Е.Алданов

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебное пособие

Нур-Султан 2019

ББК 22.172
УДК 621.391:519.27
А 45

Рецензенты:

Байбеков С.Н. - доктор технических наук, профессор.

Габбасов М.Б. - кандидат физико-математических наук, доцент.

Алданов Е.С.
А 45 Математическая статистика: учебное пособие / Е.Алданов. Изд.: Университета «Туран-Астана», - Нур-Султан, 2019.-136 с.

ISBN 978-601-7817-62-8

Рассмотрены все основные понятия, используемые при применении современных статистических методов. Изложены основные понятия и задачи современной теории математической статистики, служащие теоретической и практической базой для статистических исследований.

Пособие предназначено для обучающихся специальностей "6М070300-Информационные системы" факультета «Бизнес и информационные технологий» ТАУ.

ББК 22.172
УДК 621.391:519.27

Рекомендована к печати Ученым Советом университета «Туран-Астана»

ISBN 978-601-7817-62-8

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Под математической статистикой понимают «раздел математики, посвященный математическим методам сбора, систематизации, обработки и интерпретации статистических данных, а также использование их для научных или практических выводов. Правила и процедуры математической статистики опираются на теорию вероятностей, позволяющую оценить точность и надежность выводов, получаемых в каждой задаче на основании имеющегося статистических данных.

При этом *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.

Исторически первой появились некоторые области статистики объектов нечисловой природы (в частности, задачи оценивания доли брака и проверки гипотез о ней) и одномерная статистика. Математический аппарат для них

проще, поэтому на их примере обычно демонстрируют основные идеи математической статистики.

Лишь те методы обработки данных, т.е. математической статистики, являются доказательными, которые опираются на вероятностные модели соответствующих реальных явлений и процессов. Речь идет о моделях поведения потребителей, возникновения рисков, функционирования технологического оборудования, получения результатов эксперимента, течения заболевания и т.п. 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.

Если по результатам проведенных экспериментов требуется проверить некоторое предположение относительно генеральной совокупности и сделать обоснованный вывод, то используется статистическая проверка гипотез. Например, если сравниваются различные способы лечения, или разные варианты инвестиций, измерений, технологических процессов, рассматриваются вопросы об эффективности нового метода обучения, управления, о значимости математической модели и т.д. Практической реализации эксперимента предшествует этап, на котором исследователь должен четко сформулировать предположение, подлежащее проверке. Предположительное утверждение относительно генеральной совокупности, проверяемое по выборочным данным, называется статистической гипотезой. Далее осуществляется проверка фактического соответствия реальных результатов экспериментов предполагаемой гипотезе. Различают простую и сложную статистические гипотезы.

Простой называют гипотезу, содержащую только одно предположение. *Сложной* называют гипотезу, которая состоит из конечного или бесконечного числа простых гипотез. Простая гипотеза, в отличие от сложной, полностью

определяет теоретическую функцию распределения случайной величины. Например, гипотезы «вероятность появления события в схеме Бернулли равна 1/2», «закон распределения случайной величины – нормальный с параметрами a = 0, σ = 1 » - являются простыми, а гипотезы «вероятность появления события в схеме Бернулли заключена между 0,3 и 0,6», «закон распределения не является нормальным» - сложными.

Итак, статистика – это наука о том, как обрабатывать данные а статистические методы основаны на вероятностных моделях. Они активно применяются в технических исследованиях, экономике, теории и практике управления. А также в социологии, медицине, геологии, истории и т.д. С обработкой результатов наблюдений, измерений, испытаний, опытов, анализов имеют дело специалисты во всех отраслях практической деятельности, почти во всех областях научных исследований.

Инженеры, менеджеры, экономисты, социологи, врачи, психологи, историки успешно применяют интеллектуальные инструменты принятия решений, основанные на вероятности и статистике.

В специальных дисциплинах часто используются вероятностно-статистические методы и модели. Значит, надо уметь в них разобраться.

Цель этого курса – кратко, но на современном научном уровне рассказать об основных вероятностно-статистических понятиях и фактах.

1. ПРЕДМЕТ И МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

1.1. Вводные определения

Цель любого метода научного познания - это изучение внутренних закономерностей, лежащих в основе того или иного процесса, явления, прогнозирование дальнейшего их развития и выбор наиболее рационального способа поведения. В основе научного знания, как правило, лежат наблюдения. При этом наблюдения разделяют на *единичные* и *многократные* (повторные). Единичное наблюдение не вполне учитывает общие закономерности явления, поскольку всегда подвержено воздействию множества случайных факторов.

Пример 1. Анализ качества работы персонального компьютера (ПК) дает большой разброс параметров от одного образца к другому в пределах даже одной заводской партии.

Для изучения случайных явлений используют многократные наблюдения в примерно одинаковых условиях. Под одинаковыми условиями понимают наблюдения при неизменных значениях контролируемых параметров анализируемого процесса или системы.

Отдельного внимания требует вопрос об интерпретации данных наблюдений.

Пример 2*:* По результатам работы партии компьютеров на испытательном стенде выявлены данные о времени безотказной работы: ПК №1 - 600 часов, ПК №2 - 1200 часов, ПК №3 - 1000 часов. Можно ли, основываясь на этих данных, утверждать, что время безотказной работы данной серии не превышает 1200 часов? Очевидно, что нет. Для получения достоверных сведений необходимо существенно увеличить объем наблюдений или расширить анализируемую выборку. Как организовать такие наблюдения? Сколько их должно быть произведено? Какова надежность выводов по серии многократных наблюдений? На все эти вопросы исследователь получает исчерпывающие ответы, пользуясь аппаратом теории вероятностей и математической статистики.

Предметом исследования указанного раздела высшей математики являются стохастические, или случайные, явления, системы и процессы. Природа такой случайности (непредсказуемости) большинства явлений связана с влиянием большого числа не учитываемых факторов, которые меняются от одного наблюдения к другому. Поэтому в любом реальном явлении всегда наблюдаются определенные отклонения в поведении или развитии от существующих общих закономерностей. В указанном смысле закономерность случайного явления значительно беднее самого явления и поэтому не может служить полной характеристикой явлений во всем их многообразии.

Случайные явления подчиняются, в свою очередь, собственным (их называют статистическими) закономерностям. *Статистические закономерности* характеризуют случайные явления в некотором усредненном, статистическом смысле.

Выявление и изучение статистических закономерностей - главная цель теории вероятностей и математической статистики. При этом теория вероятностей оперирует почти исключительно с абстрактными математическими моделями статистических связей и закономерностей. А математическая статистика, как метод научного познания реальных явлений и систем, оперирует с данными эмпирических (опытных) наблюдений, которые впоследствии согласуются с той или иной абстрактной статистической моделью.

1.2. Понятие вариационного ряда

Установление статистических закономерностей случайных процессов и явлений основано на анализе статистических данных - сведений о том, какие значения принял в результате многократных наблюдений контролируемый параметр или признак процесса.

Рассмотрим следующий пример. Наблюдатель, анализирующий квалификацию рабочих сборочного цеха производства ПК, в результате опроса группы из 100 рабочих получил сведения об их тарифных разрядах в виде

последовательности чисел: 5,1,4,5,4,5,2,5,6,... Здесь признаком производственной системы является тарифный разряд, а полученная последовательность чисел образует статистические данные (статистический ряд). Располагая полученные данные в порядке возрастания значения данного признака, получаем видоизмененную последовательность чисел: 1,1,1,1 (4 раза), 2,2,...,2 (6 раз), 3,3,...,3 (12 раз), 4,4,...,4 (16 раз), 5,5,...,5 (44 раза), 6,6,...,6 (18 раз). Такая операция называется *ранжированием* статистических данных, а полученная последовательность чисел образует ранжированный ряд.

Различные значения признака в ранжированном ряде называют *вариантами ряда* (обозначают *x*), а под *варьированием признака* понимают изменение его значения в ранжированном ряде. При этом различают дискретно варьирующие признаки и непрерывно варьирующие признаки. Тарифный разряд - дискретно варьирующий признак. В примере он принимает шесть различных значений. Число, которое показывает, сколько раз встречается вариант *x* в ряде наблюдений, называется *частотой варианта* и обозначается $m_{x.}$ Ранжированный ряд удобно представить в виде таблицы (табл.1.1)

Таблица 1.1

Вариант *x* (тарифный разряд)	Частота m_x (количество рабочих)	Относительная частота w_x
1	4	0.04
2	6	0.06
3	12	0.12
4	16	0.16
5	44	0.44
6	18	0.18
Всего	100	1.00

Здесь относительная частота (частость) варианта *x* определяется по формуле

$$W_x = \frac{m_x}{\sum_x m_x} = \frac{m_x}{n},$$

где n = 100 - объем наблюдений, или длина статистического ряда.

Полученная последовательность частот вариантов x на множестве проведенных наблюдений образует дискретный вариационный ряд (ДВР) как общую форму отображения дискретно варьирующих признаков случайных явлений и систем. Для описания непрерывно варьирующих признаков используется интервальный вариационный ряд (ИВР). Его определение схоже с определением ДВР. Различие связано лишь с интервальным группированием признаков в варианты. Понятие вариационного ряда является ключевым в задачах статистического анализа случайных процессов и систем.

2. КОЛИЧЕСТВЕННЫЕ ПОКАЗАТЕЛИ ВАРИАЦИОННЫХ РЯДОВ

Наряду с вариационным рядом при анализе случайных явлений и систем широко применяются количественные показатели вариаций контролируемого признака, такие, как средняя статистическая величина, эмпирическая дисперсия и др.

2.1. Средняя статистическая величина.

Средняя *статистическая величина*, или среднее выборочное значение, характеризует усредненное на множестве наблюдений значение контролируемого признака. Различают следующие разновидности средних статистических величин: средняя арифметическая величина (САВ); средняя квадратическая (квадратичная) величина и другие.

Из них наиболее применяется САВ, определяемая по формуле

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

Здесь x_i - значение контролируемого признака в i-м наблюдении для $\forall i = 1,..., n.$ В практических расчетах в ориентации на ДРВ определение САВ заменяют на более простое в виде весовой суммы или взвешенного среднего

$$\overline{X} = \frac{\sum_{X} x m_X}{\sum_{X} m_X} = \sum_{X} x W_X,$$

Где символ $\sum_{X}$ обозначает суммирование по всем вариантам признака x; m_x-частота соответствующего варианта; W_x-относительная частота. Применительно к ИВР последнее выражение является упрощенным приближением к определению (2.1), а не точным его эквивалентом.

