Глава 15. Теория вероятностей и математическая статистика

Задачи, решаемые средствами теории вероятностей и математической статистики, имеют огромную практическую важность, связанную, прежде всего, с контролем качества продукции на промышленных предприятиях. Решение такого рода проблем влечет за собой применение весьма сложного математического аппарата с внушительным объемом вычислительных работ. Поэтому с самых первых дней рождения вычислительной техники ЭВМ активно начали применять для статистической обработки данных.

Система Mathcad имеет огромные возможности в сфере решения задач математической статистики. В Mathcad имеется значительное количество специальных статистических функций, позволяющих сократить до минимума время решения любой поставленной проблемы. Разнообразие генераторов случайных чисел, используя которые, можно создавать последовательности, распределенные по любому из важных статистических законов, дает возможность моделировать всевозможные случайные процессы.

Также, используя Mathcad, можно предельно просто строить гистограммы высоких художественных качеств, проводить обработку выборки, проверять статистические гипотезы благодаря наличию встроенных функций практически всех теоретических распределений. В общем, вам вряд ли встретится задача, которую при правильном подходе вы не сможете решить, обратившись к возможностям Mathcad.

15.1. Комбинаторика

Комбинаторика — это раздел математики, изучающий способы подсчета количества элементов в конечных множествах. В теории вероятности она применяется в случае решения простейших задач, когда вероятность того или иного события вычисляется непосредственно.

Так как в Mathcad на панели Calculator (калькулятор) имеется оператор вычисления факториала (сочетание клавиш Shift+1), то наиболее просто можно решать задачи комбинаторики, непосредственно задавая соответствующие выражения. Однако с учетом того, что при этом могут получаться весьма громоздкие формулы, в некоторых случаях лучше все-таки использовать специальные встроенные функции Mathcad.

- permut(n,k) — от англ. permutation (размещение). Функция вычисляет количество возможных размещений из n по k. В случае непосредственного задания формул ей соответствует выражение:

$$\text{permut}(n,k) = \frac{n!}{(n-k)!}$$

- combin(n,k) — от англ. combination (сочетание). Эта встроенная функция служит для вычисления количества сочетаний из n элементов по k позициям. Вместо нее можно использовать и известную формулу:

$$\text{combin}(n,k) = \frac{n!}{k!\,(n-k)!}$$

Количество перестановок множества из n элементов можно подсчитать непосредственно как n!.

Приведем пример использования функций комбинаторики для решения следующей задачи.

Пример 15.1. В ящике находятся 5 красных, 8 синих, 2 желтых, 13 белых и 45 черных кубиков. Какова вероятность, что среди вытянутых наудачу 5 кубиков все окажутся разного цвета

$$C(n,k) := \text{combin}(n,k)$$

$$\frac{C(5,1)\cdot C(8,1)\cdot C(2,1)\cdot C(13,1)\cdot C(45,1)}{C(73,5)} = 3.116\times 10^{-3}$$

Вероятность успеха в поставленной задаче оказалось мизерной — всего 0.3116 %.

Обратите внимание, что мы переопределили функцию сочетаний, чтобы вид решения был ближе к классическому. Кстати, подобную задачу вы не найдете в учебниках, так как расчет полученного выражения слишком сложен для человека (лишь один 73! чего стоит!).

15.2. Определение характеристик непрерывной случайной величины

Если вы студент технического или естественного факультета, то на практических занятиях и дома вам придется решать множество задач, связанных с определением функции распределения, ее плотности вероятности, математического ожидания, дисперсии, центральных и начальных моментов.

В случае большинства задач, приведенных в сборниках, невозможно использовать встроенные функции Mathcad, так как они рассчитаны для статистической обработки данных и не могут быть применены для решения многих абстрактных примеров. Так что, чтобы облегчить себе расчетную работу, вам придется вспомнить принципы проведения различных вычислений в Mathcad и создать алгоритм самостоятельно. В случае дискретной случайной величины это совсем не сложно и требует лишь корректного использования оператора суммы или ранжированных переменных. Более интересно

решаются задачи, связанные с непрерывными случайными величинами. Рассмотрим конкретный пример.

Пример 15.2. Плотность распределения непрерывной случайной величины X задана на всей числовой оси:

$$P(X) = \frac{2 \cdot C}{X^2 + C}$$

Найти значение параметра C, функцию распределения, вероятность попадания случайной величины на отрезок от –3 до 3, математическое ожидание, дисперсию, моду, медиану, центральные моменты

Так как заданий в задаче поставлено довольно много, разберем ее по пунктам.

Для начала зададим плотность распределения как функцию:

$$P(X) := \frac{2 \cdot C}{X^2 + C}$$

Найти значение коэффициента можно, вспомнив, что плотность распределения является нормированной функцией, и площадь, которую она ограничивает, должна быть равна значению полной вероятности, то есть 1. А это означает, что, подсчитав эту площадь аналитически, как несобственный интеграл (вполне понятно, что для этого нужно использовать символьное интегрирование), и выразив из полученного выражения C (для этого можно применить оператор символьного решения уравнений solve), мы сможем получить для него корректное значение. Объединив два преобразования в одно выражение, имеем:

$$\int_{-\infty}^{\infty} P(X)\,dX = 1 \text{ solve}, C \rightarrow \frac{1}{4 \cdot \pi^2}$$

Выполним подстановку полученной величины C в исходное выражение. Наиболее просто это можно сделать, использовав оператор символьной замены переменных substitute:

$$P(X) \text{ substitute}, C = \frac{1}{4 \cdot \pi^2} \rightarrow \frac{1}{2 \cdot \pi^2 \cdot \left(X^2 + \frac{1}{4 \cdot \pi^2} \right)}$$

Переопределив функцию распределения плотности, проверим, соблюдается ли условие нормировки, а также построим ее график (рис. 15.1):

$$P(X) := \frac{1}{2 \cdot \pi^2 \cdot \left(X^2 + \frac{1}{4 \cdot \pi^2} \right)} \qquad \int_{-\infty}^{\infty} P(X)\,dX \rightarrow 1$$

На данном этапе логичным будет определить моду случайной величины — значение X, которому соответствует наибольшее значение плотности вероятности. Наиболее просто это можно сделать, использовав встроенную функцию численного определения максимума Maximize:

$$X := 1$$

$$\text{Maximize}(P, X) = 0$$

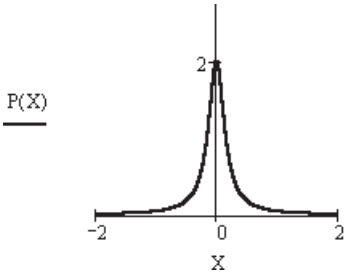

Рис. 15.1. График плотности вероятности

Глядя на график (см. рис. 15.1), можно было предположить, что максимум плотности смещен немного вправо относительно оси ординат. Однако визуальное впечатление на этот раз оказывается обманчивым: максимум лежит в точке 0, и рассматриваемая функция абсолютно симметрична относительно вертикальной оси. Кстати, к такому же выводу можно было прийти, просто вычислив аналитически производную:

$$\frac{d}{dX}P(X) \rightarrow \frac{-1}{\pi^2 \cdot \left(X^2 + \frac{1}{4 \cdot \pi^2}\right)^2} \cdot X$$

Далее найдем, исходя из определения (см. подразд. 15.3.1), функцию распределения и построим ее график (рис. 15.2):

$$\int_{-\infty}^{x} P(X)\,dX \rightarrow \frac{\operatorname{atan}(2 \cdot x \cdot \pi)}{\pi} + \frac{1}{2} \qquad F(X) := \frac{\operatorname{atan}(2 \cdot X \cdot \pi)}{\pi} + \frac{1}{2}$$

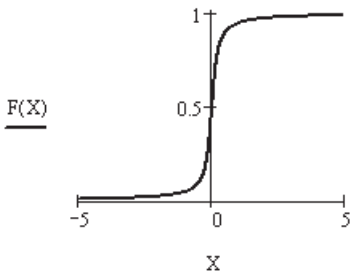

Рис. 15.2. График функции распределения

Как видно из графика, функция распределения у нас получилась корректной по всем показателям и очень похожей на соответствующие функции «настоящих» распределений.

Зная функцию распределения, очень просто можно найти теперь вероятность того, что случайная величина попадет в интервал от –3 до 3:

$$F(3) - F(-3) \rightarrow 2 \cdot \frac{\operatorname{atan}(6 \cdot \pi)}{\pi} \qquad F(3) - F(-3) = 0.966$$

Учитывая то, что наше распределение симметрично, заранее можно утверждать, что медиана (медиана — это точка, функция вероятности в которой принимает свое среднее значение, то есть 0.5) должна быть равна 0. Проверим это, решив уравнение, составленное в соответствии с определением медианы:

$$\frac{\operatorname{atan}(2 \cdot X \cdot \pi)}{\pi} + \frac{1}{2} = \frac{1}{2} \text{ solve}, X \rightarrow 0$$

Далее вычислим математическое ожидание. Для этого придется подсчитать несобственный интеграл от произведения случайной величины на плотность ее распределения. При попытке сделать это символьно система возвращает само выражение интеграла:

$$\int_{-\infty}^{\infty} P(X) \cdot X \, dX \rightarrow \int_{-\infty}^{\infty} \frac{1}{2 \cdot \pi^2 \cdot \left(X^2 + \frac{1}{4 \cdot \pi^2} \right)} \cdot X \, dX$$

Это означает, что аналитическое решение не найдено Mathcad. Следовательно, нужно попробовать использовать численный метод:

$$M := \int_{-\infty}^{\infty} P(X) \cdot X \, dX$$

$$M = 0$$

Впрочем, результат, с учетом симметричности плотности распределения и значения моды, вполне очевиден и без проведения каких-либо специальных вычислений.

Более интересный результат получается при вычислении дисперсии:

$$\int_{-\infty}^{\infty} P(X) \cdot X^2 \, dX \rightarrow \infty$$

Нужно отметить, что в противоположность вычислению математического ожидания, численным методом интеграл дисперсии не посчитается. Связано же это с тем, что численные методы интегрирования нельзя применять к вычислению расходящихся интегралов.

Наиболее удачным при вычислении центральных моментов будет задать функцию общего вида:

$$\mu(k) := \int_{-\infty}^{\infty} X^k \cdot P(X) \, dX$$

$$\mu(3) = 0 \qquad \mu(4) \rightarrow \infty$$

Приведенный пример решения обычной задачи по теории вероятностей демонстрирует, как значительно может помочь в этом Mathcad. Такая объемная задача решается в нем за несколько минут, а если бы мы просчитывали все это на бумаге, то понадобилось бы никак не меньше часа. И необходимо учитывать, что вероятность допущения ошибки была бы при этом значительно выше.

15.3. Теоретические распределения

Законов, описывающих распределение случайных величин или их характеристик, в статистике существует довольно много, впрочем, практический интерес представляют лишь несколько из них. И для всех в Mathcad имеются специальные встроенные функции. Причем для большинства распределений в систему встроены не только функции плотности, но и функции, позволяющие вычислять сопутствующие величины — функции распределения и квантили. Кроме того, для всех распределений в Mathcad имеются генераторы случайных величин, позволяющие создавать векторы из нужного количества псевдослучайных чисел.

В этом разделе мы рассмотрим, каким образом можно вычислить основные характеристики теоретических распределений, а также подробно разберем наиболее практически важные распределения.

15.3.1. Основные характеристики распределений

Главной характеристикой непрерывно распределенной случайной величины является плотность вероятности. В общем случае она равна производной функции распределения и понимается как отношение вероятности попадания случайной величины в узкую окрестность определенного значения к размеру этой окрестности. С помощью плотности вероятности выводятся все важнейшие характеристики непрерывной случайной величины, такие как дисперсия или математическое ожидание.

Все функции теоретических плотностей в Mathcad именуются по следующему принципу: в начале пишется приставка d (от англ. density — плотность), а затем вводится соответствующий корень слова. Например, плотность для нормального распределения задается функцией dnorm(x,μ,σ), а t-распределения Стьюдента — dt(x,f).

