

Н.М. Новикова С.Л. Подвальный

**ПРИКЛАДНАЯ МАТЕМАТИЧЕСКАЯ
СТАТИСТИКА
Часть 1**

Учебное пособие



Воронеж 2012

УДК 681.3

Новикова Н.М. Прикладная математическая статистика учеб. пособие / Н.М. Новикова, С.Л. Подвальный Воронеж: ФГБОУ ВПО «Воронежский государственный технический университет», 2012. Ч.1. 164 с.

В учебном пособии рассматриваются методы прикладной математической статистики, которые реализуются в виде алгоритмов программного обеспечения обработки экспериментальных данных, приводятся задачи с решениями.

Издание соответствует требованиям Федерального государственного образовательного стандарта высшего профессионального образования по направлению 230100 «Информатика и вычислительная техника» (магистерская программа подготовки «Распределенные автоматизированные системы»; профиль подготовки бакалавров «Вычислительные машины, комплексы, системы и сети») дисциплине «Обработка экспериментальных данных».

Табл. 3. Ил. 15. Библиогр.: 12 назв.

Рецензенты: кафедра цифровых технологий Воронежского государственного университета

(зав. кафедрой д-р физ.-мат. наук,
проф. С.Д. Кургалин);

д-р физ.-мат. наук, проф. В.В. Провоторов

© Новикова Н.М., Подвальный С.Л., 2012

© Оформление. ФГБОУ ВПО «Воронежский
государственный технический универ-
ситет», 2012

ВВЕДЕНИЕ

Прикладная математическая статистика – математическая дисциплина, родственная теории вероятностей. Под прикладной математической статистикой понимают «раздел математики, посвященный математическим методам сбора, систематизации, обработки и интерпретации статистических данных, а также использование их для научных или практических выводов. Фундаментом прикладной статистики является математическая статистика. Правила и процедуры математической статистики опираются на теорию вероятностей, позволяющую оценить точность и надежность выводов, получаемых в каждой задаче на основании имеющегося статистического материала». Такое определение дает Гнеденко Б.В. Статистическими данными называются сведения о числе объектов в какой-либо более или менее обширной совокупности, обладающих теми или иными признаками.

По типу решаемых задач математическая статистика обычно делится на три раздела: описание данных, оценивание и проверка гипотез.

По виду обрабатываемых статистических данных математическая статистика делится на четыре направления:

- одномерная статистика (статистика случайных величин), в которой результат наблюдения описывается действительным числом;
- многомерный статистический анализ, где результат наблюдения над объектом описывается несколькими числами (вектором);
- статистика случайных процессов и временных рядов, где результат наблюдения – функция;
- статистика объектов нечисловой природы, в которой результат наблюдения имеет нечисловую природу, например, является множеством (геометрической фигурой), упорядочением или получен в результате измерения по качественному признаку.

Сначала появились некоторые области математической статистики, исследующие объекты нечисловой природы (в частности, задачи оценивания доли брака и проверки гипотез о ней), и одномерная математическая статистика. Математический аппарат для них проще, поэтому на их примере обычно демонстрируют основные идеи математической статистики. Следует отметить, что лишь те методы обработки данных, т.е. математической статистики, являются доказательными, которые опираются на вероятностные модели соответствующих реальных явлений и процессов. Вероятностную модель реального явления следует считать построенной, если рассматриваемые величины и связи между ними выражены в терминах теории вероятностей. Соответствие вероятностной модели реальности, т.е. ее адекватность, обосновывают, в частности, с помощью статистических методов проверки гипотез.

Вероятностные и статистические методы применимы всюду, где удастся построить и обосновать вероятностную модель случайного явления или процесса. Их применение обязательно, когда сделанные на основе выборочных данных выводы переносятся на всю совокупность (например, с выборки на всю партию продукции).

Методы математической статистики находят широкое применение. В конкретных областях используются как вероятностно-статистические методы широкого применения, так и специфические. Например, при статистических методах управления качеством продукции используют прикладную математическую статистику (включая планирование экспериментов). С помощью ее методов проводится статистический анализ точности и стабильности технологических процессов и статистическая оценка качества. К специфическим методам относятся методы статистического приемочного контроля качества продукции, статистического регулирования технологических процессов, оценки и контроля надежности технических устройств. Широко применяются такие прикладные вероятностно-статистические дисциплины,

как теория надежности и теория массового обслуживания. Содержание первой из них ясно из названия, вторая занимается изучением систем типа телефонной станции, на которую в случайные моменты времени поступают вызовы - требования абонентов, набирающих номера на своих телефонных аппаратах. Длительность обслуживания этих требований, т.е. длительность разговоров, также моделируется случайными величинами.

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

В конце XIX и в начале XX века крупный вклад в математическую статистику внесли английские исследователи, прежде всего К. Пирсон (1857-1936) и Р.А. Фишер (1890-1962). В частности, К.Пирсон разработал критерий проверки статистических гипотез «хи-квадрат». Р.А. Фишер предложил и использовал дисперсионный анализ для обработки результатов агрономических опытов, а также разработал теорию планирования экспериментов и метод максимального правдоподобия оценки параметров.

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

развитие теории надежности и теории массового обслуживания внесли член-корреспондент АН СССР А.Я. Хинчин (1894-1959), академик АН УССР Б.В. Гнеденко (1912-1995) и другие отечественные ученые. Математическая статистика бурно развивается и в настоящее время. Так, за последние 40 лет можно выделить четыре принципиально новых направления исследований:

- разработка и внедрение математических методов планирования экспериментов;
- развитие статистики объектов нечисловой природы как самостоятельного направления в прикладной математической статистике;
- развитие статистических методов, устойчивых по отношению к малым отклонениям от используемой вероятностной модели;
- широкое развертывание работ по созданию компьютерных пакетов программ, предназначенных для проведения статистического анализа данных.

Вероятностно-статистическая модель и задачи математической статистики

Математическая статистика – это раздел математики, который занимается методами обработки статистических данных с целью построения или уточнения вероятностной модели случайного явления. В некотором смысле задачи математической статистики обратны задачам теории вероятностей. Математические модели случайных явлений, изучаемых в теории вероятностей, основываются на понятии вероятностного пространства (Ω, \mathcal{A}, P) , где $\Omega = \{\omega\}$ – непустое множество, называемое пространством элементарных событий (элементы ω интерпретируются как взаимно исключающие исходы изучаемого случайного явления);

\mathcal{A} – некоторая выделенная совокупность подмножеств множества Ω , называемых событиями (при этом требуется,

чтобы \mathcal{A} было σ -алгеброй, т.е. чтобы \mathcal{A} содержало Ω и было замкнуто относительно операции взятия противоположного события и объединения событий в не более чем счетном числе);

P – вероятность, заданная на событиях $A \in \mathcal{A}$. В каждой конкретной ситуации вероятность P считается заданной и основной задачей теории вероятностей является нахождение вероятностей сложных событий, исходя из известных вероятностей более простых событий для данной вероятностной модели.

В теории вероятностей при заданной вероятностной модели находятся те или иные статистические характеристики.

На практике вероятность P редко известна полностью. Априори можно утверждать, что P является элементом класса вероятностей \mathcal{P} , $P \in \mathcal{P}$. Этот класс \mathcal{P} может включать в себя все вероятности, которые можно задать на \mathcal{A} (ситуация полной неопределенности). В других же случаях класс \mathcal{P} представляет узкое семейство вероятностей, заданное в той или иной явной форме (ситуация, когда имеется определенная априорная информация). В любом случае \mathcal{P} – это совокупность допустимых для описания данного эксперимента вероятностей P .

Если задан класс \mathcal{P} , то говорят, что имеется **вероятностно-статистическая модель** (или просто статистическая модель), понимая под этим набор $(\Omega, \mathcal{A}, \mathcal{P})$.

Итак, статистическая модель описывает такие ситуации, когда в вероятностной модели эксперимента имеется неопределенность в задании вероятности P . Задача математической статистики состоит в том, чтобы уменьшить эту неопределенность, используя информацию, получаемую в результате наблюдений исходов эксперимента (статистические данные).

Различные виды статистических данных

Методы прикладной математической статистики – это методы анализа достаточно большого количества данных. Статистические данные могут иметь различную природу. Исторически самыми ранними были два вида данных: сведения о числе объектов, удовлетворяющих определенным условиям, и числовые результаты измерений.

Первый вид данных до сих пор главенствует в статистических сборниках. Такие данные называют *категоризованными* потому, что о каждом из рассматриваемых объектов известно, в какую из нескольких заранее заданных категорий он попадает. Информация о населении страны, с разделением по возрастным категориям и полу является примером категоризованных данных.

Второй наиболее распространенный вид данных – *количественные* данные, рассматриваемые как действительные числа. Эти данные являются результатами измерений, наблюдений, испытаний, опытов, анализов. Количественные данные обычно описываются набором чисел (выборкой).

Существует весьма много различных видов статистических данных. Это связано, в частности, со способами их получения. Например, если испытания некоторых технических устройств продолжаются до определенного момента, то получаем *цензурированные* данные, состоящие из набора чисел – продолжительности работы ряда устройств до отказа, и информации о том, что остальные устройства продолжали работать в момент окончания испытания. Такого рода данные часто используются при оценке и контроле надежности технических устройств.

Описание вида данных, а также механизма их порождения, если это необходимо – начало любого статистического исследования.

В простейшем случае статистические данные – это значения некоторого признака, свойственного изучаемым объектам. Значения могут быть количественными или

представлять собой указание на категорию, к которой можно отнести объект. Во втором случае говорят о качественном признаке.

При измерении по нескольким количественным или качественным признакам в качестве статистических данных об объекте получаем вектор. Его можно рассматривать как новый вид данных. В таком случае выборка состоит из набора векторов. Если часть координат – числа, а часть – качественные (категоризованные) данные, то говорим о векторе разнотипных данных.

Одним элементом выборки, т.е. одним измерением, может быть и функция в целом. Например, электрокардиограмма больного или временной ряд, описывающий динамику показателей определенной фирмы. Тогда выборка состоит из набора функций.

Элементами выборки могут быть и бинарные отношения. Различные виды бинарных отношений (упорядочения, разбиения, толерантности), множества и нечеткие множества используют для описания экспертных исследований.

Итак, математическая природа элементов выборки в различных задачах прикладной статистики может быть самой разной. Однако можно выделить два класса статистических данных – числовые и нечисловые. Соответственно прикладная статистика разбивается на две части – числовую статистику и нечисловую статистику.

Числовые статистические данные – это числа, векторы, функции. Их можно складывать, умножать на коэффициенты. Поэтому в числовой статистике большое значение имеют разнообразные суммы. Математический аппарат анализа сумм случайных элементов выборки – это классические законы больших чисел и центральные предельные теоремы.

Нечисловые статистические данные – это категоризованные данные, векторы разнотипных признаков, бинарные отношения, множества и нечеткие множества. Их нельзя складывать и умножать на коэффициенты, поэтому не имеет смысла говорить о суммах нечисловых статистических

данных. Они являются элементами нечисловых математических пространств (множеств). Математический аппарат анализа нечисловых статистических данных основан на использовании расстояний между элементами (а также мер близости, показателей различия) в таких пространствах. С помощью расстояний определяются эмпирические и теоретические средние, доказываются законы больших чисел, строятся непараметрические оценки плотности распределения вероятностей, решаются задачи диагностики и кластерного анализа.

Будем рассматривать методы прикладной математической статистики, предназначенные для обработки числовых данных.

1. ОСНОВНЫЕ ПОНЯТИЯ И ЭЛЕМЕНТЫ ВЫБОРОЧНОЙ ТЕОРИИ

Исходные статистические данные – результат наблюдения некоторой совокупности случайных величин $\vec{X}=(X_1, \dots, X_n)$, характеризующей исход изучаемого эксперимента. Эксперимент, обычно, состоит в проведении n испытаний, в которых результат i -го испытания описывается случайной величиной X_i , ($i=1, \dots, n$).

Совокупность наблюдаемых случайных величин $\vec{X}=(X_1, \dots, X_n)$ называется **выборкой**; величины X_i , ($i=1, \dots, n$) называются **элементами выборки**; их число n - **объемом выборки**.

Реализация выборки \vec{X} обозначается строчными буквами: $\vec{x}=(x_1, \dots, x_n)$. Пусть $\mathcal{X}=\{\vec{x}\}$ – множество, на котором задано распределение случайного вектора \vec{X} , т.е. множество всех возможных значений выборки \vec{X} . Множество \mathcal{X} называется **выборочным пространством**. Выборочное пространство может быть либо n -мерным евклидовым

пространством \mathbb{R}^n или его частью, (если \bar{X} - непрерывна), либо состоять из конечного или счетного числа точек в \mathbb{R}^n (если случайная величина \bar{X} - дискретна).

Под статистической моделью эксперимента в данном случае понимается набор $(\mathcal{X}, \mathcal{P})$, где \mathcal{P} - класс допустимых распределений случайных величин \bar{X} , заданных на \mathcal{X} . Распределение вероятностей любой случайной величины однозначно определяется ее функцией распределения, поэтому статистическая модель задается обычно в терминах допустимых функций распределения выборки \bar{X} .

Итак, статистическая модель определяется выборочным пространством \mathcal{X} и семейством функций распределения F , которому принадлежит неизвестная функция распределения $F_{\bar{X}}(x_1, \dots, x_n) = P(X_1 \leq x_1, \dots, X_n \leq x_n)$, $-\infty < x_1, \dots, x_n < +\infty$ выборки $\bar{X} = (X_1, \dots, X_n)$.

Часто бывает ситуация, когда компоненты X_1, \dots, X_n независимы и все распределены так же, как и некоторая случайная величина ξ . Это соответствует эксперименту, в котором проводятся повторные независимые наблюдения над случайной величиной ξ . Здесь $F_{X_i}(x_i) = F_{\xi}(x_i)$ для всех $i=1, \dots, n$ и $F_{\bar{X}}(\bar{x}) = F_{\xi}(x_1) \dots F_{\xi}(x_n)$.

Такую модель можно задать в терминах функции распределения F_{ξ} и тогда $\bar{X} = (X_1, \dots, X_n)$ - выборка из распределения случайной величины ξ . Множество возможных значений ξ с распределением F_{ξ} называют **генеральной совокупностью** (или просто **совокупностью**), а \bar{X} - выборкой из этой совокупности. Обозначение таково: $\bar{X} = (X_1, \dots, X_n)$ есть выборка из $L(\xi)$, где $L(\xi)$ - распределение ξ .

Если функции распределения из класса F заданы с точностью до значений некоторого параметра θ с множеством возможных значений Θ , то такая модель обозначается $F = \{F(x, \theta), \theta \in \Theta\}$, и называется **параметрической**.

Известен тип распределения наблюдаемой случайной величины в этом случае, но не известен параметр, от которого зависит распределение. Параметр θ может быть как скалярным, так и векторным; множество Θ называется **параметрическим**.

Пусть известно, что $L(\xi)$ - нормальное распределение с известной дисперсией и неизвестным средним. Тогда статистическая модель имеет вид $F=\{F(x,\theta), \theta \in \Theta, \Theta=(-\infty, \infty)\}$, где функция распределения $F(x,\theta)$ имеет плотность

$$f(x,\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\theta)^2}{2\sigma^2}\right], -\infty < x < \infty.$$

Если и дисперсия неизвестна, то статистическая модель имеет вид $F=\{F(x,\bar{\theta}), \bar{\theta}=(\theta_1,\theta_2) \in \Theta\}$, где $\Theta=\{(\theta_1,\theta_2): -\infty < \theta_1 < \infty, 0 < \theta_2 < \infty\}$ и $F(x,\bar{\theta})$ имеет плотность

$$f(x,\bar{\theta}) = \frac{1}{\sqrt{2\pi}\theta_2} \exp\left[-\frac{(x-\theta_1)^2}{2\theta_2^2}\right], -\infty < x < \infty.$$

Модель $F=\{F_\xi\}$ называется абсолютно непрерывной или дискретной, если таковыми являются все составляющие класс F функции распределения. Рассматриваются только эти модели.

Будем использовать единое обозначение $f_\xi(x)=f(x)$ (для параметрических моделей $f(x,\theta)$) как для плотности распределения случайной величины ξ в случае непрерывной модели, так и для вероятности $P(\xi=x)$ в случае дискретной модели.

1.1. Порядковые статистики и вариационный ряд выборки

Пусть $\bar{X}=(X_1,\dots,X_n)$ - выборка объема n из распределения $L(\xi)$ и $\bar{x}=(x_1,\dots,x_n)$ - наблюдавшееся значение \bar{X} . Каждой реализации \bar{x} выборки \bar{X} можно поставить в соответствие упорядоченную последовательность

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)} \tag{1.1}$$

где $x_{(1)} = \min(x_1, \dots, x_n)$, $x_{(2)}$ - второе по величине значение среди x_1, \dots, x_n и т.д., $x_{(n)} = \max(x_1, \dots, x_n)$.

Обозначим через $X_{(k)}$ случайную величину, которая для каждой реализации \bar{x} выборки \bar{X} принимает значение $x_{(k)}$, $k=1, \dots, n$. Так по выборке \bar{X} определяют новую последовательность случайных величин $X_{(1)}, \dots, X_{(n)}$, называемых **порядковыми статистиками** выборки; при этом $X_{(k)}$ - k -тая **порядковая статистика**, а $X_{(1)}$ и $X_{(n)}$ - **экстремальные значения выборки**.

Из определения порядковых статистик следует, что они удовлетворяют неравенствам

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)} \quad (1.2)$$

Последовательность (1.2) называют **вариационным рядом** выборки. Симметричные относительно концов элементы последовательности (2) $X_{(m)}$ и $X_{(n-m+1)}$ иногда называют соответственно m -м **наименьшим** и m -м **наибольшим** значениями выборки ($m=1, 2, \dots$); при $m=1$ получаем экстремальные значения выборки. Итак, **вариационный ряд** - это расположенные в порядке возрастания их величин элементы выборки. Отметим, что реализацией последовательности (1.2) является последовательность (1.1).

1.2. Эмпирическая функция распределения

Распределение выборки (эмпирическое распределение) - это распределение вероятностей, которое определяется по выборке для оценивания истинного распределения.

Определим для каждого действительного x случайную величину $\mu_n(x)$, равную числу элементов выборки $\bar{X} = (X_1, \dots, X_n)$, значения которых не превосходят x , т.е.

$$\mu_n(x) = \sum_{i=1}^n I(X_i \leq x), \quad (1.3)$$

где $I(A)$ - индикатор события A $\{I(A)=1, \text{ если } A \text{ имеет место, и } 0 - \text{ в противном случае}\}$. Положим $F_n(x) = \frac{\mu_n(x)}{n}$.

Функция $F_n(x)$ называется **эмпирической функцией распределения** (э.ф.р.), соответствующей выборке \bar{X} . Функцию распределения $F(x)$ наблюдаемой случайной величины ξ называют **теоретической функцией распределения**.

По своему определению эмпирическая функция распределения – случайная функция: для каждого $x \in \mathbb{R}^1$ значение $F_n(x)$ есть случайная величина, реализациями которой являются числа $0, 1/n, 2/n, \dots, (n-1)/n, n/n=1$, при этом

$$P(F_n(x)=k/n) = P(\mu_n(x)=k).$$

Из определения $\mu_n(x)$ следует, что $L(\mu_n(x)) = B_i(n, p)$, где $p = P(\xi \leq x) = F(x)$. Поэтому

$$P(F_n(x)=k/n) = C_n^k F^k(x) (1-F(x))^{n-k}, \quad k=0, 1, \dots, n. \quad (1.4)$$

Итак, эмпирическая функция распределения (как и вариационный ряд) - некоторая сводная характеристика выборки. Для каждой реализации \bar{x} выборки \bar{X} функция $F_n(x)$ однозначно определена и обладает всеми свойствами функции распределения: изменяется от 0 до 1, не убывает и непрерывна справа. Она кусочно-постоянна и возрастает только в точках последовательности (1.1). Если все компоненты вектора \bar{x} различны (в последовательности (1.1) все неравенства строгие), то $F_n(x)$ задается соотношениями

$$F_n(x) = \begin{cases} 0, & x < x_{(1)}; \\ k/n, & x_{(k)} < x < x_{(k+1)}; \\ 1, & x \geq x_{(n)}. \end{cases} \quad k=1, \dots, n-1$$

В этом случае величина скачка равна $1/n$ и типичный график функции $F_n(x)$ представлен на рис.1.1.

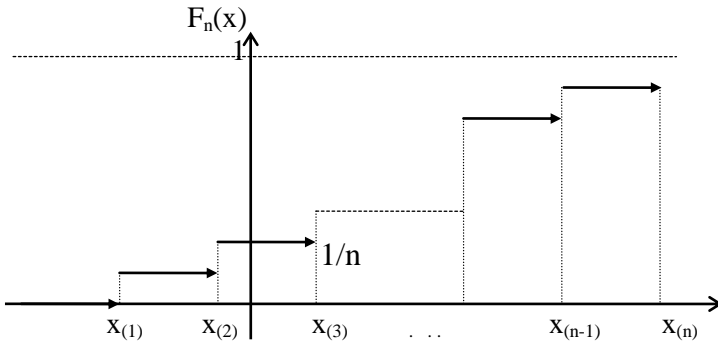


Рис. 1.1

В общем виде эмпирическую функцию распределения можно записать в виде

$$F_n(x) = \frac{1}{n} \sum_{k=1}^n I(X_{(k)} \leq x) = \frac{1}{n} \sum_{i=1}^n I(X_i \leq x). \quad (1.5)$$

В представлении (1.5) видна зависимость $F_n(x)$ от выборки \vec{X} .

Эмпирическая функция распределения играет фундаментальную роль в обработке данных. Важное свойство эмпирической функции распределения состоит в том, что при увеличении объема выборки n происходит сближение $F_n(x)$ с $F(x)$.

Теорема 1.1: Пусть $F_n(x)$ - эмпирическая функция распределения, построенная по выборке $\vec{X} = (X_1, \dots, X_n)$ из распределения $L(\xi)$, и $F(x)$ - функция распределения ξ . Тогда для любого x ($-\infty < x < +\infty$) и любого $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} P(|F_n(x) - F(x)| < \varepsilon) = 1 \quad (1.6)$$

Доказательство: Из (1.4) следует, что $F_n(x)$ - относительная частота события $\{\xi \leq x\}$ - («успеха») в n испытаниях Бернулли с вероятностью «успеха» $F(x)$. Но по теореме Бернулли [относительная частота произвольного события в n независимых испытаниях сходится по вероятности

при $n \rightarrow \infty$ к вероятности этого события], $F_n(x) \xrightarrow{P} F(x)$, т.е. имеет место равенство (1.6) \square

Замечание. Если объем выборки большой, то значение эмпирической функции распределения в каждой точке x может служить приближенным значением (**оценкой**) теоретической функции распределения в этой точке. Функцию $F_n(x)$ называют еще **статистическим аналогом** для $F(x)$.

Более глубокие свойства эмпирической функции распределения проявляются, если рассматривать ее поведение не в отдельной фиксированной точке x , а в произвольной конечной совокупности точек $x_1 < x_2 < \dots < x_n$. В частности, важно знать отклонения эмпирической функции распределения $F_n(x)$ от $F(x)$ на всей оси. Известен результат, принадлежащий Гливленко В.И.

Теорема 1.2 (Гливленко): В условиях теоремы 1.1

$$P\left(\lim_{n \rightarrow \infty} \sup_{-\infty < x < \infty} |F_n(x) - F(x)| = 0\right) = 1 \quad (1.7)$$

Другими словами, соотношение (1.7) означает, что отклонение $D_n = D_n(\bar{X}) = \sup_{-\infty < x < \infty} |F_n(x) - F(x)|$ эмпирической функции распределения от $F(x)$ на всей оси с вероятностью 1 будет сколь угодно мало при $n \rightarrow \infty$ (при достаточно большом объеме выборки).

Рассмотрим теорему, которая позволяет для больших n оценивать вероятности заданных отклонений случайной величины. D_n от 0.

Теорема 1.3 (Колмогорова): Если функция $F(x)$ непрерывна, то при любом фиксированном $t > 0$

$$\lim_{n \rightarrow \infty} P(\sqrt{n} D_n \leq t) = K(t) = \sum_{i=-\infty}^{\infty} (-1)^i e^{-2i^2 t^2} \quad (1.8)$$

Предельную функцию распределения $K(x)$ можно с хорошим приближением использовать для практических расчетов уже при $n \geq 20$. Теорему Колмогорова применяют для того, чтобы

определить границы, в которых с заданной вероятностью находится теоретическая функция распределения $F(x)$, если она неизвестна. Пусть для заданного $\gamma \in (0,1)$ число t_γ определяется уравнением $K(t_\gamma) = \gamma$.

Тогда из (1.8) имеем:

$$P(\sqrt{n}D_n \leq t_\gamma) = P\left\{ \left(F_n(x) - \frac{t_\gamma}{\sqrt{n}} \right) \leq F(x) \leq \left(F_n(x) + \frac{t_\gamma}{\sqrt{n}} \right), \right\}$$

$$\xrightarrow{n \rightarrow \infty} K(t_\gamma) = \gamma$$

Таким образом, при $n \rightarrow \infty$ с вероятностью, близкой к γ , значения функции $F(x)$ для всех x удовлетворяют неравенствам

$$F_n(x) - \frac{t_\gamma}{\sqrt{n}} \leq F(x) \leq F_n(x) + \frac{t_\gamma}{\sqrt{n}}$$

Так как $0 \leq F(x) \leq 1$, эти неравенства можно уточнить:

$$\max\left(0, F_n(x) - \frac{t_\gamma}{\sqrt{n}}\right) \leq F(x) \leq \min\left(F_n(x) + \frac{t_\gamma}{\sqrt{n}}, 1\right).$$

Область, определяемая этими нижней и верхней границами, называется **асимптотической γ -доверительной зоной** для теоретической функции распределения. Для определения числовых значений t_γ при различных γ можно воспользоваться табулированными значениями функции $K(t)$.

Теорема 1.4 (Смирнова): Пусть $F_{1n}(x)$ и $F_{2m}(x)$ - две эмпирические функции распределения, построенные на основе двух независимых выборок объемом n и m из одного и того же распределения $L(\xi)$, и $D_{n,m} = \sup_{-\infty < x < \infty} |F_{1n}(x) - F_{2m}(x)|$.

Тогда, если теоретическая функция распределения $F(x)$ непрерывна, то для любого фиксированного $t > 0$

$$\lim_{n,m \rightarrow \infty} P\left(\sqrt{\frac{nm}{n+m}} D_{n,m} \leq t\right) = K(t),$$

где функция $K(t)$ определена равенством (1.8).

Эту теорему используют для проверки гипотезы (предположения) о том, что две выборки получены из одного и того же распределения.

1.3. Гистограмма и полигон частот

Итак, эмпирическая функция распределения – удобный способ представления статистических данных (выборки \bar{X}). Он позволяет делать выводы о распределении наблюдаемой случайной величины ξ , когда оно неизвестно. По эмпирической функции распределения $F_n(x)$ при $n \rightarrow \infty$ со сколь угодно высокой точностью можно восстановить неизвестную теоретическую функцию распределения $F(x)$.

Рассмотрим другие способы представления статистических данных. Пусть наблюдаемая случайная величина ξ дискретна и принимает значения x_1, x_2, \dots . Представление о законе распределения ξ дадут частоты v_r/n , где v_r - число элементов выборки $\bar{X} = (X_1, \dots, X_n)$, принявших значение x_r :

$$v_r = \sum_{i=1}^n I(X_i = x_r).$$

В этом случае, по теореме Бернулли, при $n \rightarrow \infty$

$$v_r/n \xrightarrow{P} P(\xi = x_r), \quad r = 1, 2, \dots$$

Пусть ξ - непрерывная случайная величина и имеет непрерывную плотность распределения $f(x)$. Рассмотренную методику применим для оценивания неизвестной плотности. Это осуществляется с помощью **метода группировки наблюдений** (или метода группировки данных), который состоит в следующем.

Пусть $\{\varepsilon_r\}$ - некоторое разбиение области ε возможных значений ξ : $\varepsilon = \bigcup_r \varepsilon_r$, $\varepsilon_i \cap \varepsilon_j = \emptyset$, $i \neq j$ и $v_r = \sum_{j=1}^n I(X_j \in \varepsilon_r)$ - число

выборочных точек [элементов выборки $\vec{X}=(X_1, X_2, \dots, X_n)$], попавших в интервал ε_r . Тогда при $n \rightarrow \infty$, по теореме Бернулли,

$$\frac{v_r}{n} \xrightarrow{P} P(\xi \in \varepsilon_r) = \int_{\varepsilon_r} f(x) dx.$$

Из свойств определенного интеграла по теореме о среднем значении, последний интеграл равен $|\varepsilon_r| f(x_r)$, где x_r некоторая внутренняя точка интервала ε_r , а $|\varepsilon_r|$ - его длина. Обычно интервалы выбираются одинаковой длины, и если длина интервала мала, то в качестве x_r берут середину интервала. Поэтому можно считать $\frac{v_r}{n} \approx |\varepsilon_r| f(x_r)$ или

$$\frac{v_r}{n|\varepsilon_r|} \approx f(x). \quad (1.9)$$

Построим теперь кусочно-постоянную функцию $f(x) = \frac{v_r}{n|\varepsilon_r|}$, при $x \in \varepsilon_r$, $r=1, 2, \dots$, называемую **гистограммой**.

При $n \rightarrow \infty$ и достаточно мелком разбиении $\{\varepsilon_r\}$ гистограмма $f_n(x)$ будет оценкой $f(x)$ - теоретической плотности. Если плотность достаточно гладкая функция, то ее лучше приблизить кусочно-линейными графиками. Оценка гладких $f(x)$ основана на построении **полигона частот**. Полигон частот - это ломанная, которую строят так: если построена гистограмма, то ординаты, соответствующие средним точкам интервалов, последовательно соединяют отрезками прямых. Такой кусочно-линейный график является статистическим аналогом (оценкой) теоретической плотности (рис. 1.2.).

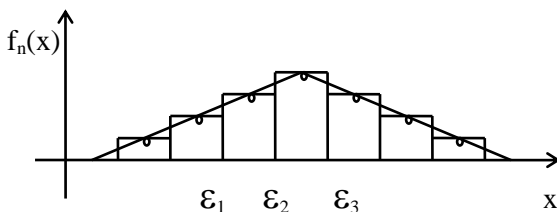


Рис. 1.2

Высота определяется формулой (1.9), а основание - длина интервала разбиения.

1.4. Определения и свойства выборочных характеристик

Пусть $\bar{X}=(X_1, X_2, \dots, X_n)$ - выборка из распределения $L(\xi)$. $F(x)$ и $F_n(x)$ - соответственно теоретическая и эмпирическая функции распределения. Точно так же, как функции $F(x)$ ставят в соответствие $F_n(x)$, любой теоретической характеристике $g = \int g(x)dF(x)$ можно поставить в соответствие ее статистический аналог $G=G(\bar{X})$, определяемый по формуле

$$G = \int g(x)dF(x) = \frac{1}{n} \sum_{i=1}^n g(X_i).$$

Случайную величину G называют **эмпирической** или **выборочной характеристикой**, соответствующей теоретической характеристике g . Таким образом, выборочная характеристика - это среднее арифметическое значение функции $g(x)$ для элементов выборки \bar{X} . Если $g(x)=x^k$, то G - **выборочный момент k -го порядка**, обозначается A_k

$$A_k^* = A_k = A_k(\bar{X}) = \frac{1}{n} \sum_{i=1}^n X_i^k, \quad (1.10)$$

(значение начального момента k -го порядка $\alpha_k^* = \frac{1}{n} \sum_{i=1}^n x_i^k$).

При $k=1$ величину A_k называют **выборочным средним** и обозначают

$$\bar{X} = A_1 = \frac{1}{n} \sum_{i=1}^n X_i.$$

Значения случайных величин A_k и \bar{X} для данной реализации \bar{x} выборки \bar{X} обозначают строчными буквами a_k и $\bar{x} = a_1$.

Выборочным центральным моментом k -го порядка называют случайную величину

$$M_k^* = M_k = M_k(\bar{X}) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^k ,$$

(значение выборочного момента $\mu_k^* = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$).

При $k=2$ величину M_k называют **выборочной дисперсией** и обозначают $S^2 = S^2(\bar{X})$:

$$S^2 = M_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 .$$

Замечания.

1. Выборочные моменты являются случайными величинами, поскольку являются функциями выборки.
2. Выборочные моменты имеют свои функции распределения и числовые характеристики.

Рассмотрим некоторые характеристики распределения среднего \bar{X} и S^2 выборки. Так как X_i - независимы и распределены так же, как и наблюдаемая случайная величина ξ , то

$$M[\bar{X}] = \frac{1}{n} \sum_{i=1}^n M X_i = M \xi = \alpha_1 ;$$

$$D[\bar{X}] = \frac{1}{n^2} \sum_{i=1}^n D X_i = \frac{1}{n} D \xi = \frac{\mu_2}{n} .$$

1.5. Асимптотическое поведение выборочных моментов. Теорема Слущкого

Рассмотрим поведение выборочных моментов A_k , определяемых равенством (1.10) при $n \rightarrow \infty$ [неограниченном возрастании n]. Чтобы подчеркнуть зависимость моментов A_k от n (объема выборки), будем использовать обозначение A_{nk} . Первые два момента случайной величины. A_{nk} определяются следующими равенствами: (предполагаем, что

соответствующие моменты наблюдаемой случайной величины ξ существуют)

$$\left\{ \begin{aligned} M[A_{nk}] &= \frac{1}{n} \sum_{i=1}^n MX_i^k = M\xi^k = \alpha_k \\ D[A_{nk}] &= \frac{1}{n^2} \sum_{i=1}^n DX_i^k = \frac{1}{n} D\xi^k = \frac{1}{n} \left(M\xi^{2k} - (M\xi^k)^2 \right) = \frac{\alpha_{2k} - \alpha_k^2}{n} \end{aligned} \right. .(1.11)$$

На основании неравенства Чебышева отсюда следует, что $A_{nk} \xrightarrow{P} \alpha_k$ при $n \rightarrow \infty$.

Таким образом, выборочный момент A_{nk} можно рассматривать в качестве приближенного значения (оценки) соответствующего теоретического момента α_k , когда число наблюдений n велико. Аналогичное утверждение справедливо и для выборочных центральных моментов и вообще для любых выборочных характеристик, которые имеют вид непрерывных функций от конечного числа величин A_{nk} .

Этот вывод является следствием общей теоремы о сходимости функций от случайных величин.