Средняя арифметическая величина обладает следующими основными свойствами:

а) сумма отклонений результатов наблюдений от САВ тождественно равна нулю, т.е.

$$\sum_{i=1}^{n}\left(x_i-\bar{x}\right)\equiv 0;$$

б) если все результаты наблюдений уменьшить (увеличить) на константу *c*, то САВ измениться ровно на *c*, так как

$$\overline{(x-c)}=\sum_{X}(x-c)W_X=\sum_{X}xW_X-c\sum_{X}W_X=\bar{x}-c;$$

в) если все результаты наблюдений увеличить (уменьшить) в *k* раз, то результирующая САВ также увеличиться (уменьшиться) в k раз:

$$\overline{(xk)}=k\sum_{X}xW_X=k\bar{x};$$

г) средняя арифметическая величина однотипного признака для суммы (разности) двух статистических рядов {x} и {y} равна сумме (разности) САВ этих рядов в отдельности, поскольку

$$\overline{(x+y)}=\sum_{i=1}^{n}\frac{(x_i+y_i)}{n}=\bar{x}+\bar{y}.$$

2.2. Дисперсия наблюдений

Дисперсия наблюдений, или эмпирическая дисперсия, определяется по выражению

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

Она характеризует среднюю интенсивность (мощность) вариаций контролируемого признака относительно его среднего арифметического значения. По аналогии с САВ строгое определение эмпирической дисперсии может быть представлено в упрощенном виде

$$S^2 = \sum_x (x - \bar{x})^2 W_x.$$

где учтено, что $c = \bar{c}$;

Основные свойства эмпирической дисперсии:

1) дисперсия постоянной величины всегда равна нулю, так как

$$S_c^2 = \sum_x (c - \bar{c})^2 W_x = 0,$$

2) если все результаты наблюдений уменьшить или увеличить на константу *c,* то значение результирующей дисперсии не изменится, поскольку

$$S_{x-c}^2 \sum_x [(x-c) - \overline{(x-c)}]^2 W_x = \sum_x (x - \bar{x})^2 W_x = S_x^2;$$

3) если все результаты наблюдений увеличить (уменьшить) в *k* раз, то дисперсия увеличится (уменьшится) в k^2 раз:

$$S_{x/k}^2 = \sum_x \left[\left(\frac{x}{k}\right) - \overline{\left(\frac{x}{k}\right)} \right]^2 W_x = \frac{1}{k^2} \sum_x (x - \bar{x})^2 W_x = \frac{1}{k^2} S_x^2;$$

4) эмпирическая дисперсия статистического ряда {x} определяется разностью вида

$$S^2 = \overline{\left(x^2\right)} - (\bar{x})^2.$$

Доказательство:

$$S^2 = \sum_x (x - \bar{x})^2 W_x = \sum_x \left[x^2 - 2x\bar{x} + (\bar{x})^2 \right] W_x = \sum_x x^2 W_x - 2\bar{x} \sum_x x W_x +$$
$$+ (\bar{x})^2 \sum_x W_x = \overline{\left(x^2\right)} - 2(\bar{x})^2 + (\bar{x})^2 = \overline{\left(x^2\right)} - (\bar{x})^2.$$

2.3. Эмпирические начальные и центральные моменты

Введенные выше понятия САВ и дисперсии наблюдений - частные случаи общего понятия "момент" статистического ряда данных.

Эмпирическим начальным моментом заданного порядка (q>0) называют взвешенную САВ *q- x* степеней вариантов:

$$\tilde{v}_q = \overline{\left(x^q\right)} = \sum_{\bar{x}} x^q W_x.$$

В частном случае при *q*=1 получаем:

$\tilde{v}_1 = \sum_{\bar{x}} x W_x = \overline{X}$ - выборочное среднее значение.

Центральным эмпирическим моментом q-го порядка называют САВ *q-x* степеней отклонений вариантов относительно их среднего значения:

$$\tilde{\mu}_q = \overline{\left[(x-\bar{x})^q\right]} = \sum_{\bar{x}} (x-\bar{x})^q W_x.$$

В частном случае при *q=2* получаем:

$\tilde{\mu}_2 = \sum_{\bar{x}} (x-\bar{x})^2 W_x = S^2$ - эмпирическая дисперсия вариационного ряда.

Таким образом, в зависимости от выбора параметра *q c* помощью введенного понятия момента вариационного ряда можно получить достаточно полный набор статистических характеристик для описания всех основных свойств анализируемого случайного явления.

3. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ И ИХ РАСПРЕДЕЛЕНИЯ ВЕРОЯТНОСТЕЙ

3.1. Понятие случайной величины

Случайной называют величину, которая в результате опыта принимает то или иное возможное значение, априори неизвестное, меняющееся от одного испытания к другому и зависящее от ряда случайных или непредсказуемых факторов. В отличие от случайного события, являющегося качественной характеристикой случайного явления, случайная величина характеризует его количественно.

Случайные величины (СВ) разделяются на дискретные и непрерывные.

Дискретной называют СВ, которая принимает счетное (как конечное, так и бесконечное) множество значений. *Непрерывной* называют СВ, возможные значения которой образуют континуальное (неперечитываемое) множество. Например, это показания термометра или вольтметра (нецифрового).

Случайная величина обозначается заглавными буквами конца латинского алфавита ***X,Y,Z,*** а соответствующие ее возможные значения -строчными буквами **x,y,z.** По результатам испытания получаем систему равенств типа *X=x, Y=y, Z=z.*

Для математического описания или задания СВ *X* недостаточно перечислить всё множество ее допустимых значений {x}. Важно также знать, как часто могут появляться в опыте различные значения из этого множества, т.е. необходимо задать вероятности их появления.

Таким образом, задать СВ означает:

1) определить для нее множество допустимых значений {x};

2) определить распределение вероятностей {p}*на* множестве {x}.

Пусть *X* - дискретная СВ, определенная на конечном множестве допустимых значений $\{x_i\}$. По результатам многократных испытаний получаем систему равенств $X = x_1$, $X = x_2$, ..., $X = x_n$. Каждое из этих равенств представляет собой некоторое случайное событие, которое характеризуется соответствующей вероятностной мерой:

$$P(X=x_1)=p_1;$$

$$P(X=x_2)=p_2;$$

……

$$P(X=x_n)=p_n.$$

Если рассматриваем полную группу событий, то для введенных вероятностей выполняется условие нормировки вида

$$\sum_{i=1}^{n} P(X = x_i) = \sum_{i=1}^{n} p_i \equiv 1.$$

Распределением, или законом, СВ X называется вполне определенное соотношение между возможными ее значениями $x_1, x_2, \ldots, x_n$ и соответствующими им вероятностями $p_1, p_2, \bullet\bullet\bullet, p_n$.

Термин "распределение" прямо связан с описанным выше свойством нормировки суммарной вероятности $\sum_{i=1}^{n} p_i = 1$: "единица" вероятности распределяется между различными значениями $x_1, x_2, \ldots, x_n$, причём в общем случае неравномерно, т.е. при $i \neq j$, $p_i \neq p_j$ разным значениям СВ X соответствуют разные вероятности.

3.2 Способы задания закона распределения случайной величины

Распределение СВ может быть задано одним из следующих способов: **1)** табличным; **2)** в виде функции распределения ; **3)** в виде плотности распределения вероятности. Рассмотрим эти способы подробно.

3.2. 1. Таблица распределения случайной величины

Таблица распределения СВ (табл. 3.1) содержит две строки, в которых перечислены, во-первых, все возможные её значения и, во-вторых, соответствующие им вероятности.

Таблица 3.1

Значение	x_1	x_2	…	x_n
Вероятность	p_1	p_2	…	p_n

Табличный способ практически применяется лишь для дискретных СВ с конечным множеством возможных значений $\{x_i\}$. Соответствующее им табличное распределение вероятностей *называют рядами распределения.*

Область применения табличного способа на практике весьма ограничена.

3. 2. 2. Функция распределения

Функция распределения является наиболее универсальной формой задания закона СВ. Она распространяется как на дискретные величины, так и на непрерывные и определяется по следующей формуле:

$$F(x)=P(X<x).$$

Таким образом, функция распределения зависит от конкретного значения *x,* которое в данном случае является аргументом. Функция распределения является интегральной формой задания СВ, поэтому ее иногда называют *интегральной функцией распределения* или *интегральным законом распределения.* Для дискретных СВ функция распределения преобразуется в эквивалентный вид

$$F(x)=\sum_{x_i<x} P(X<x).$$

Здесь $\sum(\ldots)$ обозначает суммирование на множестве значений $\{x_i\}$ при условии, что элементы этого множества $x_i<x$.

В соответствии с табл. 3.1 в общем случае имеем систему равенств

при $x<x_1$: $F(x)=P(X<x_1)=0$;

при $x \geq x_1$, $x < x_2$: $F(x)=P(X=x_1)=p_1$;

при $x_2 \leq x < x_3$: $F(x)=P(X=x_1)+P(X=x_2)=p_1+p_2$;

при $x_{n-1} \leq x < x_n$: $F(x)=p_1+p_2+\ldots+p_{n-1}$:

при $x>x_n$: $F(x)=1$

В рассматриваемом случае дискретной СВ функция F(x) имеет вид ступенчатой линии. При этом высота каждой ступеньки в точке x_t равна p_{it}, т. е. соответствующей вероятности из ряда распределения (табл. 5.1).

Аналогичный по тенденциям (но не по форме) вид имеет функция распределения непрерывной СВ .

На основании рассмотренных общих закономерностей сформулируем следующие *основные свойства* функции распределения:

а) функция распределения есть ограниченная сверху и снизу безразмерная величина:

$$0 \leq F(x) \leq 1$$

б) $\lim_{x\to\infty} F(x)=1$, $\lim_{x\to-\infty} F(x)=0$:

в) функция распределения - неубывающая функция своего аргумента *x,* т. е.

$$\forall x_2 > x_1; F(x_2) \geq F(x_1).$$

Суммируя эти свойства, можно утверждать, что в общем случае функция распределения является неубывающей неотрицательной функцией, удовлетворяющей граничным условиям $F(+\infty)=1$, $F(-\infty)=0$. И, наоборот, любая функция аргумента *x,* удовлетворяющая всем перечисленным свойствам, в принципе может рассматриваться как функция распределения некоторой гипотетической СВ.

Теорема 3.1. Вероятность попадания СВ *X* в заданный конечный интервал значений (α,β) равна

$$P(\alpha \leq X < \beta) = F(\beta) - F(\alpha).$$

Доказательство. Рассмотрим следующие случайные события А:

$x < \beta$;

B: $x < \alpha$;

C: $\alpha \leq x < \beta$.

По формуле сложения событий получаем $A=B+C$ или $C=A-B$. События *B* и C - несовместны. Тогда можно записать

$$P(C)=P(A)-P(B)=F(\beta) - F(\alpha) = P(\alpha<X<\beta),$$

что и требовалось доказать.

Следствие 1. Вероятность любого отдельного значения *x* непрерывной СВ равноа нулю, так как

$$P(X=x)=F(x+\Delta x)-F(x)\,\big|_{\Delta x \to 0}=0$$

Таким образом, нулевой вероятностью здесь характеризуется в принципе возможное событие $X=x$. Этак называемый парадокс теории вероятностей. На основании последнего равенства преобразуем формулу в общепринятый

$$P(\alpha<x<\beta)= F(\beta)-F(\alpha).$$

3.2.3. Плотность распределения

Рассмотрим элементарный интервал $(x, x + x)$ из области определения $\{x\}$ СВ X при $\Delta x \to 0$. В соответствии с теоремой 3.1 вероятность попадания СВ X в данный интервал равна

$$P\{x < X < x + \Delta x\} = F(x + \Delta x) - F(x) .$$

При $\Delta x \to 0$ эта вероятность $P\{*\} \to 0$, однако относительная величина $P\{*\}/\Delta x$ при $\Delta x \to 0$ в общем случае не равна нулю. Полагая рассматриваемую функцию распределения $F(x)$ дифференцируемой по всей ее области определения (случай непрерывной СВ X), в пределе получаем равенство

$$\lim \Delta x \to 0 \frac{P\{x < X < x + \Delta x\}}{\Delta x} = \lim \Delta x \to 0 \frac{F(x + \Delta x) - F(x)}{\Delta x} = F'(x),$$

где $F'(x)^{\Delta} = dF(x)/dx$ Введем обозначение $f(x)^{\Delta} = F'(x)$. Величина f(x) определяет *плотость распределения* СВ X или, говорят, плотность её вероятности.

Плотность вероятности наряду с функцией распределения является наиболее общим, универсальным способом задания закона СВ. Между двумя этими понятиями имеется взаимно однозначная связь: плотность вероятности определяется первой производной от функции распределения, и, наоборот функция распределения определяется по заданной плотности вероятности посредством интеграла следующего вида:

$$F(x) = \int_{-\infty}^{x} f(x)dx.$$

Теорема 3.2. Вероятность попадания непрерывной СВ X в интервал (α,β) определяется следующим выражением:

$$P\{\alpha < X < \beta\} = \int_{\alpha}^{\beta} f(x)dx.$$

Доказательство. С использованием теоремы 5.1 можно записать

$$P\{\alpha < X < \beta\} = F(\alpha) - F(\beta) = \int_{-\infty}^{\beta} f(x)dx - \int_{-\infty}^{\alpha} f(x)dx = \int_{-\infty}^{\alpha} f(x)dx + \int_{\alpha}^{\beta} f(x)dx - \int_{-\infty}^{\alpha} f(x)dx = \int_{\alpha}^{\beta} f(x)dx.$$

Рассмотрим следующие основные свойства плотности вероятности:

а) f(x) ≥ 0 для любого *x* из области определения, как производная от неубывающей функции распределения;

б) $\int_{-\infty}^{+\infty} f(x)dx \equiv 1$ - как вероятность достоверного события *A:* $-\infty<X<+\infty$ Геометрически свойство б) означает, что площадь под кривой распределения тождественно равна единице вне зависимости от формы кривой, т. е. закона распределения.

4. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ СЛУЧАЙНЫХ ВЕЛИЧИН

Во многих прикладных задачах статистический анализ сводится к определению простых числовых характеристик СВ, которые в отличие от закона распределения легко оцениваются по конечной выборке наблюдений. При этом следует учитывать, что закон распределения в любом виде (плотность вероятности, функция распределения) является наиболее полной, исчерпывающей формой задания СВ. А любой набор числовых характеристик определяет СВ лишь в некотором узком , конкретном аспекте. Основными числовыми характеристиками СВ являются *математическое ожидание , дисперсия и моменты различных порядков.*

4.1. Математическое ожидание, мода и медиана

В общем случае *математическое ожидание* (МО) произвольной СВ *X* определяется в соответствии с выражением

$$M(X)^{\Delta} = \int_{-\infty}^{+\infty} xf(x)dx.$$

В случае дискретной СВ с конечным числом *n* возможных значений определение МО сводится к виду

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

где $\sum_{i=1}^{n} p_i \equiv 1.$

Перепишем это выражение в эквивалентной форме

$$M(X) = \frac{p_1 x_1 + p_2 x_x + ... + p_n x_n}{p_1 + p_2 + ... + p_n}.$$

На основании последнего равенства можно утверждать, что МО является аналогом физического понятия центра массы некоторого тела, распределенной по закону $p_1, p_2, \ldots, p_n$ в точках с координатами $x_1, x_2, \ldots, x_n$.

В соответствии с установленной аналогией наглядно проявляются особенности МО с точки зрения описания СВ в усредненном, статистическом смысле. Отсюда следует известное определение МО как средней статистической величины.

Основные свойства МО:

1) *M(c) =c,* где c=const;

2) M(cX)=cM(X);

3) *M(X+Y)=M(X)+M(Y)*

4) *M(XY)=M(X)M(Y)* - только при независимых .X, *Y;*

5) *M(X-M(X))=0* , где *X- M(X) = X*- центрированная СВ.

Модой M_0 дискретной СВ *X* называется ее возможное значение x_i, имеющее максимальную вероятность (p_i=max) из всех возможных значений.

Аналогично, *модой непрерывной* **СВ** называется ее значение x^* из области определения (-∞;+∞) которое соответствует максимальному значению плотности вероятности:

$$f(x)_{x=x^*} = \max_x$$

Мода может быть одна (один максимум плотнос ти вероятности) или несколько. В зависимости от этого соответствующие распределения СВ распределения СВ, которые вообще не имеют моды (равномерный закон распределения).

Медианой M_e СВ *X* называют такое ее возможное значение x^*, для которого справедливо следующее равенство:

$$P\left\{X \le x^*\right\} = P\left\{X > x^*\right\} = 0{,}5,$$

т.е. с равной вероятностью 0,5 данная СВ X по результатам наблюдения оказывается либо меньшей, либо большей x^*. Если распределение СВ одновременно симметрично и одномодово, то медиана совпадает как с МО, так и с модой СВ.

4.2. Дисперсия и среднеквадратическое отклонение

Дисперсия произвольной СВ определяется по формуле

$$D(X)^{\Delta} = M\left[\left(X - M(X)\right)^2\right]$$

или, что эквивалентно,

$$D(X)^{\Delta} = \int_{-\infty}^{+\infty}\left[x - M(X)\right]^2 f(x)dx.$$

В случае дискретной СВ последнее равенство принимает следующий вид:

$$D(X)^{\Delta} = \sum_{i=1}^{n}\left[x_i - M(X)\right]^2 p_i.$$

Понятие дисперсии СВ используется для характеристики флюктуации (отклонений) СВ X относительно ее МО. Размерность дисперсии - квадрат размерности СВ. Поэтому говорят, что дисперсия характеризует мощность или интенсивность случайных флюктуации. В отличие от дисперсии среднеквадратическое отклонение (СКО)$\sigma_i = \sqrt{D(X)}$ имеет размерность СВ и служит характеристикой в усредненном смысле ее абсолютных значений флюктуации.

Основные свойства дисперсии СВ:

а) $D(X) \ge O$, причем $D(X) = 0$ в том и только в том случае, если $X = c$-const - детерминированная величина;

б) $D(cX) = c^2 D(X)$;

в) $D(X \pm Y) = D(x) + D(Y)$, если X, Y- независимые СВ;

г) $D(X) = M(X^2) - M^2(X)$.

Понятия дисперсии СВ и МО тесно связаны с понятиями средней выборочной величины и выборочной дисперсии вариационного ряда данных Указанная связь проявляется в том, что выборочные характеристики

вариационного ряда являются в некотором смысле оценками соответсвенно МО и дисперсии СВ, если последние нам известны.

4.3. Моменты случайной величины

Различают моменты *начальные* и *центральные.*

Начальным моментом q–го порядка (q=1,2..) называют величину

$$v_q \overset{\Delta}{=} M(X^q) = \int_{-\infty}^{+\infty} x^q f(x)dx$$

или в дискретном виде

$$v_q = \sum_{=1}^{n} p_i x_i^q.$$

Центральным моментом q - го порядка называют МО *q* - й степени центрированной СВ

$$\mu_q \overset{\Delta}{=} M((X - M(X))^q),$$

или для дискретной СВ:

$$\mu_q = \sum [x_i - M(X)]^q p_i.$$

Рассмотрим следующие частные случаи моментов СВ *X.*

Начальный момент первого порядка

$$v_i = M(X^1) = \int_{-\infty}^{+\infty} xf(x)dx = M(X)$$

совпадает с МО СВ.

Центральный момент второго порядка

$$\mu_2 = \int_{-\infty}^{+\infty} [x - M(X)]^2 f(x)dx \overset{\Delta}{=} D(X)$$

определяет дисперсию СВ.

Нормированный центральный момент третьего порядка

$$A = \mu_3 / \sigma^3$$

служит характеристикой *скошенности* или **асимметрии** плотности распределения *f(x).*

Нормированный центральный момент четвертого порядка

$$E = \mu_4 \Big/ \sigma^4 - 3$$

характеризует плотность распределения *f(x)* с точки зрения *островершинности* или *плосковершинности,* т.е. **эксцесса.**

5. ПРИМЕРЫ РАСПРЕДЕЛЕНИЙ СЛУЧАЙНЫХ ВЕЛИЧИН

5.1. Равномерное распределение

Равномерное распределение СВ *X,* определенной на конечном интервале значений [a,b], задается следующим видом ее плотности вероятности:

$$f(x) = \begin{cases} 0, x < a; \\ c, a \le x \le b; \\ 0, x > b, \end{cases}$$

где c=const, находиться из условия нормировки $\int_{-\infty}^{+\infty} f(x)dx = 1$ или $\int_a^b cdx = cx\Big|_a^b = c(b-a) = 1,$ в результате c=1/(b-a). Таким образом, плотность вероятности в итоге равна

$$f(x) = \begin{cases} 0, x < a; \\ 1/(b-a), a \le x \le b; \\ 0, x > b. \end{cases}$$

Функция распределения равна

$$F(x) = \int_{-\infty}^{x} f(x)dx = \begin{cases} 0, x < a; \\ \int_a^x (b-a)^{-1} dx = \dfrac{x-a}{b-a}, a \le x \le b; \\ 1, x > b. \end{cases}$$

Основные числовые характеристики равномерного распределения:

1) математическое ожидание

$$M(X)=\int_a^b \frac{x}{b-a}dx=\left.\frac{x^2}{2(b-a)}\right|_a^b=\frac{a+b}{2}-\text{середина отрезка [a, b]; 2)}$$

медиана равна $M_e=M(X)=(a+b)/2$ - в силу симметричности кривой распределения;

3) мода-отсутствует;

4) дисперсия равна

$$D(X)=\mu_2=\int_a^b (x-\frac{a+b}{2})^2\frac{1}{(b-a)}dx=\frac{(b-a)^2}{12};$$

5) среднеквадратическое отклонение равно $\sigma=(b-a)/2\sqrt{3}.;$

6) коэффициент эксцесса равен $E=\frac{\mu_4}{\sigma^4}-3=-1{,}2<0$ - в силу плосковершинности кривой распределения

$$\left(\mu_4=\int_a^b (x-\frac{a+b}{2})^4\frac{1}{(b-a)}dx=\frac{(b-a)^4}{80}\right);$$

7) коэффициент асимметрии равен $A=\frac{\mu_3}{\sigma_3}=0$ - *силу* симметричности кривой распределения.

Вероятность попадания СВ в интервал $(\alpha,\beta)\subset[b,a]$ равна

$$P(\alpha\le X\le\beta)\overset{\Delta}{=}\int_\alpha^\beta\frac{dx}{b-a}=\frac{\beta-\alpha}{b-a}.$$

5.2 Нормальное, или гауссово, распределение

Нормальный закон распределения СВ занимает особое положение в современной теории вероятностей. Если представить гипотетически все множество возникающих на практике статистических задач, то более 90% от их суммарного числа решаются с помощью модели нормально распределенной СВ.

В общем виде нормальный закон распределения СВ X задается следующим выражением для плотности ее вероятности:

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

где m,σ^2-параметры распределения.

В статистической радиотехнике наряду с определением «нормальная СВ» используют (в качестве синонима) определение «гауссовская (гауссова) СВ».

Математическое ожидание нормальной СВ равно

$$M(X)=\int_{-\infty}^{+\infty}xf(x)dx=\frac{1}{\sqrt{2\pi\sigma^2}}\int x\exp\left[-\frac{1}{2\sigma^2}(x-m)^2\right]dx$$

Пусть $t=\frac{x-m}{\sigma}$, тогда

$$M(x)=\frac{\sigma}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{+\infty}(\sigma t+m)\exp\left[-\frac{t^2}{2}\right]dt=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\sigma t\exp\left[-\frac{t^2}{2}\right]dt+$$

т.е.

$$+\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}m\exp\left[-\frac{t^2}{2}\right]dt=\frac{m}{\sqrt{2\pi}}\sqrt{2\pi}=m,$$

параметр нормального распределения m однозначно определяет МО. Часто в расчетах вместо обозначения *M(X)* используют более компактное обозначение *m,* причём не только при нормальном распределении СВ.

Аналогично, дисперсия нормальной СВ равна

$$D(X)=\frac{1}{\sqrt{2\pi\sigma^2}}\int_{-\infty}^{+\infty}(x-m)^2\exp\left[-\frac{1}{2\sigma^2}(x-m)^2\right]dx=\sigma^2,$$

т.е. второй параметр распределения σ^2 однозначно определяет дисперсию нормальной СВ. Поэтому компактное обозначение σ^2 в практических расчетах используется вместо строгого обозначения *D(X).*

Особое положение нормального распределения в теории вероятностей во многом продиктовано рядом присущих ему исключительных свойств. Рассмотрим их подробнее.

1. Кривая распределения нормальной СВ асимптотически стремится к нулевому уровню, т.е.

$$\lim_{x\to\pm\infty} f(x) = 0.$$

2. Максимальное значение кривой распределения равно $1/\sqrt{2\pi\sigma^2}$ (зависит только от дисперсии) и достигается в точке $x=m$.

3. Мода и медиана в данном случае равны МО в силу симметричности кривой распределения, т.е.

$$M_o = M_e = m.$$

4. Центральные моменты нормальной СВ произвольного порядка $(q \geq 2)$ вычисляются по рекуррентной формуле

$$\mu_q=(q-1)\sigma^2\mu_{q-2},$$

где q=2,3,..($\mu_0\equiv1$)/

5. В силу предыдущего свойства все нечетные центральные моменты нормальной СВ равны нулю,т.е.

$$\mu_1=\mu_3=\mu_5=\ldots\equiv0$$

Соответственно, четные центральные моменты нормальной СВ пропорциональны ее дисперсии:

$$\mu_2=\sigma^2;.\mu_4=3\sigma^4;\ \mu_6=15\sigma^6,\ldots$$

6. Коэффициенты асимметрии и эксцесса равны нулю:

$$A = E = 0.$$

Свойство одновременного равенства нулю коэффициентов асимметрии и эксцесса во многих задачах является отличительным признаком нормального закона распределения.

7. При изменении параметра m кривая распределения по форме не меняется, а меняется лишь точка положения ее максимума.

8. При изменении параметра σ^2 форма кривой существенно меняется: при увеличении σ^2 распределение "размывается" по горизонтальной оси *x*.

При изучении нормального распределения важную роль выполняет нормированная СВ

$$T = \frac{X - m}{\sigma} \sim N(0,1),$$

имеющая нормальный закон распределения с параметрами $m_T=0, \sigma_T^2 = 1.$

Для данного случая нормированная плотность вероятности определяется выражением

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

Значения этой плотности подробно табулированы.

Отталкиваясь от введенного понятия, определим вероятность попадания исходной нормальной величины *X* в некоторый ограниченный интервал возможных значений [α,β]:

$$P\{\alpha \le X \le \beta\} = P\left\{ \frac{\alpha - m}{\sigma} \le \frac{X - m}{\sigma} \le \frac{\beta - m}{\sigma} \right\} = P\{t_1 \le T \le t_2\},$$

где t_1, t_2-константы, Т-нормированная СВ.

С учетом интегральной функции распределения нормированной СВ.

$$F(t) \stackrel{\Delta}{=} P\{T \le t\},$$

в результате имеем

$$P\{\alpha \le X \le \beta\} = F(t_2) - F(t_1).$$

По определению интегральной функции распределения нормальной СВ можно записать

$$F(t)=\int_{-\infty}^{t} f(t)dt=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t}\exp\left[-\frac{t^2}{2}\right]dt=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{o}\exp\left[-\frac{t^2}{2}\right]dt+$$

$$+\frac{1}{\sqrt{2\pi}}\int_{0}^{t}\exp\left[-\frac{t^2}{2}\right]dt=\frac{1}{2}+\frac{1}{2}\Phi(t),$$

где $\Phi(t)\overset{\Delta}{=}\frac{2}{\sqrt{2\pi}}\int_{0}^{t}\exp\left[-\frac{t^2}{2}\right]dt$ - интеграл вероятности, или функция Лапласа.

Таким образом, в окончательном виде получаем

$$P\{\alpha\le X\le\beta\}=\frac{1}{2}\left[\Phi\left(\frac{\beta-m}{\sigma}\right)-\Phi\left(\frac{\alpha-m}{\sigma}\right)\right]..$$

Имея в распоряжении достаточно подробные т аблицы интеграла вероятности и следуя полученной формуле, легко вычислить искомую вероятность случайного события A:$\alpha\le X\le\beta$ при любом наборе констант α и βдля любой нормальной СВ $X\propto N(m,\sigma^2)$, заданной произвольными значениями своих параметров m и σ^2.

На основании последнего равенства для частного случая $\beta=m+t_0\sigma$, $\alpha=m-t_0\sigma$ можно записать

$$P\{\alpha\le X\le\beta\}=P\left\{\frac{|X-m|}{\sigma}\le t_0\right\}=\frac{1}{2}\Phi(t_0)-\frac{1}{2}\Phi(-t_0)=\Phi(t_0).$$

Таким образом, значение интеграла вероятности в произвольной точке $t=t_0$ определяет вероятность того, что нормальная СВ $X\propto N(m,\sigma^2)$ отклоняется от своего МО m по абсолютной величине не более чем на $t_0\sigma$, где σ-среднеквадратическое отклонение. Например, при $t_0=1$ имеем

$$P\{|X-m|\le\sigma\}=0{,}6827$$

При $t_0=3$ имеем

$$P\{|X-m|\le 3\sigma\}=0{,}9973$$

Последний результат широко используется в задачах практического анализа: если СВ *X* имеет нормальный закон распределения, то ее отклонение по абсолютной величине относительно МО, как правило, не превосходит порогового уровня 3σ. Это известное правило " трех сигма" для гауссовских случайных процессов.

5.3. Биномиальное распределение

Понятие биномиального распределения (БР) впервые было введено при рассмотрении схемы независимых испытаний Бернулли. Биномиальное распределение определяется выражением вида

$$P_{kn} = C_n^k p^k q^{n-k}$$

где p - вероятность наступления события *A* в каждом отдельном испытании; *q* - вероятность не наступления события *A* в каждом отдельном испытании;

$C_n^k = \frac{n!}{k!(n-k)!}$ - число сочетаний из n элементов по k≤n.

Если обозначить $P_{kn} = P(k)$, где $k = 0,1,2,...,n < \infty$, то $P(k)=P(X=k)$ –это дискретный аналог плотности распределения. Биномиальное распределение применяется для описания дискретных СВ X.

Основные числовые характеристики БР:

1) математическое ожидание равно

$$M(X)\Big|_{X \subset \{k\}} = np;$$

2) дисперсия равна

$$D(X) = npq;$$

3) коэффициент асимметрии равен

$$A = \frac{q-p}{\sqrt{npq}};$$

4) коэффициент эксцесса равен

$$E = \frac{1-6pq}{npq}$$

Таким образом, при увеличении объема наблюдений *n* до бесконечности одновременно устремляются к нулю коэффициент асимметрии и коэффициент эксцесса. Это характерное свойство БР лежит в основе следующего важного теоретического положения.

Локальная теорема Муавра - Лапласа. В условиях повторных наблюдений по схеме независимых испытаний Бернулли выполняется асимптотическое равенство

$$\lim_{n\to\infty} P_{kn} = \frac{1}{\sqrt{2\pi D(x)}} \exp\left[-\frac{1}{2D(x)} (k - M(x)) \right],$$

т.е. пределом БР является нормальный закон распределения с параметрами $m=M(X)=np$, $\sigma^2=D(X)=npq$.

В соответствии с данной теоремой при вычислении вероятностных характеристик БР пользуются таблицами нормального распределения, если число испытаний n**»**1**.**

5.4. Распределение Пуассона

Распределение Пуассона - это предельный случай БР при $n\to\infty$ равенстве *a=np=const* О распределении Пуассона (РП) говорят как о распределении вероятностей редких событий *(p = a/n* « 1). Другой отличительной особенностью РП является равенство МО *M(X)* и дисперсии *D(X)* друг другу:

$$M(X)=D(X)=a.$$

Параметр *a* определяет, таким образом, все основные числовые характеристики РП, общий вид которого задается выражением

$$P_k = \frac{a^k e^{-a}}{k!}, k = 0,1,2,...$$

Здесь k-целочисленный аргумент, принимающий неотрицательные значения.

Распределение Пуассона широко применяется при анализе бракованных изделий в малых заводских партиях и т. д.

5.5. Показательный закон распределения

Это аналог РП для непрерывных СВ. Показательный закон распределения характеризуется плотностью вероятности вида

$$f(x) = \begin{cases} \lambda \exp[-\lambda x], x \geq 0 \\ 0, x < 0. \end{cases}$$

Показательное, или экспоненциальное, распределение используют для описания непрерывных СВ, таких, как время безотказной работы радиоаппаратуры.

Параметр λ - это интенсивность отказов, которая определяет все основные числовые характеристики экспоненциального распределения:

1) математическое ожидание равно

$$M(X) \stackrel{\Delta}{=} \int_0^{\infty} \lambda x \exp[-\lambda x] dx = \frac{1}{\lambda};$$

2) дисперсия равна

$$D(X) = M(X^2) - M^2(X) = \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.$$

Обе характеристики обратно пропорциональны значению λ

Интегральная функция распределения имеет вид

$$F(x) \stackrel{\Delta}{=} P(X < x) = \int_{-\infty}^{x} f(x) dx = \int_0^x \lambda x \exp[-\lambda x] dx = 1 - \exp[-\lambda x], x \geq 0.$$

Вероятность случайного события А: $X \subset (\alpha,\beta)$ вычисляется по формуле

$$P(\alpha < X < \beta) = F(\beta) - F(\alpha) = \exp[-\lambda\alpha] - \exp[-\lambda\beta].$$

6. СИСТЕМЫ СЛУЧАЙНЫХ ВЕЛИЧИН

При статистическом анализе сложных случайных явлений возникают задачи со многими СВ. В таком случае говорят о системе СВ или о многомерной СВ $\vec{U} = (X, Y, Z, ..)$. Статистические свойства СВ $\vec{U}$ определяются не только свойствами ее составляющих, но и взаимодействием этих составляющих. Рассмотрим систему из двух СВ *X* и *Y*. По аналогии с одномерной СВ возможны три способа задания этой системы.

6.1. Табличный способ задания

Если *X* и *Y* -дискретные СВ с конечными множествами возможных значений $\{x_i\}$ и $\{y_i\}$ соответственно, то система СВ может быть задана в виде таблицы $p_{i,j}$ размера $n \times m$, где *m* - число значений Y (j=0,l,…,m); *n* - число значений X (i=0,l,2,...,n); $p_{i,j}=P(X=x_i;Y=y_i)$ вероятность совместного события А: X=Xi;Y=y_j. При этом выполняется условие нормировки $\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij} \equiv 1$

Основной недостаток данного способа - ограничения, связанные с дискретностью каждой СВ.

6.2. Многомерная функция распределения

Функция распределения двумерной системы СВ определяется выражением

$$F(x,y) = P(X < x\ Y < y).$$

Если **y—>∞,** имеем соотношение

$$F(x,\infty) \stackrel{\Delta}{=} P(X < x, Y < \infty) = P(X < x) = F(x)$$

где *F(x)* - одномерная функция распределения.

Полученный результат определяет характер взаимосвязи между одномерной и многомерной функциями распределения. Аналогично, при x=∞ выполняется равенство

$$F(+\infty, y) = F(y).$$

По определению, двумерная функция распределения обладает следующими свойствами:

1) $F(x_2,y) \geq F(x_1,y)$ при $x_2 \geq x_1$

$F(x,y_2) \geq F(x,y_1)$ при $y_2 \geq y_1$.

В общем случае многомерная функция распределения является монотонно неубывающей функцией каждого аргумента в отдельности;

2) $F(-\infty,y) = F(x,-\infty) = F(-\infty,-\infty) = 0$;

3) $F(\infty,\infty) = 1$.

Вероятность попадания системы СВ в определенную область двумерного пространства равна

$$P(\alpha_x < x < \beta_x; \alpha_y < x < \beta_y) = F(\beta_x, \beta_y) - F(\alpha_x, \beta_y) - F(\beta_x, \alpha_y) + F(\alpha_x, \alpha_y).$$

6.3. Многомерная плотность распределения

Если $F(x,y)$ непрерывна и дифференцируема по каждому аргументу , то ее вторая смешанная производная по x и y определяет двумерную плотность распределения:

$$f(x,y) = \frac{d^2 F(x,y)}{dxdy} = \lim_{\substack{\Delta x \to 0 \\ \Delta y \to 0}} \frac{P(x < X < x + \Delta x; y < Y < y + \Delta y)}{\Delta x \Delta y}.$$

По аналогии с понятием «элемент вероятности» для одномерной СВ $f(x)dx$ получаем для двумерной СВ выражение

$$f(x,y)dxdy = P\{x < X < x + dx; y < Y < y + dy\},$$

с помощью которого определяется взаимосвязь между F(x,y) и f(x,y):

$$F(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x,y)dxdy.$$

Двумерная плотность распределения характеризуется следующими свойствами:

1) $f(x,y) \geq 0$ - свойство неотрицательной определенности;

2) $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy \equiv 1$-свойство нормировки;

3) $\int_{-\infty}^{+\infty} f(x,y)dy = f(x);$

$\int_{-\infty}^{+\infty} f(x,y)dx = f(y);$

4) $P(\alpha_x < X < \beta_x; \alpha_y < Y < \beta_y) = \int_{\alpha_x}^{\beta_y}\int_{\alpha_y}^{\beta_x} f(x,y)dxdy;$

5) в общем случае двумерная плотность распределения отвечает умножения плотностей:

$$f(x,y) = f(x)f(y/x);$$

$$f(x,y) = f(x)f(x/y);$$

где *f(x/y)* и *f(y/x)* - условные плотности распределения. Отсюда следует, что *X и Y* взаимозависимы друг от друга. В частном случае независимых СВ *X и Y* будем иметь *f(y/x) = f*(y) и *f(x/y) = f(x).* Следовательно, выполняется равенство *f(x,y)=f(x)f(y). Последнее равенство есть признак статистической независимости двух СВ.*

6.4. Понятие смешанных моментов и коррелированность

Начальный смешанный момент порядка *(k+s)* системы двух СВ определяется выражением

$$\nu_{k,s} \overset{\Delta}{=} M\{X^k Y^s\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} x^k y^s f(x,y)dxdy.$$

Центральный смешанный момент порядка *(k+s)* системы двух СВ определяется выражением

$$\mu_{k,s} \overset{\Delta}{=} M\left\{ \overset{0}{X}^k \overset{0}{Y}^s \right\} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} (x-m_x)^k (y-m_y)^s f(x,y)dxdy.$$

Здесь $\overset{0}{X} = X - m_x; \overset{0}{Y} = Y - m_y$ – центрированные СВ.

В ряду введённых понятий особая роль отводится центральному смешанному моменту второго порядка $\mu_{1,1} \overset{\Delta}{=} M\left\{\overset{0}{X}\overset{0}{Y}\right\}$, который имеет собственное название - момент взаимной корреляции СВ X,Y или, проще, корреляционный момент. Его обычное обозначение

$$\mu_{1,1} = K_{x,y}$$

Переходя к нормированной величине корреляционного момента, получаем

$$\rho = \rho_{x,y} = \frac{K_{x,y}}{\sigma_x \sigma_y}$$

-коэффициент взаимной корреляции. Здесь $\sigma_x = \sqrt{D(x)}; \sigma_y = \sqrt{D(x)}$ – среднеквадратические отклонения СВ X *и* Y соответственно.

Утверждение 6.1. Корреляционный момент $K_{x,y}$ и коэффициент взаимной корреляции ρ одновременно равны нулю в случае независимых СВ X,Y.

Доказательство. В самом деле, в рассматриваемом случае двумерная плотность распределения равна

$$f(x,y) = f(x) f(y)$$

Поэтому получаем

$$K_{x,y} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x-m_x)(y-m_y) f(x) f(y) dxdy = \int_{-\infty}^{+\infty}(x-m_x) f(x) dx \int_{-\infty}^{+\infty}(y-m_y) f(y) dy = 0$$

так как $M\left\{\overset{0}{X}\right\} = 0, M\left\{\overset{0}{Y}\right\} = 0$ по определению.

Если X и *Y*- независимы, то они одновременно и некоррелированы, т.е. из равенства

$$f(x,y)=f(x)f(x)$$

автоматически следует двойное равенство

$$K_{x,y}=\rho=0.$$

Обратное утверждение в общем случае неверно, т.е. две некоррелированные СВ могут быть статистически зависимыми. Понятие "коррелированность" СВ близко к понятию "зависимость" (в статистическом смысле), но не тождественно ему. Единственное исключение из этого правила - случай двух нормальных СВ X и Y, когда понятия коррелированности и зависимости тождественны друг другу.

6.5. Двумерное нормальное распределение

Двумерный закон нормального распределения определяется следующим выражением для плотности вероятности системы СВ X и Y:

$$f(x,y)=\frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2\left(1-\rho^2\right)}\left[\frac{(x-m_x)^2}{\sigma_x^2}-2\rho\frac{(x-m_x)(y-m_y)}{\sigma_x\sigma_y}+\frac{(y-m_y)^2}{\sigma_y^2}\right]\right\}.$$

В частном случае (р = 0) двумерная плотность нормального распределения преобразуется к упрощенному виду

$$f(x,y)=\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left[-\frac{1}{2\sigma_x^2}(x-m_x)^2\right]\frac{1}{\sqrt{2\pi\sigma_y^2}}\exp\left[-\frac{1}{2\sigma_y^2}(y-m_y)^2\right]=f(x)f(y),$$

т.е. понятия «коррелированность» и «зависимость» оказываются эквивалентными друг другу.

7. ФУНКЦИОНАЛЬНЫЕ ПРЕОБРАЗОВАНИЯ СЛУЧАЙНЫХ ВЕЛИЧИН

7.1. Преобразования одномерной случайной величины

В общем виде функциональное преобразование СВ X (исходная величина) в СВ Y (результирующая величина) описывается зависимостью Y=y(X), т.е. множество X отображается во множество Y по вполне определенному закону. Задача формулируется следующим образом: зная вид функционального преобразования y(X) и закон распределения исходной СВ X, требуется определить закон распределения результирующей СВ Y.

Пусть y(x) - непрерывная дифференцируемая функция, не имеющая точек разрыва по всей области её определения. И пусть существует обратное преобразование $x(y)=y^{-1}(x)$. Рассмотрим следующие два события:

событие A : $x_0 < X < x_0 + dx$;

событие B : $y_0 < Y < y_0 + dy$.

Если между используемыми переменными установлено соответствие $y_0 = y(x_0) = y(x)\Big|_{x=x_0}$, $y_0 + dy = y(x_0 + dx)$ или $dy = y(x_0+dx) - y(x_0)$, то нетрудно убедиться, что события A и B эквивалентны друг другу, т.е. из факта свершения события A автоматически следует свершение события B . Поэтому их вероятности равны друг другу: P(A)=P(B) .

Полагая $dx \to 0$ и $dy \to 0$, запишем

$$f_1(x)dx = f_2(y)dy \quad \text{или} \quad f_2(y) = f_1(x)\frac{dx}{dy}.$$

Здесь $f_1(x), f_2(y)$-одномерные плотности СВ X и Y соответственно. Учитывая, что $f_1(x) \geq 0$, $f_2(y) \geq 0$, окончательно будем иметь

$$f_2(y) = f_1(x)\left|\frac{d}{dy}x(y)\right|.$$

Полученный результат определяет искомое решение поставленной задачи.

Пример 1. Пусть Y=aX+b-линейное преобразование и, следовательно, обратная функция имеет вид

$$x(y) = \frac{y-b}{a}$$

Тогда $\frac{dx}{dy} = \frac{1}{a}$.Поэтому $f_2(y) = \frac{1}{|a|} f_1(x) = \frac{1}{|a|} f_1\left(\frac{y-b}{a}\right)$. Пусть, в частности, исходная СВ X - это нормальная СВ с МО m_x *и* дисперсией σ_x^2.

Тогда СВ Y имеет распределение с плотность

$$f_2(y) = \frac{1}{|a|}\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left[-\frac{1}{2\sigma_x^2}\left(\frac{y-b}{a} - m_x\right)^2\right] = \frac{1}{\sqrt{2\pi\sigma_x^2 a^2}} \times$$

$$\times \exp\left[-\frac{1}{2\sigma_x^2 a^2}\left(y - (am_x + b)\right)^2\right].$$

Таким образом, линейное преобразование нормальной СВ не изменяет закон ее распределения, а лишь меняет его параметры..

В общем случае неоднозначного обратного преобразования *x(y)* решение задачи имеет принципиальные особенности. Событие $B: y_0<Y<y_0+dy$ эквивалентно здесь сумме $R\geq 1$ несовместных событий вида

$$A_1: x_1 < X < x_1 + dx,$$
$$A_2: x_2 < X < x_2 + dx,$$
$$----------$$
$$A_k: x_k < X < x_k + dx.$$

Поэтому в соответствии с теоремой сложения вероятностей имеем

$$P(R) = \sum_{i=1}^{k} P(A_i).$$

Или, используя результат предыдущего решения, получаем

$$f_2(y) = \sum_{i=1}^{k} f_1(x)\left|\frac{dx_i(y)}{dy}\right|,$$

где $x_i(y)$-*i-й* участок функции обратного преобразования для со свойством взаимно-однозначного соответствия между множествами *{x}* и *{y}*

Пример 2. Пусть $Y=X^2$-квадратичное преобразование. Для этого случая можно записать:

$$x_{1,2} = \pm\sqrt{y};$$
$$\frac{dx_{1,2}}{dy} = \pm\frac{1}{2\sqrt{y}};$$
$$f_2(y) = \frac{1}{2\sqrt{y}}\left[f_1(+\sqrt{y}) + f_1(-\sqrt{y})\right].$$

Рассмотрим случай нормальной СВ $X\propto N\left(m_x,\sigma_x^2\right)$.Тогда выполняется равенство

$$f_2(y)=\frac{1}{2\sqrt{y}\sqrt{2\pi\sigma_x^2}}\left[\exp\left[-\frac{1}{2\sigma_x^2}\left(\sqrt{y}-m_x\right)^2\right]+\exp\left[-\frac{1}{2\sigma_x^2}\left(\sqrt{y}+m_x\right)^2\right]\right].$$

Предположим, что МО исходной СВ X равно нулю. Тогда в окончательном виде будем иметь

$$f_2(y)=\frac{1}{\sqrt{2\pi\sigma_x^2 y}}\exp\left[-\frac{y}{2\sigma_x^2}\right],\quad y\geq 0.$$

Таким образом, нелинейное преобразование нормальной СВ меняет её закон распределения.

7.2. Функциональное преобразование двумерной случайной величины

Рассмотрим систему двух СВ (X_1, X_2), совместное распределение которых описывается двумерной плотностью распределения $f\backslash (x_1,x_2)$. Данная система СВ подвергается двумерному преобразованию вида

$$\begin{cases} y_1 = y_1(x_1,x_2); \\ y_2 = y_2(x_1,x_2), \end{cases}$$

т.е. исходная система СВ (X_1,X_2) преобразуется в систему СВ (Y_1,Y_2) с неизвестным распределением $f_2 = f_2(y_1, y_2)$ = ?

Найдем это распределение. Используя логику предыдущего рассуждения, введем понятие двух эквивалентных событий с равными вероятностями

$$f_1(x_1,x_2)dx_1dx_2 = f_2(y_1,y_2)dy_1dy_2$$

Учитывая свойство положительной определенности плотностей в левой и правой частях последнего равенства, можно записать

$$f_2(y_1,y_2)=f_1(x_1,x_2)\left|\frac{d^2(x_1,x_2)}{d^2(y_1,y_2)}\right|=f_1(x_1,x_2)\frac{1}{\left|\dfrac{d^2(y_1,y_2)}{d^2(x_1,x_2)}\right|}.$$

Здесь $\left|\frac{d^2(y_1,y_2)}{d^2(x_1,x_2)}\right| = \begin{vmatrix} \frac{dy_1}{dx_1} & \frac{dy_1}{dx_2} \\ \frac{dy_2}{dx_1} & \frac{dy_2}{dx_2} \end{vmatrix} = \frac{dy_1}{dx_1}\frac{dy_2}{dx_2} - \frac{dy_1}{dx_2}\frac{dy_2}{dx_1}$ – якобиан двумерного преобразования.

Рассмотрим частный случай двумерного функционального преобразования $y = y(x_1,x_2)$. Введем для этого случая систему обозначении

$$\begin{cases} y_1 = x_1; \\ y_2 = y_2(x_1,x_2) = y \end{cases}$$

и после этого воспользуемся предыдущим результатом. Якобиан рассматриваемого вида функционального преобразования равен

$$J = 1\frac{dy_2}{dx_2} - \frac{dy_2}{dx_1}0 = \frac{dy_2}{dx_2} = \frac{dy}{dx_2}.$$

Поэтому совместная (двумерная) плотность распределения системы преобразованных величин имеет вид

$$f_2(y_1,y_2) = f_1(x_1,x_2)\frac{1}{\left|\frac{dy}{dx_2}\right|}$$

Или применительно к результирующей величине Y (случай одномерной СВ) получаем

$$f_2(y) = \int_{-\infty}^{+\infty} f_2(y_1,y_2)dy_1 = \int_{-\infty}^{+\infty} \frac{f_1(x_1x_2)}{\left|\frac{dy}{dx_2}\right|}dy_1 = \int_{-\infty}^{+\infty} \frac{f_1(x_1x_2)}{\left|\frac{dy}{dx_2}\right|}dx_1.$$

Рассмотрим сумму СВ X_1,X_2 заданных совместной плотностью распределения $f(x_1,x_2)$. Для этого случая будем иметь

$$y = x_1 + x_2;$$
$$-x_2 = y - x_1;$$
$$-\frac{dy_2}{dx_2} = \frac{dy}{dx_2} = 1;$$
$$-f_2(y) = \int_{-\infty}^{+\infty} f(x_1, x_2)\left|\frac{dy}{dx_2}\right|^{-1} dx_1 = \int_{-\infty}^{+\infty} f(x_1, y - x_1)dx_1.$$

Для двух независимых СВ X_1, X_2, когда выполняется равенство

$$f(x_1, x_2) = f(x_1) f(x_2),$$

полученный результат преобразуется к следующему виду:

$$f_2(y) = \int_{-\infty}^{\infty} f(x_1) f(y - x_1)dx_1$$

Пример. Рассмотрим две нормальные независимые СВ

$$X_1 \propto N(0,\sigma^2_x);\ \ X_2 \propto N(0,\sigma^2_x),$$

где $\sigma_x^2 = D(X_1) = D(X_2)$ – их дисперсия.

Определим распределение вероятностей суммарной СВ $Y=X_1+X_2$. Учитывая, что в рассматриваемом случае совместная плотность распределения исходных СВ равна

$$f(x_1, x_2) = \frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left[-\frac{1}{2\sigma_x^2}x_1^2\right]\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left[-\frac{1}{2\sigma_x^2}x_2^2\right],$$

в соответствии с предыдущим выражением запишем

$$f_2(y)=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left[-\frac{1}{2\sigma_x^2}x_1^2\right]\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left[-\frac{1}{2\sigma_x^2}(y-x_1)^2\right]dx_1=$$

$$=\int_{-\infty}^{+\infty}\left(\frac{1}{2\pi\sigma_x^2}\right)\exp\left[-\frac{1}{2\sigma_x^2}\left[x_1^2+(y-x_1^2)\right]\right]dx_1=$$

$$=\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left[-\frac{1}{4\sigma_x^2}y_1^2\right]\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left[-\frac{1}{\sigma_x^2}\left[x_1-\frac{y}{2}\right]^2\right]dx_1=$$

$$=\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left[-\frac{1}{4\sigma_x^2}y_1^2\right]\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{t^2}{2}\right]\frac{1}{\sqrt{2}}dt=\frac{1}{\sqrt{2\pi\sigma_x^2 2}}\exp\left[-\frac{1}{4\sigma_x^2}y_1^2\right].$$

На основании полученного результата можно сделать следующие выводы.

1. Суммирование двух независимых нормальных СВ с одинаковой дисперсией σ_x^2 даёт нормально распределённую СВ с дисперсией $\sigma_y^2=2\sigma_x^2$.

2. В общем случае суммирования n независимых нормальных СВ с соответствующими дисперсиями $\sigma_{x_i}^2$ и МО $m_{x_i}, i=\overline{1,n}$, имеем также нормальное распределение вероятностей с параметрами

$$\sigma_y^2=\sum_{i=1}^{n}\sigma_{x_i}^2;$$

$$m_y=\sum_{i=1}^{n}m_{x_i}^2.$$

8. ПРЕДЕЛЬНЫЕ ТЕОРЕМЫ ТЕОРИИ ВЕРОЯТНОСТЕЙ

Основным объектом теории вероятностей являются случайные явления, точнее, массовые случайные явления, многократно наблюдаемые в примерно одинаковых условиях.

Задача теории вероятностей состоит в установлении статистических закономерностей, лежащих в основе массовых случайных явлений.

Предельные теоремы являются одним из наиболее мощных инструментов для изучения таких закономерностей. Предельные теоремы теории вероятностей по своему содержанию разделяются на две группы. Первую группу составляют теоретические утверждения, образующие закон больших чисел (лемма Маркова, неравенство Чебышева и теоремы Чебышева и Бернулли), назначение которого состоит в установлении жестких взаимосвязей между статистическими (выборочными) данными и основными понятиями классической теории вероятностей. Вторую группу образуют теоремы, объединенные общим понятием центральной предельной теоремы.

8.1. Закон больших чисел

Лемма Маркова. Если X - произвольная, положительно определенная СВ, то для любого числа $a = \text{const} > 0$ справедливо неравенство

$$P\{X \le a\} > 1 - \frac{M(X)}{a}.$$

Доказательство. Пусть множество возможных значений СВ $X = \{x_i\}$ упорядочено так, что выполняется система соотношений $x_i > x_{i+1} \forall i = \overline{1,n}$, где n - суммарное число возможных значений. При этом каждому значению поставим в соответствие его вероятность $p_i > 0 \forall i = \overline{1,n}$. Причем $\sum_{i=1}^{n} p_i \equiv 1$. Тогда для любого целого числа r=1,2...≤ n имеем неравенство

$$x_1 p_1 + x_2 p_2 + x_3 p_3 + \ldots + x_r p_r \le x_1 p_1 + x_2 p_2 + x_3 p_3 + \ldots + x_n p_n = M(X)$$

Пусть также задана некоторая константа a такая, что $x_r > a > x_{r+1}$. Тогда можно записать

$$x_1 p_1 + x_2 p_2 + x_3 p_3 + \ldots + x_r p_r \ge a \sum_{i=1}^{r} p_i.$$

На основании полученных результатов имеем

$$a\sum_{i=1}^{r} p_1 \le M(X) \quad \text{или} \quad \sum_{i=1}^{\kappa} p_i \le M(X)\frac{1}{a}.$$

Нетрудно убедиться, что левая часть этого выражения определяет сумм вероятностей значений СВ X , которые превышают константу a. Поэтому, пользуясь теоремой о сумме вероятностей 4.1, окончательно получим

$$P\{X > a\} = \sum_{i=1}^{r} p_i \le M(X)\frac{1}{a}$$

или, переходя к противоположному событию,

$$P\{X \le a\} > 1 - \frac{M(X)}{a},$$

что и требовалось доказать.

Неравенство Чебышева. Если X- произвольная СВ, то для любого числа a >0 справедливо следующее соотношение:

$$P\{|X - M(X)| \le a\} > 1 - \frac{D(X)}{a^2},$$

где *M(X), D(X)* - математическое ожидание и дисперсия СВ.

Доказательство. Рассмотрим преобразованную СВ

$$Z = (X - M(X))^2 > 0$$

В соответствии с леммой Маркова имеем неравенство

$$P\{z \le a^2\} > 1 - \frac{M(Z)}{a^2}.$$

Учтём при этом, что $M(Z) = M\left\{\left(X - M(X)\right)^2\right\} \stackrel{\Delta}{=} D(X)$ и, кроме того, выполняется равенство

Последние соотношения в совокупности и доказывают неравенство Чебышева.

Теорема Чебышева. Пусть X_i -СВ, представленная I-м случайным независимым наблюдением над исходной СВ X с заданным МО *M(X)* =const и

дисперсией *D(X)* - const . Кроме того, пусть $\overline{X_n} \overset{\Delta}{=} \frac{1}{n}\sum_{i=1}^{n} X_i$ - среднее по n наблюдениям (или среднее выборочное) значение СВ X. Тогда для любого положительного числа *a* выполняется следующее соотношение:

$$P\left\{\left|\overline{X_n} - M(X)\right| > a\right\} \le \frac{D(X)}{a^2 n}.$$

Доказательство. Пусть $Z = \overline{X_n}$, тогда

$$M(Z) = \frac{1}{n}\sum_{i=1}^{n} M(X_i) = \frac{1}{n} n M(X) = M(X);$$

$$D(Z) = M\left\{\left(Z - M(Z)\right)^2\right\} = M\left\{\left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i - M(X_i)\right)\right]^2\right\}.$$

Учитывая независимость и, следовательно, некоррелированность различных наблюдений между собой, имеем $M\left(\overset{0}{X}i\ \overset{0}{X}\,j\right) = 0$ для всех i ≠ j. Поэтому запишем

$$D(Z) = \frac{1}{n^2}\sum_{i=1}^{n} D\left[X_i\right] = \frac{1}{n} D(X).$$

В совокупности с неравенством Чебышева полученный результат и доказывает теорему Чебышева.

Важным **следствием** доказанной теоремы является равенство $P\left\{X_n - M(X)\right| > a\} = 0$ при n → ∞. При этом говорят, что среднее выборочное значение сходится по вероятности к МО СВ, т.е.

$$\overline{X_n} \Rightarrow M(X).$$

Теорема Бернулли. Для классической схемы независимых испытаний Бернулли со случайными исходами A и $\overline{A}$ с вероятностями p и q = 1 - p соответственно при любом значении константы a >0 справедливо следующее соотношение:

$$P\left\{\left|\frac{m}{n}-p\right|>a\right\}\le\frac{pq}{na^2},$$

где n - суммарное число независимых испытании, m - число испытании с исходом А.

Доказательство. Пусть $X_i=\begin{cases}1\\0\end{cases}$, (1 - при исходе А, 0 – при исходе $\overline{A}$).

Тогда

$$\sum_{i=1}^{n}X_i=m;\ \frac{1}{n}\sum_{i=1}^{n}X_i=\overline{X}=\frac{m}{n};$$

$$M(X_i)=M(\overline{X})=1p+0q=p;$$

$$D(X_i)=pq.$$

Полученные выражения совместно с теоремой Чебышева доказывают теорему Бернулли.

Следствие. Для любого числа

$$a>0: P\left\{\left|\frac{m}{n}-p\right|>a\right\}\to 0 \quad \text{при } \mathrm{n}\to\infty \text{ или } \frac{\mathrm{m}}{\mathrm{n}}\Rightarrow p.$$

Данный результат раскрывает физический смысл вероятности случайного события как предельного значения относительной частоты события *A* в схеме независимых последовательных испытаний.

Таким образом, основное назначение теоретических утверждений, образующих закон больших чисел, связано с интерпретацией таких важнейших теоретических понятий, как МО СВ и вероятность случайного события.

8.2. Центральная предельная теорема (ЦПТ)

Вторая группа предельных теорем теории вероятностей, объединённых общим понятием ЦПТ, имеет предметом своего внимания распределение суммы СВ $Y=X_1+X_2+\ldots -+X_n$

Существует целый ряд формулировок ЦПТ, отличающихся условиями, накладываемыми на отдельные слагаемые. Наиболее распространенное определение принадлежит А.М. Ляпунову.

Теорема Ляпунова. Распределение суммы *n* независимых СВ $X_1, X_2, \ldots, X_n$ асимптотически сходится к нормальному закону при неограниченном увеличении числа слагаемых $n \to \infty$ и ограниченных первых двух моментах каждого слагаемого.

При решении многих прикладных задач ЦПТ в формулировке Ляпунова преобразуется в теоретическое утверждение относительно средней выборочной величины

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Легко убедиться, что здесь справедливы оба дополнительных условия теоремы Ляпунова, а именно:

математическое ожидание $m_i = M(X_i) = M(X) = \text{const};$

дисперсия $\sigma_i^2 = D(X_i) = D(X) = const.$

Поэтому в пределе при $n \to \infty$ имеем

$$P\left\{\overline{X} < b\right\} \to \frac{1}{\sqrt{2\pi\sigma_x^2}} \int_{-\infty}^{b} e^{-\frac{1}{2\sigma_x^2}(x-m_x)^2} dx$$

вне зависимости от распределения СВ *X*. Здесь $m_x = |M|(X)$, $\sigma_x^2 = D(X)/n$.

Следует отметить, что ЦПТ справедлива не только для непрерывных, но и для дискретных СВ, т.е. имеет весьма общий характер. В общем случае можно записать

$$P\{a < \overline{X} < b\} \to \frac{1}{\sqrt{2\pi\sigma_x^2}} \int_{a}^{b} e^{-\frac{1}{2\sigma_x^2}(x-m_x)^2} dx = \frac{1}{2}\left[\Phi(t_2) - \Phi(t_1)\right],$$

где Ф(*) - интеграл Лапласа; $t_1 = \frac{a - m_x}{\sigma_x}$; $t_2 = \frac{b - m_x}{\sigma_x}$.

Практическое значение ЦПТ огромно:

1. Центральная предельная теорема представляет собой реальный механизм теоретического описания суммарных СВ в терминах нормального закона распределения.

2. Опыт показывает, что распределение конечной суммы произвольно распределенных СВ X_i уже при n = 10 ... 15 хорошо аппроксимируется нормальным законом распределения.

3. Роль суммарных СВ особенно важна в задачах статистической радиотехники, где целый класс преобразователей напряжения (фильтры, узкополосные усилители) характеризуется существенной инерционностью и, следовательно, реализует эффект накопления.

9. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

9. 1. Понятие статистических оценок

Классическая задача статистического анализа состоит в определении (оценивании) неизвестного закона распределения наблюдаемой СВ X по имеющейся повторной (многократной) выборке $\{X_i\}$ конечного объема n. Здесь X_i - i-е наблюдение СВ X, рассматриваемое далее как СВ. Ее распределение совпадает с распределением исходной СВ X.

Решение классической задачи в общем случае представляет значительные трудности. Один из возможных способов их преодоления - параметрический подход.

Пусть X - случайная величина, заданная плотностью распределения параметрического вида $f(x) = f(x, \Theta)$, где Θ - неизвестный параметр. Если $\hat{\Theta}_n = \hat{\Theta}(X_1, X_2, \ldots, X_n)$ - оценка неизвестного параметра по выборке, то $\hat{f}(x) = \hat{f}(x, \hat{\Theta})$ – результирующая оценка распределения.

Пример. Пусть $X \sim N(m, \sigma^2)$ - нормальная СВ с заданной дисперсией $\sigma^2 = \text{const}$ и неизвестным МО m. Воспользуемся оценкой МО по формуле выборочного среднего

$$\hat{m}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\Delta}{=} \overline{X}_n .$$

В соответствии с параметрическим подходом получим оценку исходного распределения вида

$$\hat{f}_n(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(x - \overline{X}_n\right)^2 \right].$$

В зависимости от точности используемых оценок неизвестных параметров в результате получаем более или менее точное представление о распределении СВ.

Различные методы статистического оценивания являются предметом изучения специального раздела математической статистики, а именно теории статистического оценивания. При этом термин “статистическое оценивание” учитывает принципиально случайный характер (результаты меняются от опыта к опыту) формируемых оценок параметров $\hat{\Theta}_n\left(X_1, X_2, \ldots, X_n\right)$, который подчеркивается в обозначении случайных наблюдений $X_1, X_2, \ldots, X_n$.

9. 2. Основные свойства статистических оценок

Оценку $\hat{\Theta}_n$ неизвестного параметра Θ по повторной выборке наблюдений $\left\{X_i\right\}$ называют *несмещенной*, если выполняется равенство

$$M\left(\hat{\Theta}_n\right) = \Theta .$$

Пример. Пусть нами используется оценка МО по формуле выборочного среднего

$$\hat{m}_n = \frac{1}{n}\sum_{i=1}^{n} X_i .$$

Тогда ее (оценки) МО равно

$$M\left(\overset{\wedge}{m_n}\right) = M\left\{\frac{1}{n}\sum_{i=1}^{n} X_i\right\} = \frac{1}{n}\sum_{i=1}^{n} M\left(X_i\right) = \frac{1}{n}\sum_{i=1}^{n} M\left(X\right) = \frac{1}{n} nm = m,$$

т.е. формируемая оценка является несмещённой.

Аналогично, если оценка дисперсии равна

$$\overset{\wedge}{\sigma}{}^2_n = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 \overset{\Delta}{=} \mu_2,$$

то $M\left(\overset{\wedge}{\sigma}{}^2_n\right) \neq \sigma^2$.

В рассматриваемом случае оценка дисперсии является смещённой.

В общем случае один и тот же параметр Θ может иметь несколько или даже множество несмещённых оценок $\overset{\wedge}{\Theta}{}^{(1)}_n, \overset{\wedge}{\Theta}{}^{(2)}_n, \ldots, \overset{\wedge}{\Theta}{}^{(k)}_n$ (κ - индекс оценки Θ_n). Например, при неизвестном МО $M(X) = m$ набор несмещённых оценок может быть представлен следующим образом :

$$\overset{\wedge}{\Theta}{}^{(j)}_n = \frac{1}{n}\sum_{i=1}^{n} X_i + \overset{\circ}{Y}_j,$$

где $\overset{\circ}{Y} = Y_j - \overset{\Delta}{M\left(Y_j\right)}$, $\forall j = \overline{1,k}$ - некоторая центрированная СВ.

Из множества несмещённых оценок параметра $\overset{\wedge}{\Theta}{}^{(1)}_n, \overset{\wedge}{\Theta}{}^{(2)}_n, \ldots, \overset{\wedge}{\Theta}{}^{(k)}_n$ оценка $\overset{\wedge}{\Theta}{}^{(*)}_n$ называется *эффективной*, если она имеет минимальную дисперсию:

$$D\left(\overset{\wedge}{\Theta}{}^{(*)}_n\right) = \min D\left(\overset{\wedge}{\Theta}{}^{(j)}_n\right).$$

Пример. Основываясь на результатах предыдущего примера в отношении множества несмещённых оценок одного и того же параметра Θ, можно утверждать, что эффективной оценкой МО является оценка по формуле выборочного среднего, так как

$$D\left(\overset{\wedge}{\Theta}_n^{(j)}\right) = \frac{1}{n}D(X) + D(Y_j) \geq \frac{1}{n}D(X) \overset{\Delta}{=} D\left(\overline{X}_n\right).$$

Оценка $\overset{\wedge}{\Theta}_n$ неизвестного параметра Θ называется *состоятельной*, если она подчиняется закону больших чисел, т.е. для любого $\beta > 0$ выполняется соотношение

$$P\left\{\left|\overset{\wedge}{\Theta}_n - \Theta\right| > \beta\right\} \leq \alpha_n > 0 \ ,$$

причем $\lim\limits_{n\to\infty} \alpha_n = 0$. Иначе говоря, состоятельная оценка $\overset{\wedge}{\Theta}_n$ сходится по вероятности к неизвестному истинному значению параметра: $\overset{\wedge}{\Theta}_n \Rightarrow \Theta$.

Пример. Оценка неизвестного МО по формуле выборочного среднего $\hat{m}_n = \frac{1}{n}\sum\limits_{i=1}^{n} X_i$ является состоятельной , поскольку

$$\overline{X}_n \overset{\Delta}{=} \frac{1}{n}\sum_{i=1}^{n} X_i \Rightarrow M(X)$$

(по теореме Чебышева).

Аналогично можно доказать состоятельность оценок неизвестных моментов распределения ν_q и μ_q $\forall q \geq 1$ по формулам соответственно начальных и центральных выборочных моментов.

9.3. Метод моментов (метод Пирсона)

Рассмотрим некоторую СВ *X*, заданную параметрическим семейством распределения $X \sim f\left(x, \Theta_1, \Theta_2, \ldots, \Theta_q\right)$ на множестве допустимых значений $\Theta_1, \Theta_2, \ldots, \Theta_q$ q неизвестных параметров. Пусть $X_1, X_2, \ldots, X_n$ - выборка объема *n* из заданной генеральной совокупности. Поставим задачу оценивания *q* неизвестных параметров по имеющейся конечной выборке наблюдений. При этом предполагается, что *q* не превышает значений *n* ($q \leq n$).

Рассмотрим следующий вариант решения поставленной задачи. Запишем выражение для q первых начальных моментов СВ:

$$\begin{cases} \nu_1 \overset{\Delta}{=} \int\limits_{-\infty}^{+\infty} x f\left(x,\Theta_1,\Theta_2,\ldots,\Theta_q\right)dx = \nu_1\left(x,\Theta_1,\Theta_2,\ldots,\Theta_q\right); \\ \nu_2 \overset{\Delta}{=} \int\limits_{-\infty}^{+\infty} x^2 f\left(x,\Theta_1,\Theta_2,\ldots,\Theta_q\right)dx = \nu_2\left(x,\Theta_1,\Theta_2,\ldots,\Theta_q\right); \\ \text{– –} \\ \nu_q \overset{\Delta}{=} \int\limits_{-\infty}^{+\infty} x^q f\left(x,\Theta_1,\Theta_2,\ldots,\Theta_q\right)dx = \nu_q\left(x,\Theta_1,\Theta_2,\ldots,\Theta_q\right). \end{cases}$$

Рассмотрим также q первых эмпирических (выборочных) моментов анализируемой СВ (см. тему 2):

$$\begin{cases} \tilde{\nu}_1 \overset{\Delta}{=} \sum\limits_x x W_x = \frac{1}{n}\sum\limits_{i=1}^{n} X_i = \tilde{\nu}_1\left(X_1, X_2,\ldots, X_n\right); \\ \tilde{\nu}_2 \overset{\Delta}{=} \sum\limits_x x^2 W_x = \frac{1}{n}\sum\limits_{i=1}^{n} X_i{}^2 = \tilde{\nu}_2\left(X_1, X_2,\ldots, X_n\right); \\ \text{– –} \\ \tilde{\nu}_q \overset{\Delta}{=} \sum\limits_x x^q W_x = \frac{1}{n}\sum\limits_{i=1}^{n} X_i{}^q = \tilde{\nu}_q\left(X_1, X_2,\ldots, X_n\right). \end{cases}$$

Основная идея метода моментов состоит в приравнивании значений эмпирических моментов их теоретическим значениям:

$$\begin{cases} \tilde{\nu}_1\left(X_1, X_2,\ldots, X_n\right) = \nu_1\left(\Theta_1,\Theta_2,\ldots,\Theta_q\right); \\ \tilde{\nu}_2\left(X_1, X_2,\ldots, X_n\right) = \nu_2\left(\Theta_1,\Theta_2,\ldots,\Theta_q\right); \\ \text{– –} \\ \tilde{\nu}_q\left(X_1, X_2,\ldots, X_n\right) = \nu_q\left(\Theta_1,\Theta_2,\ldots,\Theta_q\right). \end{cases}$$

Решая полученную систему q уравнений относительно q неизвестных параметров $\Theta_1,\Theta_2,...,\Theta_q$, приходим к окончательному результату

$$\begin{cases} \hat{\Theta}_1 = \hat{\Theta}_1(X_1, X_2, \ldots, X_n); \\ \hat{\Theta}_2 = \hat{\Theta}_2(X_1, X_2, \ldots, X_n); \\ - - - - - - - - - - - - - - - - \\ \hat{\Theta}_q = \hat{\Theta}_q(X_1, X_2, \ldots, X_n). \end{cases}$$

Предложенная идея, или собственно метод моментов, обосновывается тем, что эмпирические моменты распределения сходятся в асимптотике к их теоретическим значениям, т.е.

$$\forall q \geq 1: \qquad \tilde{\nu}_q \underset{n\to\infty}{\Rightarrow} \nu_q$$

или в соответствии с теоремой Чебышева

$$P\left\{\left|\tilde{\nu}_q - \nu_q\right| > a\right\} \to 0.$$

Основным положительным свойством статистических оценок по методу моментов является их состоятельность, т.е.

$$\forall q \geq 1: \qquad \tilde{\Theta}_q \underset{n\to\infty}{\Rightarrow} \Theta_q.$$

Вторым важным свойством метода моментов является его инвариантность к закону распределения анализируемой СВ для широкого круга практических задач, например, при оценивании МО и дисперсии по выборке наблюдений. При этом сложность вычислений незначительна.

Пример. Пусть $X \sim N\left(m,\sigma^2\right)$- нормальная СВ с неизвестными параметрами m и σ^2. Оценим их по выборке конечного объёма $X_1, X_2, \ldots, X_n$. В соответствии с методом моментов имеем систему двух уравнений (q =2)

$$\begin{cases} \int\limits_{-\infty}^{+\infty} xf\left(x,\hat{m},\hat{\sigma^2}\right)dx = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i; \\ \int\limits_{-\infty}^{+\infty} x^2 f\left(x,\hat{m},\hat{\sigma^2}\right)dx = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i^2, \end{cases}$$

левые части которых равны :

$$\int\limits_{-\infty}^{+\infty} xf\left(x,\hat{m},\hat{\sigma^2}\right)dx = \int\limits_{-\infty}^{+\infty} x\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(x-m)^2\right]dx = m = M(X);$$

$$\int\limits_{-\infty}^{+\infty} x^2 f\left(x,\hat{m},\hat{\sigma^2}\right)dx = \int\limits_{-\infty}^{+\infty} x^2\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2\sigma^2}(x-m)^2\right]dx = \sigma^2 + m^2.$$

(*Примечание***.** Из свойства дисперсии $D(X) = M\left(X^2\right) - \left[M(X)\right]^2$ следует равенство

$$\nu_2 = D(X) + \left[M(X)\right]^2 = \sigma^2 + m^2.)$$

Поэтому можно записать

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} X_i;$$

$$\hat{\sigma^2} + \hat{m^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Решая эту систему уравнений относительно оценок неизвестных параметров, окончательно получаем

$$\hat{m} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{\Delta}{=} \overline{X}_n;$$

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\overline{X_n}\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 = \tilde{\mu}_2.$$

Данный результат определяет две классические формулы математической статистики, а именно средней выборочной величины и средней квадратичной величины.

9.4. Метод максимального правдоподобия (метод Фишера)

Пусть $X_i, i=\overline{1,n}$ - результат n независимых наблюдений над СВ X, а $f(x,\Theta)$- параметрическое семейство ее плотностей распределения, где Θ – неизвестный параметр, требующий оценивания. Идея метода максимального правдоподобия неразрывно связана с понятием функции правдоподобия $L(\Theta)$, которое, в свою очередь, связано с понятием многомерной плотности распределения n независимых случайных наблюдений $f(x_1,x_2,\ldots,x_n)=f(x_1)f(x_2)\ldots f(x_n)$. Учитывая, что для любого $i=\overline{1,n}$ выполняется равенство

$$f(x_i)=f(x_i,\Theta)=f(x,\Theta),$$

т.е. закон распределения зависит от параметра Θ, имеем

$$f(x_1,x_2,\ldots,x_n)=f(x_1,\Theta)f(x_2,\Theta)\ldots f(x_n,\Theta).$$

Рассматривая последнее выражение при равенствах

$$x_1=X_1;\quad x_2=X_2,\ldots,x_n=X_n$$

(аргументы фиксируются и приравниваются выборочным значениям) в функции неизвестного параметра, приходим к определению функции правдоподобия

$$L(\Theta)=f(X_1,\Theta)f(X_2,\Theta)\ldots f(X_n,\Theta).$$

Метод максимального правдоподобия заключается в определении оптимального вида оценки неизвестного параметра Θ исходя из следующего требования:

$$\Theta_n=\mathrm{Arg}(\max L(\Theta)).$$

При этом гарантируется максимально правдоподобный в статистическом смысле результат. На практике чаще используется эквивалент последнего равенства вида

$$\Theta_n = \mathrm{Arg}\left(\max \ln L(\Theta)\right)$$

(логарифм рассматривается как монотонная функция, сохраняющая положение максимума).

10. СТАТИСТИЧЕСКИЕ ОЦЕНКИ МАТЕМАТИЧЕСКОГО ОЖИДАНИЯ И ДИСПЕРСИИ СЛУЧАЙНОЙ ВЕЛИЧИНЫ

10.1. Оценивание математического ожидания и дисперсии по выборке

Рассмотрим СВ $X \sim N\left(m, \sigma^2\right)$, распределенную по нормальному закону с неизвестными параметрами $m = M(X) = ?$ и $\sigma^2 = D(X) = ?$. Для оценивания МО и дисперсии по заданной выборке $X_1, X_2, \ldots, X_n$ $(n \geq 1)$ воспользуемся методом максимального правдоподобия.

В соответствии с этим методом плотность распределения

$$f(x) = f\left(x, m, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma}(x-m)^2 \right];$$

функция правдоподобия

$$L\left(m, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - m)^2 \right];$$

натуральный логарифм функции правдоподобия

$$\ln L\left(m, \sigma^2\right) = -\frac{n}{2} \ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - m)^2 \ .$$

Для отыскания точки максимума функции правдоподобия находим первую производную по каждому из рассматриваемых параметров :

$$\frac{d \ln L\left(m, \sigma^2\right)}{dm} = + \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - m);$$

$$\frac{d \ln L\left(m, \sigma^2\right)}{d\sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^{n} (X_i - m)^2 .$$

В итоге приходим к системе двух уравнений

$$\begin{cases} \dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n}\left(X_i - \hat{m}_n \right) = 0; \\ -n\hat{\sigma}_n^2 + \sum\limits_{i=1}^{n}\left(X_i - \hat{m}_n \right)^2 = 0, \end{cases}$$

решая которую относительно неизвестных параметров, получим :

$\hat{m}_n = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i \overset{\Delta}{=} \overline{X}_n$ – средняя арифметическая величина ;

$\hat{\sigma}_n^2 = \dfrac{1}{n}\sum\limits_{i=1}^{n}\left(X_i - \overline{X}_n \right)^2$ – средняя квадратичная величина.

Этот результат по виду повторяет выражения, полученные методом моментов. При этом можно доказать, что в данном случае функция правдоподобия достигает своего максимального значения.

Совпадение оценок МО и дисперсии, полученных методом моментов и методом максимального правдоподобия, подтверждает универсальность и высокую степень значимости выражений для средней арифметической и средней квадратичной величин, используемых в качестве вышеуказанных оценок.

10.2. Основные свойства оценок математического ожидания и дисперсии

Утверждение 10.1. Средняя арифметическая величина $\overline{X}_n = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i$, вычисленная по *n* независимым наблюдениям $X_1, X_2, \ldots, X_n$ из произвольной генеральной совокупности X , является одновременно несмещённой состоятельной оценкой неизвестного МО.

Доказательство. Несмещённость оценки вытекает из равенства её МО истинному значению МО $M\left(\overline{X}_n\right) = M(X)$, которое нами было установлено в одном из примеров предыдущей темы. Кроме того , из теоремы Чебышева следует свойство состоятельности оценки , так как

$$P\left\{\left|\overline{X}_n - M(X)\right| > a\right\} \le \frac{D(X)}{na^2} \underset{n\to\infty}{\to} 0.$$

Утверждение 10.2. Средняя квадратичная величина (СКВ) $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$ при любом конечном объёме выборки $n < \infty$ не является несмещённой оценкой дисперсии *D(X)*.

Доказательство. Запишем следующее выражение для СКВ :

$$S_n{}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - m + m - \overline{X}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\left(X_i - m\right) - \left(\overline{X}_n - m\right)\right]^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - m\right)^2 - \frac{2}{n}\sum_{i=1}^{n}\left(X_i - m\right)\left(\overline{X}_n - m\right) + \frac{1}{n}\sum_{i=1}^{n}\left(\overline{X}_n - m\right)^2 =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - m\right)^2 - 2\left(\overline{X}_n - m\right)^2 + \left(\overline{X}_n - m\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - m\right)^2 - \left(\overline{X}_n - m\right)^2.$$

Ее МО равно

$$M\left(S_n{}^2\right) = \frac{1}{n}\sum_{i=1}^{n} M\left(X_i - m\right)^2 - M\left(\overline{X}_n - m\right)^2 = \frac{1}{n}n\sigma^2 - D\left(\overline{X}_n\right) =$$

$$= \sigma^2 - \frac{1}{n}\sigma^2 = \frac{n-1}{n}\sigma^2 \neq \sigma^2,$$

т.е. оценка является смещённой.

Следствие. Несмещённая состоятельная оценка неизвестной дисперсии произвольной СВ *X* определяется следующим выражением :

$$\hat{S}{}_n^2 = \frac{1}{n-1}\sum\left(X_i - \overline{X}_n\right)^2.$$

Доказательство :

$$M\left(\hat{S}{}_n^2\right) = M\left(\frac{n}{n-1}S_n^2\right) = \frac{n}{n-1}M\left(S_n^2\right) = \sigma^2,$$

что и требовалось доказать .

Рассмотренная оценка называется *исправленной выборочной дисперсией*. При этом дробь $n/(n-1)$ называется *поправкой Бесселя*. При $n >> 1$ поправка Бесселя стремится к единице, следовательно, исчезают различия между исправленной и первоначальной оценками дисперсии. На практике считается достаточным для устранения указанных различий объём выборки $n \geq 50$. В соответствии с этим говорят, что первоначальное выражение для выборочной дисперсии по методу максимального правдоподобия гарантирует асимптотическую несмещённость и состоятельность формируемой оценки.

Утверждение 10.3. Средняя квадратичная величина, вычисленная по *n* независимым наблюдениям из заданной генеральной совокупности с произвольным законом распределения, является несмещённой дисперсией в асимптотике (при $n \to \infty$).

Доказательство основывается на асимптотическом равенстве для поправки Бесселя $n/(n-1) \to 1$.

Утверждение 10.4. В частном случае нормальной генеральной совокупности $X \sim N(m, \sigma^2)$ с неизвестными параметрами *m* и σ^2 САВ определяет одновременно состоятельную несмещённую и эффективную оценки неизвестного МО.

Утверждение 10.5. В частном случае нормальