Воронежский государственный университет

Методическое пособие по математической статистике (к практикуму на ЭВМ)

Составитель Новикова Н.М.

ВОРОНЕЖ
2006

Утверждено научно-методическим советом факультета прикладной математики, информатики и механики: протокол № 4 от 12 декабря 2005 г.

Пособие подготовлено на кафедре технической кибернетики и автоматического регулирования факультета прикладной математики, информатики и механики Воронежского государственного университета. Рекомендуется для студентов 3 курса дневного отделения.
Для специальности 510200 ОПД.Ф.04

1. ИСПОЛЬЗУЕМЫЕ ИНСТРУМЕНТЫ MATHCAD

Познакомимся с основными функциями Mathcad, предназначенными для решения задач математической статистики, а также с методами ввода данных для последующей статистической обработки. Попутно будут рассматриваться основные понятия математической статистики, постановка задач, алгоритмы и методы их решения.

Ввод и вывод файлов данных

При решении практических задач статистического анализа данных чаще всего приходится иметь дело с большими объемами исходной информации. Зачастую они представляют собой заранее введенные в файл аппаратными средствами экспериментальные данные, подготовленные специальными приложениями и сохраненные в файле таблицы чисел. Ниже будет рассказано о том, как в Mathcad можно генерировать последовательности случайных чисел. Такие последовательности позволяют имитировать результаты реальных измерений той или иной случайной величины.

Mathcad предоставляет пользователю специальные функции ввода данных из файла на диске и вывода данных в файл, т.е. функции доступа к файлам — READ, WRITE, APPEND, READPRN, WRITEPRN, APPENDPRN. Подробное описание этих функций и правила работы с ними можно найти в литературе по пакету, во встроенном в систему справочнике, в руководстве пользователя.

Познакомимся подробнее с функциями READ(file) и WRITE(file), предназначенными соответственно для чтения и записи числового значения. Файл данных для Mathcad — это файл чисел, записанных в формате ASCII, разделенных пробелом, запятой или символом конца строки. Числа могут быть целыми или с плавающей запятой, записанными с десятичной точкой или в экспоненциальной форме. При обращении к файлу Mathcad по умолчанию обращается в ту папку (каталог, директорию), из которой загружался рабочий документ или в которую документ последний раз загружался. Однако можно работать с файлами из любых папок, указывая полное имя файла. В приведенных ниже примерах всегда будет указываться полное имя файла.

Функция READ (file) считывает значение из файла и присваивает его переменной. Поскольку чаще всего читаются массивы чисел, обращение к функции записывается следующим образом: Xi := READ(file).

Предположим, что на диске **c:** в папке **tmp** в файле с именем **data.txt** записаны 20 различных чисел, подготовленных текстовым процессором и сохраненных в указанном файле. Ниже представлен фрагмент рабочего документа Mathcad, в котором этот файл прочитан.

$i := 0..19$ $\qquad$ x_i := READ("c:\tmp\data.txt")

$x^T =$

	0	1	2	3	4	5	6	7	8
0	4.591	4.326	4.868	6.018	5.181	5.513	5.74	4.984	5.279

Указание. Поскольку в Mathcad массив _–_ это вектор-столбец, запишите в рабочем документе x^T , чтобы вывести массив x в виде строки. Для того чтобы просмотреть все данные, щелкните по полю вывода элементов массива и просмотрите содержимое массива с помощью линеек прокрутки.

Функция WRITE(file) записывает в файл на диске числовое значение переменной. Поскольку, как правило, записываются массивы чисел, то чаще всего она указывается следующим образом: WRITE(file) := x_i. Если файла с указанным именем не существует, то он будет создан; если такой файл есть, то при записи предыдущая информация будет потеряна.

Ниже приведен фрагмент рабочего документа Mathcad, в котором массив, сформированный в предыдущем фрагменте, записан в файл с именем **datal.txt** в папке **tmp** на диске **c:**, а затем (для проверки) прочитан и выведен в рабочий документ.

$i := 0..19$

WRITE("c:\tmp\data1.txt") := x_i y_i := READ("c:\tmp\data1.txt")

$y^T =$

	0	1	2	3	4	5	6	7	8
0	4.591	4.326	4.868	6.018	5.181	5.513	5.74	4.984	5.279

Указание. Здесь продемонстрирована работа двух функций доступа к файлам, позволяющих простейшим способом читать и записывать файлы. Для того чтобы аккуратно работать с файлами, необходимо обязательно ознакомиться с полным описанием, как с помощью этих функций происходит обращение к файлу.

Моделирование выборок из стандартных распределений

Mathcad обладает богатой библиотекой встроенных функций, предназначенных для генерирования выборок из генеральных совокупностей с наиболее распространенными стандартными распределениями. Например, для генерации нормального распределения предназначена функция rnorm(k,μ,σ), значением которой является вектор, содержащий *k* выборочных значений нормально распределенной случайной величины с математическим ожиданием $M\xi=\mu$ и дисперсией $D\xi=\sigma^2$. Ниже приведен список функций Mathcad, генерирующих выборки:

- Бета-распределение: rbeta(k,s_1,s_2).
- Биномиальное распределение: rbinom(k,n,p).
- Распределение Коши: rcauchy(k,l,s).
- χ^2 - распределение: rchisq(k,d).
- Экспоненциальное распределение: rexp(k,r).
- Распределение Фишера (F-распределение): rF(k,m,n).
- Гамма-распределение: rgamma(k,s).
- Геометрическое распределение: rgeom(k,p).
- Логнормальное распределение: rlnorm(k,μ,σ).
- Логистическое распределение: rlogis(k,l,s).
- Отрицательное биномиальное распределение: rnbinom(k,n,p).
- Нормальное распределение: rnorm (k,μ,σ).
- Распределение Пуассона: rpois(k,λ).
- Распределение Стьюдента: rt(k,d).
- Равномерное распределение: runif(k,a,b).
- Распределение Вейбулла: rweibull(k,s).

2. ОСНОВНЫЕ ЗАДАЧИ СТАТИСТИКИ. ВЫБОРКИ. ГИСТОГРАММЫ. ПОЛИГОНЫ ЧАСТОТ

Математическая статистика в основном занимается изучением случайных величин и случайных событий по результатам наблюдений. Ее главная задача – извлечь максимум информации из эмпирических данных.

Важнейшими понятиями математической статистики являются генеральная совокупность и выборка.

Генеральная совокупность — это вероятностное пространство с определенной на нем случайной величиной ξ. Функцию распределения этой случайной величины $F_\xi(x)$ часто называют *теоретической функцией распределения,* хотя более правильным представляется другой термин — *истинная функция распределения,* в отличие от эмпириеской (экспериментальной, приближенной) функции распределения, которая будет определена ниже.

В результате проведения n экспериментов со случайной величиной ξ получаем n выборочных значений x_i ,i = 1,2,..,n. Вся совокупность этих значений называется *выборкой.*

Выборка — это, вообще говоря, случайный вектор: если в одной серии из n испытаний получена выборка ($x_1,x_2,..,x_n$), то в другой серии будет получена, скорее всего, другая выборка ($x'_1,x'_2,..,x'_n$).

2.1 Эмпирические распределения и числовые характеристики

Выборка из генеральной совокупности является основным источником информации о случайной величине. По выборке оценивается класс распределений, к которому принадлежит распределение исследуемой случайной величины, устанавливаются интервалы, в которых лежат истинные значения параметров распределения, проверяются гипотезы об этой случайной величине и формулируются выводы о других ее свойствах.

Чтобы использовать аппарат математической статистики, нужно прежде всего уметь находить некоторые числовые характеристики выборок и строить эмпирические распределения, с помощью которых в дальнейшем можно делать соответствующие выводы.

Рассмотрим некоторые правила предварительной обработки выборочных данных. Представленная ниже таблица выборки объема n = 250 будет использоваться далее во всех вычислениях, а также станет источником построения выборок для индивидуальных вариантов заданий.

145.61	143.206	145.267	140.485	133.143	150.435	148.794	155.564	171.918
158.087	159.851	158.622	159.156	156.73	139.557	150.691	142.444	156.967
148.181	143.556	142.769	144.834	155.58	147.552	150.895	162.618	142.945
150.019	161.076	158.926	120.991	128.429	152.06	143.842	138.023	150.99
157.708	153.059	150.113	142.355	145.909	143.262	148.678	160.181	151.805
155.133	157.398	149.837	152.788	151.622	154.285	145.248	143.045	180.482
147.135	137.201	157.594	146.073	137.964	139.631	149.807	150.32	152.649
154.915	152.383	143.155	133.852	164.113	159.715	138.44	151.437	166.972
146.797	129.688	135.888	136.747	144.829	150.621	144.042	146.693	155.391
152.186	154.05	138.441	138.949	138.966	145.927	136.867	121.596	162.762
157.911	151.429	139.937	140.73	141.22	152.777	145.978	163.02	136.219
153.803	154.377	167.603	143.527	155.51	165.465	131.784	163.079	139.511
154.591	139.478	137.579	154.241	130.834	148.761	154.132	164.656	137.711
146.154	154.763	151.862	151.96	155.206	158.229	159.314	158.972	152.601
143.066	154.656	148.493	141.368	171.144	137.64	133.062	153.865	135.711
145.891	158.742	144.311	140.903	141.323	160.971	139.771	137.484	156.247
142.623	155.409	156.641	155.196	151.459	149.488	153.16	152.488	148.294
145.475	152.937	151.507	140.659	157.925	157.163	160.438	158.11	156.17
147.549	149.142	156.848	157.911	153.578	147.887	148.445	151.36	158.639
169.584	150.688	155.646	155.572	168.911	164.788	127.059	156.623	145.593
145.263	150.889	143.012	153.472	141.25	169.001	122.741	158.702	171.791
160.849	161.757	140.286	134.241	154.64	164.744	161.654	142.365	155.094
154.96	141.977	143.729	144.466	146.54	145.355	152.509	146.266	147.269
162.895	151.941	170.865	134.377	150.79	154.205	166.274	156.198	132.828
136.274	173.96	157.332	149.975	141.54	139.826	133.692	139.462	161.159
159.455	157.597	139.385	145.867	166.069	150.237	146.685	145.436	153.969
154.961	149.211	150.83	154.224	142.28	148.655	135.371	152.018	166.807
140.923	157.864	148.745	138.823	157.239	152.912	141.182		

Объемом выборки называют количество наблюдений или количество значений случайной величины.

Первичная обработка данных состоит обычно в отыскании максимального x_{max} и минимального x_{min}значений выборки (в Mathcad они вычисляются соответственно функциями max(ξ) и min(ξ)), а также размаха варьирования $R = x_{max} - x_{min}$. Для приведенной выше выборки эти величины равны: $x_{max} = 180.482$, $x_{min} = 120.991$, $R = 59.49$.

Следующий этап первичной обработки – группировка и ее графическое представление. Группировка выборки объема *n* состоит в следующем. Промежуток $[x_{min}, x_{max}]$ разбивают на m интервалов группировки (чаще всего одинаковой длины) и подсчитывают число n_j выборочных значений, которые попали в j-й интервал. Обычно выбирают m = 7- 20. Теперь каждый интервал группировки $\Delta_j = (a_j, b_j)$представлен своими левой a_j и правой b_j границами и числом n_j элементов выборки, принадлежащих ему. Каждый интервал удобно представлять не двумя границами, а одним числом – срединным значением.

Наиболее наглядная форма графического представления группировки – гистограмма.

Если $\delta_1,\dots,\delta_m$ – длины интервалов группировки, а $x_1,\dots,x_m$– их середины и $h_j= =n_j/n$ – относительные частоты попадания наблюдений в j-й интервал группировки, то можно построить график ступенчатой функции:

$f(x)= h_j /\delta_j , x\in\Delta_j, j=1,2,..,m.$

Этот график называется гистограммой. В Mathcad для построения гистограмм предназначена функция hist(Δ,ξ).

Указание. Прежде чем приступать к группировке выборок, нужно их упорядочить с помощью функции sort. Перед обращением к функции hist следует вычислить середины интервалов группировки и присвоить их значения элементам массива *x*.

Очевидно, что величина интервала группировки существенно влияет на вид гистограммы. При малой их ширине в каждый интервал попадает незначительное число наблюдений или даже не попадает ни одного, в результате гистограмма становится сильно «изрезанной» и плохо передает основные особенности изучаемого распределения. Другая крайность – большие интервалы группировки; в этом случае скрадываются характерные черты распределения.

Иная форма графического представления группированных данных – полигон частот. *Полигон частот* – это ломаная линия, соединяющая точки с координатами (x_i,h_i), т.е. с абсциссами, равными серединам интервалов группировки, и ординатами, равными соответствующим частотам. Можно также построить *полигон накопленных частот* – график ломаной, соединяющей точки с координатами

$$\left(b_j, \sum_{k=1}^{j} n_k\right) \text{ или } \left(b_j, \sum_{k=1}^{j} \frac{n_k}{n}\right),$$

т.е с абцисcами, равными правым границам интервалов группировки, и ординатами, равными соответствующим накопленным частотам или относительным накопленным частотам.

Ниже приведен фрагмент рабочего документа Mathcad с вычислением x_{min} ,x_{max} и $R= x_{max}-x_{min}$ для исследуемой выборки, а также с гистограммами и полигонами частот для различных интервалов группировки.

При первичной обработке выборочных данных можно рекомендовать несколько общих правил:

1. Перед началом группировки следует упорядочить выборочные значения в порядке возрастания. Такая упорядоченная в порядке возрастания выборка называются *вариационным рядом.*

2. При выборе числа интервалов группировки следует ориентироаться на 10-20 интервалов.
3. Предпочтительнее использовать интервалы одинаковой длины.
4. При анализе охватывайте всю область данных.
5. Избегайте полуоткрытых промежутков.
6. Интервалы группировки не должны перекрываться.

ЗАДАНИЕ 1

Вычислите максимальное, минимальное значения и размах для заданной части приведенной выше выборки. Выполните группировку для заданных значений m, постройте соответствующие гистограммы, полигоны частот и полигоны накопленных частот.

Порядок выполнения задания

1. Определите и введите вектор-столбец выборочных значений.
2. Упорядочите выборку в порядке возрастания выборочных значений.
3. Вычислите минимальное значение и размах для полученной выборки.
4. Определите число интервалов группировки и их длину.
5. Определите вектор-столбец, содержащий середины интервалов группировки.
6. Определите с помощью функции hist(x,ξ) вектор-столбец частот для полученных интервалов группировки.
7. Определите вектор-столбец накопленных частот.
8. Постройте гистограмму, полигон частот.
9. Постройте полигон накопленных частот и полигон относительных накопленных частот.
10. Выполните вычисления пп. 6-9 для всех заданных значений *m*.
11. Сохраните рабочий документ в файле на диске.

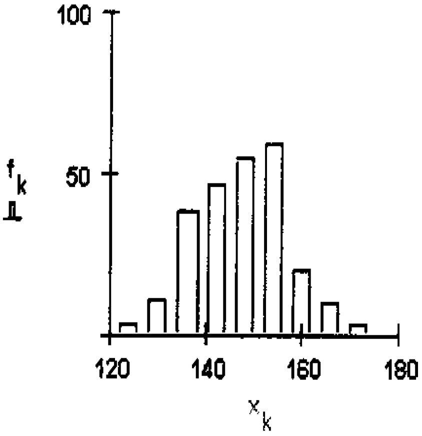
ORIGIN := 1

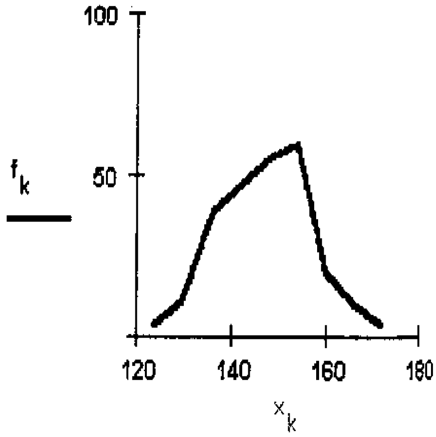
xmax := max(ξ)   xmin := min(ξ)   R := xmax − xmin

xmax = 180.482   xmin = 120.991   R = 59.491   n := 250

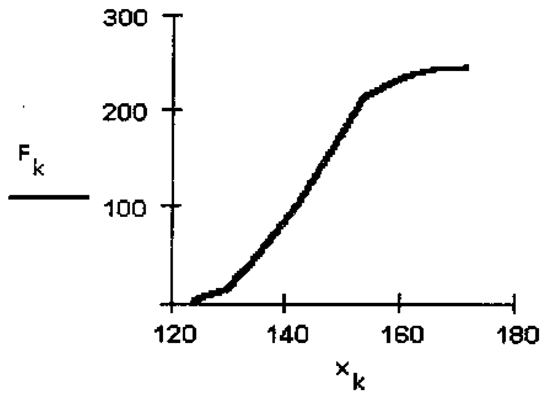
$m := 10 \qquad \Delta := \frac{R}{m} \qquad j := 1..m \qquad k := 1..m-1$

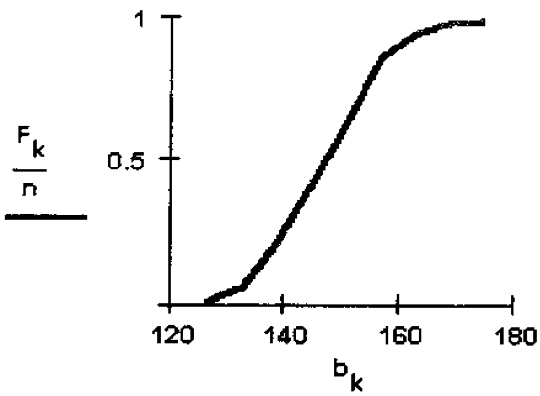
$x_j := xmin + \frac{\Delta}{2}\cdot(2\cdot j - 1) \qquad f := hist(x,\xi) \qquad \Delta = 5.949$





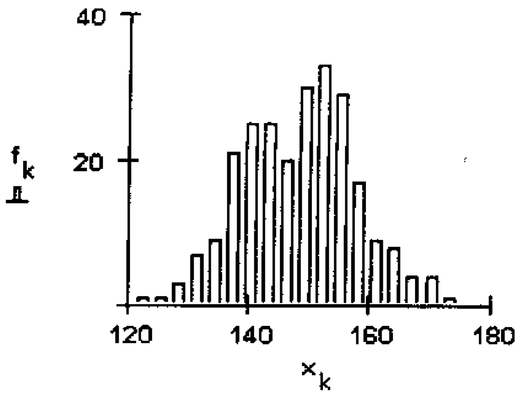
$$a_j := xmin + \Delta\cdot(j-1) \qquad b_j := a_j + \Delta \qquad F_k := \sum_{j=1}^{k} f_j$$

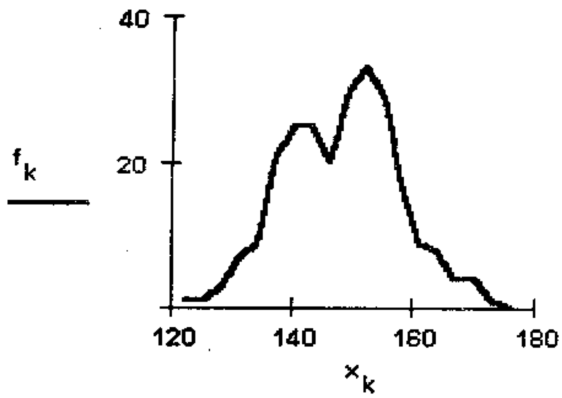




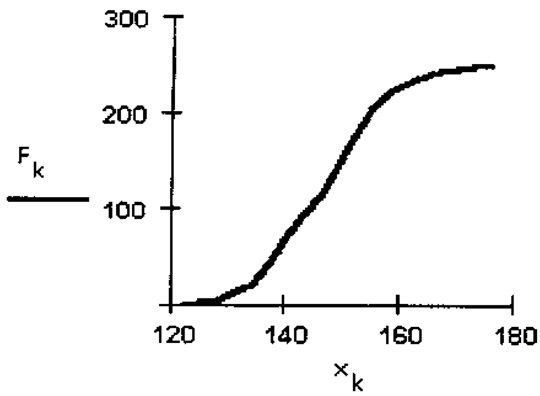
$$m := 20 \qquad \Delta := \frac{R}{m} \qquad j := 1..m \qquad k := 1..m-1$$

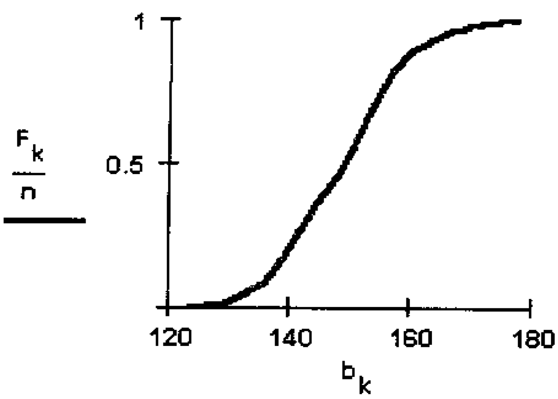
$$x_j := xmin + \frac{\Delta}{2}\cdot(2\cdot j - 1) \qquad f := hist(x, \xi) \qquad \Delta = 2.975$$





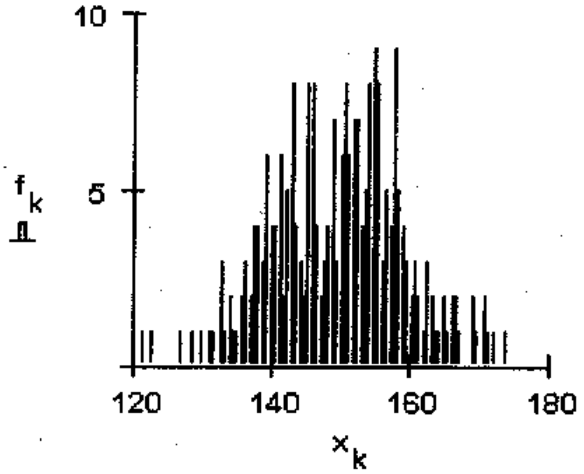
$$a_j := xmin + \Delta\cdot(j-1) \qquad b_j := a_j + \Delta \qquad F_k := \sum_{j=1}^{k} f_j$$

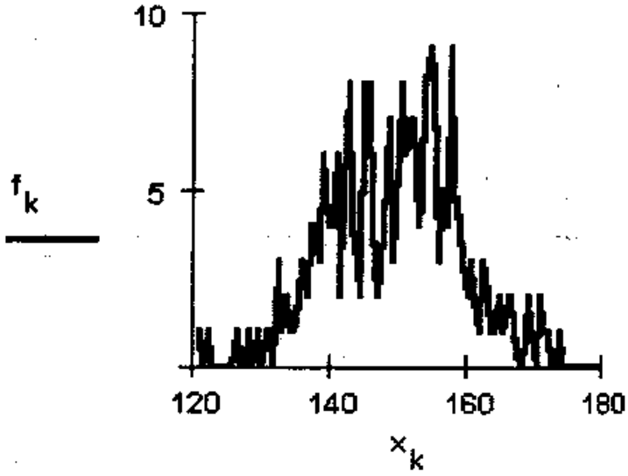




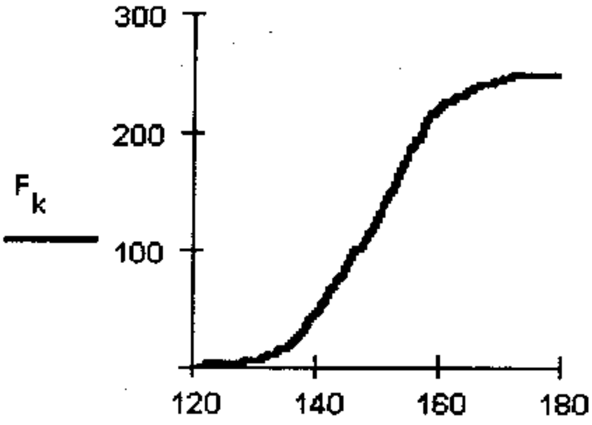
$m := 100 \qquad \Delta := \dfrac{R}{m} \qquad j := 1 .. m \qquad k := 1 .. m - 1$

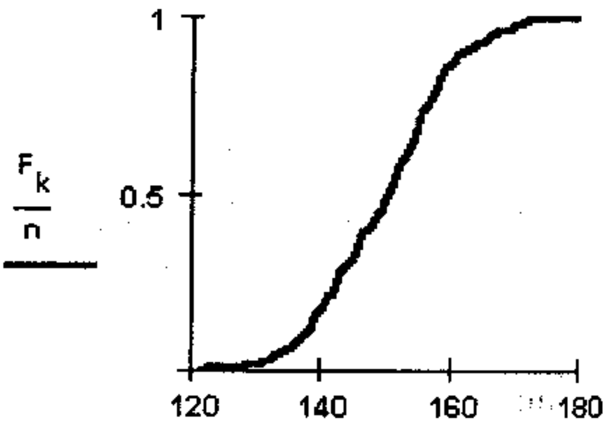
$x_j := xmin + \dfrac{\Delta}{2} \cdot (2 \cdot j - 1) \qquad f := hist(x, \xi) \qquad \Delta = 0.595$





$a_j := xmin + \Delta \cdot (j - 1) \qquad b_j := a_j + \Delta \qquad F_k := \displaystyle\sum_{j=1}^{k} f_j$





Указание. В приведенном фрагменте 250 выборочных значений сохранены в массиве с именем ξ. Прежде чем приступать к группировке выборки, необходимо упорядочить выборочные значения в порядке их возрастания. Эту операцию выполняет функция sort(ξ). Группировка производится с помощью функции hist(x, ξ), где x – массив, содержащий значения середин интервалов группировки. Прежде чем обратиться к функции hist(x,ξ), необходимо вычислить середины интервалов группировки и присвоить их значения элементам массива x. Значения функции hist(x,ξ) – вектор, компоненты которого равны количеству элементов массива ξ, попавших в интервал

Пример выполнения задания

Примерный вариант выполнения задания для всей выборки для m = 10, 20, 100 приведен выше.

2.2 Числовые характеристики выборки

Показатели положения. Среднее значение выборки вычисляется по формуле

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

В Mathcad для вычисления выборочного среднего значения выборки, сохраненной в матрице *A,* предназначена функция mean(A). *Выборочной квантилью* уровня *p* называется решение уравнения

$$F_n(x) = p,$$

где $F_n(x)$ — выборочная функция распределения.

В частности, *выборочная медиана* есть решение уравнения $F_n(x) = 0.5$, т.е. выборочная медиана — это выборочная квантиль уровня 0.5. Выборочная медиана разбивает выборку пополам: слева и справа от нее оказывается одинаковое число элементов выборки. Если число элементов выборки четно, n = 2k, то выборочную медиану определяют по формуле:$(x_k+x_{k+1})/2$, где x_k, x_{k+1} – k-е и (k+1)-е выборочные значения из вариационного ряда. При нечетном объеме выборки (n=2k+1) в качестве значения медианы принимают величину x_{k+1}.

В Mathcad для вычисления выборочной медианы выборки, сохраненной в матрице *A,* предназначена функция median(A).

К показателям положения относятся минимальный и максимальный элементы выборки, а также верхняя и нижняя квартили (они ограничивают зону, в которой сосредоточены 50% элементов выборки)

Для вычисления минимального и максимального элементов выборки, размещенной в матрице *A,* в Mathcad предназначены соответственно функции min(A) и max(A).

Показатели разброса. К показателям разброса относятся дисперсия выборки (выборочная дисперсия), стандартное отклонение, размах выборки, межквартильный размах, коэффициент эксцесса (выборочный эксцесс). *Выборочной дисперсией* называется величина

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

Однако в статистике чаще в качестве выборочной дисперсии используется величина

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

В Mathcad для определения дисперсии выборки, сохраненной в матрице A, предназначена функция var(A), а величину s^2 можно вычислить по формуле:

$$s^2 = \frac{1}{n-1}\mathrm{var}(A).$$

Стандартное отклонение рассчитывается по формуле:

$$\hat{\sigma} = \sqrt{s^2}.$$

Размах выборки вычисляется по формуле $R = x_{max} - x_{min}$

Межквартильный размах равен $x_{0.75} - x_{0.25}$ где $x_{0.75}$ – 75%-ная квартиль, решение уравнения $F_n(x_{0.75}) = 0.75$, $x_{0.25}$ – 25%-ная квартиль, решение уравнения $F_n(x_{0.25}) = 0.25$.

Выборочный эксцесс определяется следующим образом. Сначала отыскивается величина выборочного центрального момента 4-го порядка

$$\hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4.$$

А затем по следующей формуле вычисляется выборочный эксцесс:

$$\hat{E} = \hat{\mu}_4 (s^2)^{-2} - 3.$$

Показатели асимметрии. На основании этих показателей изучают информацию о симметрии распределения выборочных данных около центра выборки. Сюда в первую очередь относится коэффициент асимметрии, который вычисляется по формуле

$$\hat{a} = \frac{\hat{\mu}_3}{\hat{\sigma}^3},$$

где $\hat{\mu}_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3$ — выборочный центральный момент 3-го порядка, а σ – стандартное отклонение, формула для вычисления которого приведена выше.

ЗАДАНИЕ 2

Для выборки, сформированной в предыдущем задании, вычислите все описанные выше выборочные характеристики.

Порядок выполнения задания

1. Прочтите сохраненный ранее файл, содержащий выборку.
2. Вычислите максимальный и минимальный элементы и размах выборки.
3. Рассчитайте выборочное среднее.
4. Найдите медиану.
5. Вычислите выборочную дисперсию и стандартное отклонение.
6. Найдите выборочные моменты 3-го и 4-го порядков.
7. Вычислите выборочный эксцесс.
8. Определите коэффициент асимметрии.

Пример выполнения задания

Ниже представлен фрагмент рабочего документа Mathcad, содержащий вычисление характеристик выборочных данных, приведённых в начале раздела.

$$n := 250$$

$$xmax := max(\xi) \qquad xmin := min(\xi) \qquad R := xmax - xmin$$

$$xmax = 180.482 \qquad xmin = 120.991 \qquad R = 59.49$$

$$mean := mean(\xi) \qquad s2 := \frac{n}{n-1}\cdot var(\xi) \qquad \sigma := \sqrt{s2}$$

$$mean = 149.849 \qquad s2 = 98.174 \qquad \sigma = 9.908$$

$$\mu 3 := \frac{1}{n}\cdot\sum_{i=1}^{n}\left(\xi_i - mean\right)^3 \qquad \mu 4 := \frac{1}{n}\cdot\sum_{i=1}^{n}\left(\xi_i - mean\right)^4$$

$$median := median(\xi) \qquad E := \frac{\mu 4}{s2^2} - 3 \qquad \alpha := \frac{\mu 3}{\sigma^3}$$

$$median = 150.69 \qquad E = 0.136 \qquad \alpha = -0.055$$

Указание. В Mathcad нет встроенных функций для вычисления выборочных моментов. Для определения среднеквадратичного отклонения в Mathcad предназначена функция

$$stdev(A) = \sqrt{var(A)}.$$

Рассчитываемое с ее помощью значение среднеквадратичного отклонения отлично от определенного выше, поэтому среднеквадратичное отклонение следует вычислять как

$$\sqrt{s^2}.$$

2.3 Оценка функции распределения

Рассмотрим методы оценивания функции распределения $F_\xi(x)$ случайной величины, о которой известно, что она непрерывна.

Пусть $x=\{x_1,x_2,..,x_n\}$ – совокупность выборочных значений случайной величины ξ, т.е. выборка из случайной величины ξ. Расположим наблюдения $x_1,x_2,..,x_n$ в порядке их возрастания. Обозначим новую упорядоченную последовательность – вариационный ряд - $x`_1,x`_2,..,x`_n$, $x`_1<x`_2<..<x`_n$. По этому вариационному ряду построим следующую неубывающую ступенчатую функцию:

$$\hat{F}_n(x)=\begin{cases}0, x\le x'_1\\ \dfrac{k}{n}, x'_{k-1}\le x\le x'_k, k=1,2,...,n,\\ 1, x>x'_n\end{cases}$$

Из приведенной выше формулы видно, что функция $\hat{F}_n(x)$ претерпевает в каждой точке вариационного ряда скачок, равный по величине 1/n. Если какая-нибудь точка вариационного ряда повторяется *m* раз (*m* точек вариационного ряда совпадают), то скачок функции $\hat{F}_n(x)$ в этой точке равен *m/n.*

Функция $F_n(x)$ называется *эмпирической функцией распределения.*

Замечание. Эмпирическая функция распределения $\hat{F}_n(x)$ зависит не только от x, но и от всей выборки x. Чтобы обратить внимание на этот факт, будем обозначать эмпирическую функцию распределения через $F_n(x,\hat{x})$. Именно $F_n(x,\hat{x})$ принимают за оценку теоретической функции распределения *F(x).*

Остается выяснить, насколько хорошо эмпирическая функция распределения аппроксимирует теоретическую функцию распределения.

Если $F_\xi(x)$ – теоретическая функция распределения, а $F_n(x)$ — эмпирическая функция распределения, построенная по заданной выборке $\hat{x}$ значений случайной величины ξ, то в качестве меры расхождения теоретической и эмпирической функций распределения возьмем величину:

$$D_n(\hat{x})=\sup_x\left|F_n(x)-F_\xi(x)\right|.$$

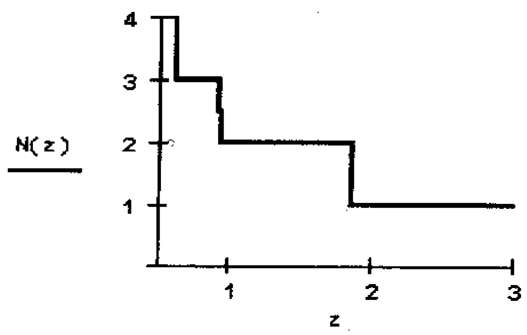
Эта функция от выборочных значений x называется *статистикой Колмогорова.* Следует помнить, что $D_n(\hat{x})$ случайная величина и что ее распределение не зависит от неизвестной теоретической функции распределения $F_\xi(x)$, если она непрерывна. Более того, справедлива теорема Колмогорова: если функция распределения $F_\xi(x)$ случайной величины ξ непрерывна, а $F_n(x)$ – ее выборочная функция распределения, то при $n\to\infty$

$$P\left(\sup_x\left|\hat{F}_n(x)-F_n(x)\right|<\frac{z}{\sqrt{n}}\right)\to K(z)=\begin{cases}0, z\le 0,\\ \sum\limits_{k=-\infty}^{\infty}(-1)^k e^{-2k^2z^2}, z>0.\end{cases}$$

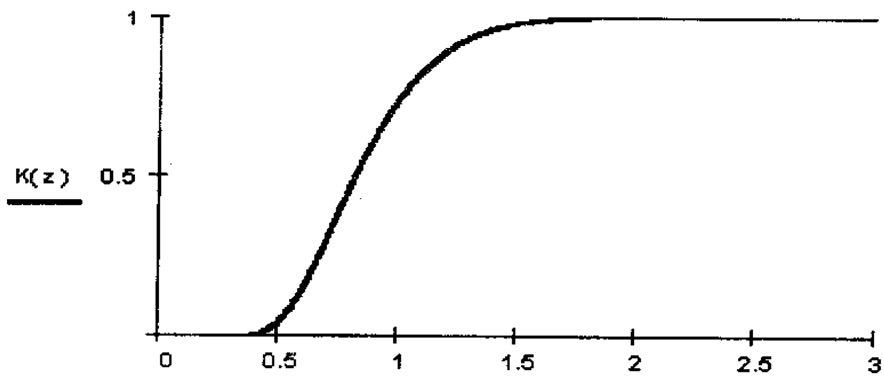
Функция K(z) представляет собой функциональный ряд, который следует протабулировать. Сразу обратим внимание на то, что этот ряд сходится абсолютно для всех z>0, но неравномерно на промежутке [0, +∞]. Это означает, что для достижения заданной точности при вычислении K(z) число N членов в соответствующей частичной сумме зависит от z. Если ε – требуемая точность вычисления К(z), то число N вычисляется по формуле $N=\left[\frac{1}{z}\sqrt{\frac{1}{2}\ln\frac{1}{\varepsilon}}\right]+1$, где символом [ ] обозначена целая часть числа.

Ниже приведен фрагмент рабочего документа Mathcad, содержащий приближенное определение функции K(z) для ε = 0.001, N = 3, и соответствующие графики.

$$\varepsilon := 0.001 \qquad N(z) := \mathrm{floor}\left(\frac{1}{z}\cdot\sqrt{\frac{1}{2}\cdot\ln\left(\frac{1}{\varepsilon}\right)}\right)+1$$



$$N := 3 \qquad K(z) := \left|\begin{array}{l}0 \text{ if } z\le 0\\ \sum\limits_{k=-N}^{N}(-1)^k\cdot\exp(-2\cdot k^2\cdot z^2) \text{ if } z>0\end{array}\right.$$



Из приведенных в документе графиков видно, что для малых z величину K(z) можно положить равной нулю, а для $z > 2$ можно считать K(z) равной единице.