В случае дискретных случайных величин в Mathcad также существуют функции (с приставкой d). Они служат для вычисления вероятности того, что случайная величина примет определенное конкретное значение. Однако, естественно, говорить при этом о плотности распределения некорректно, поскольку само это понятие применимо только к непрерывным распределениям.

Чтобы ввести нужную функцию плотности вероятности, удобно использовать окно Insert Function (вызывается сочетанием клавиш Ctrl+Shift+F). Нужные встроенные функции располагаются в нем в списке Probability Density (Плотность вероятности).

Второй важнейшей характеристикой теоретического распределения является так называемая функция распределения. В общем случае она определяет, какова вероятность того, что случайная величина примет значение, меньшее X:

$$F(x) = P(x \le X)$$

В случае непрерывных случайных величин функция распределения определяется интегрированием плотности вероятности от левой границы области определения до X:

$$F(X) = \int_{-\infty}^{X} p(x)\,dx$$

Для дискретных случайных величин функция распределения задается как соответствующая сумма (где floor(X) означает, что суммируются вероятности значений меньших X):

$$F(X) = \sum_{k=0}^{\text{floor}(X)} p(k)$$

Важным свойством функции распределения является то, что она позволяет находить вероятность попадания случайной величины в числовой интервал без применения интегрирования:

$$P(A < x \le B) = F(B) - F(A)$$

В Mathcad функции распределения отличаются от соответствующих плотностей только тем, что их имена начинаются с приставки p (probability — вероятность). Так, например, для биноминального распределения функция распределения — это pbinom(k,n,p), а для нормального — pnorm(x,μ,σ).

Чтобы ввести функцию распределения с помощью окна Insert Function (Вставить функцию), обратитесь к списку Probability Distribution (Распределение вероятности). Здесь же расположены и соответствующие квантили.

Функция, обратная к функции распределения, называется квантилью. Необходимость введения этого понятия возникла в связи с широкой потребностью в приложениях самого разного рода отвечать на вопрос, чему равняется X, если $F(X)=\alpha$.

В Mathcad определены обратные функции всех важнейших распределений и, в общем, они весьма неплохо справляются с численным определением квантилей. Однако всегда следует предельно внимательно подходить к заданию встроенных функций, предназначенных для вычисления квантилей, поскольку, перепутав последовательность введения параметров, вы получите неверный результат. Помните, что вероятность a всегда определяется на первом месте.

В случае дискретных распределений, для которых не существует обратной функции вероятности, в качестве квантили Mathcad возвращает наибольшее целое число, для которого значение функции распределения меньше либо равно α.

Задаются функции квантилей в Mathcad добавлением к соответствующим корням слова приставки q.

Важнейшей возможностью системы Mathcad в области статистики является то, что она позволяет создавать выборки случайных величин, распределенные по любому из теоретических законов с произвольными параметрами. Эта возможность весьма широко используется прежде всего для модуляции всевозможных случайных процессов. Таким образом можно создать коррелированный вектор данных, в которых генератор нормального распределения создаст эффект случайной ошибки нужной величины стандартного отклонения, что позволит беспристрастно испытать, например, встроенные функции регрессии.

Задаются функции случайных величин добавлением приставки r (от англ. random — случайный) к корню термина соответствующего распределения. При этом первый параметр всегда определяет количество величин в случайном векторе. Так, например, чтобы задать вектор нормально распределенных случайных значений с математическим ожиданием 3 и среднеквадратичным отклонением 2, образованный 1000 величинами, нужно ввести rnorm(1000,3,2).

Подробнее о задании и использовании функций случайных величин мы поговорим в подразделе 15.9.

15.3.2. Дискретные распределения

Биноминальное распределение

Биноминальным называется закон для вычисления вероятностей, определяемый формулой Бернулли:

$$P_n(k) = \frac{n!}{k! \cdot (n-k)!} \cdot p^k \cdot (1-p)^{n-k}$$

Термин «биноминальный» применяется к данному закону распределения вероятностей в связи с тем, что его формула выражает общий член разложения бинома Ньютона.

Биноминальное распределение служит для вычисления вероятности того, что некоторое событие наступит в n испытаниях k раз, если вероятность его наступления постоянна при каждом испытании и равна p.

В Mathcad приведенной выше формуле соответствует функция dbinom(k,n,p). Приведем пример решения наиболее характерной задачи с помощью этой функции.

Пример 15.3. Найти вероятность того, что при 10 бросках монеты количество выпадений орла и решки совпадет, ту же вероятность просчитать для 100, 1000, 100 000 бросков

Вероятность выпадения орла в каждом испытании постоянна и равна 0,5. Тогда вероятность того, что при 10 бросках орел выпадет 5 раз, равна

$$\mathrm{dbinom}(5, 10, 0.5) = 0.246$$

Этот же результат можно получить и при непосредственном подсчете вероятности по формуле Бернулли:

$$\frac{10!}{5! \cdot 5!} \cdot \left(\frac{1}{2}\right)^5 \cdot \left(\frac{1}{2}\right)^5 = 0.246$$

Вычислим аналогичную вероятность для большего количества бросков монеты:

$$\mathrm{dbinom}(50, 100, 0.5) = 0.08 \qquad \mathrm{dbinom}(500, 1000, 0.5) = 0.025$$

$$\mathrm{dbinom}(50000, 100000, 0.5) = 2.523 \times 10^{-3}$$

С помощью встроенной функции pbinom(k,n,p) можно предельно просто решать множество интересных задач, найти ответ для которых обычным подсчетом на бумаге было бы весьма проблематично.

Пример 15.4. Какова вероятность того, что при 1000 бросках орел выпадет от 450 до 550 раз?

$$\mathrm{pbinom}(550, 1000, 0.5) - \mathrm{pbinom}(450, 1000, 0.5) = 0.998$$

Сравним результат с вероятностью того, что из 10 бросков орел выпадет 4, 5, или 6 раз:

$$\mathrm{pbinom}(6, 10, 0.5) - \mathrm{pbinom}(4, 10, 0.5) = 0.451$$

Расчет показывает, что при большом количестве испытаний орел и решка выпадут приблизительно одно и то же количество раз, чего не наблюдается при небольшом количестве бросков. Это правило обобщено в одном из принципов известного закона больших чисел.

Иногда бывает полезной и функция, вычисляющая квантиль биноминального распределения (особенности употребления этого термина в случае дискретных распределений были оговорены выше). Задается она как qbinom(α,n,p), где α — вероятность наступления события.

Пример 15.5. Сколько раз выпал орел при 10 бросках, если вероятность этого события равна 0.34 (0.95)?

$$\text{qbinom}\left(0.95, 10, \frac{1}{2}\right) = 8 \qquad \text{qbinom}\left(0.34, 10, \frac{1}{2}\right) = 4$$

Полученные результаты можно интерпретировать следующим образом: при 10 бросках с вероятностью 0.34 орел выпадет 4 или меньше раз (при вероятности 0.95 соответственно будет 8 или меньше выпадений).

Генератором случайных чисел, распределенных по биноминальному закону, является в Mathcad функция rbinom(N,n,p), где N — количество элементов случайного вектора.

Распределение Пуассона

Распределение Пуассона является частным случаем биноминального распределения и описывается как:

$$P_n(k) = \frac{\lambda^k \cdot e^{-\lambda}}{k!},$$

где λ=n·p (это произведение считается постоянной величиной). Приведенная формула применяется для облегчения расчетов в случае большого количества испытаний и малой вероятности появления события.

Для задания четырех характеристических функций распределения Пуассона используйте соответствующие приставки к корню pois(k(α),λ).

Пример 15.6. Завод отправил потребителю 6000 доброкачественных изделий. Вероятность повреждения в пути равна 0.03 %. Какова вероятность того, что будет испорчено 10 изделий?

$$n := 6000 \qquad p := 0.0003 \qquad k := 10$$

Определяем параметр λ и подсчитываем вероятность:

$$\lambda := n \cdot p$$

$$\text{dpois}(k, \lambda) = 1.626 \times 10^{-5}$$

Раньше для вычисления вероятностей по формуле Пуассона использовали специальные таблицы. В настоящее же время, в связи с интенсивным развитием компьютерной математики, подобные расчеты теряют свое значение.

Другим распространенным обобщением формулы Бернулли является известная теорема Муавра–Лапласа, позволяющая вычислять вероятности при больших количествах испытаний.

К сожалению, встроенных функций, реализующих подсчет исходя из локальной и интегральной формул Лапласа, в Mathcad нет. Однако при необходимости вы можете их задать и самостоятельно.

Пример 15.7. Вероятность рождения мальчика равна 0,51. Найти вероятность того, что среди 100 новорожденных окажется 50 мальчиков

$$n := 100 \qquad k := 50 \qquad p := 0.51 \qquad q := 1 - p$$

Воспользуемся локальной теоремой Лапласа, поскольку n=100 — достаточно большое число.

$$x := \frac{k - n \cdot p}{\sqrt{n \cdot p \cdot q}}$$

$$P(k) := \frac{1}{\sqrt{n \cdot p \cdot q}} \cdot \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-x^2}{2}}$$

$$P(k) = 0.0782$$

Пример 15.8. Вероятность появления события в каждом из 2100 независимых испытаний равна 0,7. Найти вероятность того, что событие появится не менее 1470 и не более 1500 раз

$$n := 2100 \qquad k_1 := 1470 \qquad k_2 := 1500 \qquad p := 0.7 \qquad q := 1 - p$$

Для решения задачи воспользуемся интегральной теоремой Лапласа:

$$x' := \frac{k_1 - n \cdot p}{\sqrt{n \cdot p \cdot q}} \qquad x'' := \frac{k_2 - n \cdot p}{\sqrt{n \cdot p \cdot q}}$$

$$P(x) := \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-x^2}{2}}$$

$$\int_{x'}^{x''} P(x)\, dx = 0.423$$

Геометрическое распределение

Если вероятность наступления события описывается формулой:

$$P(X = k) = (1 - p)^{k-1} \cdot p,$$

то считается, что случайная величина распределена по геометрическому закону. Определяет же геометрическое распределение вероятность наступления некоторого события на k-м испытании, если вероятность его наступления одинакова при каждом опыте. Обратите внимание, что формула вероятности является, по сути, общим членом убывающей геометрической прогрессии, откуда и название рассматриваемого распределения.

В Mathcad функцией, служащей для вычисления вероятности наступления события, подчиненного геометрическому закону, является `dgeom(k,p)`, где `k` — количество испытаний, `p` — вероятность наступления события в одном испытании.

Наиболее известной задачей, в которой применяется формула геометрической вероятности, является задача об орудии.

Пример 15.9. Вероятность попадания в цель из пушки равна 0.11, какова вероятность того, что цель будет поражена на 2-м (5-м, 10-м, 20-м) выстреле?

$$\text{dgeom}(2, 0.11) = 0.087 \qquad \text{dgeom}(10, 0.11) = 0.034$$

$$\text{dgeom}(5, 0.11) = 0.061 \qquad \text{dgeom}(20, 0.11) = 0.011$$

В том же случае, если поставить вопрос: какова вероятность того, что цель будет поражена до k-го выстрела, то для решения этой задачи придется использовать функцию распределения `pgeom(k,p)`.

Пример 15.10. Определение вероятности попадания в цель до k-го выстрела

$$\text{pgeom}(2, 0.11) = 0.295 \qquad \text{pgeom}(10, 0.11) = 0.722$$

$$\text{pgeom}(5, 0.11) = 0.503 \qquad \text{pgeom}(20, 0.11) = 0.913$$

Очень часто задача ставится следующим образом: сколько выстрелов нужно сделать, чтобы попасть в цель с вероятностью α? В случае подобных задач нужно использовать встроенную функцию квантилей `qgeom(α,p)`.

Пример 15.11. Определение количества попаданий в цель

$$\text{qgeom}(0.25, 0.11) = 2 \qquad \text{qgeom}(0.75, 0.11) = 11$$

$$\text{qgeom}(0.5, 0.11) = 5 \qquad \text{qgeom}(0.999, 0.11) = 59$$

Существует в Mathcad и генератор случайных чисел, распределенных по геометрическому закону `rgeom(N,p)`, где N — количество элементов в векторе.

