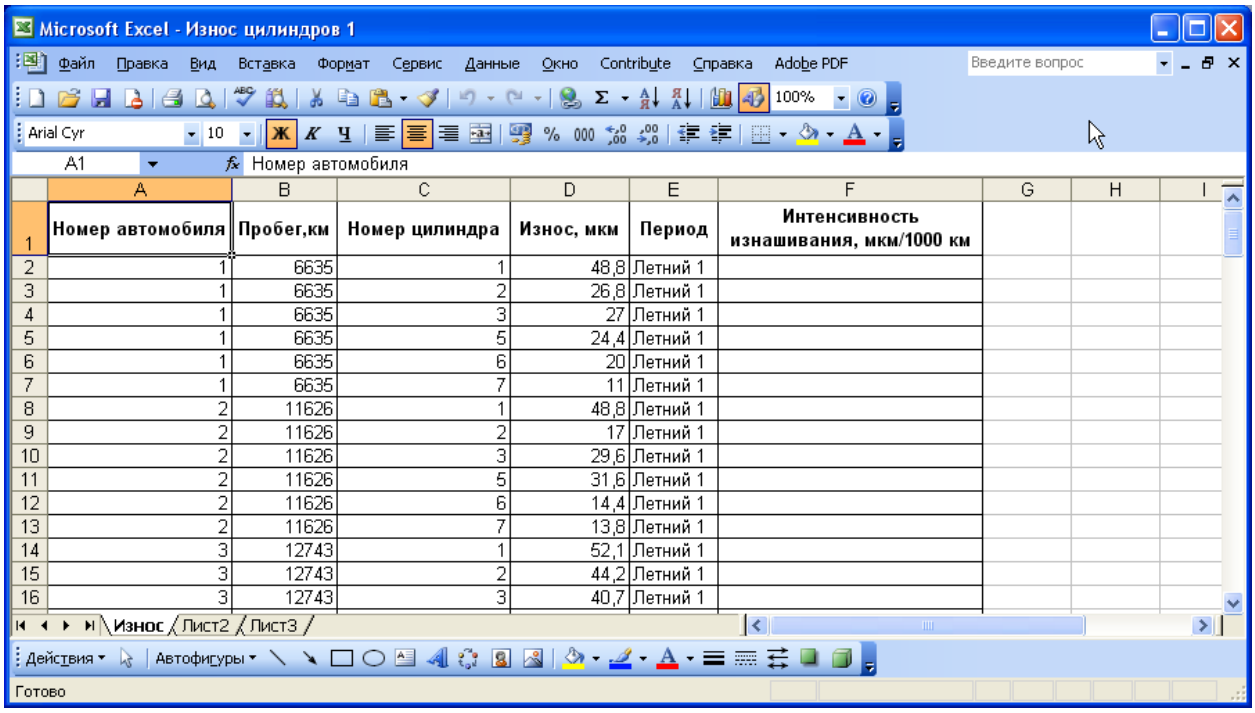
Для изучения основ описательной статистики нужно открыть файл «Износ цилиндров.xls». В этом файле содержится рабочая книга с данными об изнашивании цилиндров двигателей.

Для открытия рабочей книги «Износ цилиндров.xls» выполните перечисленные ниже действия:

а) запустите программу Excel;

б) найдите файл «Износ цилиндров.xls» и откройте его;

в) выберите команду меню «Файл ⇒ Сохранить как…» и сохраните книгу в файле «Износ цилиндров 1.xls». Открытая книга будет выглядеть так, как на рисунке 1.



	A	B	C	D	E	F	G	H	I
1	Номер автомобиля	Пробег,км	Номер цилиндра	Износ, мкм	Период	Интенсивность изнашивания, мкм/1000 км			
2	1	6635	1	48,8	Летний 1				
3	1	6635	2	26,8	Летний 1				
4	1	6635	3	27	Летний 1				
5	1	6635	5	24,4	Летний 1				
6	1	6635	6	20	Летний 1				
7	1	6635	7	11	Летний 1				
8	2	11626	1	48,8	Летний 1				
9	2	11626	2	17	Летний 1				
10	2	11626	3	29,6	Летний 1				
11	2	11626	5	31,6	Летний 1				
12	2	11626	6	14,4	Летний 1				
13	2	11626	7	13,8	Летний 1				
14	3	12743	1	52,1	Летний 1				
15	3	12743	2	44,2	Летний 1				
16	3	12743	3	40,7	Летний 1				

Рисунок 1 — Рабочая книга «Износ цилиндров 1.xls»

2.2 Создание таблицы частот

Прежде всего, проанализируем распределение значений в приведенном выше наборе данных.

Для начала рассчитаем величину интенсивности изнашивания на 1000 км:

а) введите в ячейку F2 формулу «=D2/B2*1000».

б) скопируйте ячейку F2 в ячейки F3:F139.

в) выделите ячейки F2:F139. Выберите команду меню «Формат ⇒ Ячейки». Установите числовой формат с двумя знаками после запятой.

Далее построим таблицу частот для данных об интенсивности изнашивания. Для этого выполните следующие действия:

а) отсортируйте информацию об интенсивности изнашивания:

1) щелкните на любой ячейке первой строки диапазона (строке с подписями, например по ячейке A1);

2) выберите команду меню «Данные ⇒ Сортировка» (рисунок 2);

3) выберите сортировку по столбцу «Интенсивность изнашивания».

б) посчитайте накопленную частоту, т.е. общее количество цилиндров, интенсивность изнашивания которых меньше данной или равна ей:

1) щелкните на ячейке G2, введите число 1 и нажмите «Enter»;

2) введите в ячейку G3 формулу «=G2+1»;

3) подведите курсор к нижнему правому углу ячейки J3 (курсор примет форму крестика) и дважды щелкните левой кнопкой мышки.

в) посчитайте процентную долю цилиндров с интенсивностью изнашивания меньшей данной или равной ей:

1) щелкните на ячейке H2 и введите формулу «=H2/МАКС(G:G)» (функция МАКС() возвращает максимальное значение из указанного диапазона, в нашем случае столбца G);

2) подведите курсор к нижнему правому углу ячейки K2 и дважды щелкните левой кнопкой мышки.

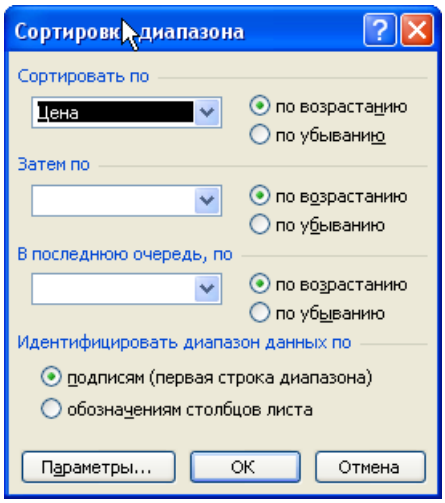


Рисунок 2 — Сортировка цилиндров по интенсивности изнашивания

Эта таблица существенно упрощает интерпретацию данных. Например, позволяет легко посчитать количество домов в заданном ценовом диапазоне (если покупатель определился с ценовым диапазоном, то ему наверняка будет интересно знать, сколько домов находится в нем).

2.3 Группирование данных и построение гистограмм

Таблицы частот очень удобно использовать для представления специальной информации о распределениях, но им не хватает наглядности. Анализируя числа, очень трудно представить себе, насколько плотно сгруппированы значения в таблице частот. Многие специалисты статистического анализа предпочитают использовать визуальную картину распределения в виде гистограммы. Гистограмма — это столбиковая диаграмма, в которой каждый столбик представляет интервал (карман), а его высота пропорциональна количеству значений в интервале. Гистограммы используются для представления относительных частот[3], накопленных частот, процентных долей и накопленных процентных долей.

В Excel гистограммы можно создавать стандартными средствами или с помощью входящего в состав Excel подключаемого модуля (надстройки) «Analysis ToolPak» («Пакет анализа»).

Для загрузки подключаемого модуля «Analysis ToolPak» выполните перечисленные ниже действия:

а) выберите команду меню «Сервис ⇒ Надстройки».

б) установите флажок элемента «Analysis ToolPak» и щелкните на кнопке ОК.

При необходимости укажите путь до установочного пакета Excel.

После инсталляции и загрузки подключаемый модуль Пакет анализа «Analysis ToolPak» будет доступен из меню Excel в виде новой команды меню «Сервис ⇒ Анализ данных…».

Для доступа к функциям модуля «Analysis ToolPak» выполните перечисленные ниже действия:

а) выберите команду меню «Сервис ⇒ Анализ данных…».

б) на экране появится диалоговое окно «Анализ данных», которое показано на рисунке 3.

в) в списке «Инструменты анализа» диалогового окна «Анализ данных» представлены команды модуля «Analysis ToolPak». Для запуска любой команды нужно выделить ее и щелкнуть на кнопке ОК.