Зададимся вероятностью α такой, что событие, происходящее с вероятностью 1-α, представляется практически достоверным. Вычислим корень z_α уравнения $1 - K(z) = \alpha$, тогда неравенство:

$$\hat{F}_n(x) - \frac{z_\alpha}{\sqrt{n}} < F_\xi(x) < \hat{F}_n(x) + \frac{z_\alpha}{\sqrt{n}}$$

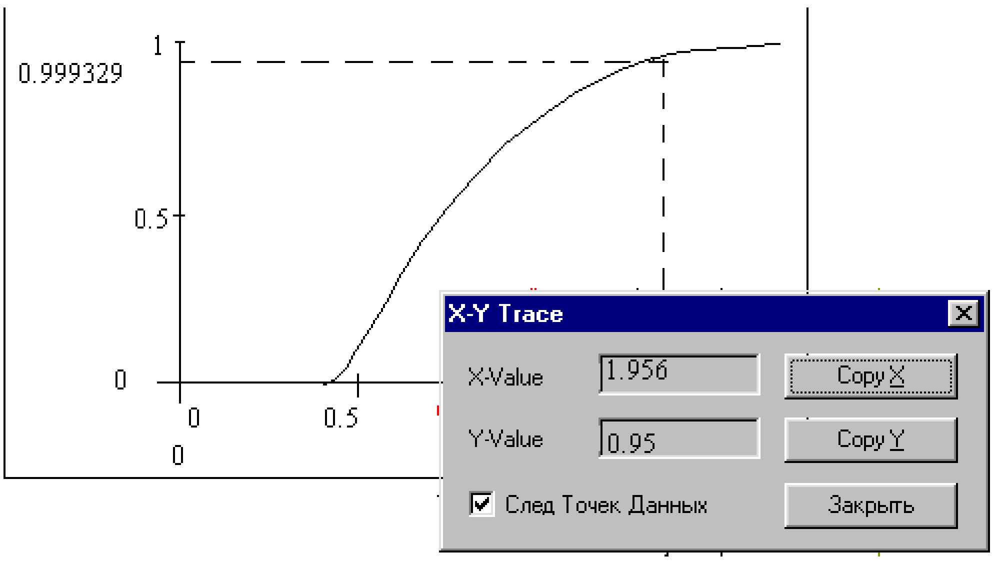
выполняется для всех действительных x с вероятностью, близкой к 1-α.

Таким образом, в окрестности эмпирической функции распределения построен "коридор", в котором лежит истинная, теоретическая функция распределения $F_\xi(x)$. С ростом *n* "ширина" этого коридора стремится к нулю.

Вместо эмпирической функции распределения будем использовать функцию накопленных относительных частот, поскольку $\hat{F}_n(x) = F_k$, для $x \in (\xi_{k-1}, \xi_k]$ и значения функций совпадают вне промежутка $[x_{min}, x_{max}]$.

На следующей странице приведен фрагмент рабочего документа Mathcad с построением 95%-ного "коридора" для функции распределения случайной величины по приведенной выборке.

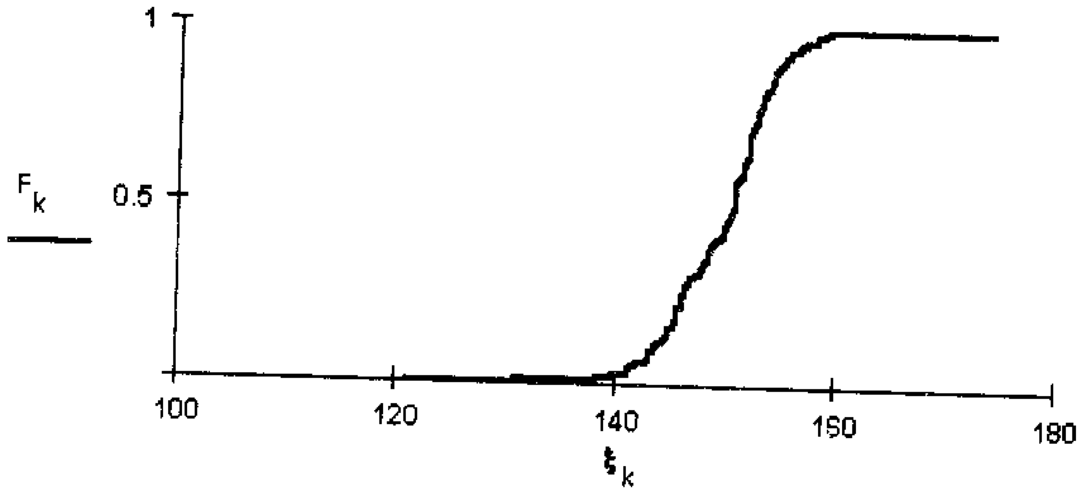
Указание. Как уже отмечалось выше, в качестве эмпирической функции распределения использована эмпирическая функция накопленных частот. Заметим, что Mathcad вместо графика ступенчатой функции строит ломаную линию, соединяя "ступеньки" вертикальными отрезками прямых. Корень уравнения $1 - K(z) = \alpha$ проще всего найти графически, используя операцию **Trace** пункта **Graph** меню Format), как точку пересечения графика K(z) и прямой $y = 1-\alpha$. Ниже приведен фрагмент окна Mathcad с окном отображения координат точки пересечения.



$m := 250 \qquad \Delta := \frac{R}{m} \qquad j := 1..m \qquad k := 1..m-1$

$x_j := xmin + \frac{\Delta}{2}\cdot(2\cdot j - 1) \qquad f := hist(x, \xi) \qquad \Delta = 0.238$

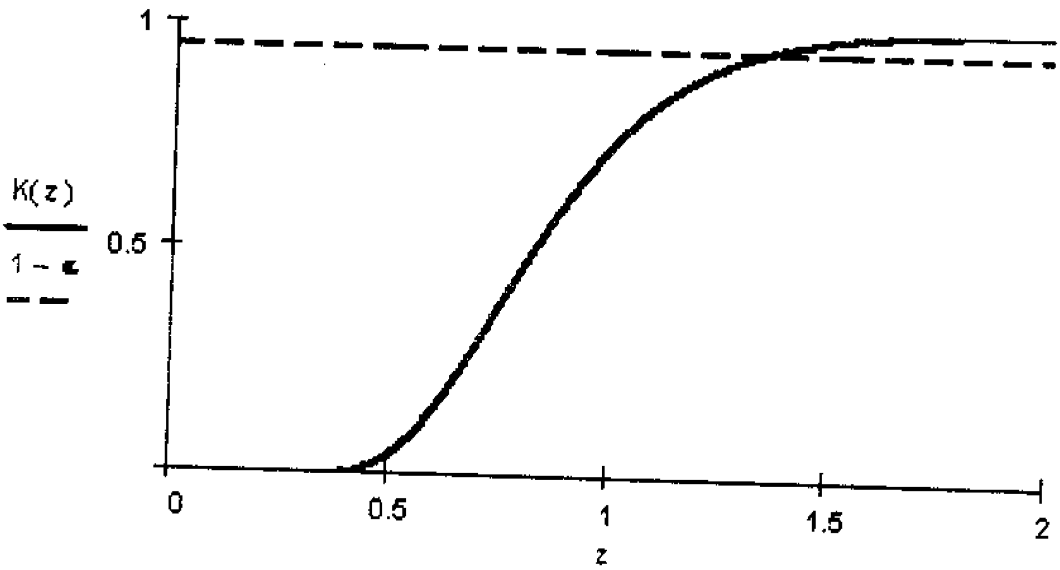
$a_j := xmin + \Delta\cdot(j-1) \qquad b_j := a_j + \Delta \qquad F_k := \sum_{j=1}^{k} \frac{f_j}{n}$

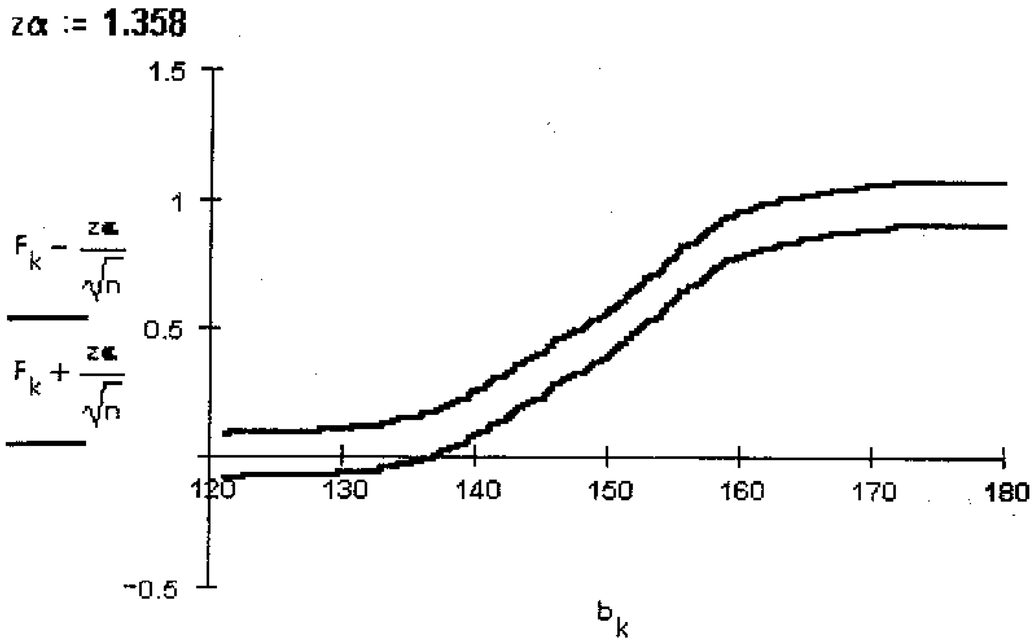


$\alpha := 0.05$

$p := 1 - \alpha$

$$K(z) := \begin{cases} 0 & \text{if } z \le 0 \\ \sum_{k=-3}^{3} (-1)^k \cdot \exp(-2\cdot k^2 \cdot z^2) & \text{if } z > 0 \end{cases}$$





Для оценки плотности распределения случайной величины можно воспользоваться полигоном частот, который представлен выше. При не очень строгих ограничениях доказано, что выборочная плотность вероятностей, т.е. полигон частот, с ростом объема выборки до бесконечности стремится к истинной, теоретической плотности распределения исследуемой случайной величины.

ЗАДАНИЕ 3

Постройте для выборки, сформированной в задании 1, 95%-ный "коридор" для функции распределения исследуемой случайной величины.

Порядок выполнения задания

1. Прочитайте файл, сохраненный при выполнении задания 1.
2. Определите статистику Колмогорова — функцию K(z) и постройте ее график.
3. Определите значение величины α.
4. Решите графически уравнение $1 - K(z) = \alpha$.
5. Постройте "коридор" для теоретической функции распределения.

Пример выполнения задания

Пример построения 95%-ного "коридора" функции распределения для исследуемой во всех примерах этого раздела выборки 250 значений случайной величины приведен выше.

3. ТОЧЕЧНЫЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЙ

Предположим, что функция распределения случайной величины ξ зависит от неизвестного параметра θ: $P(\xi<X) = F_\xi(x,\theta)$. Если $x_1,x_2,..,x_n$ – выборка из генеральной совокупности случайной величины ξ, то оценкой $\hat{\theta}$ параметра θ называется произвольная функция от выборочных значений $\hat{\theta}=\hat{\theta}(x_1,x_2,..,x_n)$. Конечно, используемые на практике оценки $\hat{\theta}_n(x_1,x_2,..,x_n)$ не совсем произвольные функции: они обладают рядом свойств, которые обеспечивают в

некотором смысле оптимальное извлечение информации из выборок. Рассмотрим эти свойства.

Точечные оценки математического ожидания

Следует отметить, что значение оценки $\hat{\theta}$ меняется от выборки к выборке и, значит, $\hat{\theta}$ есть случайная величина. Естественно потребовать, чтобы значения этой случайной величины в большинстве экспериментов были близки к значению оцениваемого параметра. Этого можно достигнуть, если для любого значения n математическое ожидание величины $\hat{\theta}_n$ равно истинному (теоретическому) значению параметра $\hat{\theta}$: $M\hat{\theta}_n = \theta$.