Теорема 1.5 (Слущкого). Пусть случайные величины $\eta_1(n), \dots, \eta_r(n)$ сходятся по вероятности при $n \rightarrow \infty$ к некоторым постоянным c_1, \dots, c_r соответственно. Тогда для любой непрерывной функции $y(x_1, \dots, x_r)$ случайная величина

$$\xi(n) = y(\eta_1(n), \dots, \eta_r(n)) \xrightarrow{P} y(c_1, \dots, c_r).$$

Доказательство: Функция y непрерывна, поэтому для любого $\varepsilon > 0$ найдется $\delta = \delta(\varepsilon)$ такое, что

$$\left| y(x_1, \dots, x_r) - y(c_1, \dots, c_r) \right| < \varepsilon \quad \text{при} \quad \left| x_i - c_i \right| < \delta, \quad i = \overline{1, r}.$$

Введем события $B_i = \left\{ \left| \eta_i(n) - c_i \right| < \delta \right\}$, $i = \overline{1, r}$. Тогда событие

$B = B_1, \dots, B_r$ ($B = \bigcap_{i=1}^r B_i$) влечет событие $C(B < C)$, где

событие $C = \left\{ \xi(n) - y(c_1, \dots, c_r) < \varepsilon \right\}$ можно представить как

$$P(C) \geq P(B) = 1 - P(\bar{B}) = 1 - P(\bar{B}_1 U \dots U \bar{B}_r) \geq 1 - \sum_{i=1}^r P(\bar{B}_i) \quad (1.12)$$

Далее, из сходимости по вероятности случайной величины $\eta_i(n)$ имеем, что для данного δ и любого $\gamma > 0$ найдется

$n_i = n_i(\gamma)$ такое, что $P(\bar{B}_i) = P(|\eta_i(n) - c_i| > \delta) < \gamma/r$ при $n \geq n_i$.

Пусть $n_0 = \max(n_1, \dots, n_r)$, тогда при $n \geq n_0$ выполняются

все неравенства $\sum_{i=1}^r P(\bar{B}_i) < \gamma$. Следовательно, из формулы

$$(1.12) \text{ получим } P\left(\left|\xi(n) - y(c_1, \dots, c_r)\right| < \varepsilon\right) \geq 1 - \gamma, n \geq n_0,$$

отсюда имеем $\xi(n) \xrightarrow{P} y(c)$ при $n \rightarrow \infty$, что и требовалось доказать. ■

1.6. Асимптотическая нормальность выборочных моментов.

Введем дополнительные обозначения. Если распределение случайной величины η_n сходится при $n \rightarrow \infty$ к распределению случайной величины η и при этом $L(\eta) = N(m, \sigma^2)$, то будем писать $L(\eta_n) \rightarrow N(m, \sigma_n^2)$. Будем считать, что случайная величина η_n асимптотически нормальна

с параметрами m_n, σ_n^2 , $N(m_n, \sigma_n^2)$ и записывать это так $L(\eta_n) \sim N(m_n, \sigma_n^2)$. Это означает, что $L\left(\frac{\eta_n - m_n}{\sigma_n}\right) \rightarrow N(0, 1)$.

Исследуем распределения выборочных характеристик для больших выборок ($n \rightarrow \infty$). Каждый выборочный момент A_{nk} представляет собой сумму n независимых и одинаково распределенных случайных величин, поэтому к нему можно применить центральную предельную теорему. Имеет место следующая теорема.

Теорема 1.6: Выборочный момент A_{nk} асимптотически нормален $N(\alpha_k, (\alpha_{2k} - \alpha_k^2)/n)$

Доказательство: Так как (см. формулы (1.11)) $MX_i^k = \alpha_k$; $DX_i^k = \alpha_{2k} - \alpha_k^2$, то по центральной предельной теореме $L(\eta_n) \rightarrow N(0, 1)$, где

$$\eta_n = \frac{1}{\sqrt{n(\alpha_{2k} - \alpha_k^2)}} \left(\sum_{i=1}^n X_i^k - n\alpha_k \right) = \sqrt{\frac{n}{\alpha_{2k} - \alpha_k^2}} \cdot (A_{nk} - \alpha_k).$$

Следовательно, случайная величина A_{nk} асимптотически нормальна с параметрами α_k и $(\alpha_{2k} - \alpha_k^2)/n$.

Эта теорема позволяет оценивать для больших выборок вероятность заданных отклонений значений выборочных моментов от теоретических. Действительно, из этой теоремы имеем, что при любом фиксированном $t > 0$ и $n \rightarrow \infty$

$$P\left(\sqrt{\frac{n}{\alpha_{2k} - \alpha_k^2}} \cdot |A_{nk} - \alpha_k| < t\right) \rightarrow \frac{1}{\sqrt{2\pi}} \int_{-t}^t e^{-x^2/2} dx = 2\Phi(t) - 1.$$

В частности, из теоремы 1.6 следует, что выборочное среднее $\bar{X} = A_{n1}$ асимптотически нормально $N(\alpha_1, \mu_2/n)$.

Отметим, что если $L(\xi) = N(\alpha_1, \mu_2)$, то случайная величина \bar{X} как сумма независимых нормальных случайных величин также нормальна с параметрами α_1 и μ_2/n , т.е. в этом случае $L(\bar{X}) = N(\alpha_1, \mu_2/n)$ при любом n . Центральные

выборочные моменты M_{nk} также при $n \rightarrow \infty$ обладают свойством асимптотической нормальности.

Задачи и решения

Выборка и способы ее представления

Пусть имеется случайная величина ξ и функция распределения $F_\xi(x)$ и некоторый эксперимент. Осуществляя этот эксперимент мы наблюдаем значение x которое принимает случайная величина ξ . Осуществив n независимых испытаний эксперимента, мы получим последовательность (x_1, \dots, x_n) , называемую выборкой объема n из генеральной совокупности с функцией распределения $F_x(x)$.

Расположив величины (x_1, \dots, x_n) в порядке возрастания получим вариационный ряд $x_1 < x_2 < \dots < x_n$. Размахом выборки называется величина $w = x_n - x_1$.

Из исходной выборки сформируем выборку (z_1, \dots, z_k) , где z_i встречается в исходной выборке n_i раз ($i=1, \dots, k$). Число n_i называется частотой элемента z_i .

$$\sum_{i=1}^k n_i = n$$

Статистическим рядом называется последовательность пар (z_i, n_i) , его представляют в виде таблицы:

z_i	z_1	z_2	\dots	z_k
n	n_1	n_2	\dots	n_k
i				

При большом объеме выборки ее элементы объединяются в группы (разряды) и результаты опытов представляются в виде группированного статического ряда. Для этого интервал, содержащий все элементы выборки, разбивается на k не пересекающихся интервалов. Для упрощения, обычно длины интервалов выбирают одинаковыми. В этом случае длину

интервала полагают равной $b = \frac{w}{k}$. Далее для каждого из интервалов определяют частоты - количество элементов n_i попавших в него. При этом элемент, совпадающий с верхней границей интервала, относят к последующему интервалу. Наряду с частотами одновременно подсчитывают:

относительные частоты $\frac{n_i}{n}$, накопленные частоты $\sum_{j=1}^i n_j$ и

относительные накопленные частоты $\sum_{j=1}^i \frac{n_j}{n}$. Результаты

представляют в виде таблицы:

№ интервала	Границы интервала	Средина интервала	Частота	Накопленная частота	Относительная частота	Относительная накопленная частота
...

Гистограммой частот группированной выборки называют кусочно-постоянную функцию, построенную на интервалах

группировки и принимающую на каждом из них значения $\frac{n_i}{b}$

соответственно. Площадь ступенчатой фигуры расположенной под графиком гистограммы равна объему выборки. Аналогичным образом определяется гистограмма относительных частот. Ее площадь будет равна 1.

Полигоном частот называется ломаная линия с вершинами в

точках $(z_i, \frac{n_i}{b})$.

Эмпирической функцией распределения называется функция

$$F_n(x) = \frac{1}{n} \sum_{z_i < x} n_i, \text{ где } z_i - \text{середины интервалов группировки.}$$

Заметим что $F_n(x)=0$ при $x \leq x_1$ и $F_n(x)=1$ при $x > x_n$.

Методы статистического описания результатов наблюдений

Задание. Для каждой из приведённых ниже выборок определить размах, а также построить вариационный и статистический ряды.

Задача 1

11,15,12,0,16,19,6,11,12,13,16,8,9,14,5,11,3.

Решение: вариационный ряд:

0,3,5,6,8,9,11,11,11,12,12,13,14,15,16,16,19

$\omega=19-0=19$

$z_1=0; z_2=3; z_3=5; z_4=6; z_5=8; z_6=9; z_7=11; z_8=12; z_9=13; z_{10}=14; z_{11}=15; z_{12}=16; z_{13}=19;$

$n_1=1; n_2=1; n_3=1; n_4=1; n_5=1; n_6=1; n_7=3; n_8=2; n_9=1; n_{10}=1;$

$n_{11}=1; n_{12}=2;$

$n_{13}=1;$

$i=1, \dots, 13.$

Статистический ряд:

Z_i	0	3	5	6	8	9	11	12	13	14	15	16	19
N_i	1	1	1	1	1	1	3	2	1	1	1	2	1

Задача 2

17,18,16,16,17,18,19,17,15,17,19,18,16,16,18,18

Решение: вариационный ряд:

15,16,16,16,16,16,17,17,17,17,18,18,18,18,19,19

$$\omega = 19 - 15 = 4$$

$$i = 1, \dots, 15.$$

Статистический ряд:

Z_i	15	16	17	18	19
N_i	1	4	4	5	2

Задание. Найти размах выборки, число и длину интервалов, а также составить таблицу частот (границы первого интервала указываются).

Задача 3

Время решения контрольной задачи учениками 4-го класса
(в секундах):

38 60 41 51 33 42 45 21 53 60 68 52 47 46 49 49 14 57 54 59 77
47 28 48 58

32 42 58 61 30 61 35 47 72 41 45 44 55 30 40 67 65 39 48 43 60
54 42 59 50

Первый интервал: 14 – 23

Решение: $w = 77 - 14 = 63$; $b = 23 - 14 = 9$; $k = 63 / 9 = 7$;

Но- мер ин- тер- вала	Границы интервала	Сере- дина интер- вала	Ча- сто та	Накоп- ленная частота	Относит ельная частота	Накоп- ленная относи- тельная частота
1	14 – 23	18,5	2	2	0,04	0,04
2	23 – 32	27,5	3	5	0,06	0,1
3	32 – 41	36,5	6	11	0,12	0,22
4	41 – 50	45,5	17	28	0,34	0,56
5	50 – 59	54,5	10	38	0,2	0,76

6	59 – 68	63,5	9	47	0,18	0,94
7	68 – 77	72,5	3	50	0,06	1

Задача 4

Продолжительность работы электронных ламп одного типа (в час).

13,4 14,7 15,2 15,1 13,0 8,8 14,0 17,9 15,1 16,5 16,6 14,2 16,3
 14,6 11,7 16,4 15,1 17,6 14,1
 18,8 11,6 13,9 18,0 12,4 17,2 14,5 16,3 13,7 15,5 16,2 8,4 14,7
 15,4 11,3 10,7 16,9 15,8 16,1
 12,3 14,0 17,7 14,7 16,2 17,1 10,1 15,8 18,3 17,5 12,7 20,7 13,5
 14,0 15,7 21,9 14,3 17,7 15,4
 10,9 18,2 17,3 15,2 16,7 17,3 12,1 19,2

Первый интервал: 8,4 – 10,4

Решение: $\omega=21,9-8,4=13,5$; $k=7$; $n=65$; $l=13,5/7=2$;

Но- мер ин- тер- вала	Грани- цы интер- вала	Сере- дина интер- вала	Ча- сто- та	Накоп- лен- ная час- тота	Относи- тельная частота	Накоплен- ная относи- тельная частота
1	8,4 – 10,4	9,4	3	3	0,0462	0,0462
2	10,4 – 12,4	11,4	7	10	0,1077	0,1539
3	12,4 – 14,4	13,4	13	23	0,2	0,3538
4	14,4 – 16,4	15,4	21	44	0,3231	0,6769
5	16,4 – 18,4	17,4	17	61	0,2615	0,9385
6	18,4 – 20,4	19,4	2	63	0,0308	0,9693

7	20,4 – 22,4	21,4	2	65	0,0308	1
---	----------------	------	---	----	--------	---

Задание. Построить графики эмпирических функций распределения, гистограммы и полигоны частот для выборок, представляемых статистическими рядами.

Задача 5

Z i	15	16	17	18	19
N i	1	4	5	4	2

Решение:

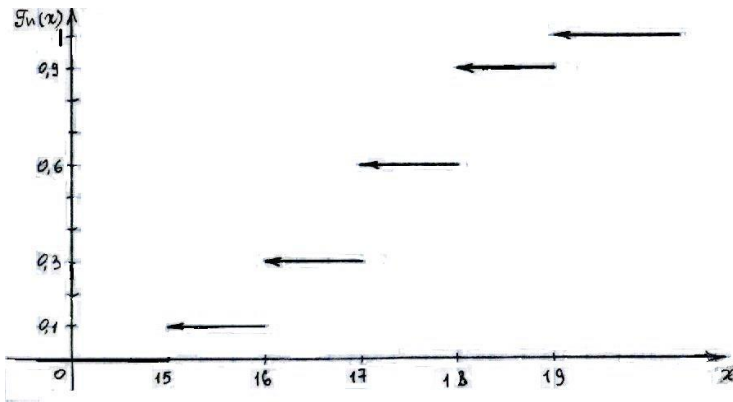
$$F_n(x) = \frac{1}{n} \sum_{z \leq x} n_i$$

$$z \leq x^{(1)} \Rightarrow F_n(x) = 0;$$

$$z > x^{(n)} \Rightarrow F_n(x) = 1$$

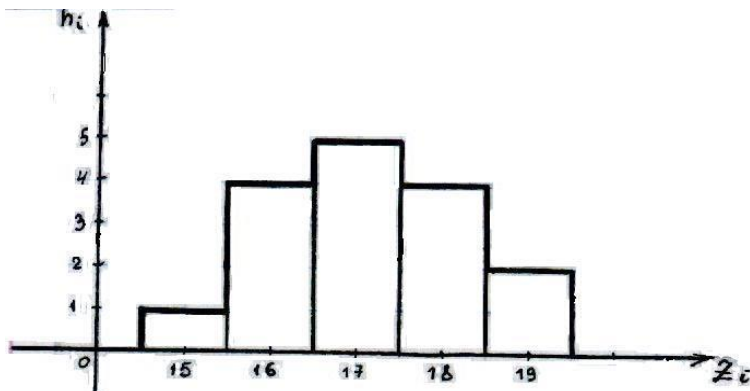
n_i	$\sum_{j=1}^i n_j$	$\frac{1}{n} \sum_{j=1}^i n_j$
1	1	0,1
4	5	0,3
5	10	0,6
4	14	0,9
2	16	1

а) ЭФР: $n=16$

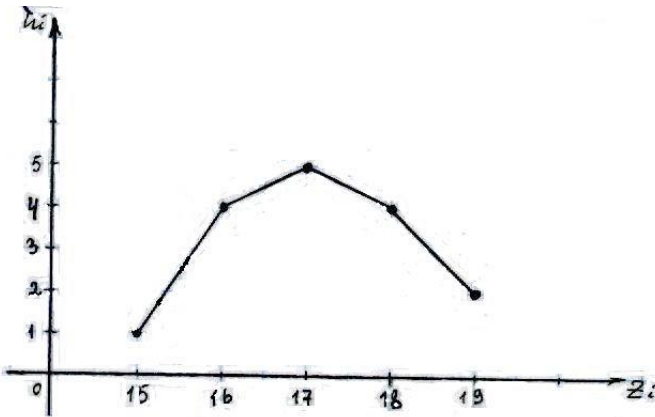


б) Гистограмма частот $l=1$;

Полуинтервал	[14,5; 15,5)	[15,5; 16,5)	[16,5; 17,5)	[17,5; 18,5)	[18,5; 19,5)
n_i	1	4	5	4	2
$h_i = \frac{n_i}{l}$	1	4	5	4	2



в) Полигон частот: $(z_i; \frac{n_i}{l})$: (15,1); (16,4); (17,5); (18,4); (19,2)



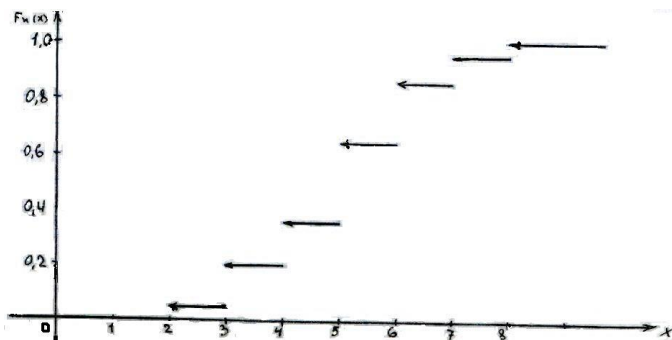
Задача 6

z_i	2	3	4	5	6	7	8
n_i	1	3	4	6	5	2	1

Решение: а) ЭФР

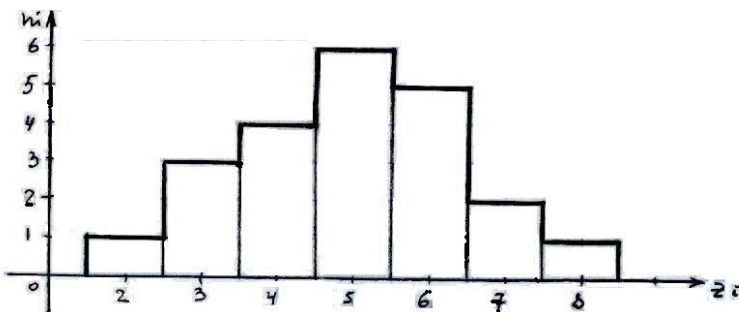
$n=22$

n_i	1	3	4	6	5	2	1
$\sum_{j=1}^i n_j$	1	4	8	14	19	21	22
$\frac{1}{n} \sum_{j=1}^i n_j$	0,045	0,2	0,36	0,64	0,86	0,95	1

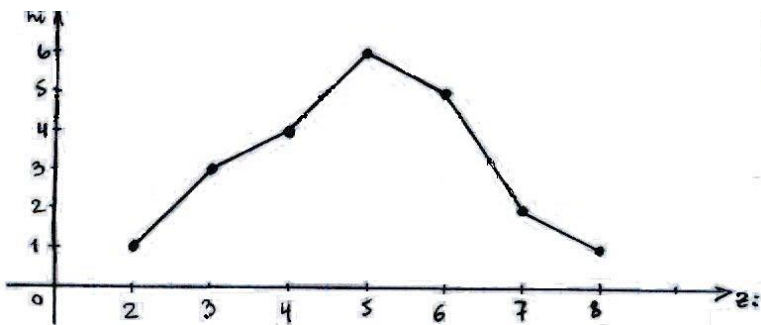


б) Гистограмма частот: $l=1$

Полуинтервал.	[1,5; 2,5)	[2,5; 3,5)	[3,5; 4,5)	[4,5; 5,5)	[5,5; 6,5)	[6,5; 7,5)	[7,5; 8,5]
n_i	1	3	4	6	5	2	1
$h_i = n_i / l$	1	3	4	6	5	2	1



в) Полигон частот: (2,1); (3,3); (4,4); (5,6); (7,2); (8,1); (6,5)



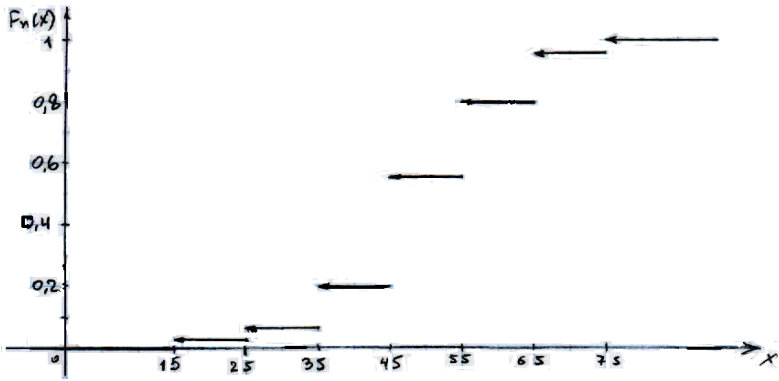
Задача 7

Границы интервала.	10	20	30	40	50	60	70
	-	-	-	-	-	-	-
	.20	30	40	50	60	70	80
Частоты	1	2	7	18	12	8	2

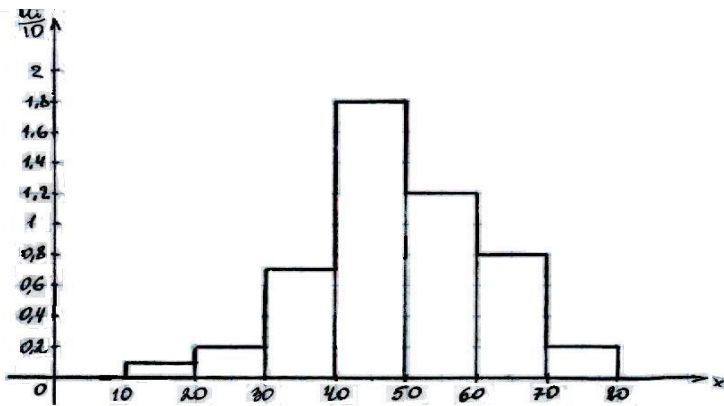
Решение:

а) ЭФР $n=50$

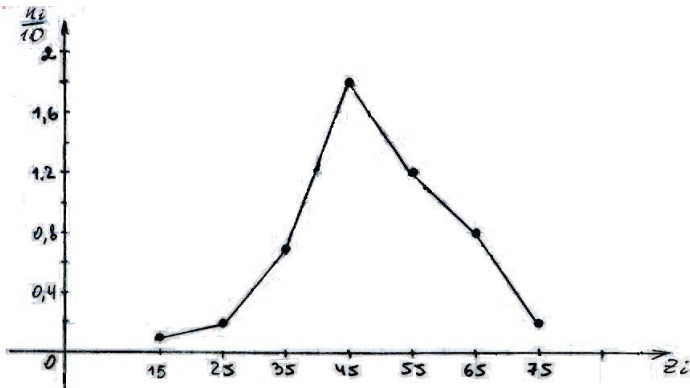
Середина интер. z_i	15	25	35	45	55	65	75
n_i	1	2	7	18	12	8	2
$\sum_{j=1}^i n_j$	1	3	10	28	40	48	50
$\frac{1}{n} \sum_{j=1}^i n_j$	0.02	0.06	0.2	0.56	0.8	0.96	1



б) гистограмма частот: $l=10$



в) ПОЛИГОН ЧАСТОТ



Задача 8

Для выборки:

11, 68, 45, 45, 54, 12, 18, 45, 12, 56, 23, 24, 36, 15, 78, 53, 29,
 45, 26, 35, 65, 14, 72, 42, 12, 26, 18, 14, 23, 17, 39, 40, 24, 29,
 65, 15, 45, 41, 28, 64

построить таблицу частот группированной выборки, полигон, гистограмму частот и эмпирическую функцию распределения, выбрав длину интервала разбиения равной 10.

Решение:

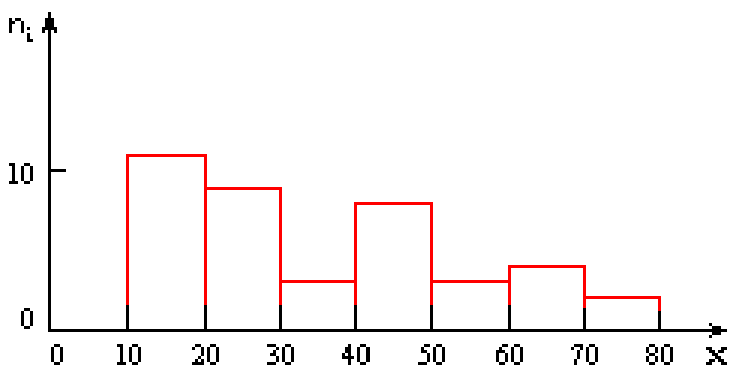
$$A = 11, \quad B = 78$$

$$W = B - A = 67 - \text{размах выборки}$$

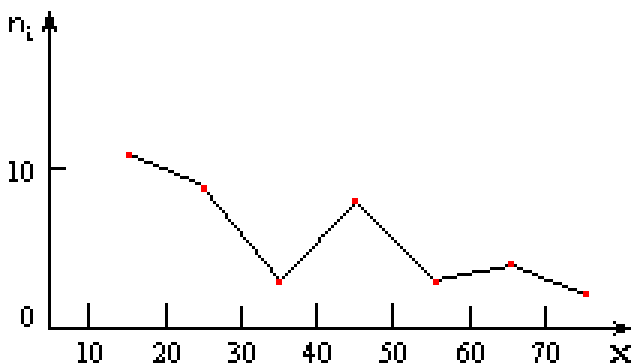
$$n = 40 - \text{объем выборки}$$

$$b = 10 - \text{длина интервала группировки}$$

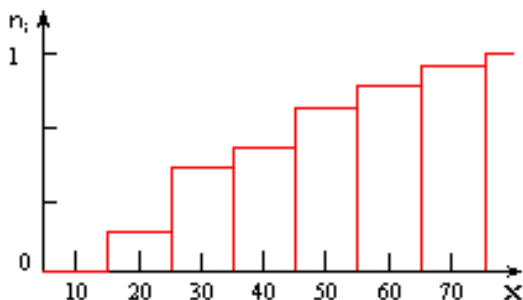
№	Границы интервала	Средина интервала	Частота	Накопленная частота	Относительная частота	Относительная накопленная частота
1	10-20	15	11	11	0,275	0,275
2	20-30	25	9	20	0,225	0,500
3	30-40	35	3	23	0,075	0,575
4	40-50	45	8	31	0,200	0,775
5	50-60	55	3	34	0,075	0,850
6	60-70	65	4	38	0,100	0,950
7	70-80	75	2	40	0,050	1,000



а) Гистограмма частот



б) Полигон частот



в) График эмпирической функции распределения

Числовые характеристики выборочного распределения

Для каждой реализации измерений $x_1(\omega), \dots, x_n(\omega)$ эмпирическая функция распределения $F_n^*(x)$ является функцией распределения некоторой дискретной случайной величины, принимающей n значений: $x_1(\omega), \dots, x_n(\omega)$ с вероятностями равными $1/n$. При различных ω соответствующие функции $F_n^*(x)$ различны. При каждом ω можно ввести различные числовые характеристики соответствующего данному ω закона распределения,

определяемого $F_n^*(x)$. Эти характеристики носят название выборочных. Выборочные моменты (выборочное математическое ожидание и дисперсия) порядка v вычисляются по формулам

$$m_v = \frac{1}{n} \sum_{k=1}^n x_k^v, \quad D_v = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^v,$$

$$\text{где } \bar{x} = a_1 = \frac{1}{n} \sum_{k=1}^n x_k$$

Выборочной модой d_X^* унимодального распределения называется элемент выборки, встречающийся с наибольшей частотой.

Выборочной медианой называется число h_X^* , которое делит вариационный ряд на две части, содержащие равное число элементов.

Если объем выборки нечетное число (т.е. $n=2*L+1$), то $h_X^* = x^{(L+1)}$, то есть является элементом вариационного ряда со средним номером.

Если же $n=2*L$ – объем выборки четное число, то $h_X^* = \frac{1}{2}(x^L + x^{(L+1)})$.

Предположим, что в результате наблюдений случайной величины ξ присутствует одна и та же систематическая погрешность или. если результаты наблюдений подвергнуть преобразованию масштаба, т.е. увеличить или уменьшить одновременно в k раз. Как изменятся выборочное среднее, мода, медиана и дисперсия? Для ответа на этот вопрос группированную выборку преобразуют следующим образом:

$$z_i = \frac{1}{b}(z_i - d_X^*), \quad i = 1, 2, \dots, k \quad (1.13)$$

где d_X^* - выборочная мода,

z_i - элемент группированной выборки,

b - длина интервала группировки.

Соотношение (1.13) показывает, что в выборку z_1, z_2, \dots, z_k внесена систематическая ошибка d_x^* , а результат подвергнут преобразованию масштаба с коэффициентом $k = 1/b$. Полученный в результате набор чисел $u_1, u_2, u_3, \dots, u_k$ можно рассматривать как выборку из генеральной совокупности $U = \frac{1}{b}(x - d_x^*)$.

Тогда выборочное среднее \bar{x} и дисперсия исходных данных связаны со средним \bar{u} и дисперсией D_u^* преобразованных данных следующими соотношениями:

$$\begin{aligned}\bar{x} &= b\bar{u} + d_x^*, \\ D_x^* &= b^2 D_u^*.\end{aligned}$$

Задача 9

Определить среднее, моду и медиану для выборки 5, 6, 8, 2, 3, 1, 1, 4.

Решение:

Представим данные в виде вариационного ряда: 1, 1, 2, 3, 4, 5, 6, 8. Выборочное среднее $\bar{x} = (1+1+2+3+4+5+6+8)/8 = 3,75$.

Все элементы входят в выборку по одному разу, кроме 1, следовательно, мода $\tilde{d}_x = 1$. Так как $n=8$, то медиану

определим так $\tilde{h}_x = \frac{1}{2}(3 + 4) = 3,5$

Задача 10

Рассчитать моду, медиану, среднее и дисперсию следующей выборки:

3,1; 3,0; 1,5; 1,8; 2,5; 3,1; 2,4; 2,8; 1,3

Решение:

Вариационный ряд: 1,3; 1,5; 1,8; 2,4; 2,5; 2,8; 3,0; 3,1; 3,1

Статистический ряд

z_i	1,3	1,5	1,8	2,4	2,5	2,8	3,0	3,1
n_i	1	1	1	1	1	1	1	2

$$d_x^* = 3,1;$$

$$n = 2k + 1 \Rightarrow 9 = 2 \cdot 4 + 1 \Rightarrow k = 4$$

$$h_x^* = x^{(k+1)} = x^{(5)} = 2,5$$

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k = \frac{1}{9} (1,3 + 1,5 + 1,8 + 2,4 + 2,5 + 2,8 + 3,0 + 3,1 + 3,1) \approx 2,39$$

$$D_x^* = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{1}{9} [(1,3 - 2,39)^2 + (1,5 - 2,39)^2 + (1,8 - 2,39)^2 + (2,4 - 2,39)^2 + (2,5 - 2,39)^2 + (2,8 - 2,39)^2 + (3,0 - 2,39)^2 + (3,1 - 2,39)^2 + (3,1 - 2,39)^2]$$

$$= \frac{1}{9} (1,1881 + 0,7921 + 0,3481 + 0,0001 +$$

$$0,0121 + 0,1681 + 0,3721 + 0,5041 + 0,5041) \approx 0,43$$

Ответ: $d_x^* = 3,1; h_x^* = 2,5; \bar{x} \approx 2,39; D_x^* \approx 0,43$

Задача 11

Доказать, что выборочные начальные и центральные моменты порядка $s = 1, 2, \dots$ для негруппированной выборки объёма n определяются следующим образом:

$$\alpha_s^* = \frac{1}{n} \sum_{j=1}^n x_j^*; \mu = \frac{1}{n} \sum_{j=1}^n (x_j - \alpha_1^*)^s$$

Доказательство: по определению $\alpha_s = \sum_{j=1}^n x_j^s p_j$,

т. к. x_j - независимые случайные величины, то

$$\alpha_s^* = \sum_{j=1}^n x_j^* \frac{1}{n} = \frac{1}{n} \sum_{j=1}^n x_j^*$$

$$\alpha_1^* = \frac{1}{n} \sum_{j=1}^n x_j$$

по определению $\mu_s = \mu(\xi - \mu\xi)^s$, если указанное математическое ожидание существует

$$\mu_s^* = \sum_{j=1}^n (x_j - Mx_j)^s p_j = \sum_{j=1}^n (x_j - \sum_{j=1}^n x_j p_j)^s \frac{1}{n} = \frac{1}{n} \sum_{j=1}^n (x_j - \alpha_1^*)^s$$

доказано.

Задача 12

Доказать, что выборочные начальные и центральные моменты порядка $s, s=1,2,\dots$ для группированной выборки объёма n определяются следующими формулами:

$$\alpha_s^* = \frac{1}{n} \sum_{i=1}^n u_i z_i^s; \mu_s^* = \frac{1}{n} \sum_{i=1}^n u_i (z_i - \alpha_i^*)^s$$

Доказательство: по определению $\alpha_s = \sum_{i=1}^n x_i^s p_i; p_i = \frac{n_i}{n}$,

где n_i - численность разряда (группы), z_i - среднее значение для разряда.

Таким образом: $\alpha_s^* = \sum_{i=1}^k z_i^s \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^k z_i^s n_i$

$$\alpha_1^* = \sum_{i=1}^k z_i n_i$$

По определению $\mu_s = \sum_{i=1}^k (x_i - Mx_i)^s p_i$ Таким образом:

$$\mu_s^* = \sum_{i=1}^k (z_i - Mz_i)^s \frac{n_i}{n} = \frac{1}{n} \sum_{i=1}^k n_i (z_i - \sum_{i=1}^k z_i \frac{n_i}{n})^s = \frac{1}{n} \sum_{i=1}^k n_i (z_i - \frac{1}{n} \sum_{i=1}^k n_i z_i)^s = \frac{1}{n} \sum_{i=1}^k n_i (z_i - \alpha_i^*)^s$$

Доказано.

Задача 13

Доказать, что для выборочной дисперсии справедлива следующая формула

$$D_x^* = \alpha_2^* - \bar{x}^2$$

Доказательство: по определению $D_x = \sum_{k=1}^n (x_k - Mx)^2 p$

$$\begin{aligned} D_x^* &= \sum_{k=1}^n (x_k - Mx)^2 \frac{1}{n} = \frac{1}{n} \sum_{k=1}^n (x_k - \frac{1}{n} \sum_{k=1}^n x_k)^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2 = \frac{1}{n} \sum_{k=1}^n (x_k^2 - 2x_k \bar{x} + \bar{x}^2) = \\ &= \frac{1}{n} \sum_{k=1}^n x_k^2 - 2\bar{x} \frac{1}{n} \sum_{k=1}^n x_k + \bar{x}^2 = \frac{1}{n} \sum_{k=1}^n x_k^2 - 2\bar{x}^2 + \bar{x}^2 = \alpha_2^* - \bar{x}^2 \end{aligned}$$

Доказано.

Задача 14

Вычислить среднее и дисперсию группированной выборки:

Границы интервалов	134-138	138-142	142-146	146-150	150-154	154-158
Частоты	1	3	15	18	14	2

Длина интервала группировки $b=4$, значение середины интервала, встречающегося с наибольшей частотой, $d_x^*=148$. Таким образом, преобразование последовательности середин интервалов выполняется по формуле: $u_i = \frac{z_i - 148}{4}$, $i=1, 2, \dots, 6$

Вычисления удобно свести в таблицу:

i	z_i	u_i	n_i	$n_i u_i$	$n_i u_i^2$	$N_i(u_i + 1)^2$
1	136	-3	1	-3	9	4
2	140	-2	3	-6	12	3
3	144	-1	15	-15	15	0
4	148	0	18	0	0	18
5	152	1	14	14	14	56
6	156	2	2	4	8	18
Σ	-	-	53	-6	58	99

Последний столбец служит для контроля вычислений при помощи тождества

$$\sum n_i (u_i + 1)^2 = \sum n_i u_i^2 + 2 \sum n_i u_i + \sum n_i$$

Подставляя в тождество данные последней строки таблицы, получим

$$58 + 2 * (-6) + 53 = 99$$

Следовательно, вычисления выполнены правильно. То находим

$$\bar{u} = \frac{-6}{53} \approx -0,133 \quad D_U^* = \frac{58 - (-6)^2 / 53}{53} \approx 1,108$$

И окончательно вычисляем:

$$\bar{x} \approx (-0.113) * 4 + 148 \approx 147,548$$

$$D_x^* \approx 4^2 * 1,103 \approx 17,728$$

Для выборок, приведённых в следующих задачах, выполнить следующие задания:

1) вычислить среднее и дисперсию, предварительно проведя группировку выборки с заданной длиной интервала, для упрощения вычислений преобразовать данные по формуле

$$u_i = \frac{1}{b}(z_i - d_x^*), i = \overline{1, k} \quad \text{где } d_x^* - \text{выборочная мода, } b - \text{длина}$$

интервала, $k = \frac{1}{b}$;

2) вычислить среднее и дисперсию негруппированной выборки, используя заданные значения.

Сравнить результаты вычислений

Задача 15

Положительные отклонения от номинального размера у партии деталей (в мм)

17 21 8 20 23 18 22 20 17 12

20 11 9 19 20 9 19 17 21 13

17 22 22 10 20 20 15 19 20 20

13 21 21 9 14 11 19 18 23 19

$$n = 40; b = 2; \sum x_i = 689; \sum x_i^2 = 12635$$

Решение: 1)

Номер интервала	Границы интервала	Середина интервала	Частота n_i	u_i	$u_i \cdot n_i$	$n_i \cdot u_i^2$
1	8 – 10	9	4	-6	-24	144
2	10 – 12	11	3	-5	-15	75
3	12 – 14	13	3	-4	-12	48
4	14 – 16	15	2	-3	-6	18
5	16 – 18	17	4	-2	-8	16
6	18 – 20	19	7	-1	-7	7
7	20 – 22	21	12	0	0	0
8	22 - 24	23	5	1	5	5

$$\alpha_x^* = 21; u_i = \frac{1}{2}(z_i - 21)$$

$$\bar{u} = \frac{1}{n} \sum_{i=1}^k u_i n_i = \frac{1}{40} (-67) = -1,675$$

$$\bar{x} = b\bar{u} + d_x^* = 2(-1,675) + 21 = 17,65$$

$$D_u^* = \frac{1}{n} \sum_{i=1}^k n_i u_i^2 - \bar{u}^2 = \frac{1}{40} \cdot 313 - 2,806 \approx 7,825 - 2,8056 \approx 5,02$$

$$D_x^* = b^2 D_u^* = 20,08$$

$$2) \bar{x} = \frac{1}{n} \sum_{i=1}^k x_i = \frac{1}{40} \cdot 689 \approx 17,225$$

$$D_x^* = \frac{1}{n} \sum x_i^2 - \bar{x}^2 = \frac{1}{40} \cdot 12635 - 296,701 \approx 19,17$$

$$\bar{x}_1 > \bar{x}_2; D_{x1}^* > D_{x2}^*$$

Ответ: 1) $\bar{x} \approx 17,7; D_x^* \approx 20,08$; 2) $\bar{x} = 17,2; D_x^* \approx 19,17$

Задача 16

Время восстановления диодов у одной партии (в наносекундах)

69 73 70 68 61 73 70 72 67 70 66 70 76 68 71 71 68

70 64 65
 72 70 70 69 66 70 77 69 71 74 62 72 72 68 70 67 71
 67 72 69
 66 75 76 69 71 67 70 73 71 74
 $n = 50; b = 3; \sum x_i = 3492; \sum x_i^2 = 244342$

Решение: 1)

Номер интервала	Границы интервала	Середина интервала	Частота n_i	u_i	$u_i \cdot n_i$	$n_i \cdot u_i^2$
1	61 – 64	62,5	2	-3	- 6	18
2	64 – 67	65,5	5	-2	-10	20
3	67 – 70	68,5	13	-1	-13	13
4	70 – 73	71,5	21	0	0	0
5	73 – 76	74,5	6	1	6	6
6	76 – 79	77,5	3	2	6	12

$$\alpha_x^* = 71,5$$

$$u_i = \frac{1}{3}(z_i - 71,5)$$

$$u_i = \frac{1}{50}(-6 - 10 - 13 + 0 + 6 + 6) = -\frac{17}{50} = -0,34$$

$$\bar{x} = 3\left(-\frac{17}{50}\right) + 71,5 = -1,02 + 71,5 = 70,48$$

$$D_u^* = \frac{1}{50}(9 + 20 + 13 + 0 + 6 + 12) - 0,1156 = 1,2644$$

$$D_x^* = 9 \cdot 1,2644 \approx 11,38$$

$$2) \bar{x} = \frac{1}{50} \cdot 3492 = 69,84$$

$$D_x^* = \frac{1}{50} \cdot 244342 - 4877,63 \approx 9,21$$

$$\bar{x}_1 > \bar{x}_2; D_{x1}^* > D_{x2}^*$$

Ответ: 1) $\bar{x} \approx 70,44; D_x^* \approx 11,38$; 2) $\bar{x} = 69,84; D_x^* \approx 9,21$

Задача 17

Время реакции (в секундах)

8,5 7,1 6,7 6,2 2,9 4,4 6,0 5,8 5,4

8,2 6,9 6,5 6,1 3,8 6,0 6,0 5,6 5,3

7,7 6,8 6,5 6,1 4,2 4,7 5,6 5,4 5,3

7,4 6,7 6,4 6,1 4,5 6,0 5,8 5,6 5,1

$$n = 36; b = 1; \sum x_i = 213,8; \sum x_i^2 = 1316,82$$

Решение: 1)

Но- мер ин- тер- вала	Границы интервала	Сере- дина интер- вала	Час- тота n_i	u_i	$u_i \cdot n_i$	$n_i \cdot u_i^2$
1	2,9 – 3,9	3,4	2	-3	-6	18
2	3,9 – 4,9	4,4	4	-2	-8	16
3	4,9 – 5,9	5,4	10	-1	-10	10
4	5,9 – 6,9	6,4	14	0	0	0
5	6,9 – 7,9	7,4	4	1	4	4
6	7,9 – 8,9	8,4	2	2	4	8

$$\alpha_x^* = 6,4$$

$$u_i = z_i - 6,4$$

$$u_i = \frac{1}{36}(-16) \approx -0,44$$

$$\bar{x} = -0,44 + 6,4 = 5,96$$

$$D_u^* = \frac{1}{36}56 - 0,1936 \approx 1,35$$

$$D_x^* = 1,35$$

$$2) \bar{x} = \frac{1}{36} \cdot 213,8 \approx 5,94$$

$$D_x^* = \frac{1}{36} \cdot 1316,82 - 35,2704 \approx 1,31$$

$$\bar{x}_1 > \bar{x}_2; D_{x1}^* > D_{x2}^*$$

Ответ: 1) $\bar{x} \approx 5,96; D_x^* \approx 1,35$; 2) $\bar{x} = 5,94; D_x^* \approx 1,31$

Лабораторная работа № 1

Целью лабораторной работы является изучение интерфейса пакета STATISTICA 6.0, методов генерации случайных чисел, процессов формирования выборки с заданным законом распределения.

Сведения о пакете STATISTICA 6.0.

Часто, закончив эксперимент и получив достаточное количество наблюдений и характеристик, экспериментатор сталкивается с необходимостью все это как-то обобщить и сделать правильные выводы из массы разрозненных данных. Статистическая обработка данных приводит порой к далеко идущим выводам и позволяет делать достаточно уверенные прогнозы, выявить закономерности в череде, казалось бы, случайных событий. Математическая статистика как наука уже давно помогала экспериментаторам отвечать на многие интересные вопросы, но с помощью электронной обработки экспериментальных данных можно избежать выполнения

колоссальной рутинной работы, поручив ее компьютеру. Пакет STATISTICA позволяет использовать новые технологии статистической обработки экспериментальных данных. К тому же STATISTICA заменяет очень много статистических таблиц по достаточно широкому спектру законов распределения и тем самым освобождает от необходимости хранить и использовать большое количество справочной литературы.

Система STATISTICA представляет собой интегрированную систему статистического анализа и обработки данных. Она состоит из пяти компонентов:

1. Электронных таблиц для ввода и задания исходных данных, а также специальных таблиц для вывода результатов статистического анализа.

2. Графической системы для визуализации данных и результатов статистического анализа.

3. Набора специализированных статистических модулей, таких как дисперсионный анализ, множественная регрессия, основные статистики и таблицы, кластерный анализ и другие модули, в которых собраны группы логически связанных между собой статистических процедур.

4. Специального инструментария для подготовки отчетов.

5. Встроенных языков программирования, позволяющих расширить стандартные возможности системы.

Пакет STATISTICA 6.0, на котором будут выполнены лабораторные работы, предназначен для статистической обработки данных. Он позволяет:

- ввести либо смоделировать данные для обработки в нужном формате;
- отредактировать данные;
- провести разнообразный статистический анализ данных, обратившись к подходящим процедурам пакета;
- графически отобразить результаты анализа;
- создать отчет всего проделанного.

Для запуска пакета, который работает под Windows, необходимо нажать кнопку [Пуск](#), и в меню [Программы](#) ► [STATISTICA6.0](#) ► [STATISTICA](#). В результате открывается главное окно пакета (рис. 1.3).

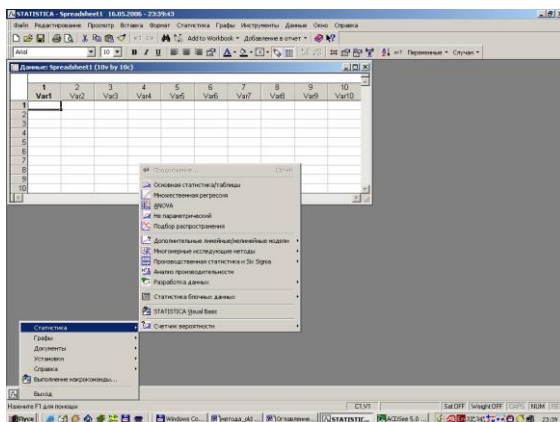



Рис. 1.3. Внешний вид пакета [STATISTICA 6.0](#) после запуска

На рис. 1.3, для более продуктивной работы с пакетом, нажата кнопка [Меню выбора основных модулей](#) обработки информации в программном обеспечении [STATISTICA6.0](#). Рассмотрим пункты меню в процессе выполнения лабораторных работ подробнее. Достаточно наглядно прорисовывается, всем известная из приложений Microsoft Office, *Стандартная панель инструментов*. Основное поле работы программного обеспечения [STATISTICA](#) напоминает также приложение пакета Microsoft Office – Excel, другими словами – электронная таблица. В рассматриваемом пакете эта таблица называется набором данных, в которой столбцам соответствуют обрабатываемые *переменные* (*Variables*), а строкам – *наблюдения* (*Cases*), значения переменных. В качестве переменных обычно выступают исследуемые величины, а случаи (*Cases*) - это значения, которые принимают переменные в отдельных измерениях. Для создания нового набора данных нужно, прежде всего, создать файл с

трафаретом таблицы нужных размеров.

Создание, сохранение, открытие файла данных

Открывают файлы стандартным для Windows способом. В строке меню выбирается пункт File. После щелчка левой кнопкой мыши в появившемся меню выбирают команду Open (Открыть), далее в каталоге выделяется имя файла и нажимается кнопка ОК. Можно также воспользоваться стандартной панелью инструментов, нажав на кнопку .

Электронные таблицы могут содержать как числовую, так и текстовую информацию. Они поддерживают различные типы операций с данными, такие как операции с использованием буфера обмена Windows; операции с выделенными блоками значений, в том числе с использованием метода Drag and Drop автозаполнения блоков и т.д.

Чтобы создать файл данных, находясь в основном рабочем окне системы STATISTICA, необходимо выбрать курсором в строке меню пункт File и щелкнуть левой кнопкой мыши. В появившемся меню выбрать команду New (Новый). В открывшемся диалоговом окне необходимо указать количество переменных и случаев, используемых при выполнении той или иной задачи и нажать кнопку ОК. STATISTICA откроет пустую электронную таблицу.

Чтобы сохранить файл данных, находясь в основном рабочем окне системы STATISTICA, необходимо выбрать курсором в строке меню пункт File и щелкнуть левой кнопкой мыши. В появившемся меню выбрать команду Save (Сохранить). В открывшемся диалоговом окне необходимо указать имя файла и нажать кнопку ОК. STATISTICA автоматически сохранит вашу электронную таблицу.

Размер таблицы (число строк и число столбцов) можно увеличивать и уменьшать. Число столбцов регулируется посредством контекстного меню. После нажатия правой

кнопки мыши на любой переменной (Var) вашему обозрению будет представлено следующее меню (см. рис. 1.4).

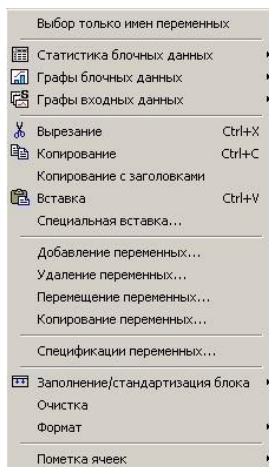


Рис. 1.4. Контекстное меню переменной Var

Посредством этого меню выбирают одну из команд: Delete Variable (Удалить переменные) или Add Variable (Добавить переменные). Аналогично вызывается контекстное меню, позволяющее оперировать с числом случаев Cases (Случаи).

Можно задавать заголовок таблицы, имена переменных и случаев. В качестве имен случаев можно использовать либо числа, либо текст, либо дату. Свойства переменной (имя, формат, код пропущенных значений, метка, формула для вычисления или связь) называются спецификацией переменной и устанавливаются с помощью команды All Specs (Все спецификации) или Current Specs (Текущие спецификации). Эти команды доступны после нажатия кнопки Vars (Переменные) на панели инструментов или правой кнопки мыши посредством контекстного меню. Переменные и случаи можно также копировать (Copy), сдвигать (Shift), ранжировать (Rank), перемещать (Move), перекодировать (Recede) и пр.

Генерация случайных чисел

Реализуем самую простую функцию в пакете STATISTICA – это генератор случайных чисел. Для начала построения выборки необходимо создать новый документ, содержащий одну переменную и некоторое количество случаев.

Ознакомимся с некоторыми внутренними функциями пакета STATISTICA. Генератор случайных чисел, распределенных равномерно на отрезке $[0;1]$, запускается формулой `=rnd(1)`. Случайные числа, распределенные равномерно на отрезке $[0;2]$, можно сгенерировать с помощью оператора `=rnd(2)`. Оператор `=rnd(b-a)+a` генерирует числа, распределенные равномерно на отрезке $[a,b]$.

Пример: Для наглядности сформируем выборку чисел, распределенных равномерно на отрезке $[0;3]$. Щелкнув дважды левой кнопкой мыши по переменной Var1, вам будет представлено окно свойств этой переменной (см. рис. 1.5).

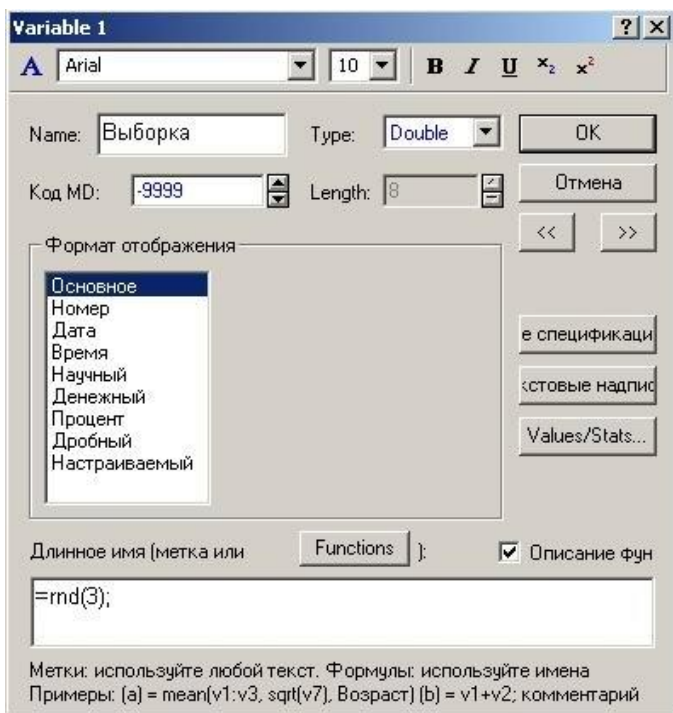


Рис.1.5. Свойства переменной Var1

После заполнения формы свойств переменной, необходимо нажать ENTER и на сообщение пакета STATISTICA (см. рис. 1.6), ответить ДА. Перед вами, в переменной Var1, которая теперь носит имя «Выборка» сформирована выборка чисел, распределенных равномерно на отрезке [0;3].

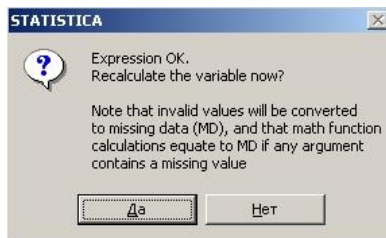


Рис. 1.6. Сообщение пакета STATISTICA6.0

Рассмотрим более подробно функции, позволяющие сформировать выборки чисел, распределенных по разным законам распределения.

1. Нормальное распределение.

$$= VNormal (rnd(1); 2; 3) \quad \text{для } N(2;3),$$

2. Экспоненциальное распределение

$$= VExpon (rnd(1); 2) \quad \text{для } E(0.5) \text{ со средним } 1/2 = 0.5,$$

3. Распределение Коши

$$= VCauchy (rnd(1); 0; 1) \quad \text{для } C(0; 1),$$

4. Логарифмическое распределение

$$= VLognorm (rnd(1); 0,5; 0,5) \quad \text{для } Lgn(0,5; 0,5),$$

5. Распределение Хи-квадрат

$$= VChi2 (rnd(1); 8) \quad \text{для } \chi^2$$

Задания к лабораторной работе

1. Запустить пакет STATISTICA 6.0. Проанализировать содержание рабочего окна построчно. Открыть файлы, имеющиеся в директории Examples, поочередно, пока вам не покажется, что вы это где-то уже видели.

2. Самостоятельно попробовать сформировать выборку, по тому или иному закону распределения, в зависимости от варианта (см. табл. 1.1). Сохранить файл в своем каталоге.

Таблица 1.1

Объем выборки и закон распределения

№	Закон	Объем	p	№	Закон	Объем	p
1	R [0; 2]	50	0.03	9	N(1;4)	60	0.03
2	N(2;0.5)	60	0.02	10	E(1)	70	0.03
3	E(3)	70	0.01	11	R[0;3]	80	0.1
4	R [1,3]	80	0.02	12	N (0; 4)	50	0.3
5	N(0; 1)	50	0.01	13	E(5)	60	0.2
6	E (2)	60	0.03	14	R [3; 6]	70	0.03
7	R [2; 3]	70	0.01	15	N (0; 9)	80	0.02
8	N (0; 4)	80	0.03	16	E (0.2)	50	0.01

3. Создать файл *bv x 15c* с результатами воздействия лекарственного препарата на кровяное давление. Исходные данные содержатся в табл. 1.2.

Таблица 1.2

Кровяное давление (в мм ртутного столба) до и после приема препарата

Но- мер паци- ента	Систолическое давление			Диастолическое давление		
	до	после	Раз ность	до	после	Раз ность
1	210	201	-9	130	125	-5
2	169	165	-4	122	121	-1
3	187	166	-21	124	121	-3
4	160	157	-3	104	106	2
5	167	147	-20	112	101	-11

Продолжение табл. 1.2

6	176	145	-31	101	85	-16
7	185	168	-17	121	98	-23
8	206	180	-26	124	105	-19
9	173	147	-26	115	103	-12
10	146	136	-10	102	98	-4
11	174	151	-23	98	90	-8




Шаг 1. Создание электронной таблицы.

Выберите команду New (Новый) из меню File (Файл). В появившейся форме создания нового документа необходимо количество переменных установить равным 6, а случаев – 15 и нажать на кнопку ОК. На экране появится пустая электронная таблица размером *6x15c*. Сохраните файл с именем *1_1.sta*.

Шаг 2. Оформление заголовка таблицы.

Дважды щелкните мышью (кликните) на белом поле, находящемся ниже панели инструментов, в таблице и введите заголовок «КРОВЯНОЕ ДАВЛЕНИЕ». Нажмите ENTER.


Шаг 3. Задание имен переменных.

Кликните дважды по переменной VAR1 электронной таблицы. В поле Name (Имя) открывшегося окна напишите: C_DO. Далее нажмите кнопку , переменной VAR2 присвойте имя C_POST, , переменной VAR4 присвойте имя D_DO, , переменной VAR5 присвойте имя D_POST.

Шаг 4. Ввод данных в электронную таблицу.

Введите данные в электронную таблицу пакета STATISTICA с клавиатуры. Значения переменных VAR3 и VAR6 не вводите в связи с тем, что они будут подсчитаны посредством внутренних функций пакета.

Шаг 5. Вычисление значений переменной по формуле.

Кликните по переменной VAR3 электронной таблицы. В поле Long Name (Длинная метка) запишите формулу для вычисления: $=v2-v1$ и нажмите на кнопку ОК. Аналогичным образом поручите системе вычислять и вводить данные в столбец VAR6. После получения результатов, не забудьте сохранить файл данных кнопкой  на панели инструментов.

Примечание: Не забывайте после ввода каждой формулы ставить в конце символ «;» и следом пробел!

4. Необходимо генерировать выборку объема $n = 50$, распределенную по показательному закону с математическим ожиданием 5 ($E(5)$).

Шаг 1. Создание электронной таблицы.

Выберите команду New (Новый) из меню File (Файл). В появившейся форме создания нового документа необходимо количество переменных установить равным 1, а случаев – 50 и нажать на кнопку ОК. На экране появится пустая электронная таблица размером $1v \times 50c$. Сохраните файл с именем ***1_3.sta***.

Шаг 2. Оформление заголовка таблицы.

Дважды щелкните мышью (кликните) на белом поле, находящемся ниже панели инструментов, в таблице и введите заголовок «Выборка по показательному закону распределения». Нажмите ENTER.

Шаг 3. Генерируем выборку.

Кликнем дважды по переменной VAR1 и введем имя Name x (например), в нижнем поле Long Name введем выражение, определяющее переменную. Ввод сделайте набором на клавиатуре или с помощью клавиши Functions, выбирая в меню Category и Name требуемую функцию и вставляя клавишей Insert. Для задания закона распределения $E(5)$ введите:

=VExpon (rnd(1); 1/5)

Здесь вместо выражения $1/5$ можно набрать значение параметра $\lambda = 0.2$.

Шаг 4. Построение выборки графически.

В пункте меню Graph (Графы) выбрать пункт 2D Графы ► Гистограммы. После правильного выполнения операции перед вами будет представлена форма, изображенная на рис. 1.5.

В пункте меню Graph (Графы) выбрать пункт 2D Графы ► Гистограммы. После правильного выполнения операции перед вами будет представлена форма, изображенная на рис. 1.7.

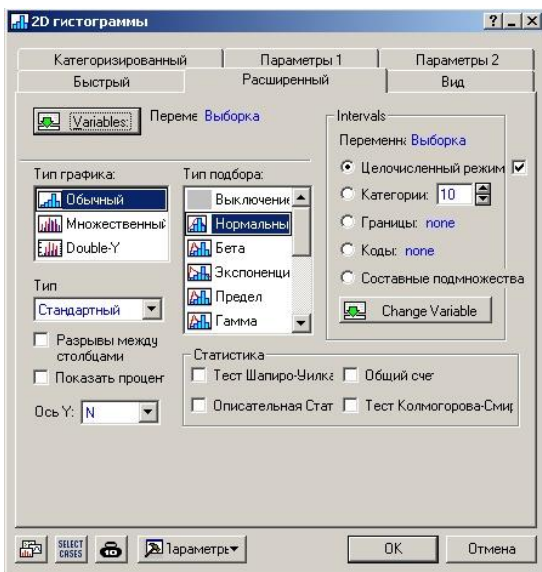


Рис. 1.7. Построение 2D-Гистограммы

Нажав на кнопку Variables, выбираем переменную и нажимаем ОК. Гистограмма построена (см. рис. 1.8).

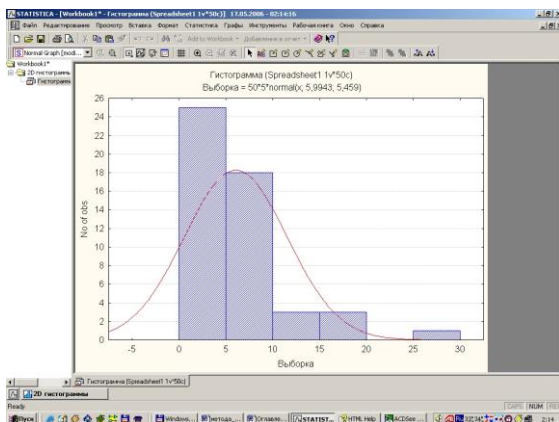


Рис. 1.8. Построенная гистограмма выборки

Составить отчет по выполненной работе.

Отчет по выполненной работе должен содержать:

- Постановку задачи.
- Сохраненные на переносном носителе информации файлы *1_1.sta*, *1_2.sta*, *1_3.sta*.
- Описание процесса формирования выборки, по тому или иному закону распределения, в зависимости от варианта (см. табл. 1.1).
Конечный файл должен быть сохранен на том же носителе.
- Вывод о проделанной лабораторной работе.

2. ОЦЕНИВАНИЕ НЕИЗВЕСТНЫХ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЙ. ТОЧЕЧНЫЕ ОЦЕНКИ И ИХ СВОЙСТВА

2.1. Понятие статистической оценки

Пусть задана статистическая модель $F=\{F\}$ для схемы повторных независимых наблюдений над случайной величиной ξ и $\bar{X}=(X_1,\dots,X_n)$ - выборка из распределения $L(\xi)\in F$. Рассматриваются различные функции $T=T(\bar{X})$ от выборки, которые являются случайными величинами (т.е. для которых при всех t определены вероятности $F_T(t)=P\{T(\bar{X})\leq t\}$). Если при этом функция T не зависит от неизвестного распределения наблюдений, то её называют **статистикой**. Часто по выборке \bar{X} требуется сделать выводы об истинном значении g_0 неизвестной теоретической характеристики $g=g(F)$ наблюдаемой случайной величины. Под этим понимается задача оценить это значение g_0 , т.е. построить такую статистику $T(\bar{X})$, значение которой $t=T(\bar{X})$ при наблюдавшейся реализации \bar{x} выборки можно считать разумным приближением (в каком-то смысле) для g_0 : $t \approx g_0$. В этом случае, говорят, что статистика $T(\bar{X})$ есть оценка g . Так формируется задача точечного оценивания неизвестных параметров распределений.

Для оценивания характеристики g можно использовать различные оценки. Чтобы выбрать лучшую, надо иметь критерий сравнения качества (точности) оценок. Критерии могут быть разными, но любой критерий определяется выбором меры точности оценок (меры близости оценки к истинному значению оцениваемой характеристики). Класс оценок ограничивают некоторыми дополнительными требованиями.

Если определён некоторый класс оценок T_g и выбрана мера точности, то оценка $T\in T_g$, оптимизирующая эту меру, называется оптимальной (в классе T_g).

Если модель F параметрическая: $F = \{F(x; \theta), \theta \in \Theta\}$, то любая теоретическая характеристика является функцией от параметра θ , т.е. речь идёт об оценивании параметрических функций, которые обозначаются: $\tau(\theta)$. Стремятся, чтобы области значений оценок $T(\bar{X})$ и оцениваемой функции совпадали.

2.2. Состоятельность и несмещённость оценок

Качество оценок характеризуется следующими свойствами:

- 1) состоятельность,
- 2) несмещённость,
- 3) эффективность.

1. Состоятельность.

Оценка $\tilde{\theta} \equiv \theta^* \equiv \tilde{\theta}_n = \tilde{\theta}(x_1, \dots, x_n)$ называется состоятельной оценкой параметра θ , если она сходится по вероятности к параметру θ при $n \rightarrow \infty$

$$\tilde{\theta}_n \xrightarrow[n \rightarrow \infty]{P} \theta$$

Следующая теорема устанавливает достаточное условие состоятельности

Теорема 2.1. Если математическое ожидание оценки $\tilde{\theta}_n$ стремится к истинному значению параметра θ , дисперсия стремится к нулю при $n \rightarrow \infty$ ($M[\tilde{\theta}_n] \rightarrow \theta$; $D[\tilde{\theta}_n] \rightarrow 0$, $n \rightarrow \infty$), то $\tilde{\theta}_n$ - состоятельная оценка параметра θ , т.е. $\tilde{\theta}_n \xrightarrow[n \rightarrow \infty]{P} \theta$

Доказательство: В соответствии с неравенством Чебышева имеем

$$P(|\tilde{\theta}_n - M[\tilde{\theta}_n]| > \varepsilon) \leq \frac{D\tilde{\theta}_n}{\varepsilon^2}$$

При $n \rightarrow \infty$ по условиям теоремы: $M[\tilde{\theta}_n] \rightarrow \theta$; $D[\tilde{\theta}_n] \rightarrow 0$, т.е. $P(|\tilde{\theta}_n - M[\tilde{\theta}_n]| > \varepsilon) \rightarrow 0$, а это и есть определение сходимости по вероятности, т.е. $\tilde{\theta}_n \xrightarrow[n \rightarrow \infty]{P} \theta$.

2. Несмещённые оценки.

Любая оценка $T=T(\bar{X})$ является случайной величиной. Общим требованием к построению оценок является требование сосредоточенности распределения T около истинного значения оцениваемого параметра. Чем выше степень этой сосредоточенности, тем лучше соответствующая оценка.

Предположим, что параметр θ - скалярный и введём понятие несмещённой оценки. Статистика $T(\bar{X})$ называется **несмещённой оценкой** для параметра θ , если выполняется условие

$$M_{\theta}[T(\bar{X})] = \theta \quad (2.1)$$

для любого $\theta \in \Theta$.

Для оценок, не удовлетворяющих условию (2.1), введём величину

$$b(\theta) = M_{\theta}[T(\bar{X})] - \theta,$$

называемую **смещением** оценки $T(\bar{X})$. Можно считать, что несмещённые оценки - это такие оценки, для которых смещение $b(\theta) = 0$, для любого $\theta \in \Theta$.

Величину

$$M_{\theta}[(T-\theta)^2] = D_{\theta}[T] + b^2(\theta) \quad (2.2)$$

называют **средним квадратом ошибки** или **среднеквадратической ошибкой** оценки T . Для несмещённых оценок среднеквадратическая ошибка совпадает с дисперсией оценки.

Если оценивается параметрическая функция $\tau(\theta)$, то статистика $T=T(\bar{X})$ является несмещённой оценкой для $\tau(\theta)$, если для любого $\theta \in \Theta$ выполняется соотношение

$$M_{\theta}[T] = \tau(\theta) \quad (2.3)$$

Для класса несмещённых оценок можно построить простую теорию, в которой критерием измерения точности оценки является её дисперсия. А в некоторых случаях требование несмещённости может оказаться очень "жёстким" и привести к нежелательным результатам – отсутствию оценки.

Простейший метод статистического оценивания – **метод подстановки** [или аналогии] состоит в том, что в качестве оценки той или иной числовой характеристики (среднего, дисперсии и др.) генеральной совокупности берут соответствующую характеристику распределения выборки – выборочную характеристику.

Пример: Пусть (x_1, \dots, x_n) выборка \vec{X} из генеральной совокупности с конечными математическим ожиданием m и дисперсией σ^2 . Найти оценку m . Проверить несмещённость и состоятельность полученной оценки.

Решение: По методу подстановки в качестве оценки \tilde{m} математического ожидания надо взять выборочное среднее. Таким образом, получаем

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

Чтобы проверить несмещённость и состоятельность выборочного среднего как оценки m , рассмотрим эту статистику как функцию выборочного вектора (X_1, \dots, X_n) . По определению выборочного вектора имеем: $M[X_i]=m$ и $D[X_i]=\sigma^2$, $i=1,2,\dots,n$, причём X_i - независимые в совокупности случайные величины.

Следовательно

$$M[\bar{X}] = M\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n M[X_i] = \frac{1}{n} \cdot n \cdot m = m$$

$$D[\bar{X}] = D\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \sum_{i=1}^n D[X_i] = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \sigma^2$$

Отсюда получаем, что \bar{X} - несмещённая оценка m . Так как $D[\bar{X}] \rightarrow 0$, при $n \rightarrow \infty$, то в силу теоремы 2.1 \bar{X} является состоятельной оценкой математического ожидания m генеральной совокупности.

2.3. Оптимальные оценки. Теорема об оптимальности оценок

Пусть требуется оценить параметрическую функцию $\tau = \tau(\theta)$ в модели $F = \{F(x; \theta), \theta \in \Theta\}$ по статистической информации, доставляемой выборкой $\vec{X} = (X_1, \dots, X_n)$. Пусть статистика $T = T(\vec{X})$ удовлетворяет условию (2.3). Класс несмещённых оценок обозначим T_τ . Таким образом, $T \in T_\tau$ тогда и только тогда, когда выполнено условие (2.3). Дополнительно предположим, что дисперсии всех оценок из класса T_τ конечны:

$$D_\theta[T] = M_\theta[(T - \tau(\theta))^2] < \infty$$

для любых $T \in T_\tau$ и $\theta \in \Theta$.

В этом случае точность оценок можно измерять величиной их дисперсии и мы получаем простой критерий сравнения различных оценок из класса T_τ . Если

$$D_\theta[T^*] \leq D_\theta[T], \quad \forall \theta \in \Theta \quad (2.4)$$

то по критерию минимума дисперсии оценка T^* равномерно (по параметру θ) не хуже оценки T ; если же в (2.4) строгое неравенство выполняется хотя бы при одном θ , то следует отдать предпочтение T , как более точной оценке. Если условие (2.4) выполняется для любой оценки $T \in T_\tau$, то T^* называют **несмещённой оценкой с равномерно минимальной дисперсией**. Такую оценку для краткости называют

оптимальной, и обозначают τ^* , так как она относится к функции $\tau(\theta)$.