[3] Относительная частота — отношение количества выборочных значений попавших в интервал к объему выборки. Является оценкой вероятности попадания значений случайной величины в интервал.

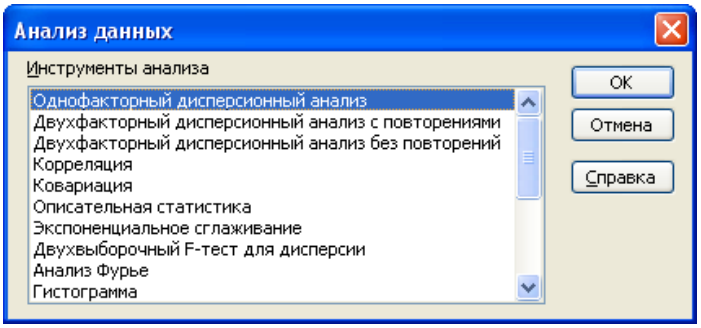


Рисунок 3 — Окно надстройки диалоговое окно «Анализ данных»

Перед построением гистограммы необходимо определиться с интервалами разбиения.

Для этого найдем максимальное и минимальное значения по интервалу данных:

г) введите в ячейку I2 формулу «=МИН(F:F)» (получим минимальную интенсивность изнашивания).

д) введите в ячейку I3 формулу «=МАКС(F:F)» (получим максимальную интенсивность изнашивания).

В нашем примере минимальная интенсивность изнашивания составляет 0,05 мкм/1000 км, максимальная 8,41 мкм/1000 км. Примем за нижнюю границу интервалов 0 мкм/1000 км. Ширину интервала (кармана) примем равной 0,5 мкм/1000 км.

Для задания границ интервалов выполните перечисленные ниже действия:

а) щелкните на ячейке J2, введите число 0 и нажмите «Enter».

б) введите в ячейку J3 формулу «=J2+0,5».

в) скопируйте ячейку J3 в ячейки от J4 до J19 (подведите курсор к нижнему правому углу ячейки J3 (курсор примет форму крестика) и нажмите левую кнопку мышки и, удерживая ее, переместите курсор до ячейки J19).

Построим гистограмму интенсивности изнашивания с использованием модуля «Analysis ToolPak».

Для этого выполните перечисленные ниже действия:

а) выберите команду меню «Сервис ⇒ Анализ данных…» и инструмент анализа «Гистограмма».

б) на экране появится диалоговое окно «Гистограмма», которое показано на рисунке 4.

в) укажите в качестве входного интервала интервал F2:A139.

г) укажите в качестве интервала карманов интервал H3:H18 (верхнюю и нижнюю границу не включаем).

д) установите флажок напротив элемента «Вывод графика» и щелкните на кнопке ОК.

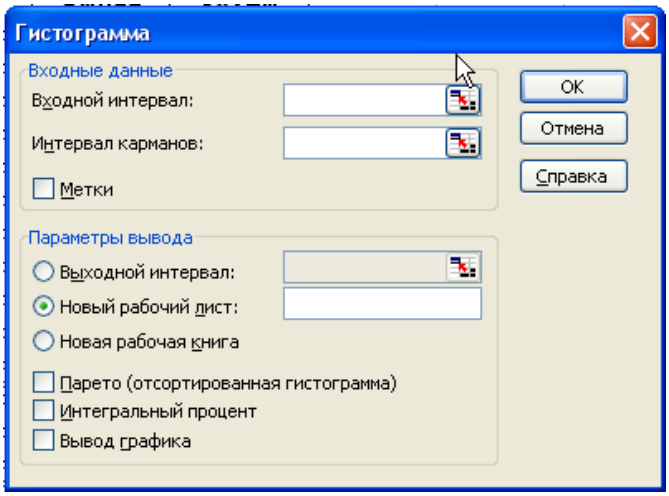
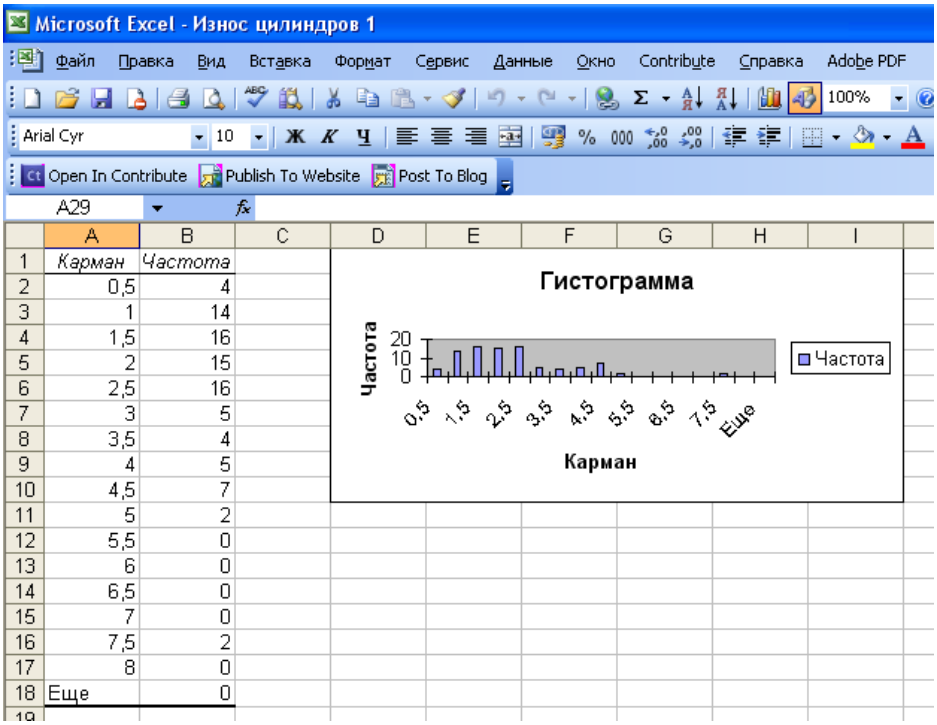


Рисунок 4 — Окно инструмента построения гистограмм модуля «Analysis ToolPak»

В результате на новом рабочем листе будет подсчитано количество попаданий в интервалы и построена соответствующая гистограмма (рисунок 5).



	A	B
1	Карман	Частота
2	0,5	4
3	1	14
4	1,5	16
5	2	15
6	2,5	16
7	3	5
8	3,5	4
9	4	5
10	4,5	7
11	5	2
12	5,5	0
13	6	0
14	6,5	0
15	7	0
16	7,5	2
17	8	0
18	Еще	0

Рисунок 5 — Гистограмма цен на недвижимость, построенная с помощью модуля «Analysis ToolPak»

Теперь построим гистограмму самостоятельно.

Для построения гистограммы в первую очередь необходимо подсчитать количество попаданий предложений недвижимости в выбранные интервалы.

Для этого выполните перечисленные ниже действия:

а) перейдите назад на лист «Недвижимость» рабочей книги (выберите вкладку «Недвижимость» слева снизу в окне программы Excel).

б) введите в ячейку K2 формулу «="<"&ТЕКСТ(J2;"0,0")» (задает условие в виде текстовой строки, «"0,0"» – формат преобразования числа в строку).

в) введите в ячейку L2 формулу «=СЧЁТЕСЛИ(F$2:F$139;K2)» (подсчет количества единиц недвижимости, удовлетворяющей указанному условию).

г) скопируйте ячейку K2 в ячейки K3:K19, а ячейку L2 в ячейки L3:L20 (выделите ячейки K2 и L2, подведите курсор к нижнему правому углу ячейки L2 (курсор примет форму крестика) и дважды щелкните левой кнопкой мышки).

д) введите в ячейку M3 формулу «=L3-L2».

е) скопируйте ячейку M3 в ячейки M4:M19.

Теперь можно построить гистограмму. Для этого:

а) выберите команду меню «Вставка ⇒ Диаграмма…».

б) выберите тип диаграммы «Гистограмма».

в) укажите диапазон ячеек M3:M19, «Ряды в столбцах».

г) перейдите на вкладку «Ряд» и в «Подписи по оси X» укажите диапазон ячеек J2:J19.

д) нажмите кнопку «Готово».

Более глубокое представление о распределении можно получить после разбивки гистограммы на категории. В данном примере интерес представляет влияние сезонных условий на интенсивность изнашивания.

Для каждого цилиндра недвижимости определим интервал интенсивности изнашивания, к которому он относится. Для этого:

а) введите в ячейку N2 формулу «=ПОИСКПОЗ(F2;J$2:J$19;1)» (функция ПОИСКПОЗ возвращает относительное положение элемента массива, который соответствует указанному значению).

б) скопируйте ячейку N2 в ячейки N3:M139.

Разделим цилиндры на категории по условию эксплуатации (эксплуатировался ли двигатель в зимний период или нет):

а) введите в ячейку O2 формулу «=ЕСЛИ(E2="Зимний";N2;"")» (если двигатель не эксплуатировался в зимний период, то интервал интенсивности изнашивания примет пустое значение).