Гипергеометрическое распределение

Гипергеометрическое распределение решает задачи, схожие с теми, для которых применяется биноминальный закон, с единственным (но иногда принципиальным) отличием — объем выборки в этом случае не является постоянным (то есть в постановке большинства примеров это означает, что извлечение элементов производится без возврата).

В Mathcad вероятность наступления события, подчиняющегося гипергеометрическому закону, вычисляется с помощью встроенной функции `dhypergeom(m,a,b,n)`. Самым сложным моментом в ее применении является предельная запутанность в ее параметрах, последовательность записи которых совершенно не очевидна. Попробуем разобраться на конкретном примере, как нужно задавать параметры `dhypergeom`.

Пример 15.12. В партии из 12 деталей имеется 8 стандартных. Найти вероятность того, что среди 6 отобранных наугад деталей 5 окажутся стандартными

Пусть:

m — количество стандартных деталей среди отобранных (5);

a — количество стандартных деталей во всей партии (8);

b — количество нестандартных деталей во всей партии (4) (обратите внимание на особенность задания этого параметра — традиционно более принято использовать формулы, в которых фигурирует непосредственно полный объем партии);

n — количество всех отобранных деталей (6).

Вероятность соответствующего события равна:

$$\text{dhypergeom}(5,8,4,6) = \frac{8}{33}$$

Аналогично всем остальным распределениям, в Mathcad для гипергеометрического распределения существуют функции phypergeom(m,a,b,n) и qhypergeom(p,a,b,n). С помощью данных функций можно решать задачи, подобные приведенной выше.

Пример 15.13. Чему равна вероятность того, что в выборке окажется меньше 5 стандартных деталей (остальные условия те же, что и в предыдущем примере) и сколько их там может быть при вероятности 0,7?

$$\text{phypergeom}(5,8,4,6) = 0.97 \qquad \text{qhypergeom}(0.7,8,4,6) = 4$$

Существует и гипергеометрический генератор случайных чисел — rhypergeom(N,a,b,n).

Отрицательное биноминальное распределение

Помимо обычного, в Mathcad имеются функции и отрицательного биноминального распределения (negative binomial distribution). Это распределение имеет лишь теоретическое значение и применяется, например, для определения количества неудачных экспериментов до n-го успешного эксперимента в ряду независимых испытаний Бернулли, если вероятность успеха равна p, либо для подсчета качественных изделий, отобранных до появления n-го поврежденного изделия.

Задаются функции отрицательного биноминального распределения с помощью корня nbinom и соответствующих приставок. Например, pnbinom(k,n,p) — функция вероятности для рассматриваемого распределения.

15.3.3. Непрерывные распределения

Равномерное распределение

Наиболее простым непрерывным распределением является равномерное распределение, то есть имеющее одинаковую плотность на всем промежутке определения.

Случайные величины, распределенные по равномерному закону, имеют конечные границы интервалов изменения. Зная их, совсем не трудно с учетом нормировки вывести формулу плотности вероятности ($X \in (a,b)$):

$$P(X) = \frac{1}{b-a}$$

В Mathcad данной формуле соответствует функция dunif(x,a,b) (от англ. uniform distribution —равномерное распределение).

Функция равномерного распределения задается в Mathcad как punif(x,a,b), где a и b — границы интервала изменения случайной величины (рис. 15.3). Общую же ее формулу можно найти, использовав символьное интегрирование:

$$\int_a^X \frac{1}{b-a}\,dx \rightarrow \frac{-X}{-b+a} + \frac{a}{-b+a}$$

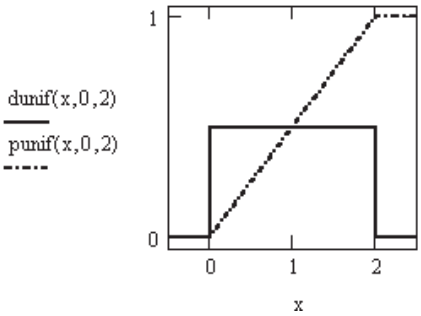

Рис. 15.3. Плотность вероятности и функция равномерного распределения

На практике равномерное распределение используется при работе с округленными величинами.

Пример 15.14. Цена деления шкалы измерительного прибора равна 0,2. Показания прибора округляют до ближайшего целого деления. Найти вероятность того, что при отсчете будет сделана ошибка: а) меньшая 0,04; б) большая 0,02

Ошибка округления есть случайная величина, равномерно распределенная на промежутке между соседними целыми делениями. Рассмотрим в качестве такого деления интервал (0; 0,2) (рис. 15.4). Округление может проводиться как в сторону левой границы — 0, так и в сторону правой — 0,2, значит, ошибка, менее либо равная 0,04, может быть сделана два раза, что необходимо учесть при подсчете вероятности:

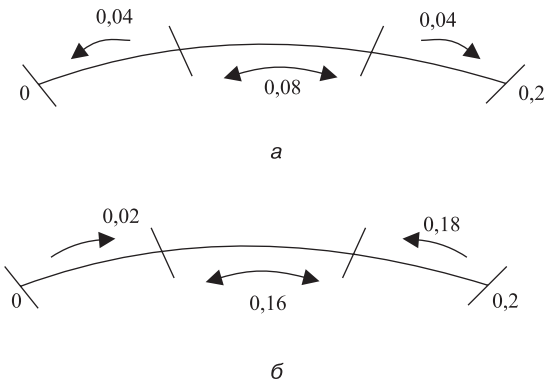

Рис. 15.4. Варианты округления показаний прибора

$$a := 0 \qquad b := 0.2$$

$$\text{punif}(0.04, a, b) + \text{punif}(0.04, a, b) = 0.4$$

Для второго случая величина ошибки может превышать 0,02 также с обеих границ деления, то есть она может быть либо больше 0,02, либо меньше 0,18. Тогда вероятность появления такой ошибки:

$$\text{punif}(0.18, 0, 0.2) - \text{punif}(0.02, 0, 0.2) = 0.8$$

С помощью символьного интегрирования можно найти математическое ожидание и дисперсию равномерно распределенной случайной величины.

Пример 15.15. Найти математическое ожидание и дисперсию случайной величины X, равномерно распределенной в интервале (a, b)

Используя хорошо знакомую формулу, определяем математическое ожидание:

$$\int_a^b \frac{x}{b-a}\,dx \text{ simplify } \rightarrow \frac{1}{2}\cdot b + \frac{1}{2}\cdot a$$

Полученное выражение можно использовать для вычисления дисперсии:

$$\int_a^b \frac{\left(x - \frac{a+b}{2}\right)^2}{b-a}\,dx \text{ factor } \rightarrow \frac{1}{12}\cdot(b-a)^2$$

Обратите внимание, мы получили общие формулы, которые вы можете использовать для расчета математического ожидания и дисперсии случайной величины, распределенной равномерно на конкретном числовом интервале.

Нормальное распределение

Согласно известной центральной предельной теореме Ляпунова, в том случае, если случайная величина определяется суммой большого количества взаимно независимых случайных величин, вклад каждой из которых в общую сумму ничтожно мал, то такая случайная величина распределена по нормальному (или близкому к нему) закону. Подобные ситуации возникают на практике очень часто: например, при измерении любой физической величины на результат влияют многие факторы (температура, влажность, особенности прибора и пр.). По этой причине нормальный закон используется значительно чаще, чем любой другой.

Плотность вероятности нормально распределенной случайной величины описывается формулой:

$$p(x) = \frac{1}{\sigma\cdot\sqrt{2\pi}}\cdot e^{\frac{-(x-a)^2}{2\cdot\sigma^2}},$$

где a — это математическое ожидание случайной величины, σ— ее среднеквадратичное отклонение.

В Mathcad плотность нормального распределения вычисляется с помощью встроенной функции dnorm(x,σ,a) (рис. 15.5).

Интегрирование плотности вероятности дает функцию распределения. В Mathcad вычисления, связанные с ее применением, можно производить благодаря наличию специальной встроенной функции pnorm(x,σ,a). Используя функцию распределения, можно определять, какова вероятность того, что случайная величина примет значение из определенного промежутка.

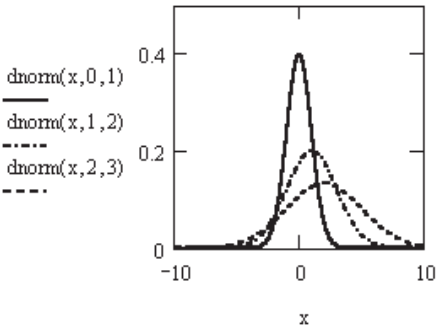

Рис. 15.5. Кривые плотностей вероятности нормального распределения при различных величинах параметров

Пример 15.16. Производится взвешивание некоторого вещества без систематических ошибок. Случайные ошибки взвешивания подчинены нормальному закону со среднеквадратичным отклонением y=20 г. Найти вероятность того, что взвешивание будет произведено с ошибкой, не превышающей по абсолютной величине 10 г

В сущности, нам необходимо определить вероятность попадания случайной величины (в данном случае ошибки взвешивания) в интервал (–10; 10). По аналогии с приведенной выше классической формулой, запишем выражение для вероятности с использованием функции pnorm (в качестве параметра x укажем границы интервала)

$$a := 0 \qquad \sigma := 20 \qquad x := 10$$

$$P := \text{pnorm}(x, a, \sigma) - \text{pnorm}(-x, a, \sigma)$$

$$P = 0.383$$

Раньше для вычисления вероятности попадания в числовой интервал нормально распределенной случайной величины использовали функцию Лапласа — нормированную (a=0, σ=1) функцию вероятности. Эта функция была табулирована, и ее узловые значения приводились в любом учебнике по статистике. При проведении вычислений в Mathcad, в принципе, нет необходимости использовать нормированные функции вероятности, так как в систему встроены довольно мощные алгоритмы численного интегрирования, позволяющие просчитать функцию распределения и в ее стандартном виде. Однако если вам необходимо решить задачу традиционным способом, то вы можете применить специальную встроенную функцию cnorm(x), в основе которой лежит формула Лапласа.

Если вам нужно подсчитать с помощью нормированной функции распределения вероятность попадания случайной величины в интервал между A и B при величинах математического ожидания и дисперсии равных соответственно a и σ, то используйте для этого следующую формулу:

$$P(A < X \le B) = \Phi\left(\frac{B - a}{\sigma}\right) - \Phi\left(\frac{A - a}{\sigma}\right)$$

Реально применение функций pnorm и cnorm математически абсолютно идентично, так что вы можете использовать из них ту, которая вам более удобна.

Пример 15.17. Использование функции cnorm(x) для проверки правила «трех сигм»

Правило «трех сигм» гласит, что для того, чтобы быть практически уверенным в корректности используемого результата, погрешность должна быть учтена на уровне трех стандартных отклонений. Проверим это утверждение.

$$a := 1 \qquad \sigma := 3$$

$$\text{pnorm}(3\sigma + a, a, \sigma) - \text{pnorm}(-3\sigma + a, a, \sigma) = 0.997$$

$$A := -3\sigma \qquad B := 3\sigma$$

$$\text{cnorm}\left(\frac{B - a}{\sigma}\right) - \text{cnorm}\left(\frac{A - a}{\sigma}\right) = 0.996$$

Очень близкой, отличающейся только условием нормирования, к функции cnorm является функция ошибки erf(x). В основе ее лежит следующий интеграл вероятности:

$$\sqrt{\frac{2}{\pi}} \cdot \int_0^x e^{-t^2} dt$$

Никаких различий в результате при правильном задании условий эти функции не дают. Однако о функции ошибки erf(x) нужно иметь четкое представление, так как она часто используется в символьных ответах.

Пример 15.18. Использование функции ошибки erf(x) для проверки правила «трех сигм»

$$a := 1 \qquad \sigma := 3 \qquad A := -3\sigma \qquad B := 3\sigma$$

$$\frac{1}{2} \cdot \left(\text{erf}\left(\frac{B - a}{\sqrt{2} \cdot \sigma}\right) - \text{erf}\left(\frac{A - a}{\sqrt{2} \cdot \sigma}\right)\right) = 0.996$$

Помимо простой, в Mathcad имеется и так называемая дополняющая (complementary) функция ошибки erfc(x), определяемая как 1–erf(x).