Оценки $\hat{\theta}_n$, удовлетворяющие условию $M\hat{\theta}_n = \theta$, называются *несмещенными.* Несмещенность оценки означает, что эта оценка не несет в себе систематической ошибки.

Еще одно важное свойство, которым должны обладать оценки, — состоятельность. Оценка $\hat{\theta}_n$ называется *состоятельной* оценкой параметра θ, если для любого ε>0 справедливо: lim $P(|\theta_n$^-θ|<ε) =1.

Поясним смысл последнего равенства. Пусть ε>0 – как угодно малое положительное число. Тогда с ростом *n* растет наша уверенность в том, что значение оценки $\hat{\theta}_n$ отличается от истинного значения параметра θ не более чем на величину ε, т.е. с ростом объема выборки увеличивается точность результата. Правда, здесь приходится отойти от традиционного понятия точности: нет гарантии, что $|\hat{\theta}_n - \theta| < \varepsilon$, однако утверждается, что, начиная с некоторого *n,* событие $|\hat{\theta}_n - \theta| < \varepsilon$ становится практически достоверным.

Вообще говоря, для оценки одного и того же параметра можно придумать много различных несмещенных и состоятельных оценок. Рассмотрим несколько примеров.

Пусть x_1, x_2, .., x_n – выборка из генеральной совокупности, соответствующей случайной величине ξ с неизвестным математическим ожиданием $M\xi=\theta$ и известной дисперсией $D\xi=\sigma^2$. Построим несколько оценок неизвестного параметра $M\xi=\theta$. Например, если $\hat{\theta}_n^{(1)} = x_1$, то $M\hat{\theta}_n^{(1)} = x_1$, т.е. рассматриваемая оценка является *несмещенной оценкой.* Поскольку значение $\hat{\theta}_n^{(1)} = x_1$ вообще не зависит от объема выборки n, то оценка $\hat{\theta}_n^{(1)}$ не является *состоятельной.* Рассмотрим другую оценку:

$$\hat{\theta}_n^{(2)} = \frac{x_1 + 2x_2 + ... + nx_n}{\frac{n(n+1)}{2}}.$$

Найдем

$$M\hat{\theta}_n^{(2)} = M\left(\frac{\xi_1 + 2\xi_2 + \ldots + n\xi_n}{\frac{n(n+1)}{2}}\right) =$$

$$= \frac{2}{n(n+1)} \cdot \theta(1 + 2 + \ldots + n) =$$

$$= \frac{2\theta}{n(n+1)} \cdot \frac{n(n+1)}{2} = \theta,$$

т.е. имеем несмещенную оценку. Проверим ее состоятельность.

$$D\hat{\theta}_n^{(2)} = D\left(\frac{\xi_1 + 2\xi_2 + \ldots + n\xi_n}{\frac{n(n+1)}{2}}\right) = \frac{4}{n^2(n+1)^2} \cdot \sigma^2(1^2 + 2^2 + \ldots + n^2) =$$

$$= \frac{4\sigma^2}{n^2(n+1)^2} \cdot \frac{n(n+1)(2n+1)}{6} = \frac{2(2n+1)}{3n(n+1)}\sigma^2.$$

Поскольку $D\hat{\theta}_n^{(2)} \to 0$ при n→∞, то, согласно закону больших чисел, для ε> 0 справедливо $\lim P(\left|\hat{\theta}_n^{(2)} - \theta\right| < \varepsilon) = 1$, т.е. имеем состоятельную оценку.

Рассмотрим еще одну оценку неизвестного математического ожидания:

$$\hat{\theta}_n^{(3)} = \frac{x_1 + x_2 + \ldots + x_n}{n},$$

у которой

$$M\hat{\theta}_n^{(3)} = \frac{1}{n}\left(M\xi_1 + M\xi_2 + \ldots + M\xi_n\right) = \frac{1}{n}\theta_n = \theta$$

$$D\hat{\theta}_n^{(3)} = \frac{1}{n^2}\left(D\xi_1 + D\xi_2 + \ldots + D\xi_n\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n} \to 0$$

при n→∞, т.е. последняя оценка также несмещенная и состоятельная. Какой же оценке отдать предпочтение? Для того чтобы ответить на этот вопрос, сравним дисперсии последних двух оценок.

Поскольку $D\hat{\theta}_n^{(3)} = \frac{\sigma^2}{n}, D\hat{\theta}_n^{(2)} = \frac{2(2n+1)}{3n(n+1)}\sigma^2 = \frac{\sigma^2}{n}\left(1 + \frac{n-1}{3(n+1)}\right),$ то $D\hat{\theta}_n^{(2)} > D\hat{\theta}_n^{(3)}$, т.е. оценка $\hat{\theta}_n^{(3)}$дает меньший разброс около значений параметра θ и, следовательно, предпочтительнее. В таком случае говорят, что оценка $\hat{\theta}_n^{(3)}$*эффективнее* оценки $\hat{\theta}_n^{(2)}$. Теперь естественно поставить вопрос о построении самой эффективной оценки – об оценке с минимальной дисперсией. Такая оценка называется *эффективной* оценкой. Доказано, что эффективной оценкой математического ожидания нормально распределенной случайной величины является оценка $\hat{\theta}_n$=($x_1+x_2+\ldots+x_n$)/n. Именно поэтому последняя оценка так широко используется в математической статистике. Итак, впредь для оценки неизвестного математического ожидания случайной величины будем использовать *выборочное среднее*, т. е:

$$\hat{\theta}_n = \overline{x} = \frac{x_{1+} + x_2 + \ldots + x_n}{n}.$$

Точечные оценки дисперсии

Для дисперсии σ^2 случайной величины ξ можно предложить следующую оценку:

$$\overline{DX} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2,$$ где $\overline{x}$ – выборочное среднее,

Доказано, что эта оценка состоятельная, но *смещенная.*

В качестве состоятельной несмещенной оценки дисперсии используют величину

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

Именно несмещенностью оценки s^2 объясняется ее более частое использование в качестве оценки величины Dξ.

Заметим, что Mathcad предлагает в качестве оценки дисперсии величину $\overline{\mathrm{DX}}$, а не s^2: функция var(x) вычисляет величину:

$$\frac{1}{n}\sum_{i=1}^{n}\left(x_i - mean(x)\right)^2$$, где mean(x) - выборочное среднее: $\frac{1}{n}\sum_{i=1}^{n} x_i$.

ЗАДАНИЕ 4

Найдите состоятельные несмещенные оценки математического ожидания Mξ и дисперсии Dξ случайной величины ξ по приведенным в задании выборочным значениям $x_1, x_2, .., x_n$.

Порядок выполнения задания

1. Прочитайте с диска файл, содержащий выборочные значения, или введите заданную выборку с клавиатуры.
2. Вычислите точечные оценки Mξ и Dξ.

Пример выполнения задания

Найдите состоятельные несмещенные оценки математического ожидания Mξ и дисперсии Dξ случайной величины ξ по выборочным значениям, заданным следующей таблицей.

X	904.3	910.2	916.6	928.8	935.0	941.2	947.4	953.6	959.8	966.0	972.2	978.4
N	1	3	1	1	1	1	2	1	1	1	2	1

Для выборки, заданной таблицей такого типа (приведено выборочное значение и число, указывающее, сколько раз это значение встречается в выборке), формулы для состоятельных несмещенных оценок математического ожидания и дисперсии имеют вид:

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

где k – количество значений в таблице; n_i – количество значений x_i в выборке, n – объем выборки.

Фрагмент рабочего документа Mathcad с вычислениями точечных оценок приведен ниже. Из приведенных вычислений видно, что смещенная оценка дает

$$\mathrm{ORIGIN} := 1$$

$$i := 1..12$$

$$D1 := \begin{pmatrix} 1 & 904.3 \\ 3 & 910.2 \\ 1 & 916.6 \\ 1 & 928.8 \\ 1 & 935.0 \\ 2 & 941.2 \end{pmatrix} \qquad D2 := \begin{pmatrix} 1 & 947.4 \\ 1 & 953.6 \\ 1 & 959.8 \\ 1 & 966.0 \\ 2 & 972.2 \\ 1 & 978.4 \end{pmatrix}$$

$$D := \mathrm{stack}(D1, D2)$$

$$n := \sum_{i=1}^{12} D_{i,1} \qquad n = 16$$

$$Mx := \frac{1}{n}\sum_{i=1}^{12} D_{i,1}\, D_{i,2} \qquad Mx = 940.456$$

$$Dx := \frac{1}{n-1}\cdot\sum_{i=1}^{12} D_{i,1}\cdot\left(D_{i,2} - Mx\right)^2 \qquad Dx = 632.811$$

$$Dx1 := \frac{1}{n}\cdot\sum_{i=1}^{12} D_{i,1}\cdot\left(D_{i,2} - Mx\right)^2 \qquad Dx1 = 593.26$$

4. МЕТОДЫ ПОЛУЧЕНИЯ ТОЧЕЧНЫХ ОЦЕНОК

Известны методы получения точечных оценок: метод моментов, метод максимального правдоподобия и метод наименьших квадратов.

Оценки, полученные методом максимального правдоподобия, обладают хорошими *асимптотическими* свойствами: при $n\to\infty$ они становятся эффективными, несмещенными, состоятельными. Познакомимся с этим методом на примерах.

4.1 Метод максимального правдоподобия для дискретной случайной величины

Пусть ξ – дискретная случайная величина, распределенная по закону Пуассона с неизвестным параметром λ, т.е. $P(\xi = m) = \dfrac{\lambda^m}{m!}e^{-\lambda}$ и m_1, m_2,.., m_n – результаты независимых наблюдений случайной величины ξ. Задача состоит в построении точечной оценки неизвестного параметра λ. Для ее решения введем в рассмотрение *функцию правдоподобия,* заданную равенством

$$L(m_1, m_2,.., m_n) = p(\xi_1 = m_1, \xi_2 = m_2,.., \xi_n = m_n),$$

где ξ_1, ξ_2,.., ξ_n - независимые случайные величины, распределенные так же, как и случайная величина ξ.

Поскольку случайные величины ξ независимы и $P(\xi_i = m_i) = \dfrac{\lambda^{m_i}}{m_i!}e^{-\lambda}$, то

$$L(m_1, m_2,.., m_n) = P(\xi_1 = m_1)P(\xi_2 = m_2)...P(\xi_n = m_n) = \frac{\lambda^{m_1+...+m_n}}{m_1!..m_n!}e^{-n\lambda}.$$

Как видно из последнего равенства, функция правдоподобия зависит только от результатов наблюдений m_1, m_2,.., m_n и от неизвестного параметра λ. За оценку неизвестного параметра λ примем такое число $\hat{\lambda}$, которое доставляет максимум функции правдоподобия.

При решении задач отыскания максимума функции правдоподобия чаще всего находят максимум функции ln*L*:

$$\ln L = \ln\left(\frac{\lambda^{m_i+...+m_n}}{m_1!\ldots m_n!}e^{-n\lambda}\right) = -n\lambda + (m_{1+} + \ldots + m_n)\ln\lambda - \ln(m_1!\ldots m_n!),$$

которая достигает максимума в той же точке, что и функция правдоподобия $L(m_1, m_2,.. m_n)$.

Из необходимого условия экстремума

$$\frac{\partial \ln L}{\partial\lambda} = -n + \frac{m_1 + \ldots + m_1}{\lambda} = 0$$

имеем искомую оценку $\hat{\lambda} = \dfrac{1}{n}(m_1, \ldots, m_n)$ неизвестного параметра λ.

Вспомним, что математическое ожидание случайной величины ξ, имеющей распределение Пуассона с параметром λ, равно Mξ = λ и что эффективной, несмещенной, состоятельной оценкой математического ожидания ξ по выборке $m_1, m_2, .. m_n$ является величина $\overline{m} = \frac{1}{n(m_1 + ... + m_n)} = \hat{\lambda}$.

В Mathcad для моделирования выборки значений случайной величины, распределенной по закону Пуассона, предназначена функция rpois(k,λ)), которая формирует вектор из *k* случайных чисел, распределенных по Пуассону с параметром λ.

ЗАДАНИЕ 5

Смоделируйте несколько выборок объема *n* значений случайной величины ξ, имеющей распределение Пуассона с параметром λ=0.1N, N – номер варианта. Для одной выборки постройте график функции правдоподобия. Найдите оценку максимального правдоподобия параметра λ как функцию объема выборки. Выполните вычисления для n = 10N, 20N,..., 50N при N < =15 и для n = N, 2N,..., 10N при N > 15. Изобразите на графике зависимость оценки от объема выборки. Сравните полученные оценки с заданным значением параметра.

Порядок выполнения задания

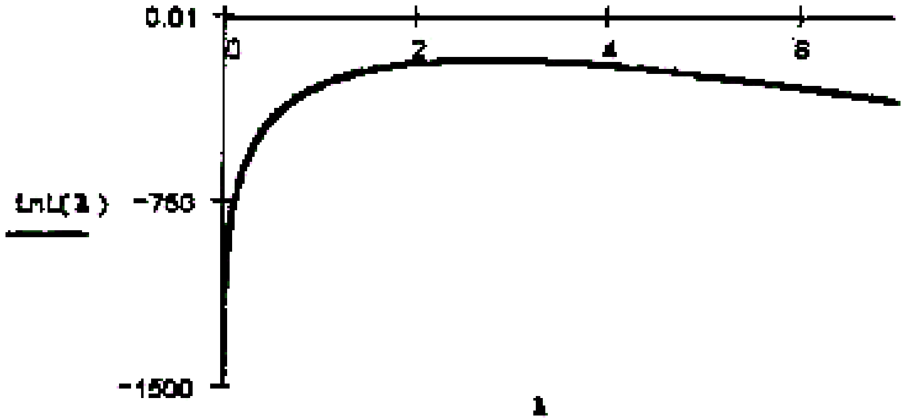
1. Смоделируйте выборку значений случайной величины, имеющей распределение Пуассона с заданным значением параметра λ.
2. Определите логарифм функции максимального правдоподобия и изобразите его график.
3. Смоделируйте несколько выборок разного объема значений случайной величины, имеющей распределение Пуассона с заданным значением параметра λ.
4. Вычислите оценку максимального правдоподобия параметра λ как функцию объема выборки.
5. Изобразите на графике зависимость оценки максимального правдоподобия от объема выборки.

Пример выполнения задания

В приведенном ниже фрагменте рабочего документа выполнены требуемые вычисления для распределения Пуассона с параметром λ = 3.

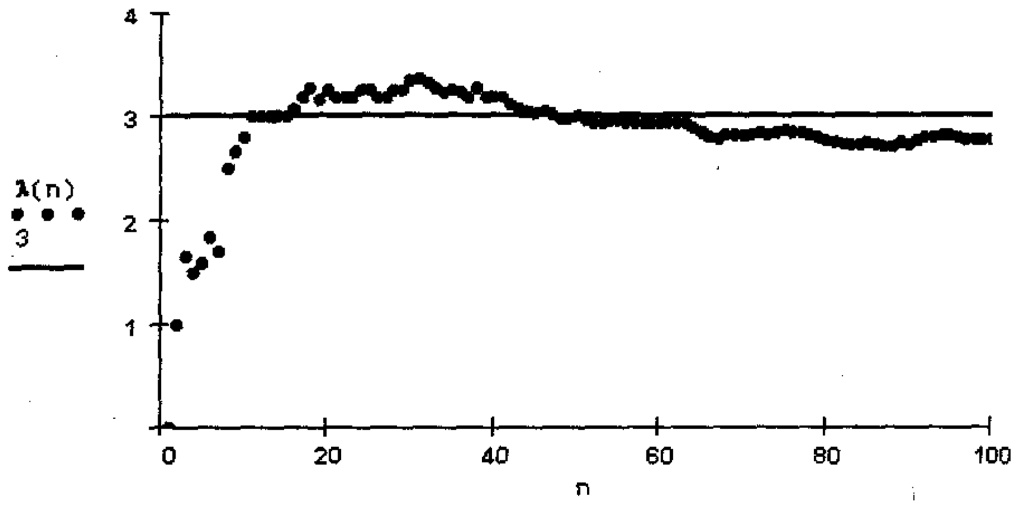
ORIGIN := 1　　　　　$P := rpois(100, 3)$

$$LnL(\lambda) := \left[-100\cdot\lambda + \left(\sum_{i=1}^{100} P_i \right)\cdot\ln(\lambda) \right] - \ln\left(\prod_{i=1}^{100} P_i ! \right)$$



$n := 1..100$　　　　　$\lambda(n) := \frac{1}{n}\cdot\sum_{i=1}^{n} P_i$

$\lambda(50) = 3$　　　　　$\lambda(100) = 2.77$



4.2 Метод максимального правдоподобия для непрерывной случайной величины

Пусть ξ – случайная величина, распределенная по показательному закону с неизвестным параметром λ:

$$P_\xi(x) = \begin{cases} 0, x < 0 \\ \lambda e^{-\lambda x}, x \ge 0 \end{cases}$$

Задача состоит в построении методом максимального правдоподобия оценки $\hat{\lambda}$ параметра λ по выборочным значениям $x_1, x_2, .., x_n$.

По аналогии с предыдущим разделом определим функцию правдоподобия равенством

$$L(x_1, x_2, .., x_n) = \prod_{i=1}^{n} p_\xi(x_i) = \lambda^n e^{-\lambda(x_1 + \ldots + x_n)}.$$

Как видно, функция правдоподобия зависит не только от выборочных значений, но и от неизвестного параметра распределения λ. Как и выше, за оценку неизвестного параметра λ примем такое число $\hat{\lambda}$, которое доставляет максимум функции правдоподобия. Снова переходим к логарифму функции правдоподобия, применяем необходимое условие экстремума и после несложных вычислений получаем:

$$\ln L = n \ln \lambda - (x_1 + \ldots + x_n)\lambda, \frac{\partial \ln L}{\partial \lambda} = \frac{n}{\lambda} - (x_1 + \ldots + x_n) = 0$$

$$\hat{\lambda} = \frac{n}{x_1 + \ldots + x_n} = \frac{1}{\overline{x}},$$

что естественно, поскольку математическое ожидание случайной величины, имеющей показательное распределение с параметром λ, равно 1/λ.

В Mathcad для моделирования выборки значений случайной величины, имеющей показательное распределение, предназначена функция rexp(k,λ), которая формирует вектор из *k* случайных чисел, распределенных показательно с параметром λ.

ЗАДАНИЕ 6

Смоделируйте несколько выборок объема n значений случайной величины ξ, имеющей показательное распределение с параметром λ == 0.1N, где N – номер варианта. Для одной выборки постройте график функции правдоподобия. Найдите оценку максимального правдоподобия параметра λ как функцию объема выборки. Выполните вычисления для n = 10N, 20N,.... 50N при N < =15 и для n = N, 2N,..., 10N при N > 15. Изобразите на графике зависимость оценки от объема выборки. Сравните полученные оценки с заданным значением параметра.

Порядок выполнения задания

1. Смоделируйте выборку значений случайной величины, имеющей экспоненциальное распределение с заданным значением параметра λ.
2. Определите логарифм функции максимального правдоподобия и изобразите его график.
3. Смоделируйте несколько выборок разного объема значений случайной величины, имеющей экспоненциальное распределение с заданным значением параметра λ.
4. Вычислите оценку максимального правдоподобия параметра λ как функцию объема выборки.
5. Изобразите на графике зависимость оценки максимального правдоподобия от объема выборки.

Пример выполнения задания

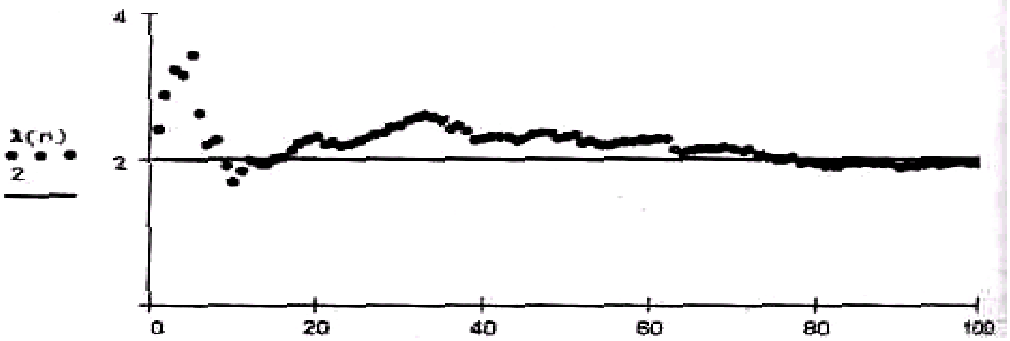
В приведенном ниже фрагменте рабочего документа выполнены требуемые вычисления для экспоненциального распределения с параметром λ=2.

$ORIGIN := 1 \qquad P := rexp(100, 2)$

$$LnL(\lambda) := 100 \cdot \ln(\lambda) - \lambda \cdot \sum_{i=1}^{100} P_i$$

$n := 1 .. 100 \qquad \lambda(n) := \dfrac{n}{\sum_{i=1}^{n} P_i} \qquad \lambda(50) = 2.33$

$\lambda(100) = 1.948$



5. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ О ПАРАМЕТРАХ НОРМАЛЬНО РАСПРЕДЕЛЕННОЙ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

Пусть дана некоторая оценка $\hat{\theta}$, построенная по выборке из случайной величины ξ. Есть основания считать, что истинное значение оцениваемого параметра равно θ_0. Однако выборочное значение $\hat{\theta}$ вряд ли будет совпадать с θ_0, поскольку $\hat{\theta}$ – случайная величина. В связи с этим возникает вопрос, при каком отклонении $\hat{\theta}$ от θ_0 и с какой степенью уверенности можно утверждать, что истинное значение оцениваемого параметра $\hat{\theta}$ отлично от θ_0. Ответом на этот вопрос может служить вероятность (вычисленная в предположении $\theta = \theta_0$) того, что $\hat{\theta} - \theta_0$ больше некоторого фиксированного числа, о величине которого будет сказано ниже. Если эта вероятность мала, то мы являемся свидетелями маловероятного события, т.е. отличие *эмпирического* значения $\hat{\theta}$ от *гипотетического* значения θ_0 представляется *значимым,* и гипотеза о том, что $\theta = \theta_0$, должна быть отвергнута. Если же эта вероятность велика, то отклонение $\hat{\theta}$ от θ_0, по-видимому, обусловлено естественной случайностью, и, следовательно, гипотеза о том, что $\hat{\theta} = \theta_0$, может быть принята.

Наша задача состоит в выработке общего подхода к процедуре, которая называется проверкой гипотез. Пусть $\hat{\theta}$ – выборочное значение оцениваемого параметра θ и пусть $p_{\hat{\theta}}$ – плотность вероятностей случайной величины $\hat{\theta}$ при условии, что $\theta = \theta_0$ На рис. 1 изображен график функции $p_{\hat{\theta}}$, на котором отмечены точки θ_{left} и θ_{right}, для которых выполнены условия: $P(\hat{\theta} \le \theta_{left}) = 0.5\alpha$, $P(\hat{\theta} > \theta_{right}) = 0.5\alpha$, где α – некоторое малое число. Это число имеет простой смысл: если вероятность события не превышает α, то событие маловероятно, т.е. если вычисленное значение $\hat{\theta}$ окажется вне промежутка $(\theta_{left}, \theta_{right})$, то есть все основания усомниться в том, что истинное значение параметра θ равно θ_0, и в этом случае гипотезу о том, что $\theta = \theta_0$, следует отвергнуть (отклонить). Если же $\hat{\theta}$ попадает в интервал $(\theta_{left}, \theta_{right})$, то гипотеза о том, что $\theta = \theta_0$, может быть принята. Вероятность α, использованная при вычислении интервала $(\theta_{left}, \theta_{right})$, называется *уровнем значимости;* области значений $\hat{\theta}$, при которых гипотеза отвергается или принимается, называются соответственно *областью отклонения (критической областью)* и *областью принятия гипотезы.*

В приведенном на рис. 1 примере критерий проверки гипотезы был *двусторонним,* поскольку значимыми были отклонения $\hat{\theta}$ от θ_0 в обе стороны. Если отклонения значимы только в одну сторону ($\hat{\theta} > \theta_0$ или $\hat{\theta} < \theta_0$), то строятся *односторонние критерии.*

Рис. 1. Области принятия и отклонения гипотез

Следует обратить внимание на то, что в рассматриваемых нами задачах принятие или отклонение гипотезы не носят категорического характера. Решение об отклонении или принятии гипотезы может оказаться ошибочным: гипотеза отклоняется, хотя она на самом деле верна (ошибка *первого рода),* и гипотеза принимается, хотя она на самом деле неверна (ошибка *второго рода)*.

5.1 Проверка гипотезы о числовом значении математического ожидания при известной дисперсии

Пусть $M\xi = a$ – неизвестная величина, а дисперсия $D\xi = \sigma^2$ известна. Сформулируем нулевую гипотезу H_0 о том, что неизвестный параметр a равен заданному числу a_0 , т.е. H_0: $a=a_0$. Альтернативную гипотезу H_1 можно сформулировать тремя способами:

1) H_1: $a \neq a_0$;

2) H_1: $a > a_0$;

3) H_1: $a < a_0$.

Рассмотрим подробно каждый из этих трех случаев.

Обратимся к первому случаю: нулевая гипотеза H_0: $a = a_0$ и альтернативная гипотеза H_1: $a \neq a_0$. Зададимся некоторым уровнем значимости α и вычислим по выборке $\hat{x} = (x_1, x_2, .., x_n)$ значение критерия $\varphi = \dfrac{\overline{x} - a_0}{\sqrt{\dfrac{\sigma^2}{n}}}$.

Если гипотеза H_0 верна, то случайная величина φ имеет стандартное нормальное распределение, и здравый смысл подсказывает, что в большинстве экспериментов величина φ будет мало отличаться от нуля. Если же ее отклонения от нуля велики, то это, скорее всего, указывает на ошибочность гипотезы H_0. Придадим приведенным соображениям более четкую форму.

Выделим для критерия φ – *критическую область,* т.е. укажем такие значения φ, при которых гипотезу H_0 следует отвергнуть. На рис. 2 изображена плотность распределения критерия φ (плотность стандартного нормального распределения).

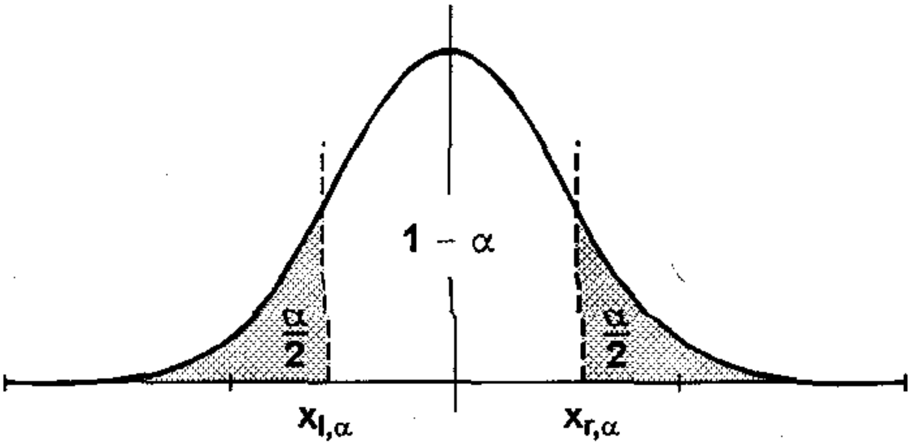


Рис. 2. Критическая область для альтернативной гипотезы $H_1: a \neq a_0$

Определим границы критической области $x_{l,\alpha}$ и $x_{r,\alpha}$ так, чтобы

$$P(\varphi < x_{l,\alpha}) = \frac{\alpha}{2}, P(\varphi > x_{r,\alpha}) = \frac{\alpha}{2}.$$

Критические точки $x_{l,\alpha}$ и $x_{r,\alpha}$ расположены симметрично относительно нуля, правая является корнем уравнения $\Phi(x_{r,\alpha}) = 1 - 0.5\alpha$, а левая вычисляется по формуле $x_{l,\alpha} = - x_{r,\alpha}$.

Когда критическая область найдена, можно вычислить по выборке значение критерия φ и проверить, попадает ли оно в критическую область.

Если $\varphi < x_{l,\alpha}$ или $\varphi > x_{r,\alpha}$, то гипотеза H_0 отвергается и принимается гипотеза H_1. Если же $x_{l,\alpha} < \varphi < x_{r,\alpha}$ то принимается гипотеза H_0.

Ниже приведен фрагмент рабочего документа Mathcad, содержащий проверку гипотезы о величине математического ожидания нормально распределенной случайной величины H_0: $a = 1$ при альтернативной гипотезе H_1: $a \neq 1$.

Рассмотрим второй случай с нулевой гипотезой H_0: $a = a_0$ и альтернативной гипотезой H_1: $a > a_0$. Снова зададимся некоторым уровнем значимости α и вычислим по выборке $\hat{x} = (x_1, x_2, .., x_n)$ значение критерия

$$\varphi = \frac{\bar{x} - a_0}{\sqrt{\frac{\sigma^2}{n}}}, \qquad \text{где} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

ORIGIN := 1

$\alpha := 0.1$ $\quad N := 100$ $\quad \xi := \mathrm{rnorm}(N,1,2)$ $\quad \sigma := 2$

$Xmean := \mathrm{mean}(\xi)$ $\quad Xmean = 0.699$

$Xright := \mathrm{qnorm}\left(1 - \frac{\alpha}{2}, 0, 1\right)$ $\quad Xleft := -Xright$

$Xright = 1.645$ $\quad Xleft = -1.645$

$$\phi 1 := \frac{Xmean - 1}{\sqrt{\frac{\sigma^2}{N}}} \qquad \phi 1 = -1.505$$

Гипотеза H_0: $M\xi=1$ принимается

Указание. Сначала с помощью функции rnorm (N,m,σ) сгенерирована выборка объема N = 100 из значений случайной величины, имеющей нормальное распределение N(1,2). Для уровня значимости α=0.1 вычислены границы критической области, Xright = 1.645, Xleft = -1.645, и оценка математического ожидания Xmean = 1.12. Высказана *нулевая гипотеза о том, что* значение параметра $M\xi$ =a равно a_0 = 1, т.е. H_0*:* a_0 = 1. Затем вычислено значение критерия φ = -1.505 и, поскольку -1.505 $\in$ (-1.645,1.645), нулевая гипотеза принимается на уровне значимости α =0.1.

В рассматриваемом случае критическая область значении критерия φ, при которых гипотеза H_0 отвергается, правосторонняя (рис. 3).

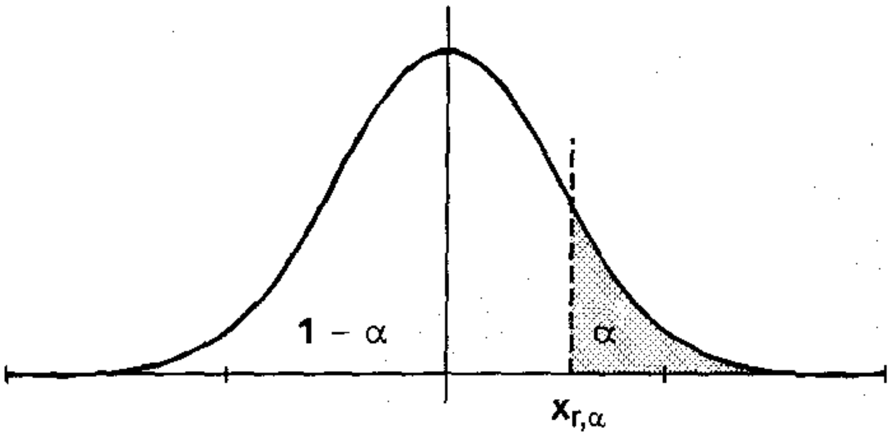


Рис. 3. Критическая область для альтернативной гипотезы H_1*:* $a>a_0$

Критическая точка удовлетворяет условию $P(\varphi > x_{r,\alpha}) = \alpha$ и находится как решение уравнения $\varphi(x_{r,\alpha}) = 1 - \alpha$.

Когда критическая область найдена, можно вычислить по выборке значение критерия φ и проверить, попадает ли оно в критическую область. Если $\varphi > x_{r,\alpha}$, то гипотеза H_0 отвергается и принимается гипотеза H_1. Если же $\varphi < x_{r,\alpha}$, то гипотеза H_0 не отвергается.

Ниже приведен фрагмент рабочего документа Mathcad, содержащий проверку гипотезы H_0: $a = a_1$ о величине математического ожидания a нормально распределенной случайной величины при альтернативной гипотезе H_1: $a > a_1$.

ORIGIN := 1

α := 0.05   N := 100   X := rnorm(N, 1, 0.2)   σ := 0.2

Xmean := mean(X)   Xmean = 0.972

Xright := qnorm(1 − α, 0, 1)   Xright = 1.645

$$\phi := \frac{\mathrm{Xmean} - 1}{\sqrt{\frac{\sigma^2}{N}}} \qquad \phi = -1.391$$

Гипотеза H_0: Mξ=1 принимается, поскольку -1.391<1.645

В третьем случае с нулевой гипотезой H_0: $a = a_1$ и альтернативной гипотезой H_1: $a < a_1$ опять зададимся некоторым уровнем значимости α и вычислим по выборке $\hat{x} = (x_1, x_2, .., x_n)$ значение критерия $\varphi = \dfrac{\bar{x} - a_0}{\sqrt{\dfrac{\sigma^2}{n}}}$,

где $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

В рассматриваемом случае критическая область значении критерия φ при которых гипотеза H_0 отвергается, левосторонняя (рис. 4).

Критическая точка удовлетворяет условию $P(\varphi < x_{l,\alpha}) = \alpha$ и находится по формуле $x_{l,\alpha} = -x_{r,\alpha}$, где $x_{r,\alpha}$ – решение уравнения $\Phi(x_{r,\alpha}) = 1 - \alpha$. Когда критическая область найдена, можно вычислить по выборке значение критерия φ и проверить, попадает ли оно в критическую область. Если $\varphi < x_{l,\alpha}$, то гипотеза H_0 отвергается и принимается гипотеза H_1. Если же $\varphi > x_{l,\alpha}$,то гипотеза H_0 не отвергается.

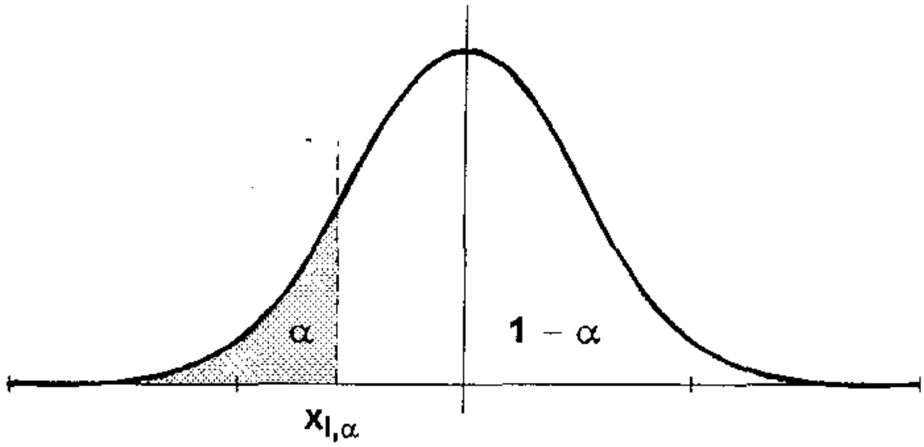


Рис. 4. Критическая область для альтернативной гипотезы $H_1{:}a<a_0$

Ниже приведен фрагмент рабочего документа Mathcad, содержащий проверку гипотезы о величине математического ожидания a нормально распределенной случайной величины H_0: $a = 6$ при альтернативной гипотезе H_1: $a < 6$.

$\mathrm{ORIGIN} := 1$

$\alpha := 0.05 \qquad N := 100 \qquad X := \mathrm{rnorm}(N, 5, 1) \qquad \sigma := 1$

$\mathrm{Xmean} := \mathrm{mean}(X) \qquad \mathrm{Xmean} = 4.85$

$\mathrm{Xright} := \mathrm{qnorm}(1 - \alpha, 0, 1) \qquad \mathrm{Xleft} := -\mathrm{Xright} \qquad \mathrm{Xleft} = -1.645$

$$\phi := \frac{\mathrm{Xmean} - 6}{\sqrt{\dfrac{\sigma^2}{N}}} \qquad \phi = -11.505$$

Гипотеза H_0: Mξ=6 отвергается, поскольку - 11.505< - 1.645

ЗАДАНИЕ 7 а

Смоделируйте выборку 100 значений нормально распределенной случайной величины с указанными параметрами. Сформулируйте нулевую гипотезу о величине математического ожидания и проверьте для заданных уровней значимости три альтернативные гипотезы.

Порядок выполнения задания

1. Смоделируйте описанную в условии выборку.
2. Найдите по выборке точечную оценку математического ожидания.
3. Сформулируйте нулевую гипотезу о значении математического ожидания H_0: $a = a_0$.
4. Вычислите значение критерия.
5. Найдите границы критической области для альтернативной гипотезы H_1: $a \neq a_0$.
6. Сравните значение критерия с границами критической области и сформулируйте соответствующее утверждение.
7. Найдите границы критической области для альтернативной гипотезы H_1: $a > a_0$.
8. Сравните значение критерия с границами критической области и сформулируйте соответствующее утверждение.
9. Найдите границы критической области для альтернативной гипотезы H_1: $a < a_0$.
10. Сравните значение критерия с границами критической области и сформулируйте соответствующее утверждение.

Пример выполнения задания

В каждом из приведенных выше трех фрагментов рабочих документов Mathcad проведены проверки для одной из альтернативных гипотез.

5.2 Проверка гипотезы о числовом значении математического ожидания при неизвестной дисперсии

Методика проверки гипотез в этом случае практически не отличается от описанной выше. Однако теперь критерий φ имеет распределение Стьюдента, а не стандартное нормальное, как в предыдущем случае.

Если $\hat{x} = (x_1, x_2, .., x_n)$ — выборка из нормального распределения и величины $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_1$ и $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$,то при проверке нулевой гипотезы H_0: $a = a_0$ используется критерий $\varphi = \frac{\bar{x} - a_0}{\sqrt{\frac{s^2}{n}}}$, который при выполнении гипотезы H_0 имеет распределение Стьюдента с числом степеней свободы n - 1.

Как и в описанной выше методике, рассмотрим три случая альтернативных гипотез при проверке гипотезы H_0: $a = a_0$:

1) H_1: $a \neq a_0$;
2) H_1: $a > a_0$;
3) H_1: $a < a_0$.

В первом из этих случаев, H_1: $a \neq a_0$, критическая область двусторонняя, и ее границы определяются из условий $P(\varphi < x_{l,\alpha}) = 0.5\alpha, P(\varphi > x_{r,\alpha}) = 0.5\alpha$. Причем в силу симметричности распределения Стьюдента достаточно вычислить только $x_{r,\alpha}$, поскольку $x_{l,\alpha} = - x_{r,\alpha}$.

Зададимся некоторым уровнем значимости α и вычислим значение $x_{r,\alpha}$ как решение уравнения $F_{n-1}(x_{r,\alpha}) = 1 - 0.5\alpha$, где $F_{n-1}(x)$ – функция распределения Стьюдента с n - 1 степенями свободы. Когда критическая область найдена, можно вычислить по выборке значение критерия φ и проверить, попадает ли оно в критическую область. Если $\varphi < x_{l,\alpha}$ или $\varphi > x_{r,\alpha}$, то гипотеза H_0 отвергается и принимается гипотеза H_1. Если же $x_{l,\alpha} < \varphi < x_{r,\alpha}$, то принимается гипотеза H_0.

Во втором случае, когда альтернативной является гипотеза H_1: $a > a_0$, критическая область значений критерия φ, при которых гипотеза H_0 отвергается, правосторонняя. Она представляет собой интервал $(x_{r,\alpha}, +\infty)$, где критическая точка удовлетворяет условию $P(\varphi > x_{r,\alpha}) = \alpha$ и находится как решение уравнения $F_{n-1}(x_{r,\alpha}) = 1 - \alpha$ при некоторым заданном уровне значимости α.

Теперь вычислим по выборке $\hat{x} = (x_1, x_2, .., x_n)$ значение критерия

$$\varphi = \frac{\bar{x} - a_0}{\sqrt{\frac{s^2}{n}}},$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, и проверим, попадает ли оно в критическую область. Если $\varphi > x_{r,\alpha}$, то гипотеза H_0 отвергается и принимается гипотеза H_1.

В третьем случае альтернативной гипотезы H_1: $a < a_0$ критическая область левосторонняя и представляет собой интервал $(-\infty, x_{l,\alpha})$. Критическая точка $x_{l,\alpha}$ удовлетворяет условию $P(\varphi < x_{l,\alpha}) = \alpha$ и находится по формуле $x_{l,\alpha} = - x_{r,\alpha}$, где $x_{r,\alpha}$ – решение уравнения $F_{n-1}(x_{r,\alpha}) = 1 - \alpha$.

Если выборочное значение критерия попадает в критическую область, т.е. $\varphi < <x_{l,\alpha}$, то гипотеза H_0 отвергается и принимается гипотеза H_1. Если же $\varphi > x_{l,\alpha}$, то гипотеза H_0 не отвергается.

ЗАДАНИЕ 7 б

Смоделируйте выборку 100 значений нормально распределенной случайной величины с указанными параметрами. Сформулируйте нулевую гипотезу о величине математического ожидания и проверьте для заданных уровней значимости три альтернативные гипотезы.

Порядок выполнения задания

1. Смоделируйте описанную в условии выборку.

2. Найдите по выборке точечную оценку математического ожидания.

3. Найдите по выборке точечную оценку дисперсии.

4. Сформулируйте нулевую гипотезу о значении математического ожидания $H_0: a=a_0$.

5. Вычислите значение критерия.

6. Найдите границы критической области для альтернативной гипотезы $H_1: a \neq a_0$.

7. Сравните значение критерия с границами критической области и сформулируйте соответствующее утверждение.

8. Найдите границы критической области для альтернативной гипотезы $H_1: a > a_0$.

9. Сравните значение критерия с границами критической области и сформулируйте соответствующее утверждение.

10. Найдите границы критической области для альтернативной гипотезы $H_1: a > a_0$.

11. Сравните значение критерия с границами критической области и сформулируйте соответствующее утверждение.

5.3 Проверка гипотезы о числовом значении дисперсии

Пусть $x = (x_1,x_2,..,x_n)$ – выборка объема n из нормального распределения. Требуется проверить нулевую гипотезу о том, что параметр нормального распределения σ^2 равен заранее заданному числу σ_0^2, т.е. $H_0: \sigma^2>\sigma_0^2$.

В процедуре проверки гипотезы используется критерий $\varphi = (n-1)\dfrac{s^2}{\sigma_0^2}$,

$s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ – несмещенная точечная оценка дисперсии, $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ – оценка математического ожидания. Поскольку рассматривается выборка из нормального распределения, то критерий φ имеет χ^2 - распределение с n - 1 степенями свободы.

Как и ранее, рассмотрим три случая альтернативных гипотез при проверке гипотезы $H_0: \sigma^2>\sigma_0^2$:

l) $H_1: \sigma^2 \neq \sigma_0^2$;

2) $H_1: \sigma^2 > \sigma_0^2$;

3) $H_0: \sigma^2 < \sigma_0^2$.

В первом из этих случаев, $H_0: \sigma^2 \neq \sigma_0^2$, критическая область двусторонняя (рис. 5) и ее границы определяются из условий $P(\varphi < x_{l,\alpha}) = 0.5\alpha$, $P(\varphi > x_{r,\alpha}) = 0.5\alpha$.

Зададимся некоторым уровнем значимости α и найдем значение $x_{l,\alpha}$ как решение уравнения $F_{n-1}(x_{l,\alpha}) = 1 - 0.5\alpha$., а $x_{r,\alpha}$ – как решение уравнения $F_{n-1}(x_{r,\alpha}) = 1 - 0.5\alpha$., где $F_{n-1}(x)$ – функция χ^2-распределения с n-1 степенями свободы.

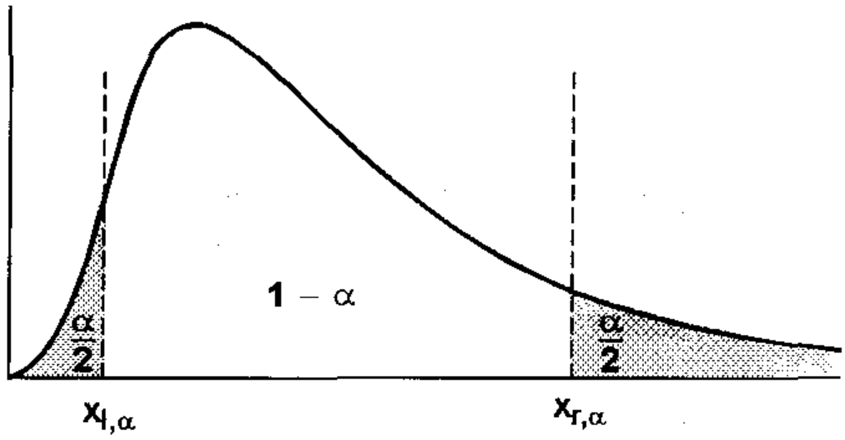


Рис. 5. Критическая область для альтернативной гипотезы H_1: $\sigma^2 < \sigma_0^2$

Когда критическая область найдена, можно вычислить по выборке значение критерия φ и проверить, попадает ли оно в критическую область. Если $\varphi < x_{l,\alpha}$ или $\varphi > x_{r,\alpha}$, то гипотеза H_0 отвергается и принимается гипотеза H_0. Если же $x_{l,\alpha} < \varphi < x_{r,\alpha}$, то принимается гипотеза H_0.

Во втором случае, когда альтернативная гипотеза H_1:$\sigma^2 > \sigma_0^2$ критическая область значений критерия φ, при которых гипотеза H_0 отвергается, правосторонняя и представляет собой интервал $(x_{r,\alpha},+\infty)$, где критическая точка удовлетворяет условию $P(\varphi > x_{r,\alpha}) = \alpha$ и находится как решение уравнения $F_{n-1}(x_{r,\alpha}) = 1 - \alpha$ при некотором заданном уровне значимости α.

Теперь вычислим по выборке $\hat{x} = (x_1, x_2, .., x_n)$ значение критерия φ и проверим, попадает ли оно в критическую область. Если $\varphi > x_{r,\alpha}$, то гипотеза H_0 отвергается и принимается гипотеза H_0.

В третьем случае альтернативной гипотезы H_1: $\sigma^2 < \sigma_0^2$ критическая область левосторонняя и представляет собой интервал $(0, x_{l,\alpha})$, где критическая точка $x_{l,\alpha}$ удовлетворяет условию $P(\varphi < x_{r,\alpha}) = \alpha$ и находится как решение уравнения $F_{n-1}(x_{l,\alpha}) = \alpha$. Если выборочное значение критерия попадает в критическую область, т.е. $\varphi < x_{l,\alpha}$, то гипотеза H_0 отвергается и принимается гипотеза H_1. Если же $\varphi > x_{l,\alpha}$, то гипотеза H_0 не отвергается.

ЗАДАНИЕ 7 в

Смоделируйте выборку 100 значений нормально распределенной случайной величины с указанными параметрами. Сформулируйте нулевую гипотезу о величине дисперсии и проверьте для заданных уровней значимости три альтернативные гипотезы.

Порядок выполнения задания

1. Смоделируйте описанную в условии выборку.
2. Найдите по выборке точечную оценку математического ожидания.
3. Найдите по выборке точечную оценку дисперсии.
4. Сформулируйте нулевую гипотезу о значении математического ожидания H_0: $\sigma^2 = \sigma_0^2$.
5. Вычислите значение критерия.

36. Найдите границы критической области для альтернативной гипотезы H_1: $\sigma^2 \neq \sigma_0^2$.
7. Сравните значение критерия с границами критической области и сформулируйте соответствующее утверждение.
8. Найдите границы критической области для альтернативной гипотезы H_1: $\sigma^2 > \sigma_0^2$.
9. Сравните значение критерия с границами критической области и сформулируйте соответствующее утверждение.
10. Найдите границы критической области для альтернативной гипотезы H_1: $\sigma^2 < \sigma_0^2$.
11. Сравните значение критерия с границами критической области и сформулируйте соответствующее утверждение.

Литература

1. Плис А.И. MathCad: математический практикум для экономистов и инженеров / А.И. Плис, Н.А. Сливина.— М.: Финансы и статистика, 2000. — 656с.
2. Боровиков В. STATISTICA: искусство анализа данных на компьютере / В. Боровиков. — СПб.: Питер, 2001.— 656с.
3. Тюрин Ю.Н. Статистический анализ данных на компьютере / Ю.Н. Тюрин, А.А. Макаров. — М.: ИНФРА-М, 1998.— 528с.

СОДЕРЖАНИЕ

Составитель: Нелля Михайловна Новикова

Редактор Золотарёва К.А.