б) скопируйте ячейку O2 в ячейки O3:O139.

в) введите в ячейку P2 формулу «=ЕСЛИ(E2<>"Зимний";N2;"")» (если двигатель эксплуатировался в зимний период, то интервал интенсивности изнашивания примет пустое значение).

г) скопируйте ячейку P2 в ячейки P3:P100.

Подсчитаем количество попаданий в выбранные интервалы (в долях единицы):

а) введите в ячейку J25 число 1.

б) подведите курсор к нижнему правому углу ячейки J25 (курсор примет форму крестика), нажмите левую кнопку мышки и клавишу «Ctrl», удер-

живая их, переместите курсор до ячейки J42 (получится список чисел от 1 до 18).

в) введите в ячейку K25 формулу «="="&ТЕКСТ(J25;"0")» (задает условие в виде текстовой строки).

г) введите в ячейку L25 формулу «=СЧЁТЕСЛИ(O$2:O$139;K25)/СЧЁТ(O$2:O$139)» (подсчет цилиндров в долях единицы, относящихся к зимнему периоду эксплуатации и удовлетворяющих указанному условию).

д) введите в ячейку M25 формулу «=СЧЁТЕСЛИ(P$2:P$139;K25)/СЧЁТ(P$2:P$139)» (подсчет цилиндров в долях единицы, не относящихся к зимнему периоду эксплуатации и удовлетворяющих указанному условию).

е) скопируйте ячейку K25 в ячейки K26:K41, ячейку L25 в ячейки L26:L41, а ячейку M25 в ячейки M25:M41.

ж) введите в ячейку L24 текст «Зимний», а в ячейку M24 текст «Летний».

Теперь можно построить гистограмму. Для этого:

а) выделите ячейки L24:M41.

б) выберите команду меню «Вставка ⇒ Диаграмма…».

в) выберите тип диаграммы «Гистограмма».

г) нажмите кнопку «Готово».

д) щелкните на любом столбике полученной гистограммы левой копкой мыши.

е) щелкните правой копкой мыши и выберите из контекстного меню команду «Формат рядов данных…».

ж) на вкладке «Параметры» установите ширину зазора равной 0 и нажмите кнопку «ОК».

2.4 Параметры распределения

После создания диаграммы распределения для дальнейшего анализа набора данных желательно иметь оценки статистических параметры распределения (*выборочные статистики*[4]).

Обычно стремятся, чтобы оценки обладали свойствами *состоятельности, эффективности* и *несмещенность*.

Оценка параметра является **состоятельной**, если при увеличении числа наблюдений до бесконечности оценка сходится к оцениваемому параметру по вероятности, т.е. вероятность отклонения оценки от оцениваемого параметра стремится к 0.

Оценка параметра является **эффективной**, если это наилучшая из возможных оценок при наложенных ограничениях.

Оценка параметра является **несмещенной**, если математическое ожидание оценки (математическое ожидание можно представить как среднее от

[4] В последующем описании перед названиями параметров для краткости опущено слово *выборочный*, однако не стоит забывать, что речь идет о выборочных статистиках, которые являются оценками аналогичных характеристик теоретического распределения.

выборки, размер которой стремится к бесконечности) совпадает с оцениваемым параметром.

2.4.1 Квантили и квартили

Одним из итоговых ориентиров является квантили. **Квантиль** порядка P одномерного распределения, есть такое значение случайной величины x_P случайной величины для которого вероятность того что случайная величина x меньше x_P равно P:

$$\mathrm{P}\{x < x_P\} = P.$$

Например, если ребенка вес новорожденного ребенка больше квантиля 0,9, то это значит, что он весит больше, чем 90% всех новорожденных детей.

Частным случаем квантиля являются:

– **квартили** $x_{1/4}$, $x_{1/2}$, $x_{3/4}$, соответствующие четвертям распределения (обычно их называют первой, второй и третьей квартилями);

– **децили** $x_{0,1}, x_{0,2}, \ldots, x_{0,9}$;

– **процентали** $x_{0,01}, x_{0,02}, \ldots, x_{0,99}$.

Квартили, децили и процентали делят область изменения x на 4, 10 и 100 интервалов, попадания в которые имеют равные вероятности.

Один из способов подсчета квантилей заключается в создании таблицы частот. В столбце накопленных вероятностей можно легко найти значения, которые соответствуют заданным значениям квантилей. Однако при работе с очень большим набором данных сделать это довольно сложно. Для экономии времени в Excel предусмотрено несколько функций для вычисления этих значений, которые перечислены в таблице 1.

Таблица 1 – Функции Excel для вычисления квантилей

Функция	Описание
ПЕРСЕНТИЛЬ (массив; *k*)	Возвращает *k*-й квантиль для значений из массива или интервала данных с числовыми значениями, который определяет относительное положение; *k* — это значение порядка квантиля в интервале от 0 до 1
ПРОЦЕНТРАНГ (массив; *x*; разрядность)	Возвращает процентное содержание значений меньших *x* в множестве данных. Разрядность — это необязательное значение, которое определяет количество значащих цифр в возвращаемой величине процентного содержания значения
КВАРТИЛЬ (массив; часть)	Возвращает квартиль множества данных для указанной части, для 1, 2 и 3 возвращается соответственно первая, вторая и третья квартили, для 0 возвращается минимальное значение, для 4 максимальное

2.4.2 Характеристики положения

Еще один способ подытожить данные — вычислить одно значение, характеризующее весь набор данных.

Например, для таблицы квантилей таким значением является квантиль $x_{1/2}$, который называется **медианой**. Медиана делит распределение на два равновероятных интервала.

Для нахождения медианы необходимо упорядочить значения выборки по возрастанию. Точный подсчет медианы зависит от количества наблюдений в наборе данных. При нечетном количестве значений медианой является центральное по порядку значение, а при четном — полусумма двух центральных значений.

Еще одной распространенной характеристикой является **среднее значение**, которое равно сумме значений, деленной на их количество. Графически оно обычно обозначается в виде черточки над именем переменной ($\bar{x}$), и именно это обозначение используется далее в книге. В целом формула вычисления среднего значения имеет вид:

$$\bar{x} = \frac{\textit{сумма всех значений}}{\textit{общее количество наблюдений}} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n},$$

где x_1, x_2, …, x_n – отдельные наблюдения в наборе n наблюдений.

Одним из недостатков среднего значения является то, что оно существенно зависит от экстремальных значений. Предположим, что в компании работают 10 работников. Девять из них имеют заработную плату 10 тыс. рублей в месяц, а начальник 200 тыс. рублей в месяц. Медианой данного распределения является зарплата 10 тыс. рублей, а средним 29 тыс. рублей. Таким образом, медиана в большей степени представляет «типичную» зарплату, а средняя зарплата почти в три раза выше медианы из-за влияния всего лишь одного крупного значений.

Из этого примера можно сделать следующий вывод: не нужно слепо доверять любой единственной итоговой характеристике распределения. Учтите, что среднее значение чувствительно к экстремальным значениям, а медиана — нет, поскольку игнорирует величину экстремальных значений. Обе характеристики обладают определенными недостатками, поэтому перед составлением итоговых характеристик распределения рекомендуется проанализировать набор данных с помощью гистограммы. Кроме того, попробуйте вычислить и сравнить разные итоговые характеристики.

Медиана и среднее являются наиболее распространенными, но не единственными итоговыми характеристиками распределения. Далее кратко описываются некоторые другие характеристики.

Для снижения влияния экстремальных значений можно использовать **усеченное среднее**, т.е. среднее для набора данных, из которого исключены несколько процентов значений с обоих концов распределения. Например, 5%-ное усеченное среднее равно среднему значению для 90% значений из

набора данных, за исключением 5% с каждого конца распределения. Усеченное среднее представляет собой компромиссный вариант итоговой характеристики по сравнению с медианой и средним.

Не менее популярна такая характеристика распределения, как **среднее геометрическое**, которое является *n*-м корнем произведения всех *n* значений набора данных:

$$среднее\ ггеометрическое = \sqrt[n]{x_1 \cdot x_2 \cdot \ldots \cdot x_n}\ .$$

Среднее геометрическое можно использовать, например, для вычисления средних темпов роста, если задан составной доход с переменными ставками. Среднее геометрическое нельзя применять для наборов данных, содержащих ноль или отрицательные значения.

Еще одна итоговая характеристика, которая редко используется в наши дни, называется **средним гармоническим**. Для определения среднего гармонического *H* используется формула:

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}.$$

Среднее гармоническое всегда меньше среднего геометрического, которое всегда меньше среднего арифметического.

Еще одной итоговой характеристикой распределения является **мода**. Для непрерывного распределения мода есть точка максимума плотности распределения вероятности. Мода дискретного распределения есть такое спектральное значение ξ_m, что предшествующее и следующее за ним спектральные значения имеют вероятности меньше чем $p(\xi_m)$. Распределение может иметь несколько мод (одномодальное, двухмодальное и многомодальное распределение).