Итак, оптимальной является оценка $\tau^* \in T_\tau$ для которой выполняется условие

$$D_\theta \tau^* = \inf_{T_\tau} D_\theta T, \quad \forall \theta \in \Theta$$

Требование равномерной минимальной дисперсии сильное и не всегда имеет место. Однако оно выделяет оптимальную оценку в классе T_τ однозначно, если такая оценка существует, о чём свидетельствует следующая теорема.

Теорема 2.2. Пусть $T_i = T_i(\bar{X})$, $i=1,2$ - две оптимальные оценки для $\tau = \tau(\theta)$. Тогда $T_1 = T_2$.

Доказательство: Рассмотрим новую оценку $T_3 = (T_1 + T_2)/2$.

Ясно, что $T_3 \in T_\tau$ и

$$D_\theta T_3 = (D_\theta T_1 + D_\theta T_2 + 2\text{cov}_\theta(T_1, T_2))/4 \quad (*)$$

Для любых случайных величин η_1, η_2 имеет место неравенство Коши-Буняковского

$$|\text{cov}(\eta_1, \eta_2)| \leq \sqrt{D\eta_1 D\eta_2},$$

причём знак равенства имеет место, если η_1 и η_2 линейно связаны. Отсюда и из равенства (*), положив

$$D_\theta T_1 = D_\theta T_2 = v = v(\theta),$$

получим

$$D_\theta T_3 \leq (v + |\text{cov}_\theta(T_1, T_2)|)/2 \leq v \quad (**)$$

Поскольку T_i ($i=1,2$) оптимальные оценки,

$$v = D_\theta T_i \leq D_\theta T_3,$$

откуда $D_\theta T_3 = v$, т.е. T_3 также оптимальная оценка.

Но так как в неравенствах (**) имеют место знаки равенства, то $\text{cov}(T_1, T_2) \geq 0$ и более того, $\text{cov}_\theta(T_1, T_2) = D_\theta T_1 = D_\theta T_2 = v$. Следовательно T_1 и T_2 линейно связаны, т.е. $T_1 = kT_2 + a$.

Из условия несмещённости оценок имеем $\tau = k\tau + a$; т.е. $a = \tau(1-k)$, и, следовательно,

$$T_1 - \tau = k(T_2 - \tau).$$

Коэффициент $k = k(\theta)$ является функцией от параметра θ и определяется цепочкой равенств

$v = \text{cov}_0(T_1, T_2) = M_0[(T_1 - \tau)(T_2 - \tau)] = k M_0[(T_1 - \tau)^2] = k D_0 T_2 = kv$
 Отсюда имеем $k \equiv 1$ и, следовательно, $T_1 = T_2$ ■■■

2.4. Критерии оптимальности оценок, основанные на неравенстве Рао-Крамера

Понятия функции правдоподобия, вклада выборки, функции информации.

Рассмотрим что такое функция правдоподобия. Пусть $f(x, \theta)$ - плотность распределения случайной величины ξ (или вероятность в дискретном случае), $\bar{X} = (X_1, \dots, X_n)$ - выборка из $L(\xi) \in F$ и $\bar{x} = (x_1, \dots, x_n)$ - реализация \bar{X} . Функция $L(\bar{x}, \theta) = f(x_1, \dots, x_n; \theta)$ является плотностью распределения случайного вектора \bar{X} . Функция $L(\bar{x}, \theta)$, рассматриваемая при фиксированном \bar{X} , как функция параметра $\theta \in \Theta$, называется **функцией правдоподобия**.

Если элементы выборки независимы, то функция правдоподобия $L(\bar{x}, \theta) = \prod_{i=1}^n f(x_i, \theta)$ [x_i распределены как и ξ].

Если имеется выборка (X_1, \dots, X_n) из ξ , имеющей дискретное распределение $P(\xi = x_i) = p_i$, то функцией правдоподобия называется вероятность того, что $L(\bar{x}, \theta) = P(X_1 = x_1, \dots, X_n = x_n; \theta)$, рассматриваемая как функция параметра θ

$$L(\bar{x}, \theta) = \prod_{i=1}^n P(X_i = x_i; \theta) = \prod_{i=1}^n P(\xi = x_i, \theta)$$

для независимых случайных величин. Функция правдоподобия показывает на сколько при фиксированных значениях выборки правдоподобно то или другое значение параметра.

Рассмотрим основные свойства функции правдоподобия

1. Функция правдоподобия неотрицательна

$$L(\bar{x}, \theta) > 0$$

при всех $\vec{x} \in \mathcal{X}$ и $\theta \in \Theta$

2. Условия нормировки

$$\int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} L(\vec{x}; \theta) d\vec{x} = 1, \quad (d\vec{x} = dx_1 \dots dx_n) \quad (2.5)$$

(для дискретных моделей условия нормировки выражаются через суммы).

Предположим, что функция правдоподобия удовлетворяет следующим условиям:

- 1) дифференцируема по параметру θ ;
- 2) условие регулярности: порядок дифференцирования по θ и интегрирования по x можно менять.

Модели, для которых выполняются эти условия, называют **регулярными**. Общее необходимое условие состоит в том, что выборочное пространство \mathcal{X} не должно зависеть от неизвестного параметра θ .

Пусть параметр θ - скалярный. Случайная величина

$$U(\vec{X}; \theta) = \frac{\partial \ln L(\vec{X}; \theta)}{\partial \theta} = \sum_{i=1}^n \frac{\partial \ln f(X_i; \theta)}{\partial \theta} \quad (2.6)$$

называется **вкладом** (или функцией вклада) выборки \vec{X} (i -е слагаемое в правой части (2.6) называется вкладом i -го наблюдения $i=1, \dots, n$). Предполагается, что $0 < M_0[U^2(\vec{X}; \theta)] < \infty$, для любого $\theta \in \Theta$.

Рассмотрим некоторые свойства вклада $U(\vec{X}; \theta)$ для регулярных моделей. Дифференцируя условие нормировки (2.5) по θ , получаем

$$0 = \int_{-\infty}^{\infty} \frac{\partial L(\vec{x}; \theta)}{\partial \theta} d\vec{x} = \int_{-\infty}^{\infty} \frac{\partial \ln L(\vec{x}; \theta)}{\partial \theta} L(\vec{x}; \theta) d\vec{x} = M_0[U(\vec{X}; \theta)]$$

Для удобства и краткости записи вместо многомерного интеграла будем писать одномерный, но иметь в виду многомерный.

Итак, для регулярной модели

$$M_0[U(\vec{X}; \theta)] = 0, \quad \forall \theta \in \Theta \quad (2.7)$$

Определим функцию информации Фишера или просто информацию Фишера о параметре θ , содержащуюся в выборке \bar{X} :

$$i_n(\theta) = D_\theta[U(\bar{X}; \theta)] = M_\theta[U^2(\bar{X}; \theta)] \quad (2.8)$$

Величину:

$$i(\theta) = i_1(\theta) = M_\theta \left(\frac{\partial \ln f(X_1; \theta)}{\partial \theta} \right)^2 \quad (2.9)$$

называют количеством (фишеровской) информации, содержащейся в одном наблюдении.

Из соотношений (2.6) - (2.9) следует, что $i_n(\theta) = n i(\theta)$, т.е. количество информации, содержащейся в выборке, возрастает пропорционально объёму выборки. Если функция $f(x, \theta)$ дважды дифференцируема по θ , то продифференцировав при $n=1$ выражение (2.7) и применив тот же приём, что и ранее, т.е. умножим и разделим на $f(x_1, \theta)$, получим эквивалентное представление для $i(\theta)$

$$0 = \int_{-\infty}^{\infty} \frac{\partial^2 \ln f(X; \theta)}{\partial \theta^2} f(X; \theta) dx + \int_{-\infty}^{\infty} \left(\frac{\partial \ln f(X; \theta)}{\partial \theta} \right)^2 f(X; \theta) dx, \text{ т.е.}$$

$$i(\theta) = - M_\theta \left(\frac{\partial^2 \ln f(X_1; \theta)}{\partial \theta^2} \right) \quad (2.10)$$

Пример: Вычислим функцию $i(\theta)$ для нормальной модели $N(\theta, \sigma^2)$. Вклад одного наблюдения

$$U(X_1; \theta) = \frac{\partial \ln f(X_1; \theta)}{\partial \theta} = \frac{X_1 - \theta}{\sigma^2},$$

$$\frac{\partial^2 \ln f(X_1; \theta)}{\partial \theta^2} = -\frac{1}{\sigma^2},$$

отсюда по формуле (2.10) получаем:

$$i(\theta) = \frac{1}{\sigma^2}.$$

2.5. Неравенство Рао-Крамера и эффективные оценки

Рассмотрим задачу оценивания заданной параметрической функции $\tau(\theta)$ в модели $\mathcal{F} = \{F(x, \theta), \theta \in \Theta\}$. Пусть модель \mathcal{F} - регулярна, $\tau(\theta)$ - дифференцируема и T_τ - класс всех несмещенных оценок $\tau(\theta)$. Тогда имеет место следующая теорема.

Теорема 2.3. (Неравенство Рао-Крамера). Для любой оценки $T = T(\bar{X}) \in T_\tau$ справедливо неравенство:

$$D_\theta T \geq \frac{[\tau'(\theta)]^2}{i_n(\theta)} \quad (2.11)$$

Равенство здесь имеет место тогда и только тогда, когда T - линейная функция вклада выборки, т.е.

$$T(\bar{X}) - \tau(\theta) = a(\theta)U(\bar{X}; \theta), \quad (2.12)$$

где $a(\theta)$ - некоторая функция от θ .

Доказательство: По условию теоремы

$$M_\theta[T(\bar{X})] = \int_{-\infty}^{\infty} T(\bar{x})L(\bar{x}; \theta)d\bar{x} = \tau(\theta), \quad \forall \theta \in \Theta.$$

Модель \mathcal{F} - регулярна, поэтому дифференцируя это тождество по θ и учитывая (2.7), получаем

$$\begin{aligned} \tau'(\theta) &= \int_{-\infty}^{\infty} T(\bar{x}) \frac{\partial L(\bar{x}; \theta)}{\partial \theta} \cdot \frac{L(\bar{x}; \theta)}{L(\bar{x}; \theta)} d\bar{x} = \int_{-\infty}^{\infty} T(\bar{x}) \frac{\partial \ln L(\bar{x}; \theta)}{\partial \theta} L(\bar{x}; \theta) d\bar{x} = \\ &= M_\theta[T(\bar{X})U(\bar{X}; \theta)] = M_\theta[T(\bar{X})]M_\theta[U(\bar{X}; \theta)] + \text{cov}_\theta[T(\bar{X})U(\bar{X}; \theta)] \end{aligned} \quad (2.13)$$

Используя неравенство Коши-Буняковского и формулу (2.8) для определения фишеровской информации, получаем

$$\tau'(\theta) \leq \sqrt{D_\theta[T(\bar{X})]D_\theta[U(\bar{X}; \theta)]}$$

Возведём в квадрат:

$$[\tau'(\theta)]^2 \leq i_n(\theta)D_\theta[T(\bar{X})] \quad (*)$$

Отсюда имеем неравенство (2.11).

Неравенство (*) обращается в равенство тогда и только тогда, когда $T(\bar{X})$ и $U(\bar{X}; \theta)$ линейно связаны. ■

Неравенство (2.11) называется **неравенством Рао-Крамера**. Оно определяет нижнюю границу дисперсии всех несмещённых оценок заданной параметрической функции $\tau(\theta)$ для регуляных моделей.

Если существует оценка $T^* \in T_\tau$, для которой нижняя граница Рао-Крамера достигается, то её называют **эффективной**. Эффективная оценка является оптимальной и согласно теореме 2.2 она единственна. Из теоремы 2.3 следует, что критерием эффективности оценки является представление (2.12). Будем называть этот критерий оптимальности **критерием Рао-Крамера**. Если для оценки T^* выполнено соотношение (2.12), то из формулы (2.13) следует

$$\tau'(\theta) = \frac{1}{a(\theta)} D_\theta[T^*]$$

т.е. для дисперсии эффективной оценки справедлива формула

$$D_\theta[T^*] = a(\theta)\tau'(\theta) \quad (2.14)$$

Отметим следующее: вклад выборки $U(\bar{X}; \theta)$ однозначно определяется моделью, поэтому представление (2.12) единственно. Следовательно, эффективная оценка может существовать только для одной параметрической функции $\tau(\theta)$ и не существует ни для какой другой функции параметра θ , отличной от $a\tau(\theta)+b$, где a и b - константы.

Замечание.

Если $T = T(\bar{X})$ оценка со смещением $b(\theta)$ и $b(\theta)$ - дифференцируема, то неравенство Рао-Крамера будет иметь вид:

$$D_\theta[T] \geq \frac{[\tau'(\theta) + b'(\theta)]^2}{ni(\theta)},$$

обобщающее (2.11).

2.6. Достаточные статистики. Теорема факторизации Неймана - Фишера

Рассмотренные ранее критерии имеют ограниченную применимость по двум причинам:

1) они требуют жёстких условий регулярности исходной модели;

2) в лучшем случае позволяют находить оптимальные оценки для отдельных параметрических функций $\tau(\theta)$. Более эффективным способом построения оптимальных оценок является использование достаточных статистик.

По определению статистика $T=T(\bar{X})$ называется **достаточной** для модели $\mathbb{F}=\{F(x,\theta),\theta\in\Theta\}$ (или для параметра θ), если условная плотность (или вероятность в дискретном случае) $L(\bar{x}|t;\theta)$ случайного вектора $\bar{X}=(X_1,\dots,X_n)$ при условии $T(\bar{X})=t$ не зависит от параметра θ .

Эквивалентным определением достаточности является следующее: для любого события $A\subset\mathbb{X}$ условная вероятность $P_\theta(\bar{X}\in A|T(\bar{X})=t)$ не зависит от θ . Это свойство статистики T означает, что она содержит всю информацию о параметре θ , имеющуюся в выборке.

Действительно, вероятность любого события, которое может произойти при фиксированном T , не зависит от θ , следовательно, оно не может нести дополнительную информацию о неизвестном параметре. Сама выборка \bar{X} , очевидно, является достаточной статистикой. Обычно стремятся найти достаточную статистику наименьшей размерности, представляющую исходные данные в наиболее сжатом виде. В этом смысле говорят о **минимальной достаточной статистике**. Такая статистика важна при обработке больших массивов статистической информации. Достаточные статистики находят на основании следующей теоремы.

Теорема 2.4. (Неймана-Фишера, критерий факторизации). Для того чтобы статистика $T(\bar{X})$ была достаточной для θ ,

необходимо и достаточно, чтобы функция правдоподобия $L(\vec{x};\theta)$ имела вид:

$$L(\vec{x};\theta)=g(T(\vec{x});\theta)h(\vec{x}) \quad (2.15)$$

где g - произвольная функция, которая зависит от параметра θ и от выборочных значений \vec{x} через статистику; h -функция выборки, от параметра θ не зависит.

Доказательство

Рассмотрим доказательство теоремы факторизации Неймана-Фишера для дискретной модели.

Если статистика T достаточна при любом t из области значений $T(\vec{x}) : \forall t \in T(\vec{x})$, то функция $L(\vec{x} | t; \theta)$ не зависит от θ и ее можно записать в виде $L(\vec{x} | t; \theta) = h(\vec{x})$.

Пусть $P_\theta(T(\vec{x}) = t) = g(t, \theta)$ и $T(\vec{x}) = t$, тогда событие $\{\vec{X} = \vec{x}\} \subseteq \{T(\vec{x}) = t\}$,

$$L(\vec{x}, \theta) = P_\theta(\vec{X} = \vec{x}) = P_\theta(\vec{X} = \vec{x}; T(\vec{x}) = t) =$$

$$P_\theta(T(\vec{x}) = t) \cdot P_\theta(\vec{X} = \vec{x} | T(\vec{x}) = t) = g(t, \theta) \cdot L(\vec{x} | t; \theta) = g(T(\vec{x}), \theta) \cdot h(\vec{x}),$$

т.е. выполняется представление (2.15). При получении данного выражения использовалась формула условной вероятности.

Верно и обратное. Пусть имеется факторизация (2.15). Тогда при любом \vec{x} , таком что $T(\vec{x}) = t$:

$$L(\vec{x} | t; \theta) = P_\theta(\vec{X} = \vec{x} | T(\vec{x}) = t) = \frac{P_\theta(\vec{X} = \vec{x}; T(\vec{x}) = t)}{P_\theta(T(\vec{x}) = t)}. \quad (2.16)$$

При $A \subseteq X : P_\theta(\vec{X} = \vec{x}; T(\vec{x}) = t) = P_\theta(A \cap \{x : T(\vec{x}) = t\})$

$$= \sum_{\vec{x} \in A, T(\vec{x})=t} g(T(\vec{x}); \theta) \cdot h(\vec{x}) = g(t; \theta) \cdot \sum_{\vec{x} \in A, T(\vec{x})=t} h(\vec{x}) \quad (2.17)$$

В формуле (2.17) используется свойство: вероятность пересечения событий равна сумме вероятностей.

При $A = X$ все P_θ обращаются в нуль на множестве $\{\vec{x} : h(\vec{x}) = 0\}$, исключая эти \vec{x} из рассмотрения и из выражения

(2.17), получаем:

$$P_{\theta}(T(\bar{x}) = t) = g(t, \theta) \cdot \sum_{T(\bar{x})=t} h(\bar{x}) \quad (2.18).$$

Подставив (2.17) и (2.18) в выражение (2.16), получим:

$$L(\bar{x} | t; \theta) = P_{\theta}(\bar{X} = \bar{x} | T(\bar{x}) = t) = \frac{\sum_{\bar{x} \in A, T(\bar{x})=t} h(\bar{x})}{\sum_{T(\bar{x})=t} h(\bar{x})},$$

т.е. статистика $T(\bar{x})$ достаточна, т.к. $L(\bar{x}, \theta)$ не зависит от θ .

Если же \bar{x} таково, что $T(\bar{x}) \neq t$, то очевидно, что $L(\bar{x} | t; \theta) = 0$.

Таким образом, в любом случае, при любом \bar{x} : $\forall \bar{x}$, условная вероятность $L(\bar{x} | t; \theta)$ не зависит от параметра θ .

Доказательство для непрерывной модели аналогично. ■

Теорема 2.5. Если существует эффективная оценка, то существует и достаточная статистика.

Доказательство: Если $\theta_{m \text{ эф}}^* = T(X)$ - эффективная оценка параметра θ , то она удовлетворяет условию Рао-Крамера, т.е.

$$\frac{\partial \ln L(X; \theta)}{\partial \theta} = a(\theta)[T(X) - \theta], \text{ т.е.}$$

$$\ln L(X; \theta) = a(\theta) \int_{-\infty}^{\infty} (T(X) - \theta) d\theta + \varphi(X)$$

$$L(X; \theta) = e^{a(\theta) \int_{-\infty}^{\infty} (T(X) - \theta) d\theta} e^{\varphi(X)}$$

в таком виде представляется функция правдоподобия, а значит выполняется условие факторизации.

Замечание.

Если эффективная оценка не существует, то достаточная статистика может существовать!

Итак, эффективная оценка существует только тогда, когда имеется достаточная статистика. Но достаточная статистика может существовать и при отсутствии эффективной оценки, т.е. условие достаточности является

менее ограничительным, чем условие существования эффективной оценки.

Пример. Случайная величина ξ имеет нормальное распределение $\xi \sim N(m, \sigma)$. Найти достаточную статистику для оценки $M\xi$.

Решение:

$$f(x, \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}$$

Запишем функцию правдоподобия и применим критерий факторизации (2.15)

$$L(\bar{x}; \theta) = \frac{1}{(2\pi\sigma)^{n/2}} \prod_{i=1}^n e^{-\frac{(x_i-\theta)^2}{2\sigma^2}} = A e^{-\frac{\sum x_i^2}{2\sigma^2}} e^{\frac{\theta \sum x_i}{\sigma^2}} e^{-\frac{n\theta^2}{2\sigma^2}}$$

$h(x) = e^{-\frac{\sum x_i^2}{2\sigma^2}}$ - зависит только от выборки;

$g(T(x); \theta) = e^{\frac{\theta \sum x_i}{\sigma^2}} e^{-\frac{n\theta^2}{2\sigma^2}}$ - зависит от выборочного значения через статистику.

$T(X) = \sum_{i=1}^n X_i \rightarrow$ сумма элементов выборки - это статистика, достаточная для оценки математического ожидания.

Задачи и решения

Точечные оценки и их свойства

Точечной оценкой $\tilde{\theta}$ неизвестного параметра θ называется приближенное значение этого параметра, полученное по выборке. Оценка θ есть некоторая функция $\theta = \theta(x_1, \dots, x_n)$. Любую функцию элементов выборки называют статистикой. $\tilde{\theta}$ называется состоятельной оценкой параметра

θ , если она сходится по вероятности к параметру θ , т.е. $\tilde{\theta}_n \xrightarrow[n \rightarrow \infty]{P} \theta$. Иначе это можно записать так $M[\tilde{\theta}_n] \rightarrow \theta$, $D[\tilde{\theta}_n] = 0$.

$T=T(x)$ называется несмещенной оценкой для параметра θ , если выполняются условия $M_\theta[T(x)] = \theta \quad \forall \theta \in \Theta$. Разность $M[\tilde{\theta}] - \theta$ называется смещением. Для несмещенных оценок систематическая ошибка оценивания равна нулю.

Несмещенная оценка с равномерно минимальной дисперсией называется оптимальной оценкой. Для дисперсии несмещенной оценки выполняется неравенство Рао-Крамера

$$D[\tilde{\theta}] \geq \frac{1}{i_n(\theta)},$$

где $i_n(\theta)$ информация Фишера, содержащаяся в выборке объема n относительно неизвестного параметра θ , и вычисляемая по следующим формулам:

для непрерывной случайной величины:

$$i_n(\theta) = nM\left[\left\{\frac{\partial}{\partial \theta} \ln(f(x, \theta))\right\}^2\right],$$

для дискретной случайной величины:

$$i_n(\theta) = nM\left[\left\{\frac{\partial}{\partial \theta} \ln(p(x, \theta))\right\}^2\right],$$

где $p(x, \theta) = P\{X=x\}$.

Эффективной оценкой называется несмещенная оценка $\hat{\theta}_0$ параметра θ , дисперсия которой достигает своего наименьшего возможного значения $D[\hat{\theta}] = \frac{1}{i_n(\theta)}$

Несмещенная оценка $\tilde{\theta} = \tilde{\theta}_n$ называется асимптотически эффективной оценкой параметра θ , если

$$\lim_{n \rightarrow \infty} \frac{1}{i_n(\theta) D[\tilde{\theta}_n]} = 1.$$

Оценка θ называется асимптотически нормально распределенной если

$$\lim_{n \rightarrow \infty} P \left[\frac{\theta - \hat{\theta}}{D[\hat{\theta}_n]} < x \right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{t^2}{2}} dt$$

Простейший метод статистического оценивания – **метод подстановки или аналогии** – состоит в том, что в качестве оценки той или иной числовой характеристики (среднего, дисперсии и др.) генеральной совокупности берут соответствующую характеристику распределения выборки – выборочную характеристику.

Задача 18

Пусть x_1, \dots, x_n выборка из генеральной совокупности с конечным математическим ожиданием m и дисперсией σ^2 . Используя метод подстановки, найти оценку m . Проверить несмещенность и состоятельность полученной оценки.

Решение: По методу подстановки в качестве оценки \tilde{m} математического ожидания надо взять математическое ожидание распределения выборки – выборочное среднее. Таким образом, получаем

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

Чтобы проверить несмещенность и состоятельность выборочного среднего как оценки m , рассмотрим эту статистику как функцию выборочного вектора (x_1, \dots, x_n) . По определению выборочного вектора имеем: $M[X_i] = m$ и $D[X_i] = \sigma^2$, $i=1, \dots, n$, причем X_i – независимые в совокупности случайные величины.

Следовательно,

$$M[\bar{X}] = M \left[\frac{1}{n} \sum_{i=1}^n X_i \right] = \frac{1}{n} \sum_{i=1}^n M[X_i] = \frac{1}{n} nm = m$$

$$D[\bar{X}] = D\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \sum_{i=1}^n D[X_i] = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

Отсюда по определению получаем, что \bar{X} - несмещенная оценка m , и так как $D[\bar{X}] \rightarrow 0$ при $n \rightarrow \infty$, то в силу теоремы является состоятельной оценкой математического ожидания m генеральной совокупности.

Задача 19

Предположим, что выборка x_1, x_2, \dots, x_n получена из генеральной совокупности с конечным математическим ожиданием m и дисперсией σ^2 . Показать, что выборочная дисперсия $D_x^* = \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2$ является смещённой оценкой дисперсии генеральной совокупности, и найти это смещение.

Решение:

$$\begin{aligned} M[D_x^*] &= M\left(\frac{1}{n} \sum_{i=1}^n (x_i - \tilde{x})^2\right) = M\left(\frac{1}{n} \sum_{i=1}^n ((x_i - \mu) - (\tilde{x} - \mu))^2\right) = \\ &= M\left(\frac{1}{n} \sum_{i=1}^n ((x_i - \mu)^2 - 2(x_i - \mu)(\tilde{x} - \mu) + (\tilde{x} - \mu)^2)\right) = \\ &= M\left(\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 - \frac{2}{n} \sum_{i=1}^n (x_i - \mu)(\tilde{x} - \mu) + (\tilde{x} - \mu)^2\right) = \end{aligned}$$

$$= M\left(\frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2 - 2(\tilde{x} - \mu) \frac{1}{n} \sum_{i=1}^n (x_i - \mu) + (\tilde{x} - \mu)^2\right) =$$

$$= \frac{1}{n} \sum_{i=1}^n M(x_i - \mu)^2 - 2M(\tilde{x} - \mu)^2 + M(\tilde{x} - \mu)^2 =$$

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^n M(x_i - \mu)^2 - M(\tilde{x} - \mu)^2 = \\ & = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2 \frac{n-1}{n} \Rightarrow M(\tilde{x} - \mu)^2 = D\tilde{x} = D\left[\frac{1}{n} \sum_{i=1}^n x_i\right] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \end{aligned}$$

Оценка D_x^* является смещённой оценкой, так как $M[D_x^*] \neq \sigma^2$

$\frac{\sigma^2}{n}$ - смещение оценки

Ответ: $\frac{\sigma^2}{n}$

Задача 20

В условиях предыдущей задачи показать, что несмещённая оценка дисперсии генеральной совокупности

задаётся статистикой: $S^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$

Доказательство: $S^2 = \frac{n}{n-1} D_x^*$

$MS^2 = \frac{n}{n-1} M[D_x^*] = \frac{n}{n-1} \cdot \frac{n-1}{n} \sigma^2 = \sigma^2 \Rightarrow S^2$ является

несмещённой оценкой дисперсии генеральной совокупности.

Она задается статистикой

$$S^2 = \frac{n}{n-1} \cdot \frac{1}{n} \sum_{j=1}^n (x_j - \bar{x})^2 = \frac{1}{n-1} \sum_{j=1}^n (x_j - \bar{x})^2$$

Доказано.

Задача 21

Показать, что оценки D_x^* и S^2 , полученные в задачах 19 и 20 соответственно, являются состоятельными оценками дисперсии генеральной совокупности.

Доказательство: По задаче 19

$$D_x^* = \frac{1}{n} \sum_{i=1}^n (x_i - \tilde{x})^2 = \frac{1}{n} \sum_{i=1}^n ((x_i - \mu) - (\tilde{x} - \mu))^2, \mu = m.$$

По теореме Чебышева:

$$\frac{1}{n} \sum_{i=1}^n (x_i - m)^2 \xrightarrow{P} M[(x_i - m)^2] = \sigma^2, n \rightarrow \infty$$

$$\tilde{x} \xrightarrow{P} m, n \rightarrow \infty \text{ (неравенство Чебышева)}$$

$$\Rightarrow (\tilde{x} - m)^2 \xrightarrow{P} 0, n \rightarrow \infty$$

Т.о. $D_x^* \xrightarrow{P} \sigma^2, n \rightarrow \infty$, т.е. D_x^* является состоятельной оценкой дисперсии.

По задаче 20:

$$S^2 = \frac{n}{n-1} D_x^*,$$

$$S^2 \xrightarrow{P} \sigma^2, n \rightarrow \infty, \text{ т.к. } D_x^* \xrightarrow{P} \sigma^2, n \rightarrow \infty,$$

$$P\left\{ \left| \frac{n}{n-1} - 1 \right| > \varepsilon \right\} = P\left\{ \left| \frac{1}{n-1} \right| > \varepsilon \right\} \xrightarrow{n \rightarrow \infty} 0 \Rightarrow \frac{n}{n-1} \xrightarrow{P} 1, n \rightarrow \infty.$$

Доказано.

Задача 22

Пусть x_1, x_2, \dots, x_n выборка из генеральной совокупности с известным средним m и неизвестной дисперсией σ^2 . Показать, что несмещённой оценкой σ^2 будет статистика

$$\sigma_0^2 = \frac{1}{n} \sum_{i=1}^n (x_i - m)^2$$

Доказательство:

$$1) M\sigma_0^2 = \frac{1}{n} \sum_{i=1}^n M(x_i - m)^2 = \sigma^2$$

$$2) M\sigma_0^2 = M\left(\frac{1}{N} \sum_{i=1}^n x_i^2 - \frac{2m}{n} \sum_{i=1}^n x_i + \frac{nm^2}{n}\right) = \frac{1}{n} \sum_{i=1}^n Mx_i^2 - \frac{2m}{n} \sum_{i=1}^n Mx_i + m^2 = \\ = Mx_i^2 - 2m^2 + m^2 = Mx_i^2 - m^2 = Dx_i = \sigma^2$$

Таким образом, σ_0^2 является несмещённой оценкой σ^2 .

Доказано.

Задача 23

Пусть $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ - выборка из двумерной генеральной совокупности. Методом подстановки найти оценку ковариации. Показать, что получаемая оценка является смещённой и состоятельной. Найти несмещённую оценку

Решение:

$$\bar{k}_{xy} = k_{xy}^* = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}),$$

$$\tilde{k}_{xy} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - \text{несмещённая оценка,}$$

$$x_i - \bar{x} = (x_i - m_x) + \frac{1}{n} \sum_{i=1}^n (x_i - m_x) = \dot{x}_i + \frac{1}{n} \sum_{i=1}^n \dot{x}_i.$$

$$\text{Аналогично } y_i - \bar{y} = \dot{y}_i + \frac{1}{n} \sum_{i=1}^n \dot{y}_i.$$

$$\text{Тогда } M[(x_i - \bar{x})(y_i - \bar{y})] = k_{xy} - \frac{2}{n} k_{xy} + \frac{1}{n} k_{xy} = \frac{n-1}{n} k_{xy},$$

так как $k_{x_i y_i} = 0$ при $i \neq j$, $k_{x_i y_i} = k_{xy}$

$$D[k_{xy}^*] = \frac{1}{n^2} \sum_{i=1}^n D[(x_i - \bar{x})(y_i - \bar{y})] \xrightarrow{n \rightarrow \infty} 0$$

Доказано.

Задача 24

Пусть $\tilde{\theta}$ - несмещённая оценка параметра θ , $D[\tilde{\theta}] < \infty$.
Показать, что $\tilde{\theta}^2$ является смещённой оценкой θ^2 , и вычислить смещение.

Решение:

$M[\tilde{\theta}^2] = D[\tilde{\theta}] + M^2[\tilde{\theta}] = D[\tilde{\theta}] + \theta^2$ - смещённая оценка θ^2 ,
смещение $\tilde{\theta}^2$ равно $D[\tilde{\theta}]$.

Доказано.

Задача 25

Показать, что выборочное среднее, вычисленное по выборке из генеральной совокупности, имеющей распределение Пуассона с параметром λ , будет несмещённой и состоятельной оценкой этого параметра.

Доказательство: $\xi \sim p_0(\lambda)$

$D\xi = \lambda, M\xi = \lambda, \lambda^* = \bar{x}, M\lambda^* = \frac{1}{n} \sum_{i=1}^n M\xi_i = \lambda$ - оценка является

несмещённой,

$$D\lambda^* = D\left(\frac{1}{n} \sum_{i=1}^n \xi_i\right) = \frac{1}{n^2} D\sum_{i=1}^n \xi_i = \frac{1}{n^2} \sum_{i=1}^n D\xi_i = \frac{\lambda}{n} \xrightarrow{n \rightarrow \infty} 0$$

оценка является состоятельной.

Доказано.

Задача 26

Показать, что выборочное среднее является эффективной оценкой параметра λ распределения Пуассона.

Доказательство: по задаче 25 выборочное среднее является несмещённой и состоятельной оценкой.

$$\lambda^* = \bar{x} = \frac{1}{n} \sum_{i=1}^n \xi_i,$$

$$f(\xi, \theta) = P\{\eta = \xi\} = \frac{\lambda^\xi}{\xi!} e^{-\lambda},$$

$$\ln f(\xi, \theta) = \xi \ln \lambda - \ln \xi! - \lambda,$$

$$\frac{\partial \ln f(\xi, \theta)}{\partial \lambda} = \frac{\xi}{\lambda} - 1 = \frac{\xi - \lambda}{\lambda},$$

$$M\left\{\left(\frac{\partial \ln f(\xi, \theta)}{\partial \lambda}\right)^2\right\} = M\left(\frac{\xi - \lambda}{\lambda}\right)^2 = \frac{1}{\lambda^2} M(\xi - \lambda)^2 = \frac{1}{\lambda^2} \lambda = \frac{1}{\lambda},$$

$$\hat{\sigma}^2 = \frac{1}{nM\left(\frac{\partial \ln f(\xi, \lambda)}{\partial \lambda}\right)^2} = \frac{\lambda}{n}.$$

Так как $D\lambda^* = \hat{\sigma}^2$, то выборочное среднее является эффективной оценкой
Доказано.

Задача 27

Пусть x_1, x_2, \dots, x_n - выборка из нормального распределения генеральной совокупности $N(m, \sigma)$. Найти информацию Фишера $In(\sigma^2)$.

Решение: $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad Dx = \sigma^2.$

$$\ln f(x) = \ln \frac{1}{\sqrt{2\pi}\sigma} + \ln e^{-\frac{(x-m)^2}{2\sigma^2}} = -\ln \sqrt{2\pi}\sigma - \frac{(x-m)^2}{2\sigma^2} = -\ln \sqrt{2\pi} - \ln \sigma - \frac{(x-m)^2}{2\sigma^2},$$

$$I(\theta) = M_{\theta} \left(\frac{\partial \ln f(x_i, \theta)}{\partial \theta} \right)^2,$$

$$\ln(\theta) = nI(\theta),$$

$$\frac{\partial \ln f(x)}{\partial \sigma^2} = \left(-\ln \sqrt{\sigma^2} - \frac{(x-m)^2}{2\sigma^2} \right)_{\sigma^2} = -\frac{1}{2\sigma^2} + \frac{(x-m)^2}{2\sigma^4},$$

$$\frac{\partial^2 \ln f(x)}{\partial (\sigma^2)^2} = \frac{1}{2\sigma^4} - \frac{(x-m)^2}{\sigma^6},$$

$$I(\sigma^2) = -M \left[\frac{\partial^2 \ln f(x)}{\partial (\sigma^2)^2} \right] = -M \left[\frac{1}{2\sigma^4} - \frac{(x-m)^2}{\sigma^6} \right] = -\frac{1}{2\sigma^4} + \frac{\sigma^2}{\sigma^6} = \frac{1}{2\sigma^4},$$

$$\ln(\sigma^2) = \frac{n}{2\sigma^4}.$$

$$\text{Ответ: } \frac{n}{2\sigma^4}.$$

Задача 28

В условиях предыдущей задачи при известном математическом ожидании m оценивается дисперсия σ^2 . Показать, что статистика

$$S_0^2 = \frac{1}{n} \sum (x_i - m)^2 \text{ является эффективной оценкой } \sigma^2.$$

Решение:

$$M[S_0^2] = \frac{1}{n} M \sum (x_i - m)^2 = \sigma^2 \Rightarrow \text{оценка является}$$

несмещённой.

$$D[S_0^2] = M[S_0^2]^2 - M^2[S_0^2](=)$$

$$\begin{aligned} M[S_0^2]^2 &= M\left[\frac{1}{n}(\sum (x_i - m)^2)\right]^2 = M\left[\frac{1}{n^2}(\sum (x_i - m)^4 + \sum_{i \neq j} (x_i - m)^2(x_j - m)^2)\right] = \\ &= \frac{1}{n^2}[\sum M(x_i - m)^4 + \sum_{i \neq j} M(x_i - m)^2 M(x_j - m)^2] = \frac{1}{n^2}(3n\sigma^4 + (n-1)n\sigma^4) = \frac{2\sigma^4}{n} + \sigma^4 \\ (=) \frac{2\sigma^4}{n} + \sigma^4 - \sigma^4 &= \frac{2\sigma^4}{n}. \end{aligned}$$

$$\widehat{S}_0^2 = \frac{2\sigma^4}{n}$$

$$\widehat{S}_0^2 = D[S_0^2] \Rightarrow \text{оценка является эффективной}$$

Доказано.

Лабораторная работа № 2

Целью лабораторной работы является получение точечных оценок параметров распределений в пакете MATHCAD.

Точечная оценка математического ожидания

Доказано, что эффективной оценкой математического ожидания нормально распределенной случайной величины является оценка $\hat{\theta}_n = (x_1 + x_2 + \dots + x_n)/n$. Именно поэтому последняя оценка так широко используется в математической статистике. Для оценки неизвестного математического ожидания случайной величины будем использовать

$$\text{выборочное среднее, т. е. } \hat{\theta}_n = \bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n}.$$

Точечные оценки дисперсии

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

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

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

В качестве состоятельной несмещенной оценки дисперсии используют величину

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

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

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

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

Задание

Найдите состоятельные несмещенные оценки математического ожидания MX и дисперсии DX случайной величины X по приведенным в задании выборочным значениям x_1, x_2, \dots, x_n .

Порядок выполнения задания

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

Пример выполнения задания

Найдите состоятельные несмещенные оценки математического ожидания MX и дисперсии DX случайной величины X по выборочным значениям, заданным следующей таблицей.

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

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

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.

$$\begin{array}{l} \text{ORIGIN} := 1 \\ i := 1..12 \\ D := \text{stack}(D1, D2) \end{array} \quad D1 := \begin{bmatrix} 1 & 904.3 \\ 3 & 910.2 \\ 1 & 916.6 \\ 1 & 928.0 \\ 1 & 935.0 \\ 2 & 941.2 \end{bmatrix} \quad D2 := \begin{bmatrix} 1 & 947.4 \\ 1 & 953.6 \\ 1 & 959.8 \\ 1 & 966.0 \\ 2 & 972.2 \\ 1 & 978.4 \end{bmatrix}$$

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

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

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

$$Dx1 = \frac{1}{n} \sum_{i=1}^{12} D_{i,1} (D_{i,2} - Mx)^2 \quad Dx1 = 593.26$$

3. МЕТОДЫ ПОЛУЧЕНИЯ ТОЧЕЧНЫХ ОЦЕНОК

3.1. Метод максимального правдоподобия (ММП)

Одним из универсальных методов оценивания параметров распределения является **метод максимального правдоподобия**. Оценку параметра θ , получаемого с помощью этого метода, будем обозначать $\hat{\theta} = \hat{\theta}(\bar{X})$, а оценку параметрической функции $\hat{\tau} = \hat{\tau}(\bar{X})$.

Пусть задана выборка $\bar{X}=(X_1, \dots, X_n)$ из распределения $L(\xi) \in \mathbb{F} = \{F(x; \theta); \theta \in \Theta\}$, и $L(\bar{x}; \theta)$ функция правдоподобия для реализации $\bar{x}=(x_1, \dots, x_n)$ выборки \bar{X} .

По определению оценкой максимального правдоподобия (о.м.п.) $\hat{\theta}$ параметра θ называется такая точка параметрического множества Θ , в которой функция правдоподобия $L(\bar{x}; \theta)$ при заданном \bar{x} достигает максимума. Таким образом

$$L(\bar{x}; \hat{\theta}) \geq L(\bar{x}; \theta), \quad \forall \theta \quad \text{или} \\ L(\bar{x}; \hat{\theta}) = \sup_{\theta \in \Theta} L(\bar{x}; \theta).$$

Замечание.

Если $L(\bar{x}; \theta_1) > L(\bar{x}; \theta_2)$, то говорят, что значение параметра θ_1 более правдоподобно, чем θ_2 . Таким образом, оценка максимального правдоподобия $\hat{\theta}$ является наиболее правдоподобным значением параметра θ .

Если для каждого \bar{x} из выборочного пространства \mathcal{X} максимум $L(\bar{x}; \theta)$ достигается во внутренней точке Θ и $L(\bar{x}; \theta)$ дифференцируема по θ , то о.м.п. $\hat{\theta}$ удовлетворяет уравнению

$$\frac{\partial L(\bar{x}; \theta)}{\partial \theta} = 0 \quad \text{или} \quad \frac{\partial \ln L(\bar{x}; \theta)}{\partial \theta} = 0.$$

Если θ векторный параметр: $\bar{\theta}=(\theta_1, \dots, \theta_r)$, то это уравнение заменяется системой уравнений

$$\frac{\partial \ln L(\bar{x}; \theta)}{\partial \theta_i} = 0, \quad i=1, \dots, r.$$

Последние уравнения называются **уравнениями правдоподобия**.

3.2. Свойства оценок максимального правдоподобия

1. Эффективность.

Теорема 3.1. Если существует эффективная оценка $T(\bar{X})$ для скалярного параметра θ , то $\hat{\theta} = T(\bar{X})$.

Доказательство: Это очевидное следствие критерия эффективности Рао-Крамера

$$\frac{\partial \ln L(\bar{x}; \theta)}{\partial \theta} = \frac{1}{a(\theta)} [T(\bar{X}) - \theta].$$

Приравняем к 0 и получим $\hat{\theta} = T(\bar{X})$.

2. Достаточность.

Теорема 3.2. Если имеется достаточная статистика $T=T(\bar{X})$, а о.м.п. существует и единственна, то она является функцией от достаточной статистики T .

Доказательство: Согласно критерию факторизации справедливо разложение:

$$L(\bar{x}; \theta) = g(T(\bar{x}); \theta) h(\bar{x})$$

$$\frac{\partial \ln L(\bar{x}; \theta)}{\partial \theta} = \frac{\partial \ln g(T(\bar{x}); \theta)}{\partial \theta} = 0.$$

Решаем уравнение относительно θ .

Получаем, $\hat{\theta} = \varphi(T(\bar{x}))$ – некоторая функция статистики, а это есть оценка МП, что и требовалось доказать.

Следовательно, $\hat{\theta}$ зависит от статистических данных через $T(\bar{x})$.

3. Инвариантность.

Полезным свойством оценок максимального правдоподобия (МП) является их инвариантность относительно преобразований параметра.

При решении уравнений правдоподобия относительно параметра θ оказывается, что их проще решать относительно функций от него, например, $\ln \theta$, $\frac{1}{\theta}$ и т.д. Обозначим эту функцию через φ и допустим, что $\varphi = t(\theta)$ – взаимно однозначная дифференцируемая функция, т.е. $\frac{d\varphi}{d\theta} \neq 0$. Тогда, если через $\hat{\theta}$ и $\hat{\varphi}$ обозначить оценки максимального правдоподобия параметров θ и φ , то $\hat{\varphi} = t(\hat{\theta})$.

Доказательство. Действительно, для регулярной модели функция правдоподобия относительно φ записывается так:

$$\lambda(\varphi) = \lambda(t(\theta)) = L(\theta), \quad \text{откуда} \quad \frac{d\lambda}{d\varphi} = \frac{dL}{d\theta} \cdot \frac{d\theta}{d\varphi}. \quad \text{Оценка}$$

максимального правдоподобия определяется как корень уравнения $\frac{d\lambda}{d\varphi} = 0$. Но $\frac{dL}{d\theta} \cdot \frac{d\theta}{d\varphi} = 0$ при $\varphi = \hat{\varphi}$, т.е. когда

$t(\theta) = \hat{\varphi}$. Поскольку по условию $\frac{d\theta}{d\varphi} \neq 0$, то последнее

уравнение эквивалентно $\frac{dL}{d\theta} = 0$, откуда следует, что

$$\hat{\varphi} = t(\hat{\theta}) \quad \blacksquare$$

3.3. Теорема об асимптотической нормальности и эффективности оценок максимального правдоподобия

1. Оценка максимального правдоподобия $\hat{\theta} \equiv \theta_m^*$ является состоятельной оценкой параметра θ ; т.е. $\hat{\theta} \equiv \theta_m^* \xrightarrow[n \rightarrow \infty]{P} \theta$.

2. При определённых условиях оценка максимального правдоподобия является асимптотически нормальной и эффективной.

Теорема 3.3.

Пусть функция правдоподобия $L(x; \theta)$

а) дважды дифференцируема по параметру θ и

б) математическое ожидание от функции вклада равно нулю $M[U(X; \theta)] = 0$,

в) кроме того $-M \frac{\partial^2 \ln L(X, \theta)}{\partial \theta^2} = i_n(\theta) \equiv R^2(\theta) \neq 0$.

Тогда оценка максимального правдоподобия стремится к случайной величине

$$\theta_m^* \xrightarrow[n \rightarrow \infty]{} \gamma \sim N \left(0; \frac{1}{|R(\theta_0)|} \right)$$

(дисперсия совпадает с дисперсией эффективной оценки).
Здесь θ_0 - истинное значение оцениваемого параметра.

Доказательство: Доказательство свойства асимптотической нормальности оценки МП (если рассматривать скалярный параметр) основывается на разложении функции вклада $U_n(\theta) = U_n(\bar{X}; \theta)$ в ряд Маклорена относительно истинного значения параметра θ_0 .

Поскольку θ_m^* состоятельная оценка параметра θ , то при достаточно большом объёме выборки ($n \gg 1$), она будет близка к истинному значению θ_0 . Поэтому функция вклада может быть представлена в виде ряда Маклорена в окрестности точки θ_0 .

$$\left. \frac{\partial \ln L(X; \theta)}{\partial \theta} \right|_{\theta_m^*} = \left. \frac{\partial \ln L(X; \theta)}{\partial \theta} \right|_{\theta_0} + (\theta_m^* - \theta_0) \left. \frac{\partial^2 \ln L(X; \theta)}{\partial \theta^2} \right|_{\tilde{\theta}},$$

где $\tilde{\theta} \in (\theta_m^*; \theta_0)$

В силу состоятельности оценки и условий теоремы первая дробь равна 0. Поэтому

$$(\theta_m^* - \theta_0) = \frac{\left. \frac{\partial \ln L(X; \theta)}{\partial \theta} \right|_{\theta_0}}{- \left. \frac{\partial^2 \ln L(X; \theta)}{\partial \theta^2} \right|_{\tilde{\theta}}}.$$

Левую и правую часть умножим на $R(\theta_0)$

$$R(\theta_0)(\theta_m^* - \theta_0) = \frac{\frac{1}{R(\theta_0)} \left. \frac{\partial \ln L(X; \theta)}{\partial \theta} \right|_{\theta_0}}{- \frac{1}{R^2(\theta_0)} \left. \frac{\partial^2 \ln L(X; \theta)}{\partial \theta^2} \right|_{\tilde{\theta}}}.$$

Вклад выборки определяется по формуле

$$U(X; \theta) = \frac{\partial \ln L(X; \theta)}{\partial \theta} = \sum_{i=1}^n \frac{\partial \ln f_{\xi}(X_i; \theta)}{\partial \theta}.$$

Рассмотрим знаменатель дроби:

$$- \frac{1}{R^2(\theta_0)} \left. \frac{\partial^2 \ln L(X; \theta)}{\partial \theta^2} \right|_{\tilde{\theta}} = - \frac{1}{R^2(\theta_0)} \sum_{i=1}^n \left. \frac{\partial^2 \ln f_{\xi}(X_i; \theta)}{\partial \theta^2} \right|_{\tilde{\theta}} \rightarrow$$

в силу закона больших чисел, если элементы выборки независимы

$$\rightarrow - \frac{1}{R^2(\theta_0)} \sum_{i=1}^n M \left. \frac{\partial^2 \ln f_{\xi}(X_i; \theta)}{\partial \theta^2} \right|_{\tilde{\theta}} = - \frac{1}{R^2(\theta_0)} M \left. \frac{\partial^2 \ln L(X; \theta)}{\partial \theta^2} \right|_{\tilde{\theta}} = \frac{R^2(\theta_0)}{R^2(\theta_0)} \rightarrow 1$$

$n \rightarrow \infty$ в виду состоятельности оценки.

Таким образом, знаменатель дроби стремится к 1.

Рассмотрим числитель дроби.

К случайной величине

$$\frac{1}{R(\theta_0)} \cdot \frac{\partial \ln L(X; \theta)}{\partial \theta} \Big|_{\theta_0} = \frac{1}{R(\theta_0)} \cdot \sum_{i=1}^n \frac{\partial \ln f(X_i; \theta)}{\partial \theta} \Big|_{\theta_0} \rightarrow$$

применима центральная предельная теорема, по которой и с учётом соотношений:

$$M \frac{\partial \ln L(X; \theta)}{\partial \theta} = 0 \quad \forall \theta \in \Theta,$$

$$i(\theta) = M \left(\frac{\partial \ln f(X; \theta)}{\partial \theta} \right)^2 \quad \text{при } n \rightarrow \infty$$

$$R(\theta_0)(\theta_m^* - \theta_0) \rightarrow \eta \sim N(0, 1).$$

Сама оценка $\theta_m^* \rightarrow \frac{1}{R(\theta)} \eta + \theta_0 = \gamma$, так как γ – линейная функция

η . ■

3.4. Метод моментов. Теоремы о свойствах оценок, полученных методом моментов

Исторически первым методом точечного оценивания неизвестных параметров является **метод моментов**, предложенный К. Пирсоном в 1894 году.

Суть метода в следующем. Пусть $\vec{X} = (X_1, \dots, X_n)$ – выборка из распределения $L(\xi) \in \mathcal{F} = \{F(x; \vec{\theta}); \vec{\theta} \in \Theta\}$, где $\vec{\theta} = (\theta_1, \dots, \theta_r)$ и $\Theta \subseteq \mathbb{R}^r$. Предположим, что у наблюдаемой случайной величины ξ существуют первые r моментов $\alpha_k = M \xi^k$, $k=1, \dots, r$. Они являются функциями от неизвестных параметров $\vec{\theta}$: $\alpha_k = \alpha_k(\vec{\theta})$. Рассмотрим соответствующие выборочные моменты $A_{nk}(\vec{X})$.

Пусть $\alpha_k = A_{nk}(\vec{x})$ – значения этих величин для наблюдавшейся реализации \vec{x} выборки \vec{X} . Тогда метод моментов состоит в приравнивании значений α_k и теоретических моментов:

$$\alpha_k(\bar{\theta}) = \alpha_k, \quad k=1, \dots, r \quad (3.1)$$

Решая эти уравнения относительно $\theta_1, \dots, \theta_r$, получаем значения оценок параметров.

Замечания.

1). Число уравнений в системе (3.1) должно совпадать с числом неизвестных параметров.

2). В системе уравнений (3.1) могут одновременно присутствовать уравнения как для начальных, так и для центральных моментов.

Рассмотрим теоретическое обоснование этого метода:

Теорема 3.4. Известно, что выборочные моменты $A_{nk}(\bar{X})$ являются несмещёнными и состоятельными оценками теоретических моментов $\alpha_k(\bar{\theta})$.

Доказательство: Проверим выполнение достаточного условия состоятельности:

$$M[A_{nk}(\bar{X})] = \frac{1}{n} \sum_{i=1}^n MX_i^k = \alpha_k$$

$$D[A_{nk}(\bar{X})] = \frac{1}{n^2} \sum_{i=1}^n DX_i^k = \frac{1}{n^2} \sum_{i=1}^n (MX_i^{2k} - (MX_i^k)^2) = \frac{\alpha_{2k} - \alpha_k^2}{n} \rightarrow 0,$$

$n \rightarrow \infty$, т.е. условие состоятельности выполнено.

Теорема 3.5. Если существует взаимно однозначное и взаимно непрерывное соответствие между параметрами $\theta_1, \dots, \theta_r$ и начальными моментами $\alpha_1, \dots, \alpha_r$, т.е. существуют непрерывные функции $\varphi_1, \dots, \varphi_r$ такие, что $\theta_i = \varphi_i(\alpha_1, \dots, \alpha_r)$, $i=1, \dots, r$. Тогда решения уравнений (31) можно записать в виде $\tilde{\theta}_i(\bar{x}) = \varphi_i(a_1, \dots, a_r)$, $i = \overline{1, r}$, а оценки $\tilde{\theta}_i(\bar{X}) = \varphi_i(A_{n1}(\bar{X}), \dots, A_{nr}(\bar{X}))$ являются состоятельными оценками соответствующих параметров.

Доказательство: В силу теоремы Слуцкого оценки метода моментов будут сходиться по вероятности к оцениваемому параметру при $n \rightarrow \infty$, т.е. статистики $\tilde{\theta}_i(\bar{X})$ являются состоятельными оценками θ_i , $i=1, \dots, r$. ■

Таким образом, метод моментов при определённых условиях приводит к состоятельным оценкам; при этом уравнения (3.1) во многих случаях просты и их решение (в отличие от метода МП) не связано с большими вычислительными трудностями.

Когда теоретические моменты нужного порядка отсутствуют (например, распределение Коши), метод моментов неприменим. Оценки метода моментов, вообще говоря, не эффективны. Их обычно используют в качестве первых приближений, на основании которых можно определять другими методами оценки с большей эффективностью..

Пример. (Модель гамма, оценивание параметров методом моментов)

Имеется выборка $\bar{X}=(X_1, \dots, X_n)$ из генеральной совокупности случайной величины ξ с плотностью распределения

$$f(x, \theta) = \frac{x^{\lambda-1} e^{-x/\theta}}{\Gamma(\lambda)\theta^\lambda}, \quad x > 0, \quad \{\theta: 0 < \theta < \infty\}$$

Рассмотрим модель гамма $\Gamma(\theta_1, \theta_2)$, когда оба параметра неизвестны. Здесь $\Theta = \{ \bar{\theta} = (\theta_1, \theta_2): \theta_i > 0, i=1, 2 \}$

Решение: Рассчитаем теоретические начальные моменты

$$\alpha_k = \int_0^\infty \frac{x^{\theta_2-1+k} e^{-x/\theta_1}}{\Gamma(\theta_2)\theta_1^\lambda} dx = \theta_1^k \frac{\Gamma(\theta_2 + k)}{\Gamma(\theta_2)} = \theta_1^k \theta_2 (\theta_2 + 1) \dots (\theta_2 + k - 1)$$

Для оценки двух параметров достаточно α_1 и α_2

$$\alpha_1 = \theta_1 \theta_2; \quad \alpha_2 = \theta_1^2 \theta_2 (\theta_2 + 1);$$

Выборочные моменты

$$A_{n1} = \frac{1}{n} \sum_{i=1}^n X_i; \quad A_{n2} = \frac{1}{n^2} \sum_{i=1}^n X_i^2.$$

Приравняем:

$$\begin{cases} A_{n1} = \theta_1 \theta_2 \\ A_{n2} = \theta_1^2 \theta_2 (\theta_2 + 1) \end{cases}.$$

Решаем систему относительно θ_i :

$$\theta_1 = \frac{A_{n2} - A_{n1}^2}{A_{n1}},$$

$$\theta_2 = \frac{A_{n1}^2}{A_{n2} - A_{n1}^2}.$$

Отсюда оценки параметров:

$$\tilde{\theta}_1(\bar{X}) = \frac{A_{n2} - A_{n1}^2}{A_{n1}} = \frac{S^2}{\bar{X}},$$

$$\tilde{\theta}_2(\bar{X}) = \frac{A_{n1}^2}{A_{n2} - A_{n1}^2} = \frac{\bar{X}^2}{S^2}$$

3.5. Цензурирование

Рассматривались методы оценивания, использующие информацию, доставляемую полной выборкой $\bar{X} = (X_1, \dots, X_n)$. Иногда возникают задачи оценивания по неполной выборке, т.е. когда некоторые наблюдения отсутствуют. В таких случаях говорят о **цензурированных** данных.

Типичными примерами цензурирования являются следующие планы испытаний на надёжность: берётся контрольная выборка из n изделий, "времена жизни" которых независимые одинаково распределённые случайные величины. Наблюдаются либо моменты отказов за заданное время t (1-й тип цензурирования), либо моменты первых r отказов, где $r < n$ (2-й тип цензурирования).

В обоих случаях достигается экономия времени на проведение эксперимента (получение исходных данных), что является важным фактором в реальных условиях.

1-й тип цензурирования определяется заданием такого интервала (t_1, t_2) , что наблюдаются лишь значения $X_i \in (t_1, t_2)$.

2-й тип определяется заданием двух целых чисел $r_1, r_2 \geq 0$ таких, что наблюдаются лишь значения порядковых статистик $X_{(k)}$ при $r_1 + 1 \leq k \leq n - r_2$.

Если соответствующее ограничение с какой-нибудь одной стороны отсутствует, то говорят о **простом** цензурировании.

Рассмотренные методы оценки параметров могут быть применены к цензурированным данным.

Задачи и решения

Методы получения точечных оценок

Метод максимального правдоподобия

Пусть X непрерывная случайная величина с плотностью распределения $f(x, \theta)$ зависящая от неизвестного параметра θ , значение которого и требуется оценить по выборке объема n .

Плотность распределения \bar{X} :

$$f_x(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f_{x_i}(x_i; \theta) \quad (3.2)$$

Функцией правдоподобия $L(\theta)$ выборки объема n называется плотность выборочного вектора (3.2), которая рассматривается при фиксированных значениях \bar{x} , т.е. функция правдоподобия – это функция неизвестного параметра θ

$$L(\theta) = \prod_{i=1}^n f_{x_i}(x_i; \theta).$$

Пусть X – дискретная случайная величина, причем $P(X=x) = p = p(x, \theta)$ есть функция неизвестного параметра θ . Для оценки параметра θ получена конкретная выборка x_1, \dots, x_n объема n . Функция правдоподобия выборки объема n равна

$$L(\theta) = \prod_{i=1}^n p(x_i; \theta) = \prod_{i=1}^n p(X_i = x_i).$$

вероятности того, что компоненты выборочного вектора \bar{X} примут фиксированные значения \bar{x} , т.е.

Метод максимального правдоподобия заключается в том, что в качестве оценки неизвестного параметра θ выбирается значение $\hat{\theta}$, достигающее максимума функции правдоподобия. Такую оценку называют максимально правдоподобной. В дискретном случае максимально правдоподобная оценка неизвестного параметра θ есть такое значение θ^* , при котором вероятность появления данной конкретной выборки максимальна. Для упрощения вычислений удобно использовать $\ln L(\theta)$.

Максимально правдоподобные оценки состоятельны, асимптотически эффективны и асимптотически нормально распределены. Если для параметра θ существует эффективная оценка, то метод максимального правдоподобия дает именно эту оценку.

Задача 29

Найти МП-оценки математического ожидания m и дисперсии σ^2 нормально распределённой генеральной совокупности.

Решение: Пусть x_1, \dots, x_n - выборка наблюдений случайной величины X с плотностью распределения

$$f_X(x, m, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-m)^2}{2\sigma^2}}.$$

Найдём функцию правдоподобия $L(m, \sigma^2)$

$$L(m, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x_i-m)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{\sum (x_i-m)^2}{2\sigma^2}}.$$

Логарифмическая функция правдоподобия равна

$$\ln L(m, \sigma^2) = -\frac{n}{2} \ln 2\pi - \frac{n}{2} \ln \sigma^2 - \frac{\sum (x_i - m)^2}{2\sigma^2}.$$

Используя условия максимума $\ln L(m, \sigma^2)$ получим систему уравнений для нахождения МП-оценок:

$$\begin{cases} \frac{\partial \ln L(m, \sigma^2)}{\partial m} = \frac{1}{\sigma^2} \sum (x_i - m) = 0 \\ \frac{\partial \ln L(m, \sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (x_i - m)^2 = 0 \end{cases}$$

Из первого уравнения системы находим $\hat{m} = \frac{1}{n} \sum x_i = \bar{x}$

Подставляя это значение во второе уравнение, получаем

$$\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = \hat{D}_X^*$$

Итак, выборочное среднее \bar{x} является несмещённой и состоятельной оценкой m , а также эффективной оценкой в случае нормального распределённой генеральной совокупности. Выборочная дисперсия \hat{D}_X^* является состоятельной и смещённой оценкой σ^2 .

Задача 30

Найти МП - оценку параметра λ распределения Пуассона.

Решение: Пусть x_1, \dots, x_n - выборка, наблюдений случайной величины X , имеющей распределение Пуассона с неизвестным параметром λ , т.е. $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$, $x=0,1,2,\dots$

Функция правдоподобия $L(\lambda)$ определяется так:

$$L(\lambda) = \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} e^{-\lambda} = \frac{\lambda^{\sum x_i}}{x_1! x_2! \dots x_n!} e^{-\lambda n}$$

Найдём логарифмическую функцию правдоподобия:

$$\ln L(\lambda) = -\ln(x_1! \dots x_n!) + \sum (x_i) \ln \lambda - \lambda n$$

Получим уравнение для определения МП - оценки:

$$\frac{d \ln L(\lambda)}{d \lambda} = \frac{\sum x_i}{\lambda} - n = 0,$$

откуда находим, что $\hat{\lambda} = \frac{1}{n} \sum x_i = \bar{x}$. Полученная МП - оценка является несмещённой, состоятельной и эффективной оценкой параметра λ .

Задача 31

Найти МП - оценку параметра σ по выборке объема n из нормального распределения генеральной совокупности с известным математическим ожиданием m . Показать, что полученная оценка является смещенной.

Решение:

$\xi \sim N(m, \sigma^2)$, m - неизвестны

$$\alpha = n f(\xi, m, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\xi_i - m)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma^2}\right)^n \exp\left(-\frac{\sum (\xi_i - m)^2}{2\sigma^2}\right)$$

$$\ln \alpha = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \left(\sum (\xi_i - m)^2\right);$$

$$\ln \alpha|_{\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum (\xi_i - m)^2;$$

$$\frac{n}{2\sigma^2} = \frac{1}{2\sigma^4} \sum (\xi_i - m)^2 \Rightarrow \sigma^2 = \hat{\rho}^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - m)^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - x)^2$$

$$M \hat{\rho}^2 = M \left(\frac{1}{n} \sum (\xi_i - \bar{x})^2 \right) = M \left(\frac{1}{n} \sum_{i=1}^n (\xi_i - m)^2 - (\bar{x} - m)^2 \right) =$$

$$= M \left(\frac{1}{n} \sum_{i=1}^n (\xi_i - m)^2 + (\bar{x} - m)^2 - n(\xi_i - m)(\bar{x} - m) \right) =$$

$$= M \left(\frac{1}{n} \sum_{i=1}^n (\xi_i - m)^2 - \frac{2}{n} (\bar{x} - m) \sum_{i=1}^n (\xi_i - m) + (\bar{x} - m)^2 \right) =$$

$$= \frac{1}{n} \sum M (\xi_i - m)^2 - 2M(\bar{x} - m) \left(\frac{1}{n} \sum (\xi_i - m) \right) + M(\bar{x} - m)^2 = \frac{1}{n} \sum \sigma^2 - D\bar{x} = \sigma^2 \left(\frac{n-1}{n} \right)$$

Задача 32

По выборке x_1, x_2, \dots, x_n объема n найти МП - оценки параметров указанного распределения. Показать, что полученная оценка является несмещенной, состоятельной и эффективной.

Показательное распределение $Ex(1/\lambda)$.

Решение:

$$\xi \sim E_x\left(\frac{1}{\lambda}\right)$$

$$f(x) = \frac{1}{\lambda} \exp\left(-\frac{x}{\lambda}\right) \int_{[0, +\infty)}(x)$$

$$L(\bar{\xi}, \lambda) = \prod_{i=1}^n f(\xi_i, \lambda) = \prod_{i=1}^n \frac{1}{\lambda} \exp\left(-\frac{\xi_i}{\lambda}\right) = \frac{1}{\lambda^n} \exp\left(-\frac{1}{\lambda} \sum_{i=1}^n \xi_i\right)$$

$$\ln L = -n \ln \lambda - \frac{1}{\lambda} \sum_{i=1}^n \xi_i$$

$$\ln L'|_n = -\frac{n}{\lambda} + \frac{1}{\lambda^2} \sum_{i=1}^n \xi_i$$

$$\frac{n}{\lambda} = \frac{1}{\lambda^2} \sum_{i=1}^n \xi_i$$

$$\lambda^* = \frac{1}{n} \sum_{i=1}^n \xi_i$$

$$M \lambda^* = \frac{1}{n} \sum M \xi_i = \frac{1}{n} n \lambda = \lambda \Rightarrow \text{несмещ}$$

$$D \lambda^* = \frac{1}{n^2} \sum D \xi_i = \frac{\lambda^2}{n} \xrightarrow{n \rightarrow \infty} 0 \Rightarrow \text{состоят}$$

Задача 33

Независимые случайные величины X_1, X_2, \dots, X_k имеют биномиальное распределение соответственно $B(n_1, p), B(n_2, p), \dots, B(n_k, p)$. Пусть x_1, x_2, \dots, x_k - значения, которые приняли эти случайные величины в некотором эксперименте. Найти МП - оценку параметра p . Показать, что

полученная оценка является несмещенной и вычислить ее дисперсию.

Решение:

$$x_1 \square B(n_1, p) \dots x_k \square B(n_k, p)$$

$$L(\xi, p) = \prod_{i=1}^n p \{ \eta = \xi_i \} = \prod_{i=1}^n \ln_i^{\xi_i} p^{\xi_i} (1-p)^{n_i - \xi_i} = \prod_{i=1}^n \ln_i^{\xi_i} \left(p^{\sum \xi_i} (1-p)^{\sum (n_i - \xi_i)} \right) =$$

$$= \prod_{i=1}^n \ln_i^{\xi_i} p^{\sum \xi_i} (1-p)^{\sum n_i - \sum \xi_i}$$

$$\ln L = \sum_{i=1}^n \ln c_{n_i}^{\xi_i} + \sum_{i=1}^n \xi_i \ln p + (\sum n_i - \sum \xi_i) \ln(1-p)$$

$$(\ln L)' p = \sum \xi_i \frac{1}{p} - (\sum n_i - \sum \xi_i) \frac{1}{1-p}$$

$$\sum \xi_i - \frac{1}{p} = \frac{\sum n_i - \sum \xi_i}{1-p}$$

$$\frac{1}{p} - 1 = \frac{\sum n_i}{\sum \xi_i} - 1 \Rightarrow p^* = \frac{\sum n_i}{\sum \xi_i}$$

$$Mp^* = \frac{\sum M \xi_i}{\sum n_i} = \frac{n_i p}{n_i} = p \Rightarrow \text{несмещ.}$$

$$Dp^* = D \left(\frac{\sum \xi_i}{\sum n_i} \right) = \frac{1}{(\sum n_i)^2} \sum D \xi_i = \frac{\sum n_i p (1-p)}{(\sum n_i)^2} = \frac{p(1-p)}{\sum n_i}$$

Задача 34

Пусть x_1, x_2, \dots, x_n - выборка из генеральной совокупности, имеющей равномерное распределение $\mathbf{R}(\mathbf{a}, \mathbf{b})$. Найти МП - оценки параметров \mathbf{a} и \mathbf{b} по выборке

Решение:

$$\xi \square R(a, b)$$

$$f(x) = \frac{1}{b-a} f|[a, b](x), x \in R$$

$$L(\bar{\xi}, a, b) = \prod_{i=1}^n p\{\eta = \xi_i\} = \prod_{i=1}^n \frac{\xi_i - a}{b-a} f|[a, b] + f|[b, +\infty)$$

$$L(m, a) = \prod_{i=1}^n F(x, a) = \left(\frac{1}{b-a}\right)^n$$

$$\ln L(x, 0) = -n \ln(b-a)$$

$$\frac{\partial \ln L(x, a)}{\partial a} = \frac{n}{b-a} = 0 \Rightarrow b - \hat{a} \rightarrow \infty \Rightarrow \hat{a} = \min_{1 \leq x \leq n} \{x_i\} = x_1$$

$$\ln L(x, b) = -n \ln(b-a)$$

$$\frac{\partial L(x, a)}{\partial b} = -\frac{n}{b-a} = 0 \Rightarrow b - a \rightarrow \infty \Rightarrow b = \max_{1 \leq L \leq n} \{x_i\} = x_n$$

$$\text{Ответ: } \hat{a} = \min_{1 \leq x \leq n} \{x_i\} = x_1; b = \max_{1 \leq L \leq n} \{x_i\} = x_n$$

Метод моментов

Метод моментов нахождения точечной оценки неизвестных параметров заданного распределения состоит в приравнивании теоретических моментов соответствующим эмпирическим (выборочным) моментам того же порядка.

Пусть $f_x(x, \theta_1, \dots, \theta_s)$ – плотность распределения случайной величины X , с помощью этой плотности определим первые s начальных моментов по формулам

$$\alpha_m(\theta_1, \dots, \theta_s) = M[x^m] = \int_{-\infty}^{\infty} x^m f_x(x, \theta_1, \dots, \theta_s) dx, \quad m=1, \dots, s$$

По выборке наблюдений случайной величины находят значения соответствующих выборочных моментов

$$\alpha_m^* = \frac{1}{n} \sum_{i=1}^n x_i^m, \quad m=1, \dots, s$$

Попарно приравнивая теоретические моменты α_m их выборочным значениям α_m^* , получаем систему s уравнений с неизвестными $\theta_1, \dots, \theta_s$

$$\alpha_m(\theta_1, \dots, \theta_s) = \alpha_m^*, m = 1 \dots s.$$

Решая полученную систему уравнений относительно $\theta_1, \dots, \theta_s$, получаем оценки неизвестных параметров $\bar{\theta}_1, \dots, \bar{\theta}_s$. Аналогично получают оценки неизвестных параметров по выборке наблюдений дискретной случайной величины.

Задача 35

Найти методом моментов оценку параметра λ распределения Пуассона.

Решение:

Плотность распределения $P\{x = k\} = \frac{\lambda^k}{k!} e^{-\lambda}$, $k = 0, 1, 2, \dots$

$$\alpha_1 = M[x] = \lambda$$

$$\alpha_m^* \Big|_{m=1} = \frac{1}{n} \sum_{i=1}^n x_i^m \Big|_{m=1} = \frac{1}{n} \sum_{i=1}^n x_i$$

Следовательно $\lambda = \bar{x}$

Задача 36

В n независимых испытаниях событие **A** произошло x раз. Методом моментов найти оценку вероятности **p** появления события **A** в одном испытании.

Решение:

$$L_1 = M[x] = \sum_{i=1}^n p[x_i] = \sum p_i = p$$

$$L_2^* = \frac{\sum x_i}{n} = \frac{x}{n} \Rightarrow \hat{p} = \frac{x}{n}$$

$$\hat{p} = \frac{x}{n}$$

Задача 37

По выборке x_1, x_2, \dots, x_n объема n найти оценки параметра λ распределение Пуассона, используя метод моментов.

Решение:

$$\xi \square P_0(\lambda)$$

$$MM : M_\xi = \lambda \Rightarrow \lambda^A = \bar{x} = \frac{1}{n} \xi_i$$

$$Ответ : \lambda^A = \bar{x};$$

Задача 38

По выборке x_1, x_2, \dots, x_n объема n найти оценки параметров нормального распределения $N(m, \sigma)$, используя метод моментов.

Решение:

$$\tilde{\xi} \square N(m, \sigma^2)$$

$$MM : M_\xi = m = \bar{x}$$

$$m^* = \frac{1}{n} \sum \xi_i$$

$$\sigma^2 = \rho^2 \rightarrow \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (\xi_i - \bar{x})^2$$

Задача 39

По выборке x_1, x_2, \dots, x_n объема n найти оценки параметра показательного распределения $E_x(\lambda)$, используя метод моментов.

Решение:

$$E_x(\lambda) \approx \hat{\xi}$$

$$f(x) = \lambda \exp(-\lambda x) \quad f \text{ на } [a, +\infty)(x), x \in R$$

$$MM : M_\xi = \bar{x} : M_\xi = \int_0^{+\infty} x \lambda \exp(-\lambda x) dx = \lambda \left(\frac{1}{\lambda} \right) \exp(-\lambda x) x \Big|_0^{+\infty} + \int_0^{+\infty} \exp(-\lambda x) dx = \frac{e}{\lambda} :$$

$$\frac{1}{\lambda} = \frac{1}{n} \sum_{i=1}^n \xi_i \rightarrow \lambda^* = \frac{n}{\sum_{i=1}^n \xi_i}$$

$$\text{Ответ: } \lambda^* = \frac{1}{\frac{1}{n} \sum_{i=1}^n \xi_i} = \frac{1}{x}$$

Лабораторная работа № 3

Цель лабораторной работы – получение точечных оценок параметров распределений дискретных и непрерывных случайных величин методом максимального правдоподобия в пакете MATHCAD.

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

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

Метод максимального правдоподобия для дискретной случайной величины

В MATHCAD. для моделирования выборки значений случайной величины, распределенной по закону Пуассона, предназначена функция $\text{prois}(k, \lambda)$, которая формирует вектор из k случайных чисел, распределенных по Пуассону с параметром λ .

ЗАДАНИЕ

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

Порядок выполнения задания

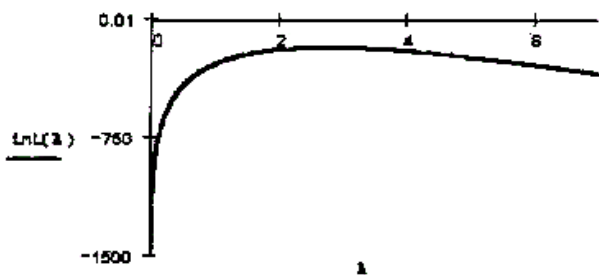
1. Смоделируйте выборку значений случайной величины, имеющей распределение Пуассона с заданным значением параметра λ .
2. Определите логарифм функции максимального правдоподобия и изобразите его график.
3. Смоделируйте несколько выборок разного объема значений случайной величины, имеющей распределение 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 $\lambda = 3$.

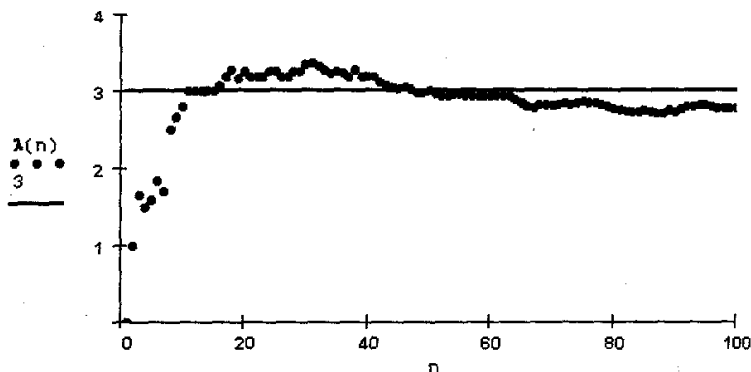
ORIGIN := 1 **P := rpois(100, 3)**

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



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

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



Метод максимального правдоподобия для непрерывной случайной величины

В MATHCAD для моделирования выборки значений случайной величины, имеющей показательное распределение, предназначена функция $\text{gehr}(k, \lambda)$, которая формирует вектор из k случайных чисел, распределенных показательно с параметром λ .

ЗАДАНИЕ

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

Порядок выполнения задания

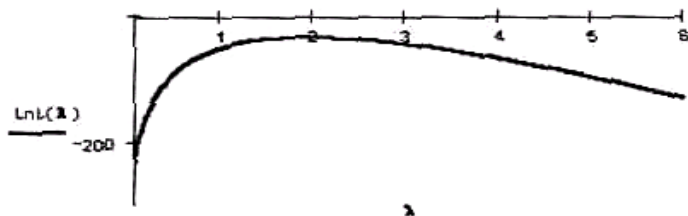
1. Смоделируйте выборку значений случайной величины, имеющей экспоненциальное распределение с заданным значением параметра λ .
2. Определите логарифм функции максимального правдоподобия и изобразите его график.
3. Смоделируйте несколько выборок разного объема значений случайной величины, имеющей экспоненциальное 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 $\lambda=2$.

ORIGIN := 1 P := геxp(100,2)

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

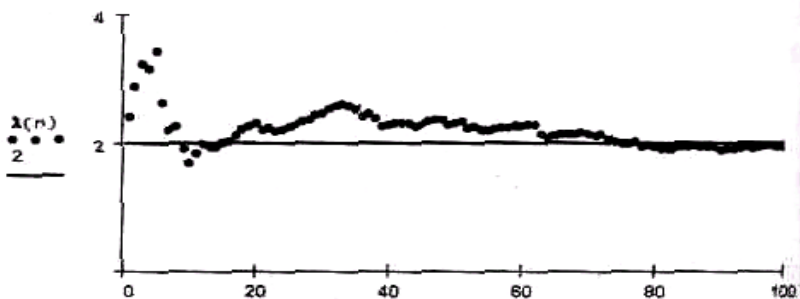


n := 1..100

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

$$\lambda(50) = 2.33$$

$$\lambda(100) = 1.948$$



4. РАСПРЕДЕЛЕНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН, ИСПОЛЬЗУЕМЫЕ В ЗАДАЧАХ ПРИКЛАДНОЙ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

4.1. Нормальное распределение

Нормальный закон распределения (закон Гаусса) играет важную роль в теории вероятностей и занимает особое положение среди других законов. Такой закон имеет место, когда на формирование случайной величины оказывает влияние множество разнообразных факторов. Например, координаты точки попадания снаряда, рост, вес человека имеют нормальный закон распределения.

Случайная величина X называется **нормальной**, если ее плотность вероятности имеет вид:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}} = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-a)^2}{2\sigma^2}\right\}. \quad (4.1)$$

$X \sim N(a, \sigma) \Rightarrow$ случайная величина распределена по нормальному закону с параметрами распределения (a, σ) .

Вычислим для нормальной случайной величины X вероятность попадания на участок (α, β)

$$P\{\alpha < X < \beta\} = \int_{\alpha}^{\beta} f(x) dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{\alpha}^{\beta} \exp\left\{-\frac{(x-a)^2}{2\sigma^2}\right\} dx. \quad (*)$$

Сделаем, в интеграле (*) замену переменной $t = \frac{(x-a)}{\sigma}$, и изменяя пределы интегрирования, получим

$$P\{\alpha < X < \beta\} = \frac{1}{\sqrt{2\pi}} \int_{\frac{\alpha-a}{\sigma}}^{\frac{\beta-a}{\sigma}} \exp\left\{-\frac{t^2}{2}\right\} dt.$$

называемую **функцией Лапласа** или **интегралом вероятностей**, для которой составлены таблицы. С помощью этой функции вероятность попадания нормальной случайной величины на участок (α, β) выражается простой формулой

$$P\{\alpha < X < \beta\} = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right). \quad (4.3)$$

Функция Лапласа $\Phi(x)$ обладает следующими свойствами:

1) $\Phi(0) = 0$

Действительно, $\Phi(0) = \frac{1}{\sqrt{2\pi}} \int_0^0 e^{-t^2/2} dt = 0$.

2) $\Phi(-x) = -\Phi(x)$ - нечетная функция.

Доказательство: $\Phi(-x) = \frac{1}{\sqrt{2\pi}} \int_0^{-x} e^{-t^2/2} dt$,

делаем замену $-t = z$, получаем

$$\Phi(-x) = \frac{-1}{\sqrt{2\pi}} \int_0^x e^{-z^2/2} dz, \text{ т.е. } \Phi(-x) = -\Phi(x). \quad \blacksquare$$

3) $\Phi(+\infty) = 0.5$; $\Phi(-\infty) = -0.5$.

Это свойство следует из того что, используя соответствующую запись можно прийти к интегралу Эйлера-Пуассона, и получаем следующее

$$\frac{1}{\sqrt{2\pi}} \int_0^{\infty} e^{-t^2/2} dt = \frac{1}{\sqrt{\pi}} \int_0^{\infty} e^{-\left(\frac{t}{\sqrt{2}}\right)^2} d\frac{t}{\sqrt{2}} = \frac{1}{\sqrt{\pi}} \cdot \frac{\sqrt{\pi}}{2} = \frac{1}{2}.$$

Интеграл Эйлера-Пуассона:

$$\int_{-\infty}^{\infty} e^{-t^2} dt = 2 \int_0^{\infty} e^{-t^2} dt = \sqrt{\pi} \quad \blacksquare$$

Через функцию Лапласа просто выражается вероятность попадания нормальной случайной величины X на участок длиной $2L$.

$$P\{a-L < X < a+L\} = P\{|X - a| < L\} = \Phi\left(\frac{a - a + L}{\sigma}\right) - \Phi\left(\frac{a - a - L}{\sigma}\right),$$

принимая во внимание нечетность функции Лапласа, получаем

$$P\{|X - a| < L\} = 2\Phi\left(\frac{L}{\sigma}\right). \quad (4.4)$$

Через функцию Лапласа выражается и функция распределения $F(x)$ нормальной случайной величины X . По формуле (4.3), полагая $\alpha = -\infty$, $\beta = x$, и учитывая, что $\Phi(-\infty) = -1/2$, получим:

$$F(x) = \frac{1}{2} + \Phi\left(\frac{x - a}{\sigma}\right). \quad (4.5)$$

При изменении параметров распределения будет изменяться кривая распределения. При изменении a $f(x)$ не изменяет своей формы, просто смещается вдоль оси абсцисс. Изменение σ равносильно изменению масштаба кривой по обеим осям (см. рис. ниже)

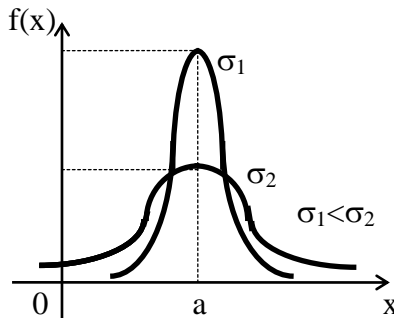


Рис. 4.1. σ - характеристика рассеивания, a - характеристика положения

4.2. Квадратичные и линейные формы от нормальных случайных величин и их свойства

Пусть $\vec{X} = (X_1, \dots, X_n)$ выборка из $L(\xi) = N(0,1)$

Рассмотрим квадратичную форму $Q = \sum_{i,j=1}^n a_{ij} X_i X_j = X^T A X$

и m линейных форм $t_k = \sum_{i=1}^n b_{ki} X_i$, $k = \overline{1, m}$ или в матричных

обозначениях $\vec{t} = BX$, где $A = \| a_{ij} \|_1^n$ - матрица, удовлетворяющая условию $A^T = A$, B - прямоугольная матрица порядка $m \times n$, а $\vec{t} = (t_1, \dots, t_m)$ - вектор.

Пусть O - матрица с нулевыми элементами, I_n - единичная матрица порядка n . Рассмотрим свойства квадратичной формы.

1. Если $BA=O$, то функции Q и t независимы.
2. Рассмотрим 2 квадратичные формы $Q_1 = X^T AX$ и $Q_2 = X^T BX$, если $AB=BA=O$, то Q_1 и Q_2 независимы.
3. Обозначим через $\text{tr } A$ след квадратной матрицы (т.е. сумму ее диагональных элементов). Имеет место утверждение. Пусть $Q = X^T AX$ и $\text{rang } A = r \leq n$. Если матрица A идемпотентна ($A^2=A$), то $Z(Q) = \chi^2(r)$ и при этом $r = \text{tr } A$.

Теорема 4.1. (теорема Фишера)

Пусть $\vec{X} = (X_1, \dots, X_n)$ - выборка из распределения $N(m, \sigma^2)$. Тогда выборочное среднее \bar{X} и дисперсия $S^2 = S^2(\vec{X})$ независимы и при этом подчиняются следующим законам распределения $L(\sqrt{n}(\bar{X} - m)/\sigma) = N(0,1)$, $L(nS^2/\sigma^2) = \chi^2(n-1)$.

Доказательство. Перейдем к новым случайным величинам $Y_i = (X_i - m)/\sigma$, $i = \overline{1, n}$, которые образуют выборку \vec{Y} из $N(0,1)$. Тогда $\bar{Y} = (\bar{X} - m)/\sigma$ и $S^2(\vec{Y}) = 1/\sigma^2 S^2(\vec{X})$.

Поэтому достаточно доказать, что \bar{Y} и $S^2(\vec{Y})$ независимы и при этом $L(\sqrt{n} \bar{Y}) = N(0,1)$, $L(nS^2(\vec{Y})) = \chi^2(n-1)$.

Рассмотрим n – мерный вектор-столбец $\vec{b} = (1/n, \dots, 1/n)^T$ и $(n \times n)$ -матрицу $B = \|\vec{b} \dots \vec{b}\|$. Заметим, что $\bar{Y} = \vec{b}^T Y$, а $nS^2(\bar{Y}) = (\bar{Y} - B\bar{Y})^T (\bar{Y} - B\bar{Y})$. Отсюда $nS^2(\bar{Y}) = \bar{Y}^T A \bar{Y}$, где матрица $A = I_n - B$ идемпотентна. Теперь $\vec{b}^T A = \vec{b}^T - \vec{b}^T B = \vec{b}^T - \vec{b}^T = 0$, и, следовательно, по свойству 1), \bar{Y} и $S^2(\bar{Y})$ – независимы.

Закон распределения \bar{Y} очевиден. Так как $\text{tr } A = \text{tr } I_n - \text{tr } B = n - 1$, то на основании свойства (3) $L(nS^2(\bar{Y})) = \chi^2(n-1)$. ■

4.3. Распределение хи-квадрат

Пусть ξ_1, \dots, ξ_n – независимые случайные величины, распределенные по стандартному нормальному закону $\xi_i \sim N(0,1)$. Распределение случайной величины

$$\chi^2 = \sum_{j=1}^n \xi_j^2 \quad (4.6)$$

назовем χ^2 – распределением с n степенями свободы.

Здесь ξ_j^2 – квадратичная форма. Число независимых слагаемых n в формуле (4.6) называется числом степеней свободы и является параметром распределения χ^2 , n – натуральное число.

Найдем плотность вероятности χ^2 – распределения с помощью характеристической функции слагаемого и ее свойств.

Характеристическая функция слагаемого будет:

$$E_{\xi} (t) = M \cdot \exp(i\xi_j^2 t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \exp(ix^2 t) \cdot \exp\left(-\frac{x^2}{2}\right) dx = \frac{1}{(1-2it)^n}$$

По свойству характеристической функции имеем:

$$E_{\chi}(t) = \prod_{j=1}^n E_{\xi}(t) = \frac{1}{(1-2it)^n} . \quad (4.7)$$

Используя следствие из теоремы обращения, имеем:

$$f_{\chi}(x) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \exp(-ixt) \cdot E_{\chi}(t) dt = \frac{x^{n-1}}{2^n \Gamma(\frac{n}{2})} \cdot \exp(-\frac{x}{2}) \quad x > 0$$

$$f_{\chi}(x) = 0 \quad x \leq 0 .$$

Числовые характеристики χ^2 -распределения находят с помощью характеристической функции (4.7)

Они имеют следующий вид:

$$\text{математическое ожидание } M[\chi_n^2] = n ,$$

$$\text{мода } [\chi_n^2]_{\text{mod}} = n - 2 ,$$

$$\text{дисперсия } D[\chi_n^2] = 2n ,$$

$$\text{асимметрия } \beta = 2^{3/2} / \sqrt{n} ,$$

$$\text{эксцесс } \beta = 12/n .$$

При $n \rightarrow \infty$ в соответствии с центральной предельной теоремой χ^2 -распределение сходится к нормальному

$$\chi_n^2 \rightarrow N(n, 2n) ..$$

При $n \geq 30$ используется аппроксимация нормальным распределением. Существуют таблицы:

$$P(\chi_n^2 \geq \chi_p^2) = 1 - F_{\chi_n^2}(\chi_p^2) = \int_{\chi_p^2}^{\infty} f_{\chi_n^2}(x) dx = p .$$

По этим таблицам при заданном n по вероятности p можно найти χ_p^2 . Иногда табулированы значения функции распределения. Квантили χ^2 -распределения определяются из таблиц или с помощью математических пакетов MATHCAD и STATISTICA,

Важным свойством χ^2 -распределения является его воспроизводимость по параметру n . Это означает, что сумма независимых случайных величин, распределенных по закону

χ^2 , распределена также по закону χ^2 с числом степеней свободы, равным сумме степеней свободы слагаемых.

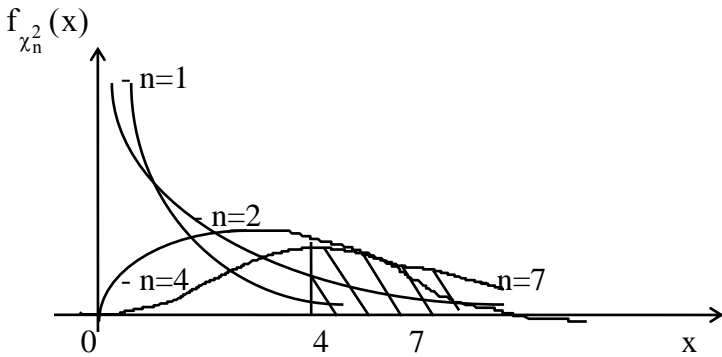


Рис. 4.2. Плотность распределения χ_p^2

Теорема 4.2. Пусть $\vec{X}=(X_1, \dots, X_n)$ - выборка из распределения $N(\mu, \sigma^2)$. Тогда выборочное среднее $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$

и дисперсия $S^2 = S^2(X) = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ независимы; при этом

$$L\left(\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma}\right) = N(0, 1);$$

$$L\left(\frac{nS^2}{\sigma^2}\right) = \chi_{(n-1)}^2.$$

4.4. Распределение Стьюдента (t – распределение)

Распределением Стьюдента с n степенями свободы $S(n)$ называется распределение случайной величины (стьюдентова отношения) $t = \frac{\xi}{\sqrt{\chi_n^2/n}}$, где случайные величины ξ и χ_n^2 независимы и при этом $L(\xi) = N(0, 1)$. Иногда это распределение называют t-распределением (с n степенями свободы).

Плотность распределения можно найти с помощью стандартного метода вычисления плотности распределения частного двух независимых случайных величин, а именно:

$$f_t(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n} \cdot \Gamma\left(\frac{n}{2}\right)} \cdot \frac{1}{\left(1 + x^2/n\right)^{(n+1)/2}}, \quad -\infty < x < \infty.$$

При $n \rightarrow \infty$, $t \rightarrow \eta \sim N(0,1)$.

При $n \geq 20$ можно считать, что $t \sim N$ (распределение Стьюдента аппроксимируется нормальным).

Существуют таблицы $F_t(x) = P(t < x)$ и $P(|t| > t_p) = 2 \int_{t_p}^{\infty} f_t(x) dx = p$.

Существуют таблицы и для плотности распределения $f_t(x)$.

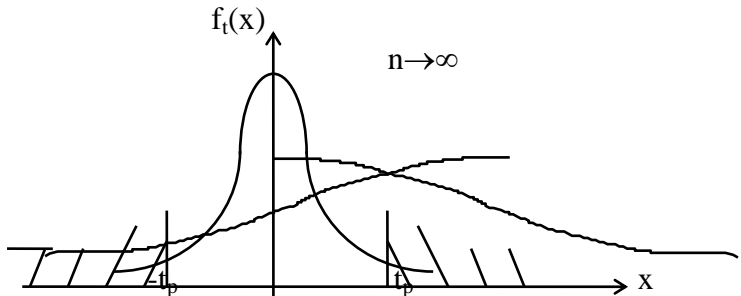


Рис. 4.3. Плотность распределения Стьюдента

Теорема 4.3. Пусть выборка $\bar{X} = (X_1, X_2, \dots, X_n)$ из генеральной совокупности $\xi \sim N(a, \sigma^2)$ и

$$t = \sqrt{n-1} \cdot \frac{\bar{x} - a}{S}. \quad (4.8)$$

(\bar{X} - выборочное среднее, S - выборочная дисперсия). Тогда при любом $\sigma^2 > 0$

$$L(t) = S(n-1).$$

т.е. случайная величина t распределена по закону Стьюдента с $(n-1)$ степенями свободы.

Тот факт, что стьюдентово соотношение t , определенное уравнением (4.8), и его распределение не зависят от σ^2 используют при получении различных статистических выводов о среднем нормального распределения, когда дисперсия неизвестна, т.е. является «мешающим» параметром. В некоторых задачах иногда нужно исключить влияние не только σ^2 , но и среднего a . В этом случае можно делать статистические выводы, не зависящие от параметров a и σ^2 , т.е. являющимися инвариантными относительно параметров модели. В таких задачах важна следующая теорема.

Теорема 4.4. Пусть $\vec{X}=(X_1, X_2, \dots, X_n)$ и $\vec{Y}=(Y_1, Y_2, \dots, Y_n)$ - две независимые выборки из одного и того же распределения $N(a, \sigma^2)$; $\bar{X}, S^2(\bar{X})$; $\bar{Y}, S^2(\bar{Y})$ - соответствующие выборочные средние и дисперсии и пусть

$$t = \sqrt{\frac{mn(m+n-2)}{m+n}} \cdot \frac{\bar{X} - \bar{Y}}{\sqrt{nS^2(\bar{X}) + mS^2(\bar{Y})}}.$$

Тогда при любых a и $\sigma^2 > 0$ $L(t) = S(m+n-2)$, t - случайная величина, распределенная по закону Стьюдента с $(m+n-2)$ степенями свободы).

4.5. Распределение Фишера – Снедекора (F – распределение)

Пусть случайные величины χ_n^2 и χ_m^2 независимы и

$$F = \frac{\chi_n^2 / n}{\chi_m^2 / m} = \frac{m}{n} \frac{\chi_n^2}{\chi_m^2}$$

Распределение случайной величины F называют распределением Снедекора с n и m степенями свободы, F -распределением или распределением дисперсионного отношения Фишера.

Плотность $f_{n,m}(x)$ распределения $S(n,m)$ имеет вид:

$$f_{n,m}(x) = \left(\frac{n}{m}\right)^{n/2} \frac{\Gamma((n+m)/2)}{\Gamma(n/2)\Gamma(m/2)} \frac{x^{n/2-1}}{\left(1 + \frac{n}{m}x\right)^{(n+m)/2}}, x > 0$$

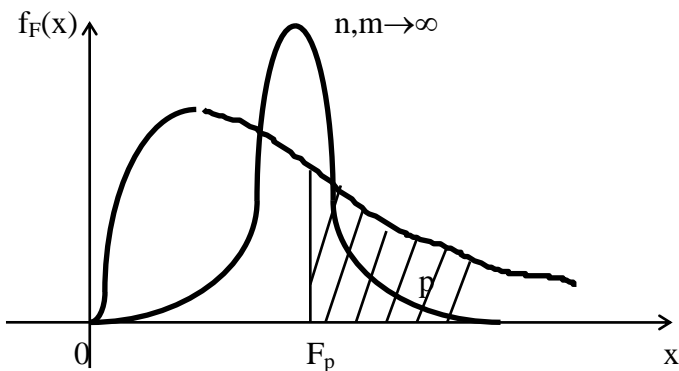


Рис. 4.4. Плотность распределения Фишера

При $n, m > 30$ - возможна аппроксимация нормальным распределением

$$S(n, m) \xrightarrow{n, m \rightarrow \infty} \eta \sim N\left(\frac{n+m}{2nm}; \sqrt{\frac{n+m}{nm}}\right).$$

Существуют таблицы функции распределения $f_F(x) = P(F < x)$ и $P(F \geq F_p) = p$. Роль F-распределения в выборочной теории раскрывает следующая теорема.

Теорема 4.5. Пусть $\vec{X} = (X_1, X_2, \dots, X_n)$ и $\vec{Y} = (Y_1, Y_2, \dots, Y_m)$ - независимые выборки из распределений $N(a_1, \sigma_1^2)$ и $N(a_2, \sigma_2^2)$. $S^2(\vec{X})$, $S^2(\vec{Y})$ - соответствующие выборочные дисперсии. Тогда при любых значениях параметров a_1 , a_2 , σ_1^2 и σ_2^2 случайная величина

$$\frac{S^2(\vec{X})n(m-1)\sigma_2^2}{S^2(\vec{Y})m(n-1)\sigma_1^2} = F(n-1, m-1).$$

распределена по закону Фишера с $(n-1)$; $(m-1)$ степенями свободы. Примем без доказательства.

Лабораторная работа № 4. Распределения непрерывных случайных величин в пакете STATISTICA.

Цель лабораторной работы – изучить параметры и свойства распределений случайных величин, используемых при анализе данных.

Теоретические сведения

В классе модельных распределений рассмотрим наиболее употребляемые в математической статистике:

- экспоненциальное (показательное);
- нормальное распределения;
- *хи-квадрат* распределение;
- распределение Стьюдента;
- распределение Фишера.

Случайная величина χ^2_k имеющая распределение *хи-квадрат*, понимается как сумма квадратов k независимых стандартных нормальных величин. Число независимых слагаемых k называется числом степеней свободы и является параметром распределения *хи-квадрат*, k - натуральное число. Математическое ожидание и дисперсия случайной величины

χ^2_k равны соответственно

$$M\chi^2_k=k, \quad D\chi^2_k=2k$$

Распределение случайной величины, представляющей собой отношение стандартной нормальной величины к корню квадратному из *хи-квадрат* распределенной величины,

$$t = \frac{\xi}{\sqrt{\frac{\chi^2}{k}}}$$

называется t – *распределением Стьюдента*. Точнее, если $\xi \sim N(0;1)$, χ^2_k распределена по закону *хи-квадрат* с числом степеней свободы k , то случайная величина $\xi/\sqrt{\chi^2/k}$ имеет t – *распределение Стьюдента* с числом степеней свободы k ,

причем, $Mt_k = 0, Dt_k = k/(k-2)$.

t – распределение важно в тех случаях, когда рассматривают оценки среднего и неизвестна дисперсия выборки.

Пусть случайные величины ξ_n и η_k независимы и распределены по закону *хи – квадрат* каждая. Распределение случайной величины

$$F_{n,k} = \frac{\frac{\xi_n}{n}}{\frac{\eta_k}{k}}$$

называют F – *распределением* Фишера с параметрами n и k или *распределением дисперсионного отношения*. Математическое ожидание и дисперсия F – *распределения* вычисляются по формулам:

$$MF_{n,k} = \frac{k}{k-2}$$
$$DF_{n,k} = \frac{2k^2(k+n-2)}{n(k-2)(k-2)(k-4)}$$

Все перечисленные распределения представляют собой параметрические семейства. Параметры, входящие в формулу каждой из функций плотности, имеют определенный геометрический и вероятностный смысл. Они (параметры) влияют на форму и расположение кривой распределения и определяют значения числовых характеристик.

Для упомянутых распределений составлены таблицы, позволяющие при разных значениях параметров определить квантили и вероятности попадания в различные интервалы. Выбор Probability Calculator (Вероятностный калькулятор) заменяет многочисленные таблицы распределений, позволяет проследить влияние параметров на форму кривой

распределения и выяснить геометрический смысл квантилей. Квантиль ζ_p порядка (p -квантиль) – это корень уравнения $F(\zeta_p) = p$.

Задания к лабораторной работе

Для запуска Вероятностного калькулятора (Probability Calculator) необходимо нажать кнопку Меню выбора основных модулей обработки информации в программном обеспечении STATISTICA6.0 и выбрать Статистика(Statistics)►Счетчик вероятности (Probability Calculator)►Распределения(Distribution). Внешнее описание и представление можно увидеть на рис.4.5.

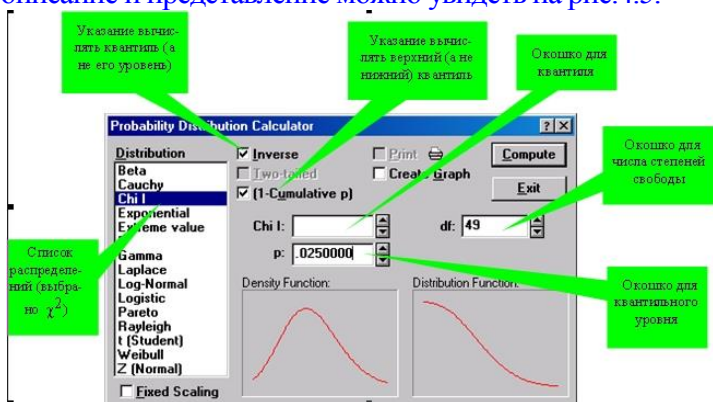


Рис.4.5. Окно χ^2 - распределения процедуры Probability Calculator

1. Для ознакомления с работой Probability Calculator необходимо реализовать любой простой алгоритм. Например, необходимо выяснить геометрический смысл параметров нормального распределения $N(\alpha; \sigma)$.

Положите $\alpha = 0$, $\sigma = 1$. В окне Probability Distribution Calculator в поле *Distribution*: выделите мышью строку Z (NORMAL), заполните поля: **mean**: 0, **sd.dev.**:1, **p**: 0.5. Поднимите флажок: Fixed Scaling и нажмите Compute. В поле **X**: открытого окна появится значение 0.0000. Это 0.5 – квантиль нормального распределения, т.е. корень уравнения $F(Z) = 0.5$. В поле

Density Function изображается кривая распределения с заштрихованной областью. Площадь отмеченной области равна указанному значению p : 0.5. Нажмите далее флажок Create Graph, для построения функции распределения, и нажмите Compute. На экране появится график плотности с отмеченным красным пунктиром квантилем. Из графика видно, что *0.5-квантиль является модой и медианой нормального распределения*. Повторяя приведенную последовательность команд для разных значений **mean** ($\alpha = 1$; 2; -2; ...), убедитесь, что значение α является точкой максимума функции плотности нормального распределения.

Меняя значение поля **st.dev** (σ) при постоянном α и p , убедитесь, что при увеличении σ плотность нормального распределения рассеивается относительно α , f_{\max} уменьшается. При уменьшении σ плотность сжимается, концентрируясь возле точки максимума, f_{\max} растет.

2. Теперь вычислим вероятность P ($137 < \zeta < 179$) случайной величины ζ , распределенной нормально с параметрами $\alpha = 149.6$, $\sigma = 12.62$.

В окне Probability Distribution Calculator заполните поля: **Distribution**: Z (NORMAL), **mean**: 149.6; **st.dev** : 12.62 ; **X**: 179 . Нажмите Compute. В поле p появится значение 0,990087. Его необходимо запомнить, потому что оно будет использовано ниже при вычислении. Измените значение **X** на 137. Нажмите Compute. Запомните новое значение поля p : 0,159039.

Теперь необходимо вычислить следующее:

$$P(137 < X < 179) = 0,990087 - 0,159039 = 0.831048 \approx 0.83.$$

Значение 0,83 или 83% является вероятностью случайной величины ζ , распределенной нормально с параметрами $\alpha = 149.6$, $\sigma = 12.62$, на интервале $137 < \zeta < 179$.

3. Вычислить **0.95** и **0.99** – квантили *хи-квадрат* распределения с 7 степенями свободы. Выяснить влияние числа степеней свободы на форму и расположение кривой распределения.

В окне Probability Distribution Calculator выделите в поле

Distribution строку *Chi I*. Заполните поля: **df**: 7; **p**: 0,95. Нажмите Compute. В поле *Chi I* число 14.067140. Это 95% -я точка (0.95 - квантиль), то есть корень уравнения $F(I) = 0.95$. Значит $P(tf < 14,068419) = 0,95$. Чтобы вычислить вероятность противоположного неравенства, поднимите флажок (*I – Cumulative p*).

Теперь поменяем значение поля **p**: на **0.99** и нажмем Compute. В поле *Chi I* появится число 18,475307. Это 99% -я точка (0.99 - квантиль).

Задавая различные значения параметра *k* в поле **df** (2; 5; 12; ...) убедитесь, что при увеличении *k* пик плотности распределения снижается и смещается вправо. График плотности становится более симметричным, приближаясь по форме к кривой Гаусса.

4. Выяснить влияние числа степеней свободы на форму и расположение кривой распределения Стьюдента.

Для этого необходимо в поле Distribution выделить строку **t** (STUDENT) и заполнить поля **p**: 0.5, **df**: 5. Поле **t** система заполнит числом 0. Для наглядности распределения необходимо пометить опцию Create Graph и нажать Compute. Рассмотрите внимательно полученный график и повторите алгоритм для **df** = 10, 35, 50, 100. Убедитесь в том, что график плотности *t*- распределения (см. рис. 2.2) симметричен относительно оси *Oy* и напоминает колокол Гаусса.

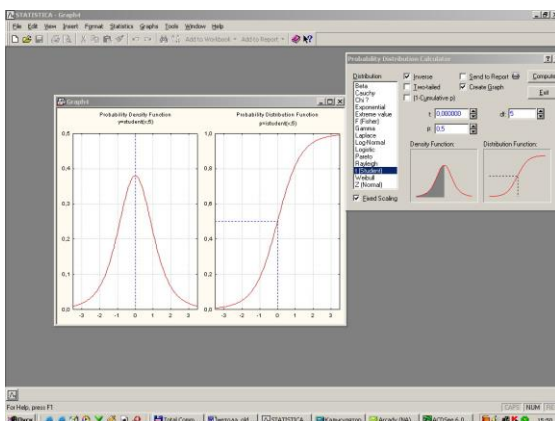


Рис. 4.6. Графическое представление плотности t -распределения

С возрастанием числа степеней свободы k максимальное значение плотности увеличивается, хвосты более круто убывают к 0. Вводя в поле **p**: значения 0.5; 0.7; 0.95; 0.99, составьте таблицу значений функции t -распределения с 10 степенями свободы (таблицу квантилей).

Теперь введите в поле **t** значение 1. Система вычислит **p**: 0,818391. Следовательно, $P(t < 1) = 0,818391$. Поднимите флажок (*1 – Cumulative p*). Содержимое поля **p**: изменится на 0,181609. Калькулятор вычислил вероятность противоположного события: $P(t \geq 1) = 0,181609$.

5. Убедиться с помощью вероятностного калькулятора, что F -распределение сосредоточено на положительной полуоси. Определить **0.5**- и **0.75**- квантили $F_{10,10}$ -распределения. Вычислить вероятности $P(F_{10,10} \leq 2)$.

Для решения этой достаточно простой задачи необходимо в поле Distribution выделить строку **F** (FISHER). Заполните поля **p**: 0.5 ; **df1**: 10; **df2**: 10 и нажимаем **Compute**. Калькулятор вычислит значение поля **F**: 1. Поменяйте значение поля **p**: 0.75. Значение поля **F**: изменится на 1,551256. Теперь измените значение поля **F**:. Поставьте сначала 2, потом 1. Калькулятор вычислит вероятности: $P(F_{10,10} \leq 2) = 0.144846$, $P(F_{10,10} \leq 1) = 0.5$.

Давая различные значения **df1** и **df2**, наблюдайте графики. Обратите внимание на то, что, в отличие от нормальной, кривая F -распределения несимметрична при небольших значениях степеней свободы (n и $k < 30$). С возрастанием n и k кривая F -распределения медленно приближается к нормальной кривой.

6. После выполнения заданий 1 - 5 нужно самостоятельно для нормального распределения с выбранными параметрами вычислить вероятность попадания в интервал, содержащий **mean** и не содержащий **mean**.

7. Составить таблицы *нормального, хи-квадрат,*

Стьюдента и *Фишера* распределений (по 10 значениям). Вычислить 0.95 и 0.99 – квантили модельных распределений для различных значений параметров.

Составить отчет по выполненной работе

Отчет по **выполненной** работе должен содержать:

- Постановку задачи.
- Графики плотностей модельных распределений при различных значениях параметров.
- Таблицы процентных точек (квантилей) по 10 значений для каждого распределения.
- Для наглядности, в процессе выполнения работы (каждого пункта) необходимо сделать как минимум один Screen Capture, которые в дальнейшем будут размещены в отчете.
- Вывод о проделанной работе.

5. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ

Любая точечная оценка параметра представляет собой функцию $T=T(\bar{X})$ выборки $\bar{X}=(X_1, X_2, \dots, X_n)$, т.е. является случайной величиной. При каждой реализации \bar{x} выборки \bar{X} эта функция определяет единственное значение $t=T(\bar{x})$ оценки, которое принимается за приближенное значение оцениваемой характеристики. Однако в каждом конкретном случае значение оценки может отличаться от значения параметра. Поэтому желательно знать и возможную погрешность, возникающую при использовании предлагаемой оценки. Например, указывая интервал, внутри которого с высокой вероятностью γ находится точное значение оцениваемого параметра. При

таким подходе говорят об интервальном или доверительном оценивании, а соответствующий интервал называют доверительным.

5.1. Понятие доверительного интервала

При статистической обработке результатов наблюдений часто необходимо не только найти оценку неизвестного параметра θ , но и охарактеризовать точность этой оценки. С этой целью вводится понятие **доверительного интервала**. Рассмотрим доверительное оценивание скалярного параметра. При интервальном оценивании ищут две такие статистики $T_1=T_1(\bar{X})$ и $T_2=T_2(\bar{X})$, что $T_1 < T_2$, для которых при заданном $\gamma \in (0,1)$ выполняется условие

$$P_{\theta}(T_1(\bar{X}) < \theta < T_2(\bar{X})) \geq \gamma, \quad \forall \theta \in \Theta \quad (5.1)$$

в этом случае интервал $(T_1(\bar{X}), T_2(\bar{X}))$ называют γ -доверительным интервалом, вероятность γ - доверительной вероятностью, а величина $q=1-\gamma$ - уровнем значимости. $T_1(\bar{X})$ и $T_2(\bar{X})$ - называются нижней и верхней доверительными границами соответственно. Таким образом γ -доверительный интервал - это случайный интервал в параметрическом множестве Θ : $T_1=T_1(\bar{X})$ и $T_2=T_2(\bar{X})$, что $T_1 < T_2$, для которых при заданном $\gamma \in (0,1)$ выполняется условие

$$P_{\theta}(T_1(\bar{X}) < \theta < T_2(\bar{X})) \geq \gamma, \quad \forall \theta \in \Theta \quad (5.1)$$

в этом случае интервал $(T_1(\bar{X}), T_2(\bar{X}))$ называют γ -доверительным интервалом, вероятность γ - доверительной вероятностью, а величина $q=1-\gamma$ - уровнем значимости. $T_1(\bar{X})$ и $T_2(\bar{X})$ - называются нижней и верхней доверительными границами соответственно.

Таким образом γ -доверительный интервал - это случайный интервал в параметрическом множестве Θ : $(T_1, T_2) \subset \Theta$, зависящий от выборки \bar{X} (но не от θ), который

содержит (накрывает) истинное значение неизвестного параметра θ с вероятностью, не меньшей γ .

Условие (5.1) означает, что в большой серии независимых экспериментов, в каждом из которых получена выборка объема n в среднем $\gamma \cdot 100\%$ из общего числа построенных доверительных интервалов содержит истинное значение параметра θ . Длина доверительного интервала, характеризующая точность интервального оценивания, зависит от объема выборки n и доверительной вероятности γ . При увеличении объема выборки длина доверительного интервала уменьшается, а с приближением доверительной вероятности к единице ($\gamma \rightarrow 1$) - увеличивается. Выбор доверительной вероятности определяется конкретными условиями. Обычно используются значения γ , равные 0.90; 0.95; 0.99. иногда рассматривают односторонние доверительные интервалы, соответственно верхний (вида $\theta < T_2(\bar{X})$) и нижний (вида $T_1(\bar{X}) < \theta$), определяемые условиями, аналогичными (5.1), в которых опускают соответствующую вторую границу

$$P(\theta < T_2(\bar{X})) = \gamma \text{ или } P(T_1(\bar{X}) < \theta) = \gamma.$$

5.2. Построение доверительного интервала с помощью центральной статистики

Общий прием, с помощью которого можно построить доверительный интервал, состоит в следующем. Пусть модель F абсолютно непрерывна и существует случайная величина $G(\bar{X}; \theta)$, зависящая от θ такая что:

- 1) распределение с.в. $G(\bar{X}; \theta)$ не зависит от θ ,
- 2) при каждом $\bar{x} \in \mathcal{X}$ функция $G(\bar{x}; \theta)$ непрерывна и строго монотонна по θ .

Такую случайную величину называют **центральной статистикой** (для θ). Будем рассматривать только случай скалярного параметра θ . Пусть для модели F построена центральная статистика $G(\bar{X}; \theta)$ и $f_G(g)$ ее плотность

распределения. Функция $f_G(g)$ от параметра θ не зависит (условие 1), поэтому для любого $\gamma \in (0,1)$ можно выбрать величины $g_1 < g_2$ (многими способами) так, чтобы

$$P_\theta(g_1 < G(\bar{X}; \theta) < g_2) = \int_{g_1}^{g_2} f_G(g) dg = \gamma. \quad (5.2)$$

Определим теперь при каждом $\bar{x} \in X$ числа $T_i(\bar{x})$, $i=1,2$, где $T_1(\bar{x}) < T_2(\bar{x})$, как решения относительно θ уравнений

$$G(\bar{x}; \theta) = g_k, \quad k=1,2 \quad (5.3)$$

(однозначность определения этих чисел обеспечивается условием 2, наложенным на функцию $G(\bar{x}; \theta)$). Тогда неравенства $g_1 < G(\bar{x}; \theta) < g_2$ эквивалентны неравенствам $T_1(\bar{x}) < \theta < T_2(\bar{x})$ (Рис. 5.1). Следовательно, формулу (5.2) можно переписать в виде

$$P_\theta(T_1(\bar{X}) < \theta < T_2(\bar{X})) = \gamma, \quad \forall \theta.$$

Таким образом, построенный интервал $(T_1(\bar{X}), T_2(\bar{X}))$ является γ -доверительным интервалом для θ .

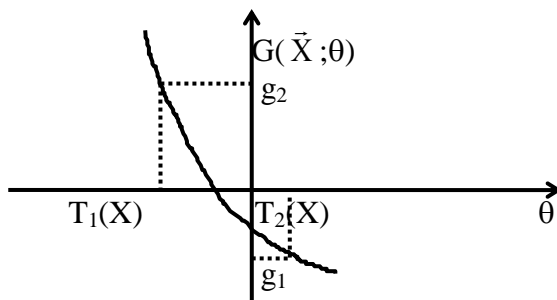


Рис.5.1

В конкретных задачах при построении центральной статистики для оцениваемой характеристики приходится учитывать специфику рассматриваемой модели. Однако, можно выделить класс моделей, для которых центральная статистика всегда существует и имеет простой вид. Именно:

если функция распределения $F(x, \theta)$ непрерывна и монотонна по параметру θ , то можно положить

$$G(\bar{X}; \theta) = -\sum_{i=1}^n \ln F(X_i; \theta) \quad (5.4)$$

Действительно, непрерывность и монотонность по θ здесь очевидны, а так как $L_\theta(F(X_i; \theta)) = R(0, 1)$ при любом θ , то распределение $G(\bar{X}; \theta)$ не зависит от θ . Из $L(\eta) = R(0, 1)$ следует, что $L(-\ln \eta) = \Gamma(1, 1)$. Таким образом, слагаемые в (5.4) независимы и каждое из них имеет распределение $\Gamma(1, 1)$. Используя свойства гамма-распределения, окончательно получаем, что плотность распределения $\Gamma(1, n)$, т.е.

$$f_G(g) = \frac{g^{n-1} e^{-g}}{\Gamma(n)}, \quad g > 0. \text{ Отсюда и из формулы (5.2) получаем}$$

следующий метод построения доверительного интервала для θ : при заданном γ выбираем числа $g_1 < g_2$ так, чтобы

$$\frac{1}{\Gamma(n)} \int_{g_1}^{g_2} g^{n-1} e^{-g} dg = \gamma$$

Решая уравнения

$$-\sum_{i=1}^n \ln F(x_i; \theta) = g_1, g_2, \quad (5.5)$$

находим корни $T_1(\bar{x}) < T_2(\bar{x})$. Тогда $(T_1(\bar{X}), T_2(\bar{X}))$ - искомый доверительный интервал для θ .

Наибольшая трудность в применении этой модели к конкретным задачам возникает при нахождении решений уравнений (5.5).

5.3. Доверительный интервал для среднего

Пусть по выборке $\bar{X} = (X_1, \dots, X_n)$ требуется построить доверительный интервал для неизвестного среднего θ , в нормальной модели $N(\theta, \sigma^2)$. Известно, что в соответствии с

теоремой Фишера $L_0\left(\sqrt{n}\frac{\bar{X}-\theta}{\sigma}\right)=N(0,1)$. Следовательно, в

данном случае центральная статистика $G(\bar{X},\theta)=\sqrt{n}\frac{\bar{X}-\theta}{\sigma}$.

Решения уравнений (5.3) имеют вид $T_1(\bar{x})=\bar{x}-\frac{\sigma}{\sqrt{n}}g_2$,

$T_2(\bar{x})=\bar{x}-\frac{\sigma}{\sqrt{n}}g_1$ поэтому γ -доверительным для θ является любой интервал

$$\Delta_\gamma(\bar{X})=\left(\bar{X}-\frac{\sigma}{\sqrt{n}}g_2,\bar{X}-\frac{\sigma}{\sqrt{n}}g_1\right), \quad (5.6)$$

где $g_1 < g_2$ - любые числа, удовлетворяются условию

$$\Phi(g_1)=\Phi(g_2)=\gamma. \quad (5.7)$$

Отметим, что хотя интервал $\Delta_\gamma(\bar{X})$ случаен, его длина постоянна и равна

$l_\gamma(g_1,g_2)=\frac{\sigma(g_2-g_1)}{\sqrt{n}}$, поэтому, чтобы среди всех интервалов

вида (5.6) выбрать кратчайший, надо минимизировать функцию $l_\gamma(g_1,g_2)$ при условии (5.7). Применяя метод Лагранжа нахождения условного экстремума, получаем следующую систему уравнений

$$\begin{cases} \lambda\varphi(g_1)=\frac{\sigma}{\sqrt{n}}; \\ \lambda\varphi(g_2)=\frac{\sigma}{\sqrt{n}}; \\ \Phi(g_2)=\Phi(g_1)=\gamma, \end{cases}$$

где λ - множитель Лагранжа; $\varphi(x)=\Phi'(x)=\frac{1}{\sqrt{2\pi}}e^{-x^2/2}$.

Отсюда находим, что $\varphi(g_1)=\varphi(g_2)$. Так как функция $\varphi(x)$ - чётная, то $g_1=g_2$. Учитывая это, а также последнее уравнение и соотношение $\Phi(-x)=1-\Phi(x)$, получаем равенство $\Phi(g_2)=(1+\gamma)/2$.

Из него находим, что $g_2=U_\gamma=\Phi^{-1}((1+\gamma)/2) - (1+\gamma)/2$ - квантиль стандартного нормального распределения $N(0,1)$

Итак, оптимальным [среди интервалов $\Delta_\gamma(\bar{X})$] γ -доверительным интервалом для параметра θ в модели $N(\theta,\sigma^2)$ является интервал

$$\Delta_\gamma^*(\bar{X}) = \left(\bar{X} - \frac{\sigma}{\sqrt{n}} U_{\frac{1+\gamma}{2}}, \bar{X} + \frac{\sigma}{\sqrt{n}} U_{\frac{1+\gamma}{2}} \right), \quad (5.8)$$

т.е. симметричный относительно случайной точки \bar{X} интервал

длины $\frac{2\sigma U_{\frac{1+\gamma}{2}}}{\sqrt{n}}$, $U_{\frac{1+\gamma}{2}}$ - квантиль стандартного нормального распределения порядка $(1+\gamma)/2$.

5.4. Доверительный интервал для дисперсии

Построим доверительный интервал для неизвестной дисперсии θ^2 в модели $N(m,\theta^2)$. Легко найти центральную статистику для $\tau=\tau(\theta)=\theta^2$:

$$G(\bar{X}; \tau) = \frac{1}{\tau} \sum_{i=1}^n (X_i - m)^2.$$

Действительно, так как $L_\theta\left(\frac{X_i - m}{\theta}\right) = N(0,1)$, то

$$L_\theta\left(\frac{(X_i - m)^2}{\theta^2}\right) = \chi^2(1) - \text{стандартное } \chi^2\text{-распределение.}$$

Следовательно, $L_\theta(G(\bar{X}; \tau)) = \chi^2(n)$. Здесь $T_1(\bar{x}) = \frac{1}{g_2} \sum_{i=1}^n (x_i - m)^2$,

$T_2(\bar{x}) = \frac{1}{g_1} \sum_{i=1}^n (x_i - m)^2$ - решения (относительно τ) уравнений

$G(\bar{x}; \tau) = g_1, g_2$. Следовательно γ -доверительным для $\tau = \theta^2$ является в данном случае любой интервал.

$$\Delta_{\gamma}(\bar{X}) = \left(\frac{1}{g_1} \sum_{i=1}^n (X_i - m)^2, \frac{1}{g_2} \sum_{i=1}^n (X_i - m)^2 \right), \quad (5.9)$$

где $g_1 < g_2$ находят из условия

$$\int_{g_1}^{g_2} f_{\chi_n^2}(x) dx = \gamma$$

[$f_{\chi_n^2}(x)$ - плотность распределения хи-квадрат].

Как правило, g_1 и g_2 надо выбирать так, чтобы выполнялись равенства

$$\int_0^{g_1} f_{\chi_n^2}(x) dx = \frac{1-\gamma}{2}, \quad \int_{g_2}^{\infty} f_{\chi_n^2}(x) dx = \frac{1+\gamma}{2}, \quad (5.10)$$

т.е. $g_1 = \chi_{\frac{1-\gamma}{2}, n}^2$; $g_2 = \chi_{\frac{1+\gamma}{2}, n}^2$, где $\chi_{p, n}^2$ - p - квантиль

распределения $\chi^2(n)$. В этом случае соответствующий доверительный интервал называют иногда **центральным**. Соотношения (5.9), (5.10) определяют правила доверительного оценивания неизвестной дисперсии в модели $N(m, \theta^2)$. Задача отыскания наикратчайшего интервала среди интервалов вида (5.9) сводится к минимизации отношения g_2/g_1 при условии

$$\int_{g_1}^{g_2} f_{\chi_n^2}(x) dx = \gamma \text{ или, если положить } g_1 = \chi_{\alpha_1, n}^2; \quad g_2 = \chi_{(1-\alpha_2), n}^2$$

(где $\alpha_1 + \alpha_2 = 1 - \gamma$) к уравнению:

$$\frac{\chi_{1-\alpha_2, n}^2}{\chi_{\alpha_1, n}^2} = \exp \left\{ \frac{\left(\chi_{1-\alpha_2, n}^2 - \chi_{\alpha_1, n}^2 \right)}{n} \right\}. \quad (5.11)$$

Значения α_1 и α_2 , удовлетворяющие (5.11), определяют оптимальный γ -доверительный интервал вида (5.9)

$$\Delta_{\gamma}^*(\bar{X}) = \left(\frac{1}{\chi_{1-\alpha_2, n}^2} \sum_{i=1}^n (X_i - m)^2, \frac{1}{\chi_{\alpha_1, n}^2} \sum_{i=1}^n (X_i - m)^2 \right). \quad (5.9')$$

Таким образом, центральный интервал в данном случае не является наикратчайшим. Нужно заметить, что центральной

статистикой для оценивания среднеквадратичного отклонения

$$\theta \text{ является, очевидно } G(\bar{X}; \theta) = \frac{1}{\theta} \left[\sum_{i=1}^n (X_i - m)^2 \right]^{1/2}.$$

5.5. Доверительные интервалы для среднего и дисперсии

Рассмотрим доверительное оценивание параметров в общей нормальной модели $N(\theta_1; \theta_2^2)$. Из теоремы Фишера следует, что

$$G(\bar{X}; \theta_2^2) = n \cdot S^2(\bar{X}) / \theta_2^2$$

– центральная статистика для оценивания дисперсии θ_2^2 . Здесь $S^2(\bar{X})$ – выборочная дисперсия. Доверительный интервал для θ_2^2 находим по схеме предыдущего параграфа. Окончательно получаем: центральным γ - доверительным интервалом для θ_2^2 является интервал

$$\left(\frac{n \cdot S^2(\bar{X})}{\chi_{\frac{(1+\gamma)}{2}, n-1}^2}, \frac{n \cdot S^2(\bar{X})}{\chi_{\frac{(1-\gamma)}{2}, n-1}^2} \right).$$

В частности, для выборки объёма $n=10$ и доверительной вероятности $\gamma=0.9$ имеем

$$\chi_{0,05;9}^2 = 3.3251; \chi_{0,95;9}^2 = 16.919.$$

Поэтому центральный 0.9 - доверительный интервал для θ_2^2 имеет вид

$$(0.5911 S^2(\bar{X}), 3.007 S^2(\bar{X})).$$

Наикратчайший в данном случае интервал

$$\left(\frac{n \cdot S^2(\bar{X})}{\chi_{1-\alpha_2, n-1}^2}, \frac{n \cdot S^2(\bar{X})}{\chi_{\alpha_1, n-1}^2} \right), \alpha_1 + \alpha_2 = 1 - \gamma, \quad (5.12)$$

где $\chi_{\alpha, n-1}^2$ и $\chi_{1-\alpha_2; n-1}^2$ определяются соотношением (5.11), в котором n заменено на $(n-1)$.

В силу теоремы 4.3 центральной статистикой для оценивания среднего θ_1 является

$$G(\bar{X}; \theta_1) = \sqrt{n-1} \cdot \frac{\bar{X} - \theta_1}{S(\bar{X})},$$

где \bar{X} - выборочное среднее, $S(\bar{X})$ - выборочная дисперсия, причём распределение этой статистики (распределение Стьюдента $S(n-1)$) симметрично относительно своей средней точки. Расчёт доверительного интервала проводится также как и в параграфе 5.3.. Окончательно получаем: γ - доверительным для θ_1 является интервал

$$\left(\bar{X} - \frac{S(\bar{X})}{\sqrt{n-1}} \cdot t_{\gamma, n-1}, \bar{X} + \frac{S(\bar{X})}{\sqrt{n-1}} \cdot t_{\gamma, n-1} \right), \quad (5.13)$$

где $t_{\gamma, n-1}$ - $(1+\gamma)/2$ - квантиль распределения $S(n-1)$.

Построенный интервал имеет минимальную длину среди всех γ -доверительных интервалов вида

$$(\bar{X} - a_1 \cdot S(\bar{X}), \bar{X} + a_2 \cdot S(\bar{X}))$$

Например, $t_{0,95;9}=2.262$, поэтому для выборки объёма $n=10$ и доверительной вероятности $\gamma=0.55$ интервал (5.13) имеет вид

$$(\bar{X} - 0.754 \cdot S(\bar{X}), \bar{X} + 0.754 \cdot S(\bar{X})).$$

Итак, получены формулы доверительных интервалов для параметров нормально распределённой генеральной совокупности. Эти формулы представлены в таблице.

Параметры	Доверительный интервал, доверительные вероятности γ уровень значимости $q=1-\gamma$
m σ^2 - известно	$\bar{X} - \frac{\sigma}{\sqrt{n}} \cdot U_{\frac{1+\gamma}{2}} < m < \bar{X} + \frac{\sigma}{\sqrt{n}} \cdot U_{\frac{1+\gamma}{2}}$
m σ^2 - неизвестно	$\bar{X} - \frac{S(\bar{X})}{\sqrt{n-1}} \cdot t_{\gamma, n-1} < m < \bar{X} + \frac{S(\bar{X})}{\sqrt{n-1}} \cdot t_{\gamma, n-1}$
σ^2 m - известно	$n \cdot S^2(\bar{X}) / \chi_{n, \frac{1+\gamma}{2}}^2 < \sigma < n \cdot S^2(\bar{X}) / \chi_{n, \frac{1-\gamma}{2}}^2$
σ^2 m - неизвестно	$n \cdot S^2(\bar{X}) / \chi_{\frac{1+\gamma}{2}, n-1}^2 < \sigma < n \cdot S^2(\bar{X}) / \chi_{\frac{1-\gamma}{2}, n-1}^2$

5.6. Построение доверительного интервала с использованием распределения точечной оценки параметра

Если имеется некоторая точечная оценка $T=T(\bar{X})$ для параметра θ и известна её функция распределения $F_T(t; \theta)$, то доверительный интервал можно построить, основываясь на этой функции.

Пусть распределение оценки T непрерывно, функция $F_T(t; \theta)$ - непрерывна и монотонна по θ . Пусть заданна доверительная вероятность γ . Определим при каждом $\theta \in \Theta$ числа $t_i = t_i(\theta)$ $i=1, 2$, где $t_1 < t_2$ и

$$P_{\theta}(t_1 < T(\bar{X}) < t_2) = F_T(t_2; \theta) - F_T(t_1; \theta) = \gamma. \quad (5.14)$$

Чтобы данная процедура была однозначной, числа выбирают так, чтобы выполнялись условия

$$F_T(t_1; \theta) = (1-\gamma)/2 \quad 1 - F_T(t_2; \theta) = (1-\gamma)/2, \quad (5.15)$$

т.е. речь идёт о построении центрального доверительного интервала. Обозначим через D_γ подмножество $\Theta \times \Theta$:

$$D_\gamma = \{(\theta, \theta') : t_1(\theta) < \theta' < t_2(\theta)\}.$$

Тогда $P_\theta((\theta, T(\bar{X})) \in D_\gamma) = \gamma$ при любом $\theta \in \Theta$. Определим теперь при фиксированном θ' сечении $D_\gamma(\theta')$ множества D_γ : $D_\gamma(\theta') = \{(\theta, \theta') \in D_\gamma\}$ и рассмотрим случайное множество $D_\gamma(T(\bar{X})) \subset \Theta$. Событие $\theta \in D_\gamma(T(\bar{X}))$ происходит тогда и только тогда, когда $(T(\bar{X})) \in (t_1(\theta), t_2(\theta))$ и следовательно, при каждом θ имеет вероятность γ . Таким образом, построено случайное множество $D_\gamma(T(\bar{X}))$, которое накрывает истинное значение параметра с вероятностью γ . Если это множество является интервалом, то построен γ -доверительный интервал для θ . Это имеет место, если кривые $\theta' = t_i(\theta)$, $i=1,2$ являются монотонными, одного типа (т.е. одновременно либо возрастают, либо убывают), что обеспечивается условием непрерывности и монотонности функции $F_T(t; \theta)$. Таким образом при сделанных предположениях множество $D_\gamma(\theta')$ при каждом θ' представляет собой интервал, следовательно, определены его концы $\theta_1(\theta') < \theta_2(\theta')$, а тем самым и соответствующий интервал $(T_1(\bar{X}), T_2(\bar{X}))$ для θ , где $(T)_i(\bar{X}) = \theta_i(T(\bar{X}))$, $i=1,2$.

Диаграмма придаёт наглядный смысл методике. Строят диаграмму по вертикали (для абсцисс θ из (5.15) находят ординаты t_1 и t_2).

Читают же её по горизонтали, т.е. для наблюдавшейся ординаты $t = T(\bar{x})$ (\bar{x} - реализация выборки \bar{X}) "считывают" две величины θ_1 и θ_2 и утверждают, что $\theta \in (\theta_1, \theta_2)$. Если "читать" диаграмму по вертикали, то границы области D_γ описываются парой переменных точек (θ, t_1) и (θ, t_2) таких, что выполняется условие (5.15). Если "читать" диаграмму по горизонтали, то границы D_γ можно описать парой переменных точек (t, θ_1) и (t, θ_2) , взяв за независимую переменную наблюдаемое значение оценки t .

Итак, алгоритм построения центрального γ - доверительного интервала для θ в случае, когда функция распределения $F_T(t; \theta)$ оценки $T = T(\bar{X})$ непрерывна и монотонна по θ , состоит в следующем. Пусть $t = T(\bar{x} :)$ - наблюдавшееся значение оценки. Решая относительно θ уравнение:

$$F_T(t; \theta) = (1 - \gamma)/2, \quad (1 + \gamma)/2 \quad (5.16)$$

найдем два числа $\theta_1 < \theta_2$. Утверждаем, что $\theta \in (\theta_1, \theta_2)$. Рассмотренная теория гарантирует, что при γ , близком к 1, вероятность ошибки равна $1 - \gamma$.

Доверительный интервал можно построить и для дискретной модели. Из-за ступенчатости функции распределения $F_T(t; \theta)$ выполняются следующие неравенства θ' :

$$P_\theta(t_1 < T(\bar{X}) < t_2) = F_T(t_2 - 0; \theta) - F_T(t_1; \theta) \geq \gamma. \quad (5.14')$$

Вместо условий (5.15) вводим условия (5.15')

$$F_T(t_1; \theta) \leq (1 - \gamma)/2, \quad 1 - F_T(t_2 - 0; \theta) \leq (1 - \gamma)/2, \quad (5.15')$$

где t_2 - наибольшее, t_1 - наименьшее значения T , удовлетворяющие этим неравенствам. Кривые $\theta' = t_i(\theta)$, $i = 1, 2$ в данном случае будут ступенчатыми. Алгоритм построения центрального γ - доверительного интервала для θ в дискретном случае тот же, что и в непрерывном, только вместо уравнения (5.16) надо решать относительно θ уравнения (5.16')

$F_T(t; \theta) = (1 - \gamma)/2, \quad 1 - F_T(t - 0; \theta) = (1 - \gamma)/2, \quad (5.16')$
где t - наблюдавшееся значение оценки T .

5.7. Асимптотические доверительные интервалы

Если имеется состоятельная и асимптотически нормальная оценка $T_n = T_n(\bar{X})$ для параметра θ , можно приближенно решить (при больших n) задачу доверительного оценивания.

Пусть при $n \rightarrow \infty$ имеет место соотношение

$$L_\theta(\sqrt{n}(T_n - \theta)) \rightarrow N(0, \sigma^2(\theta)), \quad \forall \theta \in \Theta,$$

причём $\sigma^2(\theta)$ - непрерывная функция. Тогда из теоремы об асимптотической нормальности и эффективности оценки МП следует, что при $n \rightarrow \infty$ и всех γ

$$P_{\theta} \left(\frac{\sqrt{n}|T_n - \theta|}{\sigma(T_n)} < U_{\gamma} \right) \rightarrow \Phi(U_{\gamma}) - \Phi(-U_{\gamma}) = 2\Phi(U_{\gamma}) - 1 = \gamma,$$

здесь $U_{\gamma} = \Phi^{-1} \left(\frac{1 + \gamma}{2} \right)$

Переписав это соотношение в виде:

$$P_{\theta} \left(T_n - \frac{U_{\gamma} \sigma(T_n)}{\sqrt{n}} < \theta < T_n + \frac{U_{\gamma} \sigma(T_n)}{\sqrt{n}} \right) \rightarrow \gamma,$$

получаем, что $\left(T_n \pm \frac{U_{\gamma} \sigma(T_n)}{\sqrt{n}} \right)$ - асимптотический γ -

доверительный интервал для θ . Асимптотическая дисперсия $\sigma^2(\theta) / n$ характеризует разброс распределения статистики T_n около θ .

Интервал тем уже, чем выше асимптотическая эффективность оценки (чем меньше $\sigma(\theta)$). Асимптотически кратчайший доверительный интервал будет порождаться асимптотически эффективной оценкой. Если исходная модель F регулярна, то перечисленными свойствами обладают оценки максимального правдоподобия. Таким образом, асимптотически кратчайшим γ - доверительным интервалом для θ с учётом того, что $L_{\theta}(\sqrt{n}(\theta_m^* - \theta)) \rightarrow N(0, 1/i(\theta))$,

где $i(\theta)$ -функция информации, является интервал

$$\left(\theta_m^* - \frac{U_{\gamma}}{\sqrt{ni(\theta_m^*)}}, \theta_m^* + \frac{U_{\gamma}}{\sqrt{ni(\theta_m^*)}} \right), \quad U_{\gamma} = \Phi^{-1} \left(\frac{1 + \gamma}{2} \right).$$

Если распределение генеральной совокупности не является нормальным, то в отдельных случаях по выборкам большого объёма можно построить доверительные интервалы

для неизвестных параметров приближённо, используя предельные теоремы теории вероятностей и вытекающие из них асимптотические распределения и оценки.

Пример 1. (доверительные интервалы для вероятности успеха в схеме Бернулли). Пусть в n независимых испытаниях успех наступил x раз. Найти доверительный интервал для вероятности p успеха в одном испытании.

Решение: Эффективной оценкой вероятности успеха p в одном испытании является относительная частота $\tilde{p} = h = \frac{x}{n}$. По теореме Муавра-Лапласа относительная частота h имеет асимптотически нормальное распределение $N(p, \sqrt{pq/n})$, где $q=1-p$.

Рассмотрим статистику $U = \frac{(h-p)}{\sqrt{pq/n}}$, которая, следовательно, имеет асимптотически нормальное распределение $N(0,1)$ независимо от значения p . При $n \rightarrow \infty$ имеем:

$$P \left[\left| \frac{h-p}{\sqrt{pq/n}} \right| < U_{\frac{1+\gamma}{2}} \right] \approx \gamma$$

Отсюда получаем, что с вероятностью $\approx \gamma$ выполняется неравенство

$$h - U_{\frac{1+\gamma}{2}} \sqrt{\frac{pq}{n}} < p < h + U_{\frac{1+\gamma}{2}} \sqrt{\frac{pq}{n}} \quad (5.17)$$

Заменяя значения p и q в левых и правых частях неравенства (5.17) их оценками $\tilde{p} = h$ и $\tilde{q} = 1-h$, получаем, что доверительный интервал для вероятности успеха в схеме Бернулли приближённо имеет вид

$$h - U_{\frac{1+\gamma}{2}} \sqrt{\frac{h(1-h)}{n}} < p < h + U_{\frac{1+\gamma}{2}} \sqrt{\frac{h(1-h)}{n}}. \quad (5.18)$$

Пример 2. При проверке 100 деталей из большой партии обнаружено 10 бракованных деталей.

а) Найти 95% приближённый доверительный интервал доли бракованных деталей во всей партии.

б) Какой минимальный объём выборки следует взять для того, чтобы с вероятностью 0.95 утверждать, что доля бракованных деталей во всей партии отличается от частоты появления бракованных деталей в выборке не более, чем на 1% .

Решение:

а) Оценка доли бракованных деталей в партии по выборке равна $\hat{p} = h = 10/100 = 0.1$. По таблице квантилей нормального распределения [12] находим квантиль $U_{0.975} = 1.96$. По формуле (5.18) 95-% интервал приближённо имеет вид $0.041 < p < 0.159$.

б) Представим доверительный интервал (5.18) в виде

неравенства $|h - p| < U_{\frac{1+\gamma}{2}} \sqrt{\frac{h(1-h)}{n}}$, которое выполняется с

вероятностью $\gamma = 0.95$. Так как по условию задачи $|x - p| \leq 0.01$, то для определения n получим неравенство

$U_{0.975} \sqrt{\frac{h(1-h)}{n}} \leq 0.01$, отсюда следует, что

$1.96 \sqrt{\frac{0.1 \cdot (1-0.1)}{n}} \leq 0.01$ и $n \geq (0.3 \cdot 1.96)^2 = 3457.44$. Значит,

минимальный объём выборки $n = 3458$

Задачи и решения

Интервальное оценивание. Доверительные интервалы

В рассматриваемых задачах предполагается, что выборка объёма n получена из генеральной совокупности, имеющей либо нормальное распределение, либо распределение, достаточно близкое к нормальному.

В задачах 40, 41 выборочные оценки определились по результатам n наблюдений. Используя формулы, приведенные в таблице, найти 90%- и 99%-ные доверительные интервалы

для математического ожидания (среднего) следующих характеристик:

Задача 40

Емкость конденсатора, если $\bar{x} = 20$ мкФ, $n=16$, среднеквадратичное отклонение известно и равно 4мкФ

Решение:

$$\bar{x} = 20, n = 16, \sigma^* = u;$$

$$P_1 = \frac{1+0,9}{2} = 0,95$$

$$P_2 = \frac{1+0,99}{2} = 0,995$$

$$x_j^{0,5} = 1,645$$

$$x_{0,995} = 2,576$$

$$\bar{x} - \frac{x_{0,95}\sigma}{\sqrt{n}} < M < \bar{x} + \frac{x_{0,95}\sigma}{\sqrt{n}}$$

$$20 - \frac{1,645u}{\sqrt{16}} < M < 20 + \frac{1,645u}{\sqrt{16}} \Rightarrow M \subset (18,35; 21,64)$$

$$20 - \frac{2,576u}{\sqrt{16}} < M < 20 + \frac{2,576u}{\sqrt{16}} \Rightarrow M \in (17,424; 22,576)$$

Ответ : (18,35; 21,64), (17,424; 22,576);

Задача 41

Время безотказной работы электронной лампы, если $\bar{x} = 500$, $n=100$ среднеквадратичное отклонение известно и равно 10 часов.

Решение:

$$\bar{x} = 500, n = 100, \sigma = 10$$

$$1) \quad 500 - \frac{1,645 * 10}{10} < M < 500 + \frac{1,645 * 10}{10}$$

$$498,355 < M < 501,645$$

$$2) \quad 500 - \frac{2,576 * 10}{10} < M < 500 + \frac{2,576 * 10}{10}$$

$$497,424 < M < 502,576$$

Ответ: (498,355;501,645), (497,424;502,576);

Пусть из одной генеральной совокупности получены две выборки объемов n_1 и n_2 соответственно. Выборочные оценки средних и дисперсий по этим выборкам равны $\bar{x}_1, \bar{x}_2; S_1^2, S_2^2$. Объединенные оценки среднего и дисперсии по выборке объема $n_1 + n_2$ вычисляются по формулам

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}, \quad S^2 = \frac{(n_1 - 1) S_1^2 + (n_2 - 1) S_2^2}{n_1 + n_2 - 2}.$$

Показать, что если дисперсия генеральной совокупности известна и равна σ^2 , то доверительный интервал для среднего определяется так:

$$\bar{x} - \frac{S}{\sqrt{n_1 + n_2}} t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2) < m < \bar{x} + \frac{S}{\sqrt{n_1 + n_2}} t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2). (*)$$

Для уточнения характеристик, приведенных в задачах 40, 41 были проделаны повторные эксперименты и получены новые выборочные оценки. Найти 90%- и 99%-ные доверительные интервалы для среднего, используя формулу (*).

Задача 42

Емкость конденсатора, если $n=16$, $\bar{X} = 18$ мкФ.

Решение:

$$x_1 = 20 \quad n_1 = 16$$

$$x_2 = 18 \quad n_2 = 19 \quad \sigma = 4$$

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{20 \cdot 16 + 9 \cdot 18}{25} = 19,28$$

$$19,28 - \frac{4 \cdot 1,645}{5} < M < 19,28 + \frac{4 \cdot 1,645}{5}$$

$$M \subset (17,964; 20,596)$$

$$19,28 - \frac{4 \cdot 2,576}{5} < M < 19,28 + \frac{4 \cdot 2,576}{5}$$

$$M \in (17,2192; 21,3408)$$

$$\text{Ответ: } M \subset (17,964; 20,596); \quad M \in (17,22; 21,34)$$

Задача 43

Время безотказной работы электронной лампы, если $n=64$, $\bar{X} = 480$ ч.

Решение:

$$\bar{x}_1 = 500 \quad \bar{x}_2 = 480$$

$$n_1 = 100 \quad n_2 = 64 \quad \sigma = 10$$

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{500 \cdot 100 + 480 \cdot 64}{164} = 492,19$$

$$492,19 - 10 \frac{1,645}{\sqrt{164}} < M < 492,19 + 10 \frac{1,645}{\sqrt{164}}$$

$$492,19 - 1,285 < M < 492,19 + 1,285$$

$$M \in (490,905; 493,478)$$

$$492,19 - \frac{10 \cdot 2,576}{12,8} < M < 492,19 + \frac{10 \cdot 2,576}{12,8}$$

$$M \in (490,177; 494,2025)$$

$$\text{Ответ: } M \in (490,905; 493,478); \quad M \in (490,177; 494,2025)$$

Задача 44

Диаметр вала, если $n_1=9$, $\bar{x}=30$ мм, $S^2=9$ мм², $n_2=16$,
 $\bar{x}=29$ мм, $S^2=4,5$ мм².

Решение:

$$n_1 = 9; \quad n_2 = 16; \quad \bar{x}_1 = 30; \quad \bar{x}_2 = 29; \quad S_1^2 = 9; \quad S_2^2 = 4,5.$$

$$\bar{x} - \frac{S}{\sqrt{n_1 + n_2}} t_1 - \frac{\alpha}{2} (n_1 + n_2 - \alpha) < m < \bar{x} + \frac{S}{\sqrt{n_1 + n_2}} t_1 - \frac{\alpha}{2} (n_1 + n_2 - \alpha)$$

$$1) \quad 1 - \frac{\alpha}{2} = 0,90; \quad t_1 - \frac{\alpha}{2} = 1,714;$$

$$2) \quad 1 - \frac{\alpha}{2} = 0,99; \quad t_1 - \frac{\alpha}{2} = 1,807;$$

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{9 \cdot 30 + 16 \cdot 29}{25} = 29,36$$

$$S^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{8 \cdot 9 + 15 \cdot 4,5}{23} = 6,06$$

$$28,51 < m < 30,20$$

$$27,98 < m < 30,74$$

$$\text{Ответ: } (28,51; 30,20); \quad (27,98; 30,74)$$

Задача 45

По данным задачи 44 найти 90%- и 95%-ный доверительные интервалы для дисперсии.

Решение:

$$n = 16; \quad \bar{x} = 29; \quad S^2 = 45; \quad \gamma = (0,9; 0,95);$$

$$\frac{nS^2}{\chi_{\frac{1-\gamma}{2}}^2(n)} < \sigma^2 < \frac{nS^2}{\chi_{\frac{1+\gamma}{2}}^2(n)}$$

$$\frac{1-\gamma}{2} = 0,05; \quad \frac{1+\gamma}{2} = 0,95; \quad \frac{1-\gamma}{2} = 0,025; \quad \frac{1+\gamma}{2} = 0,975;$$

$$\frac{16 \cdot 4,5}{26,2} < \sigma^2 < \frac{16 \cdot 4,5}{7,962} \Rightarrow \sigma^2 \in (2,74; 9,04)$$

$$\chi_{0,95}^2(16) = 7,962 \quad \chi_{0,05}^2(16) = 26,2;$$

$$\chi_{0,975}^2(16) = 6,67 \quad \chi_{0,025}^2(16) = 29,3;$$

$$\frac{16 \cdot 4,5}{29,3} < \sigma^2 < \frac{16 \cdot 4,5}{6,67} \quad \sigma^2 \in (2,457; 10,79)$$

Ответ: (2,7; 9,04); (2,45; 10,79);

Задача 46

С автоматической линии, производящей подшипники, было отобрано 400 штук, причем 10 оказалось бракованными. Найти 90%-ный доверительный интервал для вероятности появления бракованного подшипника. Сколько подшипников надо проверить, чтобы с вероятностью 0,9973 можно было утверждать, что вероятность появления бракованного подшипника не отличается от частоты более чем на 5%?

Решение:

$$n = 400; \quad \gamma = 0,9; \quad u_\gamma = \Phi^{-1}(0,9) = 1,67;$$

$$\hat{p} = \frac{10}{400} = \frac{1}{40}$$

$$\frac{1}{40} - 1,67 \sqrt{\frac{1}{40} \cdot \frac{39}{40}} < p < \frac{1}{40} + 1,67 \frac{\sqrt{39}}{40 \cdot 20}$$

$$p \in (0,012; 0,038) \text{ т.к. } (p^2 - p) \leq 0,05; \quad (\hat{p} - p) < u_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow$$

$$\Rightarrow 3 \sqrt{\frac{1}{10} \cdot \frac{39}{10}} \leq 0,05 \quad \sqrt{n} 3 \frac{3\sqrt{39}}{40 \cdot 0,05} \approx 9,367$$

$$n \geq 87,75$$

Ответ: $(0,012; 0,038) \quad n > 88;$

Лабораторная работа № 5

Целью лабораторной работы является изучение интервального оценивания параметров нормального распределения в пакете MATHCAD.

Точечные оценки дают приближенное значение неизвестного (оцениваемого) параметра. Сама оценка является случайной величиной, и если известно ее распределение или хотя бы дисперсия, то можно указать пределы, в которых с достаточно большой вероятностью лежит неизвестное значение параметра. Эти пределы легко вычисляются через дисперсию. Важно понимать, что пользоваться полученными значениями пределов можно, только если они не зависят от самого оцениваемого параметра.

ЗАДАНИЕ

Найдите доверительные интервалы для математического ожидания MX и дисперсии DX по заданной выборке x_1, x_2, \dots, x_n из нормального распределения.

Порядок выполнения задания

1. Определите и введите компоненты вектора выборочных значений случайной величины.

2. Вычислите точечные оценки MX и DX .
3. Вычислите 95%-ный доверительный интервал для математического ожидания при неизвестной дисперсии.
4. Вычислите 90%-ный доверительный интервал для дисперсии.

Пример выполнения задания

Найдите доверительные интервалы для математического ожидания и дисперсии приведенной выборки из нормального распределения.

x	904.3	910.2	916.6	928.8	935.0	941.2
n	3	1	2	7	8	10
x	947.4	953.6	959.8	966.0	972.2	978.4
n	4	2	4	1	1	1

Фрагмент рабочего документ MATHCAD с вычислениями доверительных интервалов представлен ниже (в приведенном фрагменте опущено определение массива DX , который во втором столбце содержит значения случайной величины, а в первом – их количество в выборке).

ORIGIN :=1

i:=1 12

$$n := \sum_{i=1}^{12} D_{i,1} \quad n = 44$$

$$Mx := \frac{1}{n} \sum_{i=1}^{12} D_{i,1} D_{i,2} \quad Mx = 938.693$$

$$Dx := \frac{1}{n-1} \sum_{i=1}^{12} D_{i,1} (D_{i,2} - Mx)^2 \quad Dx = 282.988$$

95%-ный доверительный интервал для математического ожидания

$$t := qt\left(1 - \frac{0.05}{2}, n\right) \quad t = 2.015$$

$$xl := Mx - t\sqrt{\frac{Dx}{n}} \quad xl = 933.582$$

$$xr := Mx + t\sqrt{\frac{Dx}{n}} \quad xr = 943.804$$

90%-ный доверительный интервал для дисперсии

$$hl := qchlsq\left(\frac{0.1}{2}, n - 1\right) \quad hl = 28.965$$

$$hr := qchlsq\left(1 - \frac{0.1}{2}, n - 1\right) \quad hr = 59.304$$

$$dl := Dx \frac{(n - 1)}{hr} \quad dl = 205.19$$

$$dr := Dx \frac{(n - 1)}{hl} \quad dr = 420.114$$

Таким образом, найдены 95%-ный доверительный интервал для математического ожидания (933.582, 943.804) и 90%-ный доверительный интервал для дисперсии (205.19, 420.114).

ЗАКЛЮЧЕНИЕ

Прикладная математическая статистика – это наука о том, как обрабатывать данные. Методы прикладной математической статистики активно применяются в технических исследованиях, экономике, теории и практике управления (менеджмента), социологии, медицине, геологии, истории и психологии т.д.

Часто, закончив эксперимент и получив достаточное количество наблюдений, экспериментатор сталкивается с необходимостью все это как-то обобщить и сделать правильные выводы из массы разрозненных данных. Статистическая обработка данных приводит порой к далеко идущим выводам и позволяет делать достаточно уверенные прогнозы, выявить закономерности в череде, казалось бы, случайных событий. С результатами наблюдений, измерений, испытаний, опытов, с их анализом имеют дело специалисты во всех отраслях практической деятельности, почти во всех областях теоретических исследований.

Цель учебного пособия состоит в том, чтобы рассмотреть наиболее важные идеи математической статистики и научить студентов современным методам прикладной статистики на уровне, достаточном для использования этих методов в научной и практической деятельности.

Учебное пособие посвящено основным методам современной прикладной математической статистики и состоит из двух частей. Первая часть содержит пять глав. В первой главе представлены основные понятия выборочной теории и методы обработки статистических данных. Во второй главе рассмотрено оценивание неизвестных параметров распределения, а также критерии оптимальности и эффективности оценок. Третья глава посвящена рассмотрению метода максимального правдоподобия и метода моментов нахождения точечных оценок. В четвертой главе представлены основные сведения о распределениях случайных величин, которые используются в задачах прикладной математической

статистики. Пятая глава посвящена интервальному оцениванию параметров распределений.

В конце каждой главы приводятся задачи с решениями. Задачи предназначены для активного освоения понятий и развития навыков квалифицированной статистической обработки данных. Описание лабораторных работ приводятся в конце каждой главы. Лабораторные работы необходимы для приобретения навыков работы на компьютере, алгоритмизации и программирования задач. В лабораторных работах используются два компьютерных пакета программ: MATHCAD и STATISTICA.

Данное учебное пособие предназначено для бакалавров и магистров.

ПРИЛОЖЕНИЕ 1

ВАРИАНТЫ ВЫБОРОК

Вариант 1

65,3	69,1	56,1	57,1	73,1	74,4	99,9	57,1	97,6	73,4
25,1	36,1	60,1	36,6	61,1	59,1	57,6	70,1	64,1	63,8
90,1	78,1	30,4	70,1	61,6	68,1	77,9	93,6	82,1	49,6
64,9	63,8	64,8	84,1	45,1	47,1	47,5	67,1	85,5	72,3
38,9	65,1	86,3	38,1	62,9	62,5	65,9	93,4	42,1	71,7
66,4	73,7	43,5	57,5	69,5	66,8	63,9	64,5	72,6	62,5
48,6	65,1	42,1	54,8	72,6	62,9	33,5	50,7	43,5	52,7
64,6	41,8	69,1	83,7	44,4	77,8	46,4	58,9	59,1	84,2
61,4	91,8	52,2	84,3	54,4	45,1	58,1	61,2	105,1	78,1
45,1	58,9	79,9	37,1	54,2	47,4	76,9	56,1	32,9	45,1

Вариант 2

54,2	58,0	45,0	46,0	62,2	63,3	88,8	46,0	80,5	62,3
14,0	25,0	49,0	25,5	50,0	48,0	46,5	59,0	53,0	52,7
79,0	67,0	19,3	59,0	50,5	57,0	66,8	82,5	71,0	38,5
53,9	52,8	53,7	73,0	34,0	36,0	26,4	56,0	74,4	61,2
27,8	54,0	75,2	27,0	51,8	51,4	54,8	82,3	31,0	60,6
55,3	62,6	32,4	46,4	58,4	55,7	52,8	53,4	61,5	51,4
37,5	54,0	31,0	43,7	61,5	51,8	22,4	39,6	32,4	41,6
53,5	30,7	58,0	72,6	33,3	66,7	35,2	47,8	48,0	73,1
50,3	80,7	41,1	73,2	43,3	34,0	47,0	50,1	94,0	67,0
34,0	47,8	68,8	26,0	42,8	46,3	68,8	45,0	21,8	34,7

Вариант 3

43,9	55,5	59,3	46,3	47,3	63,5	64,6	90,1	47,3	81,8
63,6	15,3	26,3	50,3	26,8	51,3	49,3	47,8	60,3	54,3

54,0	80,3	68,3	20,6	60,3	20,6	51,8	59,3	68,1	83,8
39,8	55,2	54,1	55,0	74,3	35,3	37,3	72,3	57,3	75,6
62,5	29,1	55,2	76,5	28,3	53,1	52,7	56,1	83,6	92,3
61,9	56,6	63,9	33,7	47,7	59,7	56,0	54,1	54,7	62,8
52,7	38,8	55,3	32,3	45,0	62,8	53,1	23,7	40,9	33,7
42,9	54,8	32,0	59,3	79,9	34,6	68,0	36,5	49,1	74,4
51,6	82,0	42,4	74,5	44,6	35,3	48,3	51,4	95,3	68,3
35,3	49,1	70,1	27,3	44,1	47,6	70,1	46,3	23,1	35,3

Вариант 4

63,2	67,0	54,0	55,0	71,2	72,3	97,8	55,0	89,5	71,3
23,0	34,0	58,0	34,5	59,0	57,0	55,5	68,0	62,0	61,7
88,0	76,0	28,3	68,0	75,8	91,5	80,0	59,5	66,0	47,5
62,9	61,8	62,7	82,0	43,0	45,0	35,4	65,0	83,4	70,2
36,8	63,0	84,2	36,0	60,8	60,4	63,8	91,3	40,0	69,6
64,3	71,6	41,4	55,4	67,4	64,7	61,8	62,4	70,5	60,4
46,5	63,0	40,0	52,7	70,5	60,8	31,4	48,6	41,4	50,6
62,5	39,7	67,0	81,6	42,3	75,7	44,2	56,8	47,0	82,1
59,3	89,7	50,1	82,2	52,3	43,0	56,0	59,1	103,0	76,0
43,0	56,8	77,8	35,0	52,1	55,3	77,8	54,0	30,8	43,0

Вариант 5

49,6	53,4	40,4	41,4	57,6	58,2	84,2	41,4	75,9	57,7
35,6	20,4	44,4	20,9	45,4	43,4	41,9	54,4	48,4	48,1
74,4	62,4	14,7	54,4	52,9	62,2	77,9	66,4	33,9	45,9
49,3	48,2	49,1	68,4	29,4	31,4	21,8	51,4	69,8	56,6
23,2	49,4	70,6	9,4	47,2	46,8	50,2	77,7	26,4	56,0
50,7	58,0	27,8	41,8	53,8	51,1	48,2	48,8	56,9	46,8
32,9	49,4	26,4	39,1	56,9	47,2	17,8	35,0	27,8	37,0
48,9	46,6	53,4	68,0	28,7	62,1	30,6	43,2	43,4	68,5

45,7	76,1	36,5	68,6	38,7	29,4	42,4	45,5	89,4	62,4
29,4	43,2	64,2	21,4	38,2	41,7	64,2	40,4	17,2	29,4

Вариант 6

48,9	52,7	39,7	40,7	56,9	58,0	83,5	40,7	75,2	57,0
8,7	19,7	43,7	20,2	44,7	42,7	41,2	53,7	47,7	47,4
73,7	61,7	14,0	53,8	45,6	51,7	61,5	77,2	65,7	33,2
48,6	47,5	48,4	67,7	28,7	30,7	21,1	50,7	69,1	55,9
22,5	48,7	69,9	21,7	46,5	49,5	77,0	25,7	55,3	46,1
50,0	57,3	27,1	41,1	53,1	50,4	47,5	48,1	56,2	46,1
32,2	48,7	25,7	38,4	56,2	46,5	17,1	34,3	27,1	36,3
48,2	25,4	52,7	67,3	28,0	61,4	29,9	42,5	42,7	67,8
45,0	75,4	35,8	67,9	38,0	28,7	41,7	44,8	88,7	61,7
28,7	42,7	63,5	20,7	37,5	41,0	63,5	39,7	16,5	28,7

Вариант 7

58,7	62,5	49,5	50,5	66,7	67,8	93,3	50,5	85,0	66,8
18,5	29,5	53,5	30,0	54,5	52,5	51,0	63,5	57,5	57,2
83,5	71,5	23,8	63,5	55,0	61,5	71,3	87,0	75,5	43,0
58,4	57,3	58,2	77,5	38,5	40,5	30,9	60,5	78,9	65,7
32,3	58,5	79,7	31,5	56,3	55,9	59,3	86,8	36,5	65,1
59,8	67,1	36,9	62,9	50,9	60,2	57,3	64,7	66,0	55,9
42,0	58,5	35,5	48,2	66,0	56,3	26,9	44,1	36,9	46,1
58,0	35,2	65,2	77,1	37,8	71,2	39,7	52,3	52,5	77,6
54,8	85,2	45,6	77,7	47,8	38,5	51,5	54,6	98,5	71,5
38,5	52,3	73,3	30,5	47,3	50,8	73,3	49,5	26,3	38,5

Вариант 8

58,2	62,0	49,0	50,0	66,2	67,3	92,8	50,0	84,5	66,3
18,0	29,0	53,0	29,5	54,0	51,8	50,5	63,0	57,0	56,7

90,0	71,0	23,3	63,0	54,5	61,0	70,8	86,5	75,0	42,5
57,9	56,8	57,7	77,0	38,0	40,0	30,4	60,0	78,4	65,2
31,8	58,0	79,2	31,0	55,8	55,4	58,8	86,3	35,0	64,6
59,3	66,6	36,4	50,4	62,4	59,7	56,8	57,4	65,5	55,4
41,5	58,0	35,0	47,7	65,5	55,8	26,4	43,6	36,4	45,6
57,5	34,7	62,0	76,6	37,3	70,7	39,2	51,8	52,0	77,1
54,3	84,7	45,1	77,2	47,2	38,0	51,0	54,1	98,0	71,0
38,0	51,8	72,8	30,0	46,8	50,3	72,8	49,0	25,8	38,0

Вариант 9

46,6	50,4	37,4	38,4	54,6	55,7	81,2	38,4	72,9	54,7
6,4	17,4	41,4	17,9	42,4	40,4	38,9	51,4	45,4	45,1
71,4	59,4	11,7	51,4	42,9	49,4	59,2	74,9	63,4	30,9
46,3	45,2	46,1	65,4	26,4	28,4	18,8	48,4	66,8	53,6
20,2	46,4	67,6	19,4	44,2	43,8	47,2	74,7	23,8	53,0
47,7	55,0	24,8	38,8	50,8	48,1	45,2	45,8	53,9	43,8
29,9	46,4	23,4	36,1	53,9	44,2	14,8	32,0	24,8	34,0
45,9	23,1	50,4	65,0	25,7	59,1	27,6	40,2	40,4	65,5
42,7	73,1	33,5	65,6	35,7	26,4	39,4	42,5	86,4	59,4
26,4	40,2	61,2	18,4	35,2	39,0	61,2	37,4	14,2	26,4

Вариант 10

53,8	57,6	44,6	45,6	61,8	62,9	88,4	45,6	80,1	61,9
13,6	24,6	48,6	25,1	49,6	47,6	46,1	58,6	52,6	52,3
78,6	66,6	18,9	58,6	50,1	56,6	66,4	82,1	70,6	38,1
53,5	52,4	53,3	72,6	33,6	35,6	26,0	55,6	74,0	60,8
27,4	53,6	74,8	26,6	51,4	51,0	54,4	81,9	30,6	60,2
54,9	62,2	32,0	46,0	58,0	55,3	52,4	53,0	61,1	51,1
37,1	53,6	30,6	43,3	61,1	51,4	22,0	39,2	32,0	41,2

53,1	30,3	37,6	72,2	32,9	66,3	34,8	47,4	47,6	72,7
49,9	80,3	40,7	72,8	42,9	33,6	46,6	49,7	93,6	66,6
33,6	47,4	68,4	25,6	42,4	45,9	68,4	44,6	21,4	33,6

Вариант 11

58,0	61,8	48,8	49,8	66,0	67,1	92,6	49,8	84,3	66,1
17,8	28,3	52,8	29,3	53,8	51,8	50,3	62,8	56,8	56,5
82,8	70,8	23,1	62,8	54,3	60,8	70,6	86,3	74,8	42,3
57,7	56,6	57,5	76,8	37,8	39,8	30,2	59,8	78,2	65,0
31,6	57,8	79,0	30,8	55,6	55,2	58,2	86,1	34,8	64,1
59,1	66,4	36,2	50,2	62,2	59,5	56,6	57,2	65,3	55,2
41,3	57,8	34,8	47,5	65,3	55,6	26,2	43,4	36,2	45,4
57,3	34,5	61,8	76,4	37,1	70,5	39,0	51,6	51,8	76,9
54,1	84,5	44,9	77,0	47,1	37,8	50,8	53,9	97,8	70,8
37,8	51,6	72,6	29,8	46,6	50,1	72,6	48,8	25,6	37,8

Вариант 12

47,0	50,8	37,8	38,8	55,0	56,1	81,6	38,8	73,3	55,1
6,8	17,8	41,8	18,3	42,8	40,8	39,3	51,8	45,8	45,5
71,8	59,8	12,1	51,8	43,3	49,8	59,6	75,3	63,8	31,3
46,7	45,6	46,5	65,8	26,8	28,8	19,2	48,8	67,2	54,0
20,6	46,8	68,0	19,8	44,6	44,2	47,6	75,1	23,8	53,4
48,1	55,4	25,2	39,2	51,2	48,5	45,6	46,2	54,3	44,2
30,3	46,8	23,8	36,5	44,6	54,3	15,2	32,4	25,2	34,4
46,3	23,5	50,8	65,4	26,1	59,5	28,0	40,6	40,8	65,9
43,1	73,5	33,9	66,0	36,1	26,8	39,8	42,9	86,8	59,8
26,8	40,6	61,6	18,8	35,6	39,1	61,6	37,8	14,6	26,8

Вариант 13

54,7	72,9	38,4	46,6	50,4	37,4	38,4	54,6	55,7	81,2
45,1	45,4	51,4	6,4	17,4	41,4	17,9	42,4	40,4	38,9
30,9	63,4	74,9	71,4	59,4	11,7	51,4	42,9	49,4	59,2
53,6	66,8	48,4	46,3	45,2	46,1	65,4	26,4	28,4	18,8
53,0	23,8	74,7	20,2	46,4	67,6	19,4	44,2	43,8	47,2
43,8	53,9	45,8	47,7	55,0	24,8	38,8	50,8	48,1	45,2
34,0	24,	32,0	29,9	46,4	23,4	36,1	53,9	44,2	14,8
65,5	40,4	40,2	45,9	23,1	50,4	65,0	25,7	59,1	27,6
59,4	86,4	42,5	42,7	73,1	33,5	65,6	35,7	26,4	39,4
26,4	14,2	37,4	26,4	40,2	61,2	18,4	35,2	39,0	61,2

Вариант 14

45,0	62,2	63,3	88,8	54,2	46,0	46,0	80,5	58,0	62,3
49,0	50,0	48,0	46,5	14,0	25,5	59,0	53,0	25,0	52,7
19,3	50,5	57,0	66,8	79,0	59,0	82,5	71,0	67,0	38,5
53,7	34,0	36,0	26,4	53,9	73,0	56,0	74,4	52,8	61,2
75,2	51,8	51,4	,54,8	27,8	27,0	82,3	31,0	54,0	60,6
32,4	58,4	55,7	52,8	55,3	46,4	53,4	61,5	62,6	51,4
31,0	61,5	51,8	22,4	37,5	43,7	39,6	32,4	54,0	41,6
58,0	33,3	66,7	35,2	53,5	72,6	47,8	48,0	30,7	73,1
41,1	43,3	34,0	47,0	50,3	73,2	50,1 ,	94,0	80,7	67,0
68,8	42,8	46,3	68,8	34,0	26,0	45,0	21,8	47,8	34,7

ПРИЛОЖЕНИЕ 2
ДИАМЕТРЫ 200 ВАЛОВ ПРОТОЧЕННЫХ НА СТАНКЕ,
ММ.

d1	d2	d1	d2	d1	d2	d1	d2
13.39	13.33	13.56	13.38	13.43	13.37	13.53	13.40
13.28	13.34	13.50	13.38	13.38	13.45	13.47	13.62
13.53	13.58	13.32	13.27	13.42	13.40	13.57	13.46
13.57	13.36	13.43	13.38	13.26	13.52	13.35	13.29
13.40	13.39	13.50	13.52	13.39	13.39	13.46	13.29
13.29	13.33	13.38	13.61	13.55	13.40	13.20	13.31
13.43	13.51	13.50	13.38	13.44	13.62	13.42	13.54
13.41	13.49	13.42	13.45	13.34	13.47	13.48	13.59
13.55	13.44	13.50	13.40	13.48	13.29	13.31	13.42
13.43	13.26	13.58	13.38	13.48	13.45	13.29	13.32
13.34	13.14	13.31	13.51	13.59	13.32	13.52	13.57
13.23	13.37	13.64	13.30	13.40	13.58	13.24	13.32
13.43	13.58	13.63	13.48	13.34	13.37	13.18	13.50
13.38	13.33	13.57	13.28	13.32	13.40	13.40	13.33
13.34	13.54	13.40	13.47	13.28	13.41	13.39	13.48
13.28	13.46	13.37	13.53	13.43	13.30	13.45	13.40
13.33	13.39	13.56	13.46	13.26	13.35	13.42	13.36
13.43	13.51	13.51	13.24	13.34	13.28	13.37	13.54
13.52	13.23	13.48	13.48	13.54	13.41	13.51	13.44
13.53	13.44	13.69	13.66	13.32	13.26	13.51	13.38

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Андронов А.М. Теория вероятностей и математическая статистика: учебник для вузов / А.М.Андронов, Е.А. Копытов, Л.Я Гринглаз. – СПб.: Питер, 2004. – 461 с
2. Айвазян С.А. Прикладная статистика: Исследование зависимостей / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1985. – 471 с.
- 3.Боровиков В. STATISTICA: искусство анализа данных на компьютере / В. Боровиков. – СПб.: Питер, 2001. – 656 с.
4. Бочаров П.П. Теория вероятностей. Математическая статистика / П.П. Бочаров, А.В. Печинкин. – М.: Гардарика, 1998. – 328 с.
5. Вентцель Е.С. Теория вероятностей: учебник для вузов / Е.С. Вентцель М.: Высш. шк., 1999. – 576 с.
6. Гнеденко Б.В. Курс теории вероятностей / Б.В. Гнеденко. – М.: Эдиториал УРСС, 2001. – 320 с.
7. Ивченко Г. И. Математическая статистика / Г. И. Ивченко, Ю. И. Медведев. – М.: Высш. шк., 1984. – 248 с.
8. Королев В.Ю. Теория вероятностей и математическая статистика: учебник. / В.Ю. Королев. – М.: Проспект, 2006. – 160 с.
9. Математическая статистика: учебник для вузов / под ред. В.С.Зарубина, А.П.Крищенко. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2001. – 424 с.
10. Новикова Н.М. Обработка экспериментальных данных: учеб. пособие / Н.М. Новикова. 2-е изд. Воронеж, ВГТУ, 2010. – 119 с.
11. Орлов А.И. Прикладная статистика: учебник для вузов / А.И.Орлов. – М.: Экзамен, 2004. – 656 с.
- 12 Сборник задач по математике для втузов. Ч.4. Теория вероятностей и математическая статистика: учеб. пособие / под ред. А.В.Ефимова, А.С.Поспелова. – М.: Изд-во Физ.-мат. лит-ра, 2003. – 432 с.

ОГЛАВЛЕНИЕ

Введение	3
1. Основные понятия и элементы выборочной теории	10
1.1. Порядковые статистики и вариационный ряд выборки	12
1.2. Эмпирическая функция распределения	13
1.3. Гистограмма и полигон частот	18
1.4. Определения и свойства выборочных характеристик	20
1.5. Асимптотическое поведение выборочных моментов. Теорема Слуцкого	21
1.6. Асимптотическая нормальность выборочных моментов	23.
Задачи и решения	25
Лабораторная работа № 1. Описательные статистики выборки, Построение простейших статистических графиков в пакете STATISTICA.	49
2. Оценивание неизвестных параметров распределений.	
Точечные оценки и их свойства	62
2.1. Понятие статистической оценки	62
2.2. Состоятельность и несмещенность оценок	63
2.3. Оптимальные оценки. Теорема об оптимальности оценок	66
2.4. Критерии оптимальности оценок, основанные на неравенстве Рао – Крамера	68
2.5. Неравенство Рао - Крамера и эффективные оценки	71
2.6. Достаточные статистики. Теорема факторизации Неймана - Фишера	73
Задачи и решения	76
Лабораторная работа № 2. Точечные оценки параметров распределений в пакете MATHCAD	86

3. Методы получения точечных оценок	89
3.1. Метод максимального правдоподобия (ММП)	89
3.2. Свойства оценок максимального правдоподобия:	90
3.3. Теорема об асимптотической нормальности и эффективности оценок максимального правдоподобия	92
3.4. Метод моментов. Теоремы о свойствах оценок, полученных методом моментов	94
3.5. Цензурирование	97
Задачи и решения	98
Лабораторная работа № 3. Методы получения точечных оценок в пакете MATHCAD	107
4. Распределения случайных величин, используемые в задачах прикладной математической статистики	112
4.1. Нормальное распределение.	112
4.2. Квадратичные и линейные формы от нормальных случайных величин и их свойства	114
4.3. Распределение хи-квадрат.	116
4.4. Распределение Стьюдента (t – распределение)	118
4.5. Распределение Фишера – Снедекора (F – распределение)	120
Лабораторная работа № 4. Распределения непрерывных случайных величин в пакете STATISTICA.	122
5. Интервальные оценки	128
5.1. Понятие доверительного интервала	129
5.2. Построение доверительного интервала с помощью центральной статистики	130
5.3. Доверительный интервал для среднего	132
5.4. Доверительный интервал для дисперсии	134
5.5. Доверительные интервалы для среднего и дисперсии	136

5.6. Построение доверительного интервала с использованием распределения точечной оценки параметра	138
5.7. Асимптотические доверительные интервалы	140
Задачи и решения	143
Лабораторная работа № 5. Интервальное оценивание параметров нормального распределения в пакете MATHCAD.	149
Заключение	152
Приложение 1	154
Приложение 2	160
Библиографический список	161