Очень полезной является и функция определения квантилей нормального распределения qnorm(α,a,σ), поскольку она позволяет определить ширину интервала, в который с заданной вероятностью попадет случайная величина.

Пример 15.19. Станок–автомат изготовляет валики, причем контролируется их диаметр X. Считая, что X — нормально распределенная случайная величина с математическим ожиданием a=10 мм и среднеквадратичным отклонением σ=0,1 мм, найти интервал, симметричный относительно математического ожидания, в котором с вероятностью 0,9973 будут заключены диаметры изготовленных валиков

Случайная величина может принимать значение, возвращаемое функцией qnorm(α,a,σ), либо быть меньше него при заданной вероятности. Поэтому с определением правой границы промежутка проблем не возникнет.

$$\alpha := 0.9973 \qquad a := 10 \qquad \sigma := 0.1$$

$$\text{right} := \text{qnorm}(\alpha, a, \sigma) \qquad \text{right} = 10.278$$

Интервал симметричен относительно математического ожидания. Чтобы вычислить левую границу, поменяем его знак. Случайная величина должна быть меньше полученного отрицательного значения, а следовательно, больше его модуля

$$left := |qnorm(\alpha, -a, \sigma)| \qquad left = 9.722$$

Распределение «хи-квадрат»

Если n случайных величин распределены по нормальному закону, причем для всех математическое ожидание равно 0, а среднеквадратичное отклонение — 1, то сумма их квадратов распределена по закону χ^2, плотность вероятности которого описывается следующей формулой (x>0):

$$p(x) = \frac{1}{2^{\frac{k}{2}} \cdot \Gamma\left(\frac{k}{2}\right)} \cdot e^{\frac{-x}{2}} \cdot x^{\frac{k}{2}-1}$$

Как видно из приведенной формулы, описанное распределение зависит только от одного параметра k — числа степеней свободы (k=n–1, где n — количество случайных величин). При больших k распределение «хи-квадрат» приближается к нормальному.

В статистике рассматриваемое распределение применяется для вычисления интервала, в котором может изменяться дисперсия случайной величины.
Доказано, что границы доверительного интервала для дисперсии можно определить как

$$\frac{S^2 \cdot (n-1)}{\chi^2}$$

где S^2 — выборочная дисперсия, n — количество элементов в выборке, χ^2 — коэффициент, вычисляемый исходя из «хи-квадрат» распределения. Традиционно эти коэффициенты берутся из специальных таблиц при вероятностях, равных вероятности выхода случайной величины за пределы интервала. Вычисляются же эти граничные вероятности как $(1+\alpha)/2$ и $(1-\alpha)/2$ (где α — доверительная вероятность). По сути, коэффициенты χ^2 являются квантилями соответствующего распределения, так что для того, чтобы решить задачу о доверительном интервале для дисперсии в Mathcad, совершенно не нужно обращаться к специальным таблицам. Для их вычисления просто нужно воспользоваться встроенной функцией квантилей qchisq(p,d), где p — доверительная вероятность, d — количество степеней свободы.

Пример 15.20. В результате измерения роста 20 студентов было получено значение выборочной дисперсии S=0.002. Найти 95%-ный доверительный интервал для дисперсии роста

$$S := 0.002 \qquad N := 20 \qquad \alpha := 0.95$$

Вычисляем коэффициенты $\chi 1$ и $\chi 2$:

$$\chi 1 := qchisq\left(\frac{1-\alpha}{2}, N-1\right) \qquad \chi 2 := qchisq\left(\frac{1+\alpha}{2}, N-1\right)$$

$$\chi 1 = 8.907 \qquad \chi 2 = 32.852$$

Определяем границы доверительного интервала для дисперсии:

$$I_{right} := \frac{S\cdot(N-1)}{\chi 1} \qquad I_{left} := \frac{S\cdot(N-1)}{\chi 2}$$

$$I_{right} = 4.267\times 10^{-3} \qquad I_{left} = 1.157\times 10^{-3}$$

Распределение Стьюдента

Очень важным распределением, используемым при обработке данных, является распределение Стьюдента. Это распределение было введено, прежде всего, потому, что для маленьких объемов выборок нормальное распределение давало чрезвычайно заниженное значение погрешностей. В общем случае кривая плотности распределения Стьюдента имеет более низкий и пологий максимум, чем аналогичная кривая центрированного нормального распределения (это означает, что реально при малых величинах объема выборки больших ошибок больше, а малых меньше, чем должно быть исходя из нормального распределения). В том случае, если выборка достаточно большого объема, кривые плотности вероятности ошибки, даваемые нормальным и Стьюдента распределениями, практически совпадают (рис. 15.6).

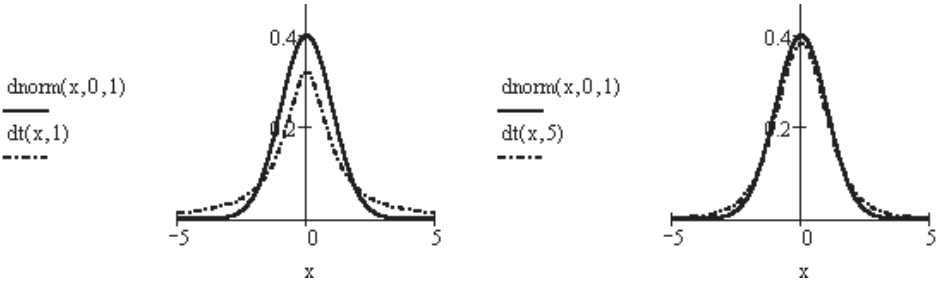

Рис. 15.6. Сравнение распределения Стьюдента с нормальным при различных объемах выборки

Распределение Стьюдента применяется для решения очень многих статистических задач, однако наиболее важно его использование для вычисления доверительного интервала математического ожидания нормально распределенных данных.

В статистике доказано, что доверительный интервал для математического ожидания можно оценить как

$$\left(\bar{x} - \frac{t_\gamma \cdot S}{\sqrt{n}}, \bar{x} + \frac{t_\gamma \cdot S}{\sqrt{n}}\right)$$

где n — объем выборки, S — «исправленное» среднеквадратичное отклонение, t — квантиль распределения Стьюдента. Таблицы с критическими точками для последнего можно найти не только в учебниках по статистике, но и в любом пособии по выполнению лабораторных работ (в которых он определяется по уровню значимости (зависит от техники выполнения измерений) и количества степеней свободы (то есть от количества параллельных измерений)).

Если вы вычисляете доверительный интервал для математического ожидания в Mathcad, то вам нет никакой необходимости обращаться к справочным таблицам для того, чтобы узнать величину коэффициента Стьюдента. Соответствующее значение можно по-

лучить и с помощью функции квантилей qt(p,d), где p — доверительная вероятность, d — количество степеней свободы случайной величины.

Имея данные независимых равноточных измерений некоторой величины, вы можете с заданной надежностью оценить математическое ожидание, а следовательно, и истинное значение измеряемой величины (поскольку математическое ожидание измеряемой величины равно ее истинному значению).

Пример 15.21. Количественный признак X генеральной совокупности распределен нормально. При выборке объема n=16 выборочная средняя равна x=20.2 и «исправленное» среднее квадратичное отклонение s=0.8. Оценить неизвестное математическое ожидание, если доверительный интервал равен 0.95

$$s := 0.8 \qquad x_{cp} := 20.2 \qquad n := 16$$

$$\alpha := 1 - 0.95$$

Находим коэффициент Стьюдента (в функции квантилей доверительная вероятность указывается для односторонней критической области)

$$t := qt\left(1 - \frac{\alpha}{2}, n - 1\right) \qquad t = 2.131$$

Определяем границы доверительного интервала для математического ожидания

$$I_{left} := x_{cp} - \frac{t \cdot s}{\sqrt{n}} \qquad I_{right} := x_{cp} + \frac{t \cdot s}{\sqrt{n}}$$

$$I_{left} = 19.774 \qquad I_{right} = 20.626$$

Показательное распределение

Показательное распределение было введено для описания процессов типа ядерного распада и имеет довольно широкое применение в некоторых областях приблизительных расчетов, например, для определения времени безотказной работы устройства. Плотность вероятности для этого распределения при x≥0

$$p(x) = \lambda \cdot e^{-\lambda \cdot x},$$

при x<0 p(x)=0. В Mathcad за вычисления по этой формуле отвечает специальная функция dexp(x,λ). Соответственно функция показательного распределения вычисляется в Mathcad с помощью pexp(x,λ). Используя встроенные функции для показательного распределения, можно решать ряд специфических задач.

Вероятность попадания случайной величины, распределенной по показательному закону, на интервал [a,b] определяется соотношением

$$P(a < X < b) = F(b) - F(a),$$

где

$$F(a) = 1 - e^{-\lambda a} \qquad F(b) = 1 - e^{-\lambda b}$$

есть значения функции распределения в конечных точках интервала, которые в Mathcad легко подсчитать с помощью функции pexp(x,λ).

Пример 15.22. Непрерывная случайная величина X распределена по показательному закону, заданному при x≥0 плотностью распределения $f(x)=0{,}04e^{-0{,}04x}$, при x<0 f(x)=0. Найти вероятность того, что в результате испытания X попадет в интервал (1, 2)

Задаем параметр λ и границы интервала:

$$\lambda := 0.04 \qquad a := 1 \qquad b := 2$$

Находим значения функции распределения на границах интервала:

$$F_a := \text{pexp}(a, \lambda) \qquad F_b := \text{pexp}(b, \lambda)$$

Вычисляем вероятность попадания X в указанный интервал:

$$P := F_b - F_a \qquad P = 0.038$$

Средствами аналитического интегрирования в Mathcad можно вычислить математическое ожидание и среднеквадратичное отклонение для показательного распределения.

Пример 15.23. Найти математическое ожидание, дисперсию и среднеквадратичное отклонение показательного распределения, заданного плотностью вероятности $f(x)=\lambda e^{-\lambda x}$ при x≥0, f(x)=0 при x<0

Вспомнив известные формулы, определяем для показательного распределения математическое ожидание:

$$M := \int_0^{\infty} x \cdot \lambda \cdot e^{-\lambda \cdot x}\, dx \qquad M \text{ assume}, \lambda > 0 \rightarrow \frac{1}{\lambda},$$

дисперсию:

$$D := \int_0^{\infty} \lambda \cdot e^{-\lambda \cdot x} \cdot \left(x - \frac{1}{\lambda}\right)^2 dx \qquad D \text{ assume}, \lambda > 0 \rightarrow \frac{1}{\lambda^2}$$

и среднеквадратичное отклонение:

$$\sqrt{D} \left| \begin{array}{l} \text{assume}, \lambda > 0 \\ \text{simplify} \end{array} \right. \rightarrow \frac{1}{\lambda}$$

Обратите внимание на одну очень важную техническую деталь, использованную при вычислении характеристик распределения в примере: чтобы получить во всех случаях корректный результат, требуется ввести ограничение на величину параметра λ с помощью оператора assume (Принять) панели Symbolic (Символьные). Иначе в качестве ответа будут выданы общие, мало что говорящие выражения с пределами.

Как и следовало ожидать, математическое ожидание и среднеквадратичное отклонение показательного распределения оказались равными между собой.

Логнормальное распределение

Если случайная величина является произведением большого количества взаимно независимых случайных величин, вклад каждой из которых в общую сумму ничтожно

мал, то такая случайная величина распределена по логнормальному закону. Логнормальное распределение задается, в общем, той же формулой, что и нормальное, единственное, вместо параметров в ней используются их натуральные логарифмы:

$$P(x) = \frac{1}{\ln(\sigma)\cdot\sqrt{2\pi}}\cdot e^{-\frac{\ln\left(\frac{x}{a}\right)^2}{2\cdot\ln(\sigma)^2}}$$

В Mathcad характеристики этого распределения вычисляются с помощью функций, имя которых задается добавлением соответствующей приставки к корню lnorm (x,ln(a),ln(σ)).

Распределение Коши

Распределение Коши относится к имеющим лишь теоретическое значение и описывается плотностью

$$p(x) = \frac{1}{\pi}\cdot\frac{\lambda}{\lambda^2 + (x-\mu)^2}$$

В Mathcad для задания имен функций характеристик распределения Коши используется корень cauchy (x,μ,λ).

Распределение Вейбулла

Это распределение также редко применяется для решения реальных проблем: может быть использовано, например, для определения времени выполнения какой-либо задачи. Распределение Вейбулла имеет следующую плотность:

$$p(x) = s\cdot x^{s-1}\cdot e^{-x^s}$$

В Mathcad оно прописано в виде встроенных функций, названия которых имеют общий корень weibull(x,s).

Гамма-распределение

Теоретическое распределение с плотностью

$$p(x) = \frac{x^{s-1}\cdot e^{-x}}{\Gamma(s)}$$

Может применяться, например, для определения времени выполнения какой-либо задачи. В Mathcad характеристики этого распределения можно вычислить с помощью функций, имена которых образованы добавлением соответствующих приставок к корню gamma(x,s) (s>0 — параметр формы).

Бета-распределение

Теоретическое распределение, плотность которого определяется формулой

$$p(x) = \frac{\Gamma(s1 + s2)}{\Gamma(s1)\cdot\Gamma(s2)}\cdot x^{s1-1}\cdot(1-x)^{s2-1}$$

Задается с помощью функций, названия которых имеют общий корень beta(x,s1,s2), где s1 и s2 — положительные параметры формы. Может быть использовано как приближенная модель в случае отсутствия данных, поскольку плотность бета-распределения может принимать различные формы в зависимости от параметров s1 и s2.

Логистическое распределение

Данное распределение широко используется в экономических исследованиях. Плотность вероятностей задается следующей формулой:

$$p(x) = \frac{1}{\beta} \cdot \frac{e^{-\frac{x-\alpha}{\beta}}}{\left(1 + e^{-\frac{x-\alpha}{\beta}}\right)^2}$$

где α — среднее распределения, β — параметр.

Для задания имен функций характеристик распределения используйте соответствующие приставки к корню logis(x,a,b). Кстати, по свойствам логистическое распределение очень схоже с нормальным.

15.4. Числовые характеристики дискретных случайных величин

В предыдущем разделе мы рассмотрели пример вычисления таких характеристик случайной величины, как дисперсия и математическое ожидание. Однако сделано это было за счет непосредственного задания соответствующих формул, что может быть для ряда параметров весьма неудобно по причине сложности таких формул. Значительно облегчить выполнение расчетов могут имеющиеся в Mathcad встроенные функции для вычисления практически всех используемых статистических характеристик выборки. Именно им мы и посвятим данный раздел.

15.4.1. Математическое ожидание

Одним из основных понятий статистики является понятие математического ожидания. Если же случайная величина принимает значения с разной вероятностью, математическое ожидание вычисляется по формуле

$$M(X) = \sum_{i=1}^{n} x_i \cdot p_i$$

Пример 15.24. Найти математическое ожидание дискретной случайной величины, закон распределения которой задан таблицей:

X	1	2	3	4	5
P	0,15	0,25	0,3	0,2	0,1

Решение данной задачи найти не сложно, если помнить принципы работы с векторами:

$$x := (1 \quad 2 \quad 3 \quad 4 \quad 5)^T \qquad p := (0.15 \quad 0.25 \quad 0.3 \quad 0.2 \quad 0.1)^T$$

$$M := \sum_{i=0}^{last(x)} x_i \cdot p_i \qquad M = 2.85$$

Если же случайная величина принимает ряд значений с равной вероятностью, то математическое ожидание определяется как среднее арифметическое значение некоторого количественного признака выборки.

В Mathcad среднее значение выборки можно подсчитать с помощью встроенной функции mean(x).

Пример 15.25. При измерении величины силы тока были получены следующие значения: 0.45, 0.41, 0.44, 0.42, 0.45, 0.41, 0.49, 0.56, 0.47, 0.48, 0.52, 0.43. Вычислить выборочное среднее

$$X := (0.45 \quad 0.49 \quad 0.44 \quad 0.42 \quad 0.48 \quad 0.41 \quad 0.44 \quad 0.56 \quad 0.47 \quad 0.45 \quad 0.52 \quad 0.43)$$

$$mean(X) = 0.463$$

При обработке экспериментальных данных среднее значение выборки считается равным истинному значению параметра. Однако такое утверждение абсолютно верно лишь в том случае, если выборка является генеральной, то есть содержит все возможные значения измеряемой величины. Естественно, что реально с генеральными совокупностями работать невозможно, а всегда приходится делать из них некоторые небольшие выборки. В зависимости от условий отбора и объема выборки она может быть репрезентативной в большей или меньшей степени — то есть передавать особенности генеральной совокупности с различной точностью. При этом такие характеристики, как среднее значение и дисперсия, приобретают случайный характер. Исследование особенностей поведения такого рода величин — это очень важная и сложная статистическая задача.

15.4.2. Дисперсия и среднеквадратичное отклонение

В статистике дисперсией называется среднее арифметическое квадратов отклонений случайной величины от ее среднего значения:

$$D = \frac{\sum_{i=1}^{N} \left(x_i - \bar{x}\right)^2}{N}$$

В общем случае дисперсия является характеристикой степени рассеяния значений выборки по сравнению с ее средней величиной.

В Mathcad простая выборочная дисперсия вычисляется с помощью функции var(x). Кроме того, существует и функция Var(x), определяющая так называемую исправленную дисперсию, используемую на практике для несмещенной оценки генеральной дисперсии при малом объеме выборки:

$$s^2 = \frac{\sum_{i=1}^{N} \left(x_i - \overline{x}\right)^2}{N-1}$$

Учитывая, что функции выборочной и исправленной дисперсий отличаются лишь форматом первой буквы, к их заданию следует подходить особенно осторожно.

На практике обычно используют не саму дисперсию, а квадратный корень из нее, называемый среднеквадратичным отклонением. В Mathcad существуют две функции для вычисления этого параметра: stdev(x) — выборочное стандартное отклонение и Stdev(x) — исправленное среднеквадратичное отклонение.

Пример 15.26. Подбрасывается игральный кубик. Случайная величина X — количество выпавших очков. Найти дисперсию и среднеквадратичное отклонение случайной величины X

$$X := (1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6)^T$$

$$\mathrm{var}(X) = 2.917 \qquad \mathrm{stdev}(X) = 1.708$$

Абсолютно аналогичные результаты получаются и при использовании известных формул:

$$D := \frac{1}{\mathrm{length}(X)} \cdot \sum_{i=0}^{\mathrm{last}(X)} \left(X_i - \mathrm{mean}(X)\right)^2 \qquad \sigma := \sqrt{\frac{1}{\mathrm{length}(X)} \cdot \sum_{i=0}^{\mathrm{last}(X)} \left(X_i - \mathrm{mean}(X)\right)^2}$$

$$D = 2.917 \qquad \sigma = 1.708$$

15.4.3. Мода и медиана

Модой в статистике называют варианту, которая встречается в выборке наиболее часто. В Mathcad подсчитать моду выборки можно с помощью встроенной функции mode(x).

В том случае, если все варианты встречаются в выборке с одинаковой частотой, система выдаст сообщение: No value occurs more than any others (Ни одна величина не встречается чаще, чем все остальные).

Медианой называется варианта, которая делит вариационный ряд (проще говоря, рассортированную выборку) на две части, равные по количеству вариант. То есть если количество элементов выборки нечетное и равно 2k+1, то, очевидно, медианой будет являться (k+1)-й элемент. В случае четного количества вариант медиана определяется как среднее арифметическое между k-м и (k+1)-м элементами выборки.

В Mathcad медиана вычисляется с помощью встроенной функции median(x).

Пример 15.27. Вычисление моды и медианы

$$X := (1 \quad 1 \quad 0 \quad 8 \quad 3 \quad 7)$$

$$\mathrm{mode}(X) = 1 \qquad \mathrm{median}(X) = 2$$

Обратите внимание на то, что статистические функции могут корректно работать не только с векторами-столбцами, но и векторами-строками.

15.4.4. Размах варьирования

Такая важнейшая характеристика рассеяния вариационного ряда, как размах варьирования может быть очень просто вычислена в Mathcad с помощью двух специальных матричных функций:

- max(x) — возвращает максимальное значение в векторе выборки;
- min(x) — функция находит минимальную величину в выборке.

Используя описанные функции, размах варьирования можно задать как

$$R = \max(x) - \min(x)$$

Пример 15.28. Вычисление размаха варьирования

Для задания вектора выборки воспользуемся генератором случайных чисел, распределенных по показательному закону:

$$X := \mathrm{rexp}(10000, 4)$$

$$\max(X) = 2.892 \qquad \min(X) = 1.58 \times 10^{-5}$$

$$\max(X) - \min(X) = 2.892$$

15.4.5. Наибольший общий делитель и наименьшее общее кратное

При решении некоторых специфических задач в статистике бывает необходимым определить, на какое максимальное целое число делятся без остатка все величины в выборке. В Mathcad можно очень просто вычислить такое число и не создавая специальных алгоритмов. Для этого следует обратиться к встроенной функции gcd(x) (от англ. Greatest common divisor — наибольший общий делитель).

Схожей с описанной является задача поиска наименьшего числа, которое делится без остатка на все значения элементов выборки. В Mathcad ее можно решить с помощью встроенной функции lcm(x) (сокращение от Least common multiple — наименьшее общее кратное).

Пример 15.29. Использование функций gcd(x) и lcm(x)

$$X := (2 \quad 4 \quad 8 \quad 16 \quad 32 \quad 64)$$

$$\gcd(X) = 2 \qquad \mathrm{lcm}(X) = 64$$

15.4.6. Геометрическое и гармоническое среднее

Чаще всего в статистике используется рассмотренное нами ранее среднее арифметическое количественного признака выборки. Однако в ряде специфических задач могут быть востребованы и функции, вычисляющие другие средние значения:

- gmean(x) — геометрическое среднее выборки;
- hmean(x) — гармоническое среднее.

Пример 15.30. Вычисление различных средних для выборки

$$X := \text{rgamma}(100, e)$$

$$N := \text{last}(X)$$

$$\frac{1}{N+1}\cdot\sum_{i=0}^{N} X_i = 3.027 \qquad \text{mean}(X) = 3.027$$

$$\left[\frac{1}{N+1}\cdot\sum_{i=0}^{N} (X_i)^{-1}\right]^{-1} = 1.952 \qquad \text{hmean}(X) = 1.952$$

$$\sqrt[N+1]{\prod_{i=0}^{N} X_i} = 2.542 \qquad \text{gmean}(X) = 2.542$$

15.5. Проверка некоторых статистических гипотез

15.5.1. Распределение Фишера. Сравнение двух дисперсий нормальных генеральных совокупностей

Отношение двух независимых случайных величин, распределенных по закону χ^2, описывается распределением Фишера–Снедекора. На практике F-критерий Фишера применяется для проверки нулевой гипотезы о равенстве дисперсий двух генеральных совокупностей. Подобная задача возникает в том случае, если требуется сравнить точность приборов, инструментов или воспроизводимость результатов измерения, полученных различными методами. Естественно, из них стоит предпочесть тот, который дает меньшую дисперсию, то есть ошибку.

Распределение Фишера зависит только от количества степеней свободы случайных величин. Напомним, количество степеней свободы для данной выборки на единицу меньше ее объема.

На практике для проверки гипотезы обычно используют таблицы критических точек распределения Фишера–Снедекора. В том же случае, если задача решается в Mathcad, можно применить встроенную функцию квантилей `qF(a,d1,d2)`, где a — доверительная вероятность, `d1` — количество степеней свободы большей исправленной дисперсии, `d2` — меньшей. Если отношение исправленных дисперсий меньше квантили распределения Фишера, то нулевую гипотезу о равенстве дисперсий принимают, в противном случае — отвергают.

Пример 15.31. Для сравнения точности двух станков взяты две пробы, объемы которых n_1=10 и n_2=8. В результате измерения контролируемого размера отобранных изделий получены следующие результаты:

X_i	1,08	1,10	1,12	1,14	1,15	1,25	1,36	1,38	1,40	1,42
Y_i	1,11	1,12	1,18	1,22	1,33	1,35	1,36	1,38	—	—

Можно ли считать, что станки обладают одинаковой точностью, если принять уровень значимости α=0,1 и в качестве конкурирующей гипотезы D(x)≠D(y)

Задаем векторы случайных величин, объемы выборок и уровень значимости:

$$x := (1.08\ \ 1.10\ \ 1.12\ \ 1.14\ \ 1.15\ \ 1.25\ \ 1.36\ \ 1.38\ \ 1.40\ \ 1.42)$$

$$y := (1.11\ \ 1.12\ \ 1.18\ \ 1.22\ \ 1.33\ \ 1.35\ \ 1.36\ \ 1.38)$$

$$n_1 := 10 \qquad n_2 := 8 \qquad \alpha := 0.1$$

Вычисляем исправленные дисперсии выборок:

$$\mathrm{Var}(x) = 0.019 \qquad \mathrm{Var}(y) = 0.012$$

Определяем отношение большей исправленной дисперсии к меньшей:

$$F := \frac{\mathrm{Var}(x)}{\mathrm{Var}(y)} \qquad F = 1.511$$

Находим критическую точку $F_{кр}$, задействовав функцию квантилей qF. Для конкурирующей гипотезы D(x)≠D(y) необходимо принять вдвое уменьшенный уровень значимости.

$$F_{kp} := qF\left(1 - \frac{\alpha}{2}, n_1 - 1, n_2 - 1\right) \qquad F_{kp} = 3.677 \qquad F < F_{kp}$$

Следовательно, нет оснований отвергать нулевую гипотезу, то есть считать точность станков различной.

15.5.2. Асимметрия и эксцесс. Проверка гипотезы о нормальном распределении

Очень часто в статистике требуется установить, является ли данное эмпирическое распределение нормальным, а если оно таковым не является, то с помощью какой-либо количественной характеристики показать меру отклонения данного распределения от нормального. В качестве таких характеристик используются асимметрия и эксцесс. Для нормального распределения эти признаки равны нулю.

Асимметрия позволяет оценить меру отклонения функции данного распределения от центра рассеяния. Для генеральной совокупности асимметрия вычисляется с учетом исправленного среднеквадратичного отклонения (которое используется для оценки среднеквадратичного отклонения генеральной совокупности):

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

Асимметрия положительна, если вытянут правый участок кривой распределения, и отрицательна, если левый.

В Mathcad асимметрию для генеральной совокупности по данным некоторой представительной выборки можно подсчитать с помощью функции skew(x) (от англ. skewness — асимметрия).

Даже если асимметрия распределений одинакова, их кривые могут значительно различаться: одни будут иметь более высокие и острые пики, другие, наоборот, будут изменяться очень плавно. Показателем остроты пика является эксцесс.

Для генеральной совокупности эксцесс рассчитывается с учетом исправленного среднеквадратичного отклонения по формуле:

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

Если эксцесс больше 0, то распределение имеет более острую вершину, чем нормальное, если же он меньше 0 — наоборот.

В Mathcad эксцесс для генеральной совокупности по данным некоторой репрезентативной выборки можно подсчитать, используя встроенную функцию kurt(x) (от англ. kurtosis — эксцесс).

Асимметрия выборочной совокупности (не представляющей пропорции некоторой генеральной совокупности) определяется как отношение центрального момента третьего порядка к кубу выборочного стандартного отклонения, а ее эксцесс — соотношением четвертого центрального момента к четвертой степени выборочного стандартного отклонения за вычетом тройки. Но следует помнить, что для малых выборок, с которыми вам чаще всего приходится встречаться при решении задач, точность оценок этих параметров невысока. В принципе, сами по себе выборочные оценки асимметрии и эксцесса еще не являются показателями того, что распределение выборки соответствует нормальному закону, поэтому к ним нужно относиться с осторожностью. В учебной литературе, как правило, приводятся упрощенные формулы для определения асимметрии или эксцесса выборки:

$$A = \frac{1}{n \cdot \sigma^3} \cdot \sum_{i=1}^{N} \left(x_i - \bar{x}\right)^3 \qquad E = \frac{1}{n \cdot \sigma^4} \cdot \sum_{i=1}^{N} \left(x_i - \bar{x}\right)^4 - 3$$

Воспользовавшись ими, вы почти наверняка получите результат, отличный от рассчитанного встроенными функциями skew(x) и kurt(x), поскольку Mathcad рассматривает любую выборку как репрезентативную для некоторой генеральной совокупности. Поэтому при вычислении данных параметров определитесь, что является для вас более важным: совпадение результата с ответом в задачнике или же максимально корректная количественная оценка отклонения распределения от нормального. В первом случае используйте привычные для вас упрощенные выражения. Со второй же проблемой функции skew(x) и kurt(x) справятся куда лучше.

В статистике асимметрия и эксцесс имеют довольно широкое применение. С их помощью можно проверять гипотезу о нормальном распределении выборки. Например, в большинстве физических и химических лабораторий проверка нормальности результатов служит непременным условием полноценной аттестации используемых методик.

Пример 15.32. Рассматривая распределение размеров обуви, проданной магазином за день, как выборку из генеральной совокупности, проверить гипотезу о том, что интересующий нас признак распределен в генеральной совокупности по нормальному закону (приняв α=0,01)

Размер обуви, x_i	36	37	38	39	40	41	42	43	44
Количество пар, y_i	1	2	3	5	10	13	9	6	1

Трудность, с которой вы можете столкнуться при решении задач по статистике в Mathcad, связана с тем, что в условиях задач, как правило, выборки представлены в виде статистических рядов распределения. В Mathcad же необходимо представлять все имеющиеся данные в виде вектора, в котором каждая варианта встречается указанное количество раз. Так, в нашем примере для корректной работы функций skew(x) и kurt(x) нужно задать вектор размеров обуви длиной равной общему количеству проданных пар.

Формируем вариационный ряд и вектор частот:

$$x := (36\ \ 37\ \ 38\ \ 39\ \ 40\ \ 41\ \ 42\ \ 43\ \ 44)^T$$

$$y := (1\ \ 2\ \ 3\ \ 5\ \ 10\ \ 13\ \ 9\ \ 6\ \ 1)^T$$

С помощью вложенного цикла задаем вектор размеров обуви:

$$\text{Vector} := \left| \begin{array}{l} n \leftarrow 0 \\ \text{for } i \in 0..\text{last}(x) \\ \quad \text{for } j \in 1..y_i \\ \quad\quad v_n \leftarrow x_i, n \leftarrow n + 1 \\ v \end{array} \right.$$

$\text{Vector}^T =$

	0	1	2	3	4	5	6	7	8	9
0	36	37	37	38	38	38	39	39	39	39

Теперь как длину полученного вектора определяем объем выборки, задействовав функцию length(x) (возвращает количество элементов вектора):

$$n := \text{length}(\text{Vector})$$

Если асимметрия и эксцесс превысят по модулю утроенные значения собственных среднеквадратичных отклонений, то гипотезу о нормальности распределения следует отвергнуть. В противном случае она должна быть принята.

Вычисляем асимметрию, эксцесс и среднеквадратичные отклонения для них:

$$A := \text{skew}(\text{Vector}) \qquad E := \text{kurt}(\text{Vector})$$

$$S_A := \sqrt{\frac{6(n-1)}{(n+1)(n-3)}} \qquad S_E := \sqrt{\frac{24 \cdot n \cdot (n-2) \cdot (n-3)}{(n+1)^2 (n+3)(n+5)}}$$

$$|A| = 0.543 \qquad |E| = 0.134$$

$$S_A = 0.35 \qquad S_E = 0.598$$

Проверяем критерии согласия:

$$3S_A = 1.051 \qquad 3S_E = 1.793$$

$$|A| < 3S_A \qquad |E| < 3S_E$$

Требуемые условия выполняются, значит, в генеральной совокупности признак распределен по нормальному закону.

15.5.3. Проверка гипотезы о показательном распределении

Если случайная величина распределена по показательному закону, математическое ожидание и среднеквадратичное отклонение распределения должны совпадать. Это утверждение используется для проверки гипотезы о показательном распределении экспериментальных данных.

Пример 15.33. В результате испытания 200 элементов на длительность работы получено следующее эмпирическое распределение

X_i-X_{i+1}	0–5	5–10	10–15	15–20	20–25	25–30
n_i	133	45	15	4	2	1

(в первой строке указаны интервалы времени в часах, во второй — количество элементов, проработавших время в пределах соответствующего интервала). Требуется при уровне значимости 0,05 проверить гипотезу о том, что время работы элементов распределено по показательному закону

Для проверки гипотезы, как и в предыдущем примере, нам необходимо составить вектор данных, длина которого равна объему выборки. Однако в нашем случае распределение случайной величины задано в виде последовательности интервалов и соответствующих им частот, поэтому в качестве «представителя» каждого интервала выберем его середину. Создадим вектор границ промежутков и путем несложного преобразования сформируем из него вектор середин интервалов:

$$n := (0 \;\; 5 \;\; 10 \;\; 15 \;\; 20 \;\; 25 \;\; 30)^T$$

$$k := 1..\,last(n)$$

$$m_{k-1} := \frac{n_k + n_{k-1}}{2} \qquad m^T = (2.5 \;\; 7.5 \;\; 12.5 \;\; 17.5 \;\; 22.5 \;\; 27.5)$$

Далее зададим вектор частот:

$$V := (133 \;\; 45 \;\; 15 \;\; 4 \;\; 2 \;\; 1)^T$$

С помощью алгоритма, описанного в предыдущем примере, представим случайную величину в удобной для анализа в Mathcad форме, а затем оценим ее математическое ожидание и среднеквадратичное отклонение:

$$Vector := \left|\begin{array}{l} n \leftarrow 0 \\ \text{for } i \in 0..\,last(m) \\ \quad \text{for } j \in 1..\,V_i \\ \quad\quad v_n \leftarrow m_i,\; n \leftarrow n + 1 \\ v \end{array}\right.$$

$$M := mean(Vector) \qquad M = 5 \qquad s := stdev(Vector) \qquad s = 4.301$$

Оценки математического ожидания и среднеквадратичного отклонения оказались довольно близкими, а это означает, что нет оснований отвергать гипотезу о распределении времени работы элементов по показательному закону. Поскольку $\lambda=1/M$, можно приближенно оценить и этот параметр:

$$\lambda := \frac{1}{M} \qquad \lambda = 0.2$$

15.6. Статистическая обработка и представление результатов измерений

15.6.1. Оценка истинного значения измеряемой величины

При проведении различных исследований зачастую приходится сталкиваться с необходимостью дать точную количественную оценку какого-либо свойства изучаемого объекта. Однако при измерении некоторой физической величины всегда нужно помнить, что ни одни экспериментальные данные не отражают ее истинного значения. На практике любое измерение или анализ всегда отягощены погрешностями различной величины и природы. Искажение результатов измерений, как правило, связано с несовершенством используемых инструментов, погрешностью методики, субъективной ошибкой исполнителя, а также с влиянием контролируемых и неконтролируемых внешних факторов. Безусловно, ряд погрешностей можно устранить: например, выбрать прибор более высокого класса точности, провести определение относительно некоторого объекта или, наконец, попытаться выполнить эксперимент максимально аккуратно. Ошибки, исключить которые невозможно, приводят к тому, что измеряемая величина принимает случайные значения, попадающие в тот или иной интервал с определенной вероятностью. Таким образом, при проведении отдельного опыта мы всегда получаем некоторое значение случайной величины, набор же из определенного количества результатов представляет собой выборочную совокупность — конечнозначную, дискретную, ограниченную случайную величину. Поэтому истинное значение измеряемой величины мы можем оценить по среднему арифметическому результатов отдельных наблюдений с помощью доверительного интервала, покрывающего неизвестный параметр с заданной надежностью (доверительной вероятностью): $(x_{ср} \pm \Delta)$.

Пример 15.34. Произведено шесть независимых равноточных измерений физической величины. Получены следующие результаты: 87,85; 88,01; 87,89; 87,56; 87,73; 87,90. Требуется оценить истинное значение измеряемой величины с надежностью 0,95

За истинное значение измеряемой величины обычно принимается среднее арифметическое результатов наблюдений, поэтому нам необходимо оценить его с помощью доверительного интервала.

$$n := 6 \qquad \alpha := 0.05$$

$$x := (87.85 \;\; 88.01 \;\; 87.89 \;\; 87.56 \;\; 87.73 \;\; 87.90)^T$$

$$x_{cp} := \text{mean}(x) \qquad x_{cp} = 87.823$$

Для определения границ доверительного интервала случайной погрешности по формуле

$$\Delta = t \cdot S_{\bar{x}}$$

нам необходимо рассчитать среднеквадратичное отклонение результата измерения:

$$S_{\bar{x}} = \frac{S}{\sqrt{n}}$$

где S — «исправленное» среднеквадратичное отклонение, t — коэффициент Стьюдента при заданной доверительной вероятности и количестве степеней свободы, который, как вы помните, в Mathcad можно найти с помощью функции квантилей распределения Стьюдента.

$$t := qt\left(1 - \frac{\alpha}{2}, n - 1\right) \qquad t = 2.571$$

$$S := Stdev(x) \qquad S = 0.157$$

$$\Delta := t \cdot \frac{S}{\sqrt{n}} \qquad \Delta = 0.165$$

Обратите внимание: применяя микростатистику Стьюдента, мы изначально предполагаем, что экспериментальные данные распределены по нормальному закону, поскольку достаточно надежного метода оценки нормальности распределения при объеме выборки меньше 15 не существует.

Числовое значение погрешности должно содержать не более двух значащих цифр. Также при записи доверительного интервала среднее арифметическое результатов измерений и их погрешность должны иметь одинаковое количество знаков после запятой. Поэтому в нашем случае конечный ответ необходимо представить в следующем виде: (87,82±0,17). Данная запись означает, что истинное значение измеряемой величины заключено в указанном интервале с доверительной вероятностью 0,95. Вообще, если быть до конца корректными, то наряду со случайной ошибкой необходимо рассчитывать и границы неисключенной систематической погрешности, а доверительные границы общей погрешности результатов измерений вычислять как

$$\Delta = \sqrt{\Delta_{случ}^{2} + \Delta_{сист}^{2}}$$

Систематическая погрешность может иметь несколько составляющих (погрешность прибора, ошибка округления, методическая ошибка и т.п.), однако при их устранении или минимизации в простейшем случае можно ограничиться учетом только случайной ошибки.

15.6.2. Обнаружение промахов и фильтрация данных эксперимента

При проведении измерений вы, конечно же, стремитесь получить результат, максимально приближенный к истинному значению определяемой величины. Однако в серии опытных данных, полученных в одинаковых условиях, часто встречаются и резко выделяющиеся результаты. Вероятнее всего, такие данные являются грубыми ошибками — промахами, допущенными вследствие небрежности или некомпетентности исполнителя. Чтобы убедиться в достоверности этого утверждения, при наличии в выборке «выскакивающих» результатов необходимо проводить фильтрацию данных

эксперимента (то есть проверять, является ли результат наблюдения, резко отличающийся от остальных, промахом). Сущность выбраковки данных заключается в установлении вероятности того, что отклонение данного результата от среднего больше или равно отклонению подозрительного результата. Если эта вероятность мала, подозрительный результат выбраковывают, в противном случае — оставляют в выборке, считая данное отклонение естественным в условиях нормального распределения. На практике вероятность не подсчитывают, а пользуются приводимыми в таблицах (табл. 15.1) для разных уровней значимости α и объемов выборки значениями τ-критерия (критического отклонения, выраженного в единицах выборочного стандартного отклонения):

$$\tau_{kp} = \frac{\left| x_{kp} - \bar{x} \right|}{S} \qquad \text{откуда} \qquad \Delta x_{kp} = \tau_{kp} \cdot S$$

Аналогичным образом рассчитывают отклонение «выскакивающего» результата. Если оно окажется больше критического, результат считают промахом и отбрасывают.

Таблица 15.1. Значения τ-критерия для оценки подозрительных результатов на промах

n	α			n	α		
	0,1	0,05	0,01		0,1	0,05	0,01
3	1,41	1,41	1,41	13	2,26	2,43	2,71
4	1,65	1,69	1,72	14	2,30	2,46	2,76
5	1,79	1,87	1,96	15	2,33	2,49	2,80
6	1,89	2,00	2,13	16	2,35	2,52	2,84
7	1,97	2,09	2,27	17	2,38	2,55	2,87
8	2,04	2,17	2,37	18	2,40	2,58	2,90
9	2,10	2,24	2,46	19	2,43	2,60	2,93
10	2,15	2,29	2,54	20	2,45	2,62	2,96
11	2,19	2,34	2,61	22	2,49	2,66	3,01
12	2,23	2,39	2,66	25	2,54	2,72	3,07

Пример 15.35. В ходе измерения некоторой физической величины получены следующие данные: 6,53; 6,43; 6,50; 6,38; 6,48; 6,49; 6,74; 6,44; 6,53; 6,38. При уровнях значимости α=0,01 и α=0,05 провести выбраковку результатов эксперимента

Определим среднее арифметическое и «исправленное» среднеквадратичное отклонение выборки:

$$x := (6.53 \;\; 6.43 \;\; 6.50 \;\; 6.38 \;\; 6.48 \;\; 6.49 \;\; 6.74 \;\; 6.44 \;\; 6.53 \;\; 6.38)^T$$

$$x_{cp} := \text{mean}(x) \qquad S := \text{Stdev}(x)$$

Рассчитаем модуль отклонения каждого результата от среднего значения:

$$i := 0 .. \text{last}(x)$$

$$\Delta x_i := \left| x_i - x_{cp} \right| \qquad \Delta x =$$

	0
0	0.04
1	0.06
2	$10 \cdot 10^{-3}$
3	0.11
4	0.01
5	0
6	0.25
7	0.05
8	0.04
9	0.11

По табл. 15.1 определим значение τ-критерия для уровня значимости 0,05 и объема выборки n=10 и рассчитаем с его помощью критическое отклонение:

$$\tau_{kp} := 2.29$$

$$\Delta x_{kp} := \tau_{kp} \cdot S \qquad \Delta x_{kp} = 0.237$$

Теперь нам необходимо сравнить каждое отклонение Δx с критическим. Заметьте, отклонение для 6,74 превышает критическое, значит, данный результат является промахом и не может принадлежать выборке. Если вам необходимо представить результат измерения с помощью доверительного интервала, то значения среднего арифметического и «исправленного» среднеквадратичного отклонения придется пересчитать для новой выборки объема n–1.

$$0.25 > \Delta x_{kp}$$

Проведем аналогичные расчеты для уровня значимости 0,01:

$$\tau_{kp} := 2.54$$

$$\Delta x_{kp} := \tau_{kp} \cdot S \qquad \Delta x_{kp} = 0.263$$

Обратите внимание, в этом случае отклонение для 6,74 не превышает критическое, следовательно, результат выбраковывать нельзя, что, в общем-то, вполне закономерно, поскольку чем выше выбранная доверительная вероятность, тем шире доверительный интервал случайной ошибки измерения.

Из всего вышесказанного следует вывод: старайтесь проводить эксперимент многократно и максимально аккуратно, а при необходимости меняйте условия измерения, которые легко поддаются учету — гораздо легче избежать промахов, чем потом исключать их методами математической статистики.

15.6.3. Планирование эксперимента

Выше мы уже отметили, что на точность результата эксперимента наряду с другими факторами влияет и количество проведенных опытов. Чем больше произведено измерений, тем меньше среднеквадратичное отклонение каждого результата, а следовательно, уже доверительный интервал. Часто на практике необходимо оценить, при ка-

ком минимальном объеме выборки истинное значение измеряемой величины окажется в строго узком интервале при заданной доверительной вероятности. Другими словами — определить, сколько нужно провести опытов, чтобы искомое значение находилось в пределах границ указанного интервала. Раздел статистики, занимающийся решением подобных задач, называется планированием эксперимента. Так как истинное значение принимается равным среднему арифметическому, то, по сути, задача сводится к оценке математического ожидания с заранее заданной точностью и надежностью.

Пример 15.36. Найти минимальный объем выборки, при котором с надежностью 0,95 точность оценки математического ожидания нормально распределенной генеральной совокупности по выборочной средней равна 0,2, если среднеквадратичное отклонение генеральной совокупности σ=2

Точность оценки вычисляется по формуле, аналогичной использованной нами для определения границ доверительного интервала погрешности в случае малой выборки:

$$\delta = \frac{t \cdot \sigma}{\sqrt{n}}$$

При работе с малыми выборками известного объема для нахождения Δ мы могли воспользоваться коэффициентом Стьюдента. В нашем примере объем выборки является искомой величиной, поэтому параметр t необходимо рассматривать как аргумент функции Лапласа (поскольку признак распределен нормально), которому соответствует вероятность 0,95/2 (ввиду симметричности функции распределения). Найти его можно в таблице значений функции вероятности, которая приводится в любом учебнике по статистике, но гораздо проще — вычислить в Mathcad.

Зададим функцию Лапласа:

$$\Phi(t) := \frac{1}{\sqrt{2\pi}} \cdot \int_0^t e^{\frac{-z^2}{2}} dz$$

Уравнение Φ(t)=0,475 (0,95/2) решим методом секущих:

$$TOL := 10^{-17}$$

$$t := 2$$

$$root(\Phi(t) - 0.475, t) = 1.96$$

Теперь все три параметра, необходимые для нахождения объема выборки, известны:

$$\delta := 0.2 \qquad \sigma := 2 \qquad t := 1.96$$

$$n := \left(\frac{t \cdot \sigma}{\delta}\right)^2 \qquad n = 384.16$$

Разумеется, полученное значение нужно округлить до большего целого числа. Так, минимальный объем выборки, удовлетворяющий приведенным в задаче условиям, равен 385.

15.7. Построение полигона и гистограммы

Гистограмма — это график, позволяющий визуализировать частоту попадания данных экспериментальной выборки в определенный интервал. При ее построении область, определяемая по размаху значений данных в выборке, разбивается на некоторое количество промежутков (как правило, равных), и затем подсчитывается количество или процент элементов, оказавшихся на каждом из них. Сама гистограмма представляет собой столбчатую диаграмму, ширина сегмента которой соответствует величине промежутка, а высота — сумме частот либо относительной частоте попадания в него данных.

Чтобы построить гистограмму в Mathcad, следует вызвать функцию histogram(n, data):

- n — количество сегментов гистограммы;
- data — вектор экспериментальных данных.

Результатом работы функции histogram является матрица размерности n×2, в первом столбце которой содержатся значения середин сегментов разбиения, во втором — количество элементов выборки, попавших на каждый из интервалов.

При построении графика-гистограммы по умолчанию система просто соединит точки, координатами которых являются середины и высоты столбцов гистограммы, ломаной линией. Полученный таким образом график называется в статистике полигоном распределения.

Чтобы построить график в форме гистограммы, выполните следующую последовательность действий.

1. Постройте по имеющимся данным полигон, настройте параметры осей и пределы графической области.
2. Дважды щелкнув левой кнопкой мыши на графике, откройте диалоговое окно Formatting Currently Selected Graph (Форматирование выбранного графика).
3. В списке Type (Тип) вкладки Traces (Ряды данных) открытого диалогового окна выберите строку solidbar (гистограмма).
4. Нажмите OK.

Пример 15.37. Возраст студентов одного потока представляется следующими данными: 17, 20, 18, 19, 18, 17, 20, 21, 24, 22, 20, 21, 20, 19, 18, 20, 21, 22, 25, 20. Построить вариационный ряд, полигон и гистограмму относительных частот по данному распределению выборки

Задаем вектор данных, количество сегментов диаграммы и статистический ряд распределения, как функцию histogram (по сути, первая строка полученной матрицы представляет собой интервальный вариационный ряд, который, правда, содержит не интервалы вариации, а их середины, однако, определив ниже шаг, вы без труда сможете самостоятельно записать ряд в требуемом виде).

$$data := (17\ \ 20\ \ 18\ \ 19\ \ 18\ \ 17\ \ 20\ \ 21\ \ 24\ \ 22\ \ 20\ \ 21\ \ 20\ \ 19\ \ 18\ \ 20\ \ 21\ \ 22\ \ 25\ \ 20)$$

$$n := 9$$

$$H := histogram(n, data) \qquad H^T = \begin{pmatrix} 17.5 & 18.5 & 19.5 & 20.5 & 21.5 & 22.5 & 23.5 & 24.5 & 25.5 \\ 2 & 3 & 2 & 6 & 3 & 2 & 0 & 1 & 1 \end{pmatrix}$$

Вычисляем относительные частоты f_i:

$$i := 0..last\left(H^{\langle 1\rangle}\right)$$

$$f_i := \frac{\left(H^{\langle 1\rangle}\right)_i}{\sum_{i=0}^{last\left(H^{\langle 1\rangle}\right)} \left(H^{\langle 1\rangle}\right)_i} \qquad f^T = (0.1 \;\; 0.15 \;\; 0.1 \;\; 0.3 \;\; 0.15 \;\; 0.1 \;\; 0 \;\; 0.05 \;\; 0.05)$$

Определяем шаг и, учитывая его длину, рассчитываем плотности относительных частот w_i:

$$h := \frac{\max\left(H^{\langle 0\rangle}\right) - \min\left(H^{\langle 0\rangle}\right)}{n - 1} \qquad h = 1 \qquad w_i := \frac{f_i}{h}$$

Строим полигон и гистограмму, отложив по оси абсцисс интервалы вариации, а по оси ординат — соответствующие плотности относительных частот w (рис. 15.7).

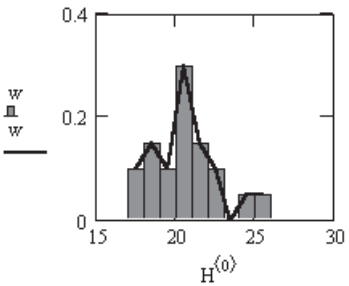

Рис. 15.7. Полигон и гистограмма относительных частот наблюдения вариант в выборке

Площадь гистограммы относительных частот должна быть равна единице. В нашем случае данное условие соблюдается, значит, задача решена верно.

$$\sum_{i=0}^{last(f)} w_i = 1$$

15.8. Статистическая обработка матриц

Все примеры, которые мы приводили выше, были связаны с обработкой векторов с помощью статистических функций. Однако большинство из этих функций может быть точно также применено и к выборкам, организованным в виде матриц. При этом соответствующая характеристика будет просчитана без какого-либо учета структурированности массива данных: то есть, независимо от того, сколько столбцов или строк содержит ваша матрица, статистическая функция увидит в ней лишь вариационный ряд из N×M элементов. И, соответственно, результатом всегда будет число, а не вектор, как логично было бы предположить.

Приведем пример определения среднего значения и дисперсии выборки, заданной как матрица. Соответствующий расчет будет проведен как с помощью встроенных функций, так и непосредственным вычислением нужной формулы.

Пример 15.38. Вычисление среднего значения и дисперсии выборки, организованной в виде матрицы

$$M := \begin{pmatrix} 3.351 & 4.348 & 8.192 \\ 1.882 & 2.03 & 4.166 \\ 6.278 & 3.718 & 5.748 \end{pmatrix}$$

$$\text{mean}(M) = 4.413$$

$$MEAN := \frac{1}{\text{rows}(M)\cdot\text{cols}(M)} \cdot \sum_{i=0}^{\text{rows}(M)-1} \sum_{j=0}^{\text{cols}(M)-1} M_{i,j} \qquad MEAN = 4.413$$

$$\text{var}(M) = 3.7$$

$$VAR := \frac{1}{\text{rows}(M)\cdot\text{cols}(M)} \cdot \sum_{i=0}^{\text{rows}(M)-1} \sum_{j=0}^{\text{cols}(M)-1} \left(M_{i,j} - MEAN\right)^2 \qquad VAR = 3.7$$

Работать с матрицами могут не только функции среднего выборочного и дисперсии, но и любые другие встроенные статистические функции.

Большинство статистических функций способно производить обработку сразу нескольких векторов, матриц или скаляров. При этом данные из них будут рассмотрены как одна выборка, поэтому, в принципе, нет особой разницы в том, в какой форме будут организованы элементы.

Пример 15.39. Вычисление среднего значения выборки, организованной в виде трех матриц

$$M := \begin{pmatrix} 3.351 & 4.348 & 8.192 \\ 1.882 & 2.03 & 4.166 \\ 6.278 & 3.718 & 5.748 \end{pmatrix} \quad N := \begin{pmatrix} 1.047 & 2.829 & 9.006 \\ 6.67 & 5.538 & 7.929 \\ 8.076 & 3.285 & 7.75 \end{pmatrix} \quad K := \begin{pmatrix} 2.217 & 5.357 & 1.58 \\ 2.9 & 5.274 & 5.303 \\ 1.912 & 3.346 & 7.781 \end{pmatrix}$$

$$\text{mean}(M, N, K) = 4.723$$

$$MEAN(X) := \frac{1}{\text{rows}(X)\cdot\text{cols}(X)} \cdot \sum_{i=0}^{\text{rows}(X)-1} \sum_{j=0}^{\text{cols}(X)-1} X_{i,j}$$

$$Mean := \frac{MEAN(M) + MEAN(N) + MEAN(K)}{3} \qquad Mean = 4.723$$

15.9. Моделирование случайных величин

Как уже отмечалось выше, Mathcad позволяет создавать выборки случайных величин, распределенные по любому из теоретических законов с произвольными параметрами. Случайные числа широко используются при моделировании всевозможных естественных явлений, в методе Монте-Карло, для решения сложных задач численного анализа. Зачастую случайные числа являются источником данных при проверке эффективности компьютерных алгоритмов, репрезентативной выборкой, на основании которой можно описать некоторое «типичное» явление. Немаловажную роль они играют и в принятии «беспристрастных» решений.

Со случайными величинами, сгенерированными компьютером, можно проводить те же преобразования, что и с экспериментальными данными.

Задаются функции случайных величин добавлением приставки r (от random — случайный) к корню термина соответствующего распределения. Число величин в случайном векторе определяется первым параметром.

Если вы не помните точного написания имени нужного вам генератора, следует обратиться к списку Random Numbers (Случайные числа) окна Insert Function (Вставить функцию).

Пример 15.40. Разыграть 100 возможных значений нормальной величины X с параметрами a=0 и σ=1. Оценить параметры разыгранной величины

Воспользуемся генератором случайных чисел, распределенных по нормальному закону. Итогом его работы является требуемый вектор

$$X := rnorm(100, 0, 1)$$

$X^T =$

	0	1	2	3	4	5	6	7
0	-0.727	0.253	0.63	0.699	-0.23	0.829	0.542	0.087

Оценим выборочное среднее и среднеквадратичное отклонение:

$$a := mean(X) \qquad \sigma := stdev(X)$$

$$a = 0.032 \qquad \sigma = 0.988$$

Результаты оценки удовлетворяют условию: a мало отлично от 0, σ близко к 1.

Кроме параметров используемого распределения на характеристики случайного вектора оказывает влияние величина системной переменной Seed Value For Random Numbers (Исходная величина для случайных чисел). В приведенном примере случайный вектор генерировался при значении Seed 601 333 713. Изменить эту переменную можно на вкладке Built-In Variables (Системные переменные) окна Worksheet Options (Параметры документа) меню Tools (Инструменты) либо используя встроенную функцию Seed(x). Данная функция меняет переменную Seed на новую величину x и возвращает ее предыдущее значение. Таким образом, для генерации последовательностей, расположенных в документе ниже или правее функции Seed, будет использоваться новая величина Seed, заданная в качестве аргумента функции Seed(x). Принимать значения переменная Seed (равно как и аргумент x) может в интервале от 1 до 2 147 483 647. Кстати, встроенную функцию Seed(x) можно с успехом применять и при создании программ.

Изменять значение Seed при создании случайных векторов необходимо в тех случаях, когда сгенерированные последовательности должны быть различными. Все дело в том,

что создать генератор действительно случайных чисел просто невозможно, так как компьютер работает исходя из точных алгоритмов, а точность и случайность несовместимы. Поэтому любая последовательность, сгенерированная на компьютере, реально является псевдослучайной. Основа же для случайных векторов всех типов всегда одна и та же — это последовательность равномерно распределенных чисел, которая выбирается соответственно заданной вами величине Seed. А это означает, что две выборки, произведенные любым генератором при равных параметрах, будут одинаковы.

Большое практическое значение имеет генератор случайных равномерно распределенных чисел. Поэтому в Mathcad он имеет две разновидности.

- Встроенная функция runif(N,a,b). Генерирует вектор из N случайных чисел, равномерно распределенных на промежутке от a до b.
- rnd(x). Функция возвращает случайно распределенное число, лежащее на промежутке от 0 до x (рис. 15.8). Данная функция используется в основном в программировании.

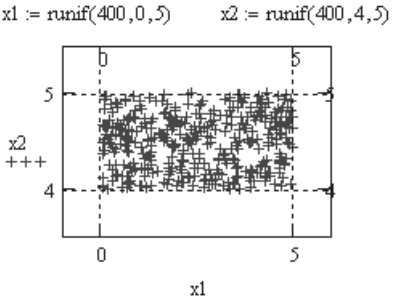

Рис. 15.8. Генерирование двумерной равномерно распределенной случайной величины

Глядя на рис. 15.8, легко понять, почему методы Монте-Карло столь эффективны. Действительно, зачем писать сложные программы для вычисления, например, интеграла, если можно просто заставить систему выбрать случайным образом несколько тысяч отрезков, что даст весьма неплохое, вполне приемлемое на практике, приближение. И времени это займет совсем немного.

Приведем пример вычисления простого интеграла и интеграла с сингулярностью на границе интегралов при помощи случайного выбора равномерно распределенных точек разбиения.

Пример 15.41. Вычисление интеграла по методу Монте-Карло

$$\begin{aligned}
&\text{Integral}(f,a,b,N,x) := \\
&\quad \text{Seed}(x) \\
&\quad R \leftarrow \text{runif}(N,a,b) \\
&\quad S \leftarrow 0 \\
&\quad \text{for } i \in 0..N-1 \\
&\quad\quad S \leftarrow S + f\left(R_i\right)\frac{b-a}{N} \\
&\quad\quad S \\
&\quad S
\end{aligned}$$

$$f(x) := \sin(x) \qquad g(x) := \frac{1}{\sqrt{x}}$$

$$\text{Integral}\left(f, 0, \frac{\pi}{2}, 5000, 1\right) = 1 \qquad \text{Integral}(g, 0, 1, 200000, 2) = 2$$

$$\int_0^{\frac{\pi}{2}} \sin(x)\,dx = 1 \qquad \int_0^1 \frac{1}{\sqrt{x}}\,dx = 2$$

Приведем еще один пример использования метода Монте-Карло для решения задачи из теории вероятностей. Используя генератор случайных, равномерно распределенных чисел rnd(x), мы определим, какова вероятность попадания при броске монеты в круг, вписанный в квадрат. Задача эта элементарно решается через нахождение отношения площадей фигур, что даст нам возможность объективно оценить точность эмпирического метода.

Пример 15.42. Задача о геометрической вероятности

```
Ver(N) := | (s ← 0  R ← 3)
          | for i ∈ 0.. N − 1
          |     (x ← rnd(R)  y ← rnd(R))
          |     s ← s + 1 if x^2 + y^2 ≤ R^2
          | Ver ← s / N
          | Ver
```

$$\frac{\pi R^2}{4 \cdot R^2} = 0.785$$

$$\text{Ver}(10) = 0.9$$

$$\text{Ver}(100) = 0.78$$

$$\text{Ver}(1000) = 0.792$$

$$\text{Ver}(10000) = 0.79$$

$$\text{Ver}(100000) = 0.784$$

$$\text{Ver}(1000000) = 0.785$$