В таблице 2 кратко описаны некоторые функции Excel, предназначенные для вычисления разных центральных мер распределения.

Таблица 2 – Функции Excel для вычисления центральных мер распределения

Функция	Описание
СРЗНАЧ (массив)	Возвращает среднее арифметическое для значений массива
СРГЕОМ (массив)	Возвращает среднее геометрическое для значений массива
СРГАРМ (массив)	Возвращает среднее гармоническое для значений массива
МЕДИАНА (массив)	Возвращает медиану для значений массива
МОДА (массив)	Возвращает наиболее часто встречающееся значение в массиве
УРЕЗСРЕДНЕЕ (массив; доля)	Возвращает среднее после отбрасывания заданной процентной доли данных с экстремальными значениями (доля должна иметь значения от 0 до 1)

2.5 Характеристики рассеяния

Среднее и медиана не полностью характеризуют распределение, так как не учитывают рассеяние данных. Рассеяние характеризует различия между данными или, что то же самое, разброс от центра.

Простейшей мерой изменчивости является **размах**, т.е. разница между максимальным и минимальным значениями распределения. Более высокому рассеянию обычно соответствует более широкий диапазон значений. Однако размах не совсем точно характеризует рассеяние распределения (рисунок 6).

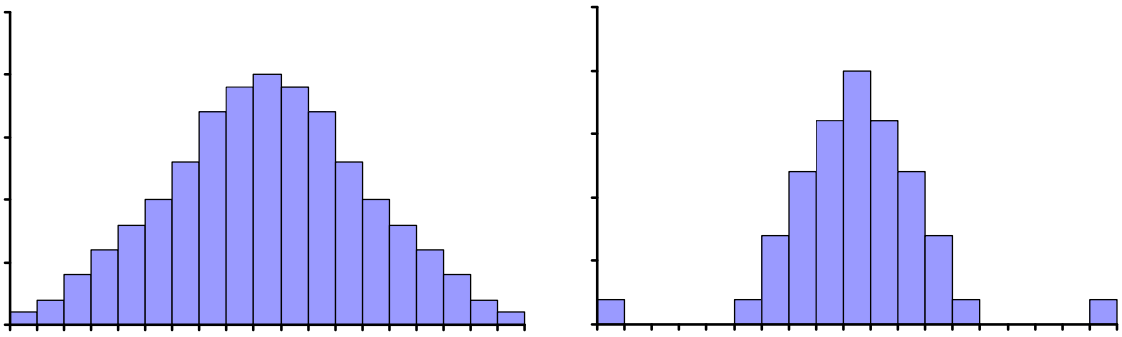


Рисунок 6 — Распределения с одинаковыми средним и диапазоном, но разным рассеянием

В качестве характеристик рассеяния часто используются **интерквартильная широта** $x_{3/4}–x_{1/4}$ (разность между третьей и первой квартилями, в этом диапазоне располагается 50% всех данных распределения) и **(10–90)-процентная широта** $x_{0,9}–x_{0,1}$ (разность между децилями 0,9 и 0,1 или процентилями 0,90 и 0,10).

Для определения рассеяния удобно использовать отклонение d_i значения наблюдения x_i от среднего $\bar{x}$:

$$d_i = x_i - \bar{x}.$$

Поскольку одни отклонения могут иметь отрицательные значения (для наблюдений, значения которых меньше среднего), а другие — положительные (для наблюдений, значения которых больше среднего), то простое суммирование отклонений ничего не даст, поскольку они будут взаимно компенсировать друг друга (сумма отклонении всегда равна нулю, а потому среднее отклонение также всегда равно нулю).

Вместо вычисления среднего для отклонений можно возвести каждое отклонение в квадрат (чтобы все значения отклонений стали положительными), а затем просуммировать их и усреднить. Эта мера изменчивости называется **дисперсией** и обозначается s^2:

$$s^2 = \frac{\textit{сумма квадратов отклонений}}{\textit{число наблюдений}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

или S^2:

$$S^2 = \frac{\textit{сумма квадратов отклонений}}{\textit{число наблюдений} - 1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \vec{x})^2.$$

Возникает вопрос, почему сумма квадратов отклонений делится на $(n-1)$, а не на n? Здесь стоит напомнить, что сумма отклонений равна нулю, а потому, зная значения $(n-1)$ отклонений, можно вычислить оставшееся отклонение. Это значит, что только $(n-1)$ отклонений имеют независимые значения, а n-е определяется остальными. В таком случае говорят, что распределение имеет $(n-1)$ степеней свободы. Из-за этого, если при расчете дисперсии поделить сумму квадратов на n, то в результате будет получена смещенная оценка. По этой причине оценка дисперсии S^2 применяется на практике чаще.

Для измерения изменчивости также используется **стандартное отклонение**, которое обозначается символом s и равняется квадратному корню дисперсии. Стандартное отклонение представляет «типичное» отклонение значений от среднего.

Еще одной характеристикой рассеяния является **среднее абсолютное отклонение**, вычисляемое по формуле:

$$\textit{среднее абсолютное отклонение} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \vec{x}|.$$

В таблице 3 перечислены функции Excel, предназначенные для определения рассеяния данных.

Таблица 3 – Функции Excel для определения рассеяния данных

Функция	Описание
СРОТКЛ (массив)	Возвращает среднее абсолютное отклонение для элементов массива
КВАДРОТКЛ (массив)	Возвращает сумму квадратов отклонений элементов массива от среднего $\left(\sum_{i=1}^{n}(x_i - \vec{x})^2\right)$
НАИМЕНЬШИЙ (массив; *k*)	Возвращает *k*-ое наименьшее значение в множестве данных (массиве)
НАИБОЛЬШИЙ (массив; *k*)	Возвращает *k*-ое наибольшее значение в множестве данных (массиве)
МИН (аргумент1; аргумент2; ...)	Возвращает наименьшее значение в списке аргументов (в качестве аргумента может выступать массив или число)
МАКС (аргумент1; аргумент2; ...)	Возвращает наибольшее значение в списке аргументов (в качестве аргумента может выступать массив или число)
СТАНДОТКЛОН (массив)	Возвращает стандартное отклонение элементов массива, вычисленное по формуле $\sqrt{\sum_{i=1}^{n}(x_i - \vec{x})^2 \Big/ (n-1)}$

Продолжение таблицы 3

Функция	Описание
СТАНДОТКЛОНП (массив)	Возвращает стандартное отклонение элементов массива, вычисленное по формуле $\sqrt{\sum_{i=1}^{n}(x_i-\vec{x})^2 \Big/ n}$
ДИСП (массив)	Возвращает дисперсию для элементов массива, вычисленную по формуле $\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\vec{x})^2$
ДИСПР (массив)	Возвращает дисперсию для элементов массива, вычисленную по формуле $\frac{1}{n}\sum_{i=1}^{n}(x_i-\vec{x})^2$

2.6 Меры формы: асимметрия и эксцесс

Асимметрия, или скос, является мерой несимметричного распределения значений данных.

Положительная асимметрия означает, что значения распределения скучены в области малых значений и распределение имеет длинный хвост в области больших значений. И наоборот: отрицательная асимметрия означает, что значения распределения скучены в области высоких значений и распределение имеет длинный хвост в области малых значений. Равное нулю значение асимметрии соответствует симметричному распределению.

Коэффициент асимметрии вычисляется по формуле:

$$\gamma_1 = \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^3,$$

где s — стандартное отклонение.

Эксцесс характеризует относительную остроконечность или сглаженность распределения по сравнению с нормальным распределением. Положительный эксцесс обозначает относительно остроконечное распределение, а отрицательный — относительно сглаженное распределение.

Коэффициент эксцесса вычисляется по формуле:

$$\gamma_2 = \left\{\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s}\right)^4\right\} - \frac{3(n-1)^2}{(n-2)(n-3)}.$$

В таблице 4 перечислены функции Excel, предназначенные для определения асимметрии и эксцесса данных.

Таблица 4 – Функции Excel определения асимметрии и эксцесса данных

Функция	Описание
СКОС (массив)	Возвращает коэффициент асимметрии распределения значений массива.
ЭКСЦЕСС (массив)	Возвращает коэффициент эксцесса распределения значений массива.

2.7 Выбросы

Как отмечалось при обсуждении средних и медиан, характеристики распределения могут существенно зависеть от экстремальных значений. Учтите, что очень трудно анализировать набор данных, в котором присутствует резко выделяющееся значение, или выброс. Выбросы порой никак не связаны с остальными данными, например, имеют очень большое или очень малое значение либо не соответствуют свойствам распределения. Как уже упоминалось ранее, выброс чрезмерно большой зарплаты может исказить анализ, так как приведет к увеличению средней зарплаты. Выброс не всегда заметен. Например, данные о 70-летней женщине с хорошей физической подготовкой могут совсем не повлиять на анализ данных об общей физической подготовке большой разновозрастной группы людей, но исказят выводы о физической подготовке ее ровесниц в этой группе.

Выбросы возникают либо из-за ошибок ввода, либо в результате необычного или уникального события. Ошибку ввода данных можно легко устранить и повторно выполнить анализ. Но если таких ошибок нет, то придется тщательно изучить выброс и выяснить его связь с остальными значениями набора данных. Например, при анализе данных о ценах на дома можно было бы исключить данные о самом дорогом доме, так как он является местной достопримечательностью и потому имеет чрезвычайно высокую стоимость.

Учтите, что выброс ни в коем случае не нужно исключать из анализа только потому, что он имеет экстремальное значение. Многие открытия в мировой науке были сделаны учеными, тщательно исследовавшими наблюдения, которые не укладывались в ожидаемое распределение. Экстремальные значения могут быть естественной составной частью набора данных (например, как в наборе данных о зарплате). Удаляя эти значения, аналитик тем самым исключит важный аспект распределения.

Одним из возможных решений проблемы выбросов является выполнение двух видов анализа: с выбросами и без них. Если сделанные выводы остались неизменными, то такие выбросы не имеют большого значения. А если полученные выводы существенно отличаются, то придется найти объяснение этим расхождениям. В любом случае экстремальные наблюдения не следует исключать из анализа, во-первых, без уважительной причины и, во-вторых, без документированного обоснования.

Какое значение можно назвать выбросом? Насколько большим (или малым) оно должно быть? Однозначного определения выброса не существу-

ет. Приведенное ниже определение выброса основано на интерквартильном диапазоне (interquartile range — IQR), который, как было ранее сказано, находится между первой и второй квартилями:

– если значение больше третьей квартили плюс l,5·IQR или меньше первой квартили минус l,5·IQR, то оно называется умеренным выбросом.

– если значение больше третьей квартили плюс 3·IQR или меньше первой квартили минус 3·IQR, то оно называется экстремальным выбросом.

Например, если первая квартиль равна 30, а третья — 80, то интерквартильный диапазон равен 50. Тогда любое значение больше 80+(1,5·50) = 155 будет считаться умеренным выбросом, а любое значение больше 80+(3·50) = 230 — экстремальным выбросом. Аналогично определяются выбросы в области очень низких значений.

2.8 Задание

Рассчитайте квартили, среднее, медиану, размах, интерквартильную широту, стандартное отклонение, дисперсию, среднее абсолютное отклонение, коэффициент асимметрии, коэффициент эксцесса для распределений стоимости недвижимости за квадратный метр (всей недвижимости, в центре и в остальных районах).

Проверьте распределения на наличие выбросов.

3 ПРОВЕРКА ГИПОТЕЗ

3.1 Метод доверительных интервалов

Как уже ранее говорилось статистики, которыми, мы до сих пор занимались, являются случайными величинами. При вычислении статистики необходимо заключение о точности оценки (о вероятность отклонения полученной оценки параметра от его истинного значения).

Р. Фишер предложил вместо функции статистики $y(x_1, x_2, \ldots, x_n)$, которая принимается за приближенное значение неизвестного параметра, указывать две функции y_1 и y_2, для которых вероятность покрытия истинного значения неизвестного параметра отрезком (y_1, y_2) равна заданной величине. Функции y_1 и y_2 называются **доверительными границами**, а (y_1, y_2) — **доверительным интервалом**.

Термин доверительность означает уверенность в том, что результат вычислен правильно, а термин 90%-ный доверительный интервал — конкретную степень (90%) уверенности в том, что истинное среднее находится в данном интервале.

Если $x_1, x_2, \ldots$ — последовательность взаимно независимых случайных величин, имеющих одно и то же распределение вероятности с конечным математическим ожиданием ξ и дисперсией σ^2, то согласно центральной предельной теореме выборочное распределение выборочного среднего $\bar{x}$ имеет

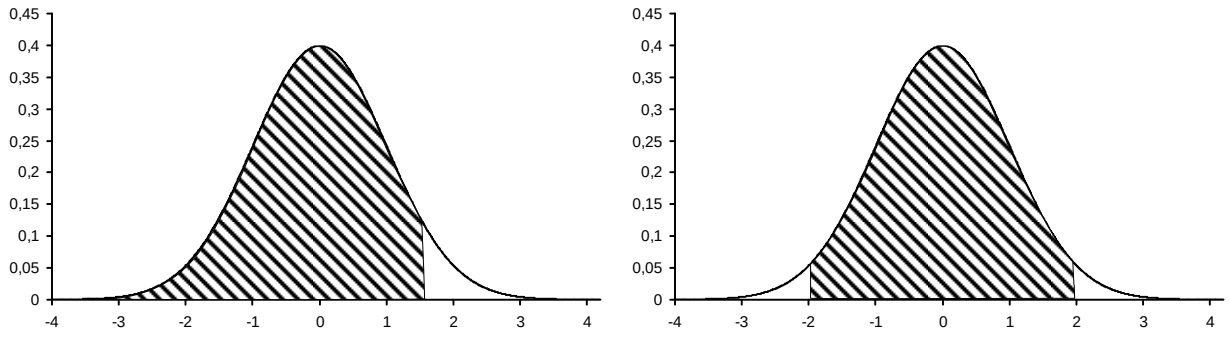
асимптотически нормальное распределение вероятностей с центром $\xi_n=\xi$ и дисперсией $\sigma_n^2 = \sigma^2/n$.

Другими словами выборочное распределение $\bar{x}$ будет распределено приблизительно нормально, несмотря на исходное распределение вероятностей. По мере возрастания размера выборки полученное выборочное распределение будет все ближе стремиться к нормальному распределению. Размер выборки и свойства исходного распределения влияют на степень близости выборочного распределения к нормальному. Крупные выборки будут очень точно соответствовать нормальному распределению, а мелкие — менее точно. Чем больше скошено распределение вероятностей, тем менее точно это соответствие. Для точного соответствия нормальному распределению достаточно иметь небольшое симметричное распределение вероятностей. Насколько небольшим оно может быть? Если исходное распределение симметрично и близко к нормальному, то достаточно 15—20 наблюдений. Для очень скошенных распределений может потребоваться выборка с размером от 40 до 50 наблюдений. Обычно центральная предельная теорема применима для выборок из 30 и более наблюдений.

Теперь можно вывести общую формулу доверительного интервала для выборочного среднего. Для этого выполним сдвиг и масштабирование выборочного среднего $\bar{x}$ таким образом, чтобы выборочное распределение измененного выборочного среднего $\bar{x}'$ имело параметры стандартного нормального распределения (математическое ожидание $\xi = 0$ и дисперсию $\sigma^2 = 1$):

$$\bar{x}' = \frac{\bar{x} - \xi}{\sigma/\sqrt{n}}.$$

Полученное значение удовлетворяет стандартному нормальному распределению и называется *z*-статистикой. Значение по оси абсцисс для кривой стандартного нормального распределения, при котором для случайной переменной Z выполняется условие $P(Z \le z_p) = p$, называется *z*-значением или z_p. Например, $z_{0,95}$ = 1,645, так как 95% площади под кривой находится слева от точки 1,645 (рисунок 7, *а*).



а *б*

Рисунок 7 — Одностороннее и двухстороннее *z*-значение

На рисунке рисунок 7 *а* показано одностороннее *z*-значение, но для получения информации о доверительном интервале нужно иметь двустороннее *z*-значение с заданной вероятностью попадания в центр распределения *p* и вероятностью попадания в один из хвостов распределения α (которая равна 1–*p*). Двусторонний диапазон с заданной вероятностью *p* ограничивается z-значениями $-z_{1-\alpha/2}$ и $z_{1-\alpha/2}$. Иначе говоря, для случайной переменной *Z* выполняется условие $P(-z_{1-\alpha/2} < Z < z_{1-\alpha/2}) = 1 - \alpha = p$. Например, для заданной вероятности $p = 0{,}95$ и $\alpha = 0{,}05$ двустороннее значение $z_{1-0{,}05/2} = z_{0{,}975} = 1{,}96$, т.е. 95% значений нормального распределения находятся в диапазоне от –1,96 до 1,96 (рисунок 7 *б*).

$$P\left(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \xi < \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha .$$

Двустороннее z-значение можно использовать для вывода общего выражения, с помощью которого определяется доверительный интервал:

Таким образом, верхняя и нижняя границы доверительного интервала для равны $\bar{x} \pm z_{1-\alpha/2}\,\sigma/\sqrt{n}$. Например, если $\alpha = 0{,}05$, то $z_{1-0{,}05/2} = 1{,}96$, а 95%-ные верхняя и нижняя границы доверительного интервала равны $\bar{x} \pm 1{,}96\,\sigma/\sqrt{n}$.

Функции Excel можно использовать для вычисления доверительного интервала на основании заданного стандартного отклонения базового распределения. Допустим, что требуется проанализировать пробег двигателей до замены деталей цилиндропоршневой группы. Предположим, что пробег двигателя удовлетворяет нормальному распределению с $\sigma = 48{,}5$, а после эксплуатации партии из 50 двигателей при одинаковых условиях эксплуатации найденное среднее выборочное равно 144 тыс. км. Чему равняется 90%-й доверительный интервал для истинного среднего ξ всего распределения значений пробега для всех двигателей при тех же условиях эксплуатации (иначе говоря, в каких границах от значения 144 тыс. км может находиться истинное среднее)?

Для вычисления 90%-ного доверительного интервала выполните перечисленные ниже действия:

а) создайте новую рабочую книгу (выберите команду меню «Файл ⇒ Создать…» и пункт «Чистая книга»).

б) введите заголовок «Среднее» в ячейку A1, заголовок «Стандартное откл.» в ячейку B1, заголовок «Нижняя граница» в ячейку C1 и заголовок «Верхняя граница» в ячейке D1.

в) введите значение выборочного среднего 144 в ячейку A2.

г) введите в ячейку B2 формулу «=48,5/КОРЕНЬ(50)».

д) введите в ячейку C2 формулу «=A2-B2*НОРМСТОБР(0,95)» (нам нужно вычислить 90%-й доверительный интервал, т.е. $1 - \alpha/2 = 1\text{-}0{,}10/2 = 0{,}95$; функция НОРМСТОБР(*x*) возвращает обратное значение стандартного нормального распределения).

е) введите в ячейку D2 формулу «=A2+B2*НОРМСТОБР(0,95)» (нам нужно вычислить 90%-ный доверительный интервал, т.е. 1 – α/2 = 1-0,10/2 = 0,95; функция НОРМСТОБР(*x*) возвращает обратное значение стандартного нормального распределения).

3.2 t-распределение Стьюдента

До сих пор предполагалось, что нам известно значение σ. А что делать, если значение σ не известно? Одно из решений заключается в том, чтобы вместо σ использовать стандартное отклонение выборки *s*. Однако *s*, являясь оценкой σ, может либо недооценивать, либо переоценивать величину σ, что может привести к ошибке определения доверительного интервала.

В начале XX века Вильям Госсет, задумался над неопределенностью, вызванной подстановкой *s* вместо σ. Он считал, что результирующая ошибка может быть особенно велика для небольших выборок. Госсет обнаружил, что при подстановке *s* вместо σ соотношение

$$\frac{\bar{x}-\xi}{s/\sqrt{n}}$$

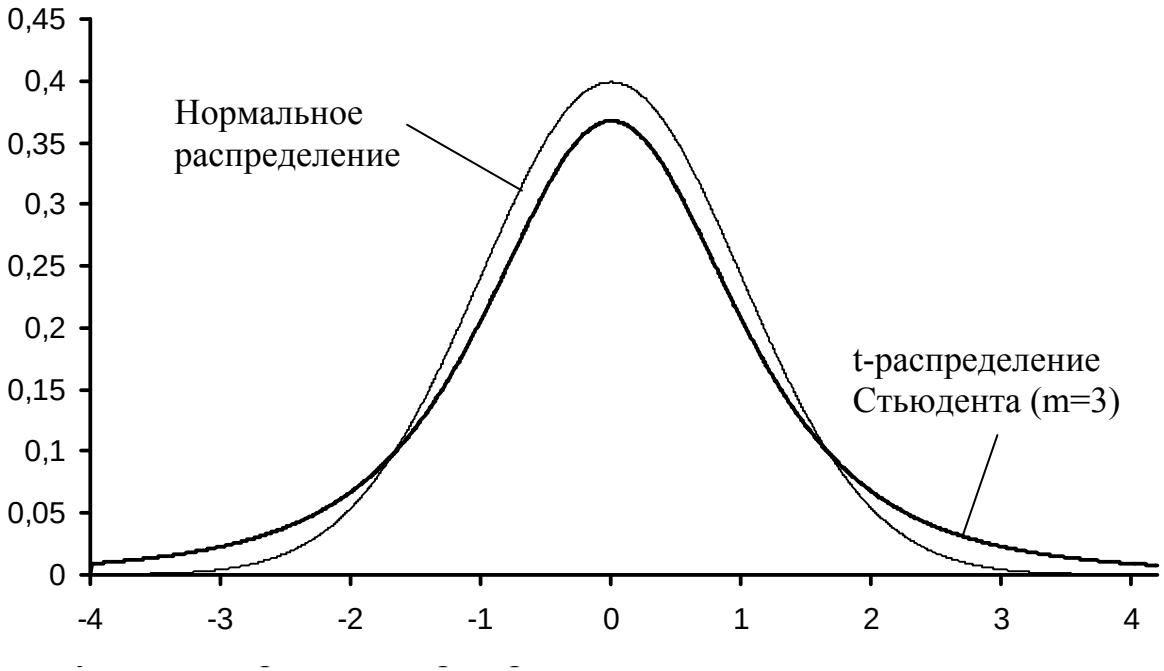
удовлетворяет не стандартному нормальному распределению, а t-распределению Стьюдента. Это распределение вероятности с центром в точке 0, которое характеризуется также наличием степеней свободы. При подстановке *s* вместо σ количество степеней свободы должно быть равно размеру выборки минус единица. Например, выборка из 20 наблюдений имеет 19 степеней свободы. За исключением более длинных хвостов, t-распределение аналогично стандартному нормальному распределению. По мере увеличения размера выборки форма t-распределения приближается к форме стандартного нормального распределения, но малые выборки весьма существенно отличаются от него.

Предположим, что в последнем примере дисперсия распределения неизвестна,

Поскольку нам не известно значение σ, использовать предложенное ранее выражение для доверительного интервала нельзя. В данном случае для определения доверительного интервала необходимо использовать t-распределение. Получаем выражение:

$$\left(\bar{x}-t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}},\bar{x}+t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right).$$

Здесь $t_{p,n-1}$ — это такая точка t-распределения с (n–1) степенями свободы при которой для случайной переменной T выполняется условие $P(T \le t_p) = p$.



свободы со стандартным нормальным распределением

Для вычисления этого значения в Excel предусмотрена функция СТЬЮДРАСПОБР. Однако в ней используется аргумент α, а не (1–α/2). Например, для вычисления значение $t_{1-\alpha/2,n-1}$ используется функция СТЬЮДРАСПОБР (α,*n*-1). Попробуйте на основе этой информации создать 90%-й доверительный интервал для предыдущего примера.

Для создания 90%-го доверительного интервала выполните перечисленные ниже действия:

а) введите в ячейку C3 формулу «=A2-B2*СТЬЮДРАСПОБР(0,1;49)».

б) введите в ячейку D3 формулу «=A2+B2*СТЬЮДРАСПОБР(0,1;49)».

Сравните этот интервал с доверительным интервалом на основе стандартного нормального распределения. Как видите, их размеры очень близки.

При использовании t-распределения для анализа данных предполагалось, что данные удовлетворяют нормальному распределению. А что произойдет, если это опущение окажется неверным? Для t-распределения характерна устойчивость (робастность), которая означает, что если предположение о нормальности незначительно нарушается, то получаемая оценка все еще будет относительно точной. Если распределение данных не очень нарушает предположение об их нормальности, то t-распределение можно использовать с определенной уверенностью.

3.3 Доверительные интервалы для неизвестной вероятности

Простейшая задача, с которой сталкиваются на практике, состоит в оценке неизвестной вероятности *p* события по наблюдаемой частоте $h = \frac{m}{n}$

его появления. Наша задача состоит в том, чтобы найти такие функции $\underline{p} = \underline{p}(h)$ и $\overline{p} = \overline{p}(h)$, чтобы вероятность попадания неизвестной вероятности p в интервал $(\underline{p}, \overline{p})$ была не меньше, чем $1–2\beta$. Иными словами, мы хотим указать такое правило, которое при большом числе его применений может привести к ошибочным заключениям не более чем в 2β части всех случаев.

Английские статистики Клоппер и Э.Пирсон указали такое правило, которое дает гарантию того, что вероятность выхода их доверительного интервала за каждую из границ не превосходит β. Предложенное ими правило состоит в следующем: пусть в n независимых испытаниях с постоянной вероятностью p наступление некоторого события было наблюдено m раз. Тогда в качестве верхней границы $\overline{p}$ доверительного интервала следует взять решение уравнения:

$$\sum_{k=0}^{m} \binom{n}{k} \overline{p}^{k} (1 - \overline{p})^{n-k} = \beta ,$$

а в качестве нижней доверительной границы $\underline{p}$ — решение уравнения:

$$\sum_{k=m}^{n} \binom{n}{k} \underline{p}^{k} (1 - \underline{p})^{n-k} = \beta ,$$

где

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

биномиальные коэффициенты.

Формула

$$\Phi(x) = \sum_{k=0}^{x} \binom{n}{k} \vartheta^{k} (1 - \vartheta)^{n-k}$$

определяет функцию биномиального распределения, а формула

$$p(x) = \binom{n}{x} \vartheta^{x} (1 - \vartheta)^{n-x}$$

плотность биномиального распределения.

В таблице 5 перечислены функции Excel, предназначенные для работы с биномиальным распределением.

Таблица 5 – Функции Excel для работы с биномиальным распределением

Функция	Описание
БИНОМРАСП (x; n; ϑ; ИСТИНА)	Возвращает значение функции биномиального распределения
БИНОМРАСП (x; n; ϑ; ЛОЖЬ)	Возвращает значение плотности биномиального распределения

Рассмотрим пример. Пусть $n = 20$, $m = 4$. Необходимо найти верхнюю и нижнюю границы доверительного интервала с вероятностью попадания в интервал равной 95% ($\beta = 0{,}025$).

Уравнение для верхней границе соответствует функции биномиального распределения с $n = 20$, $x = 4$, $\Phi(x) = 0{,}025$ и неизвестным ϑ.

Уравнение нижней границы:

$$\Phi(n) - \Phi(m) = \Phi(20) - \Phi(3) = 0{,}025,$$

при с $n = 20$ и неизвестным ϑ.

Для выполнения нахождения границ доверительного интервала выполните перечисленные ниже действия:

а) создайте новую рабочую книгу (выберите команду меню «Файл ⇒ Создать…» и пункт «Чистая книга»).

б) введите в ячейку B1 формулу «=БИНОМРАСП(4;20;A1;ИСТИНА)».

в) введите в ячейку B2 формулу «=БИНОМРАСП(20;20;A2;ИСТИНА)-БИНОМРАСП(3;20;A2;ИСТИНА))».

г) введите в ячейки A1 и A2 число 0,2 (начальную оценку нижней и верхней границ доверительного интервала — $m/n = 4/20$).

д) щелкните на ячейке B1.

е) для решения уравнения выберите команду меню «Сервис ⇒ Подбор параметра…».

ж) введите в поле «Значение» 0,025 (β), а в поле «Изменяя значение ячейки» укажите ячейку A1. Нажмите кнопку «ОК».

з) щелкните на ячейке B2.

и) для решения уравнения выберите команду меню «Сервис ⇒ Подбор параметра…».

к) введите в поле «Значение» 0,025 (β), а в поле «Изменяя значение ячейки» укажите ячейку A2. Нажмите кнопку «ОК».

После выполнения данных действий в ячейке A1 будет находиться примерное значение верхней границы диапазона, а в ячейке A2 — нижней границы.

3.4 Проверка гипотез

Доверительные интервалы — это лишь один из способов создания статистических выводов. Другой способ называется проверкой гипотез и основан на создании теории изучаемого явления и проверке ее обоснованности с помощью статистических параметров.

Статистической гипотезой H называется некоторое непротиворечивое множество предположений, относящихся к распределению случайной величины.

Гипотеза проверяется на основании некоторого критерия статистической гипотезы, который представляет собой правило, позволяющее отвергнуть или не отвергнуть гипотезу H на основании выборки.

Идея образования таких правил состоит в том, что пространство выборок, т.е. множество всех возможных результатов наблюдений (множество

всех возможных выборок), разделяют на два непересекающихся подмножества (области). Область называется критической, если гипотеза отвергается в случае принадлежности выборки этой области. Если выборка не принадлежит критической области, гипотеза не отвергается.

В качестве критической области используется некий набор (диапазон) значений статистики теста. Статистика теста — это статистика[5], вычисленная после анализа данных, которые используются для принятия или непринятия гипотезы.

Такое принятие или отбрасывание гипотезы не дает ее логического доказательства или опровержения. Здесь возможны четыре случая:

– гипотеза *H верна* и *принимается* согласно критерию;

– гипотеза *H неверна* и *отвергается* согласно критерию;

– гипотеза *H верна*, *но отвергается* согласно критерию (имеет место так называемая **ошибка первого рода**);

– гипотеза *H не верна*, *но принимается* согласно критерию (имеет место так называемая **ошибка второго рода**).

Вероятности ошибок первого и второго рода однозначно определяются выбором критической области. Вероятность возникновения ошибки первого типа обозначается греческой буквой α, а вероятность возникновения ошибки второго типа — буквой β. Само собой, что наша задача состоит в выборе такой критической области, для которой ошибки первого и второго рода минимальны. Однако оказывается, что при заданном объеме выборки невозможно одновременно сделать и α и β сколь угодно малыми. Поскольку в большинстве случаев ошибка первого рода более важна, задачу обычно приходится ставить иначе: выбрав по тем или иным соображениям α, найти такую критическую область, для которой β принимает наименьшее возможное значение.

Каждый раз, когда приходится проверять гипотезу H, имеют дело не с одной, а по меньшей мере с двумя гипотезами: H и не H.

Проверяемая гипотеза называется нулевой гипотезой H_0. С ней конкурируют альтернативные гипотезы.

Часто альтернативная гипотеза — это именно та гипотеза, которую нужно проверить и принять.

Допустим, требуется оценить способность нового присадки влиять на износ двигателя. Нулевая гипотеза состоит в том, что данное присадка не влияет на износ. Альтернативная гипотеза заключается в том, что присадка влияет на износ (в положительном или отрицательном смысле).

Возьмем в качестве примера завод по производству некоторых деталей. Согласно ранее проведенным исследованиям, количество дефектных деталей в партии в среднем равно 50.

[5] Как уже говорилось ранее статистика представляет собой некую функцию от выборки.

Допустим, что на заводе предлагается внедрить новый технологический процесс, который позволяет сократить количество дефектных деталей с экономией материалов. В результате внедрения нового технологического процесса оказалось, что после анализа выборки из 25 партий среднее количество дефектных деталей в партии равно 44, а стандартное отклонение 15. Можно ли на основании этих данных утверждать, что новый технологический процесс позволяет сократить количество дефектных деталей или число 44 является результатом допустимого случайного отклонения, а внедренный технологический процесс ни на что не влияет?

Сформулируем следующую нулевую гипотезу: среднее количество дефектных деталей в новом технологическом процессе не изменилось (равно 50). В данном случае альтернативная гипотеза — среднее количество дефектных деталей в новом технологическом процессе не равно 50.

Пусть ξ — это среднее количество деталей в нулевой гипотезе в нулевой гипотезе, тогда в ее рамках должно выполняться следующее соотношение (доверительный интервал для выборочного среднего):

$$P\left(-t_{1-\alpha/2,n-1} < \frac{\bar{x}-\xi}{s/\sqrt{n}} < t_{1-\alpha/2,n-1}\right) = 1-\alpha\,.$$

Из него можно получить соотношение:

$$P\left(\xi - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} < \bar{x} < \xi + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}\right) = 1-\alpha\,.$$

Это значит, что выборочное среднее должно находиться в диапазоне $\xi \pm t_{1-\alpha/2,n-1}\, s/\sqrt{n}$ с вероятностью 1–α, если нулевая гипотеза верна. Если α равно уровню значимости, то при выходе значений выборочного среднего за пределы данного диапазона следует отвергнуть нулевую гипотезу и принять альтернативную. Такие значения образуют упомянутую ранее критическую область гипотезы. Наоборот, значения внутри данного диапазона образуют область принятия гипотезы, т.е. при попадании значений выборочного среднего в эту область нулевая гипотеза принимается. Верхняя и нижняя границы области принятия гипотезы называются критическими значениями, так как занимают критически важное положение при определении приемлемости или неприемлемости нулевой гипотезы.

Задание. Используя Excel, проверьте достоверность нулевой гипотезы с уровнем значимости 5%.

Рассмотренный пример является типичным примером двустороннего критерия, в котором предполагается наличие критических значений с обеих сторон. Аналогично, в одностороннем критерии предполагается наличие критических значений только с одной стороны.

Для данного примера односторонний критерий может охватывать следующие две гипотезы:

– нулевая гипотеза: среднее количество дефектных деталей в новом технологическом процессе равно 50;

– альтернативная гипотеза: среднее количество дефектных деталей в новом технологическом процессе меньше 50.

Этот набор гипотез используется в том случае, если по какой-то причине в новом технологическом процессе абсолютно невозможно увеличение среднего количества дефектных деталей. В таком случае при расчете критического значения не нужно делить α пополам:

$$P\left(\xi - t_{1-\alpha,n-1}\frac{s}{\sqrt{n}} < \bar{x}\right) = 1-\alpha\,.$$

Задание. Используя Excel, проверьте достоверность нулевой гипотезы с уровнем значимости 5%.

Учтите, что «статистически значимые» результаты гораздо проще получить при использовании односторонних критериев. Поэтому их следует применять крайне осторожно и только исходя из обоснованных предположений. Помните, что альтернативную гипотезу нужно формулировать до статистического анализа и ни в коем случае не принимать решение о выборе одностороннего критерия после просмотра результатов двустороннего теста.

3.5 Критерий согласия хи-квадрат Пирсона

Для проверки статистических гипотез часто используется удобный для применения критерий согласия χ^2. Данный критерий позволяет проверять гипотезы о том, что относительные частоты $h_k=h\{E_k\}=n_k/n$ случайных событий $E_1, E_2, \ldots, E_r$ в выборке из n независимых наблюдений согласуется с гипотетическими вероятностями $p_k=\mathrm{P}\{E_k\}$. Во многих приложениях каждое событие E_k состоит в том, что некоторая случайная величина x попадает в определенный классовый интервал, так что критерий χ^2 позволяет сравнивать гипотетическое теоретическое распределение величины x с ее эмпирическим распределением.

Например, нам нужно проверить гипотезу H, состоящую в том, что наши наблюдения образуют выборку из n значений $x_1, x_2, \ldots, x_n$, случайной величины X с заданным гипотетическим распределением $\mathbf{P}(X)$. Разобьем все пространство значений X на непересекающиеся области $X_1, X_2, \ldots, X_r$. Тогда p_k будет вероятностью попадания в область X_k.

Разности между h_k и p_k могут послужить основанием для построения статистической характеристики гипотезы, что случайная наблюдаемая величина имеет распределение предполагаемого типа. Если разности малы, то мы склонны принять гипотезу; наоборот, при больших различиях между действительными и теоретическими численностями здравый смысл будет склонять нас отвергнуть гипотезу.

Согласие измеряется с помощью статистики:

$$y = \sum_{\substack{\text{по всем} \\ \text{событиям}}} \frac{(\textit{наблюдаемое} - \textit{ожидаемое количество возникновений события})^2}{\textit{ожидаемое количество возникновений события}} =$$

$$= n\sum_{k=1}^{r} \frac{(h_k - p_k)^2}{p_k} = \sum_{k=1}^{r} \frac{(n_k - np_k)^2}{np_k},$$

распределение которой при $n \to \infty$ стремится к распределению χ^2 с $m = r - 1$ степенями свободы. Если гипотетические вероятности p_k зависят от q неизвестных параметром, то сначала по выборке находят наиболее правдоподобные оценки этих параметром. В этом случае необходимом количество степеней свободы распределения χ^2 принимают равным $m = r - q - 1$.

Критерий χ^2 *отвергает* гипотетические вероятности с уровнем значимости α при $y > \chi^2_{1-\alpha}(m)$. В силу того, что распределение y асимптотически стремится с распределению χ^2, мы можем воспользоваться данным критерием только в тех случаях, когда $np_k > 5$, а лучше $np_k > 10$ (для этого возможно потребуется объединить соседние классовые интервалы).

Вернемся к ранее рассмотренному примеру по износу цилиндров. Очевидно, что интенсивность изнашивания не может быть отрицательной. Поэтому интенсивность изнашивания не может иметь симметричное распределение вероятностей. Изучение гистограммы распределения стоимости квадратного метра недвижимости, позволяет оценить общий характер плотности распределения. Немного похожий характер плотности имею несколько известных распределений вероятности (например, логарифмически нормальное (рисунок 9), гамма).

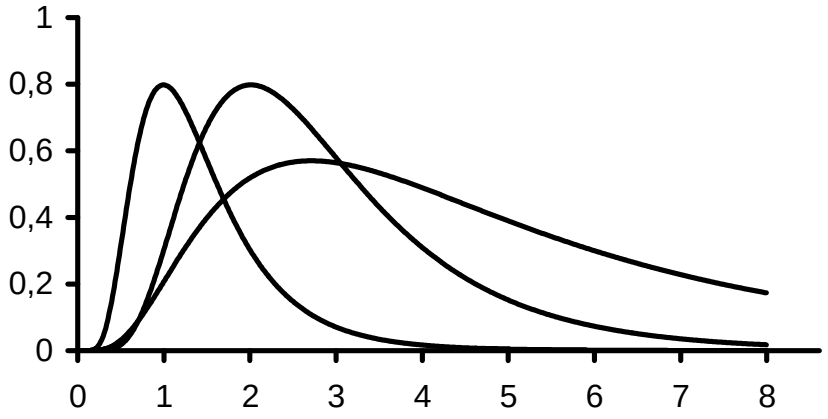


Рисунок 9 — Кривые плотности логарифмически нормального распределения при различных значениях параметров ξ и σ

Сформулируем следующую гипотезу: *стоимость квадратного метра недвижимости имеет логарифмически нормальное распределение*.

Случайная переменная Y имеет логарифмически нормальное распределение с параметрами ξ и σ, если случайная величина $X = \ln(Y)$ имеет нормальное распределение с теми же параметрами ξ и σ.

Для проверки гипотезы выполните следующие действия:

а) запустите программу Excel.

б) найдите файл «Износ цилиндров.xls» и откройте его.

в) выберите команду меню «Файл ⇒ Сохранить как…» и сохраните книгу в файле «Износ цилиндров 2.xls».

г) рассчитайте в ячейках столбца F интенсивность изнашивания каждого цилиндра.

д) рассчитайте в ячейках столбца K логарифм натуральный от стоимости квадратного метра каждой единицы недвижимости (введите в ячейку G2 формулу «=LN(F2)» и скопируйте ее в остальные ячейки). Это позволит перейти от логарифмически нормального распределения к нормальному.

е) введите в ячейку H2 формулу «=СРЗНАЧ(G:G)» (выборочное среднее для логарифма от интенсивности изнашивания), а в ячейку I2 формулу «=СТАНДОТКЛОН(G:G)» (стандартное отклонение для логарифма от интенсивности изнашивания).

ж) введите в ячейку H4 формулу «=СЧЁТ(G:G)» (размер выборки).

з) введите в ячейку I4 формулу «=ОКРУГЛВНИЗ(H4/10;0)» (количество интервалов разбиения из расчета, чтобы ожидаемое количество попаданий в интервал было не меньше 10).

и) введите в ячейку H7 число 0, в ячейку H8 формулу «=H7+1/I$4». Скопируйте ячейку H8 вниз так, чтобы последнее получившееся значение было равно (примерно равно) 1 (в нашем случае в ячейки H9:H20). В результате получаем разбиение вероятности от 0 до 1 на равные интервалы, количество интервалов находится в ячейке I4.

к) введите в ячейку I8 формулу «=НОРМОБР(H8;H$2;I$2)». Скопируйте ее вниз до конца границ интервалов вероятности (в нашем случае в ячейки I9:I20). Последние число при значении границы интервала точно равной 1 соответствует бесконечности, поэтому можем в последней ячейке получить ошибку. Чтобы избавиться от ошибки введите в ячейку (I20) число больше максимального в выборке (формулу «=МАКС(G:G)+1»). В результате получаем разбиение области значения функции на равновероятные интервалы.

л) введите в ячейку J8 формулу «="<"&ТЕКСТ(I8;"0,0000")» (задает условие верхней границы интервала в виде текстовой строки). Скопируйте ячейку J8 вниз (в ячейки J9:J20).

м) введите в ячейку K8 формулу «=СЧЁТЕСЛИ(G:G;J8)» (подсчет количества элементов выборки, удовлетворяющей указанному условию). Скопируйте ячейку K8 вниз (в ячейки K9:K20).

н) введите в ячейку L8 формулу «=(K8-K7)/H$4». Скопируйте ячейку L8 в ячейки L9:L20 (получаем относительные частоты попадания значений выборки в интервалы).

о) введите в ячейку M8 формулу «=((P8-1/M$4)^2)/(1/M$4)*L$4». Скопируйте ячейку M8 в ячейки M9:M20. В ячейке M21 подсчитайте сумму по

ячейкам M8:M20 (теперь ячейка M21 содержит статистику критерия согласия).

п) введите в ячейку M22 формулу «=1-ХИ2РАСП(M21;I4-3)» (I4-3 — количество степеней свободы, равное количеству интервалов минус количество оцениваемых параметров (среднее и стандартное отклонение) и минус один). Теперь в этой ячейке находится уровень значимости, с которым мы можем отвергнуть гипотезу.

Сделайте вывод, есть ли у нас основания отвергнуть гипотезу.

Дик Дмитрий Иванович

ОБРАБОТКА РЕЗУЛЬТАТОВ ЭКСПЕРИМЕНТА С ПОМОЩЬЮ MICROSOFT EXCEL

ЧАСТЬ 1

Лабораторный практикум
по дисциплине «Компьютерные технологии в науке и образовании» для студентов квалификации (степени) Магистр направления 190500 (552100)

Редактор Е.А. Устюгова

Подписано к печати	Формат 60×84 1/16	Бумага тип. №1
Печать трафаретная	Усл. печ. л. 2,0	Уч.–изд. л. 2,0
Заказ	Тираж 30	Цена свободная

Редакционно-издательский центр КГУ.
640069, г. Курган, ул. Гоголя, 25.
Курганский государственный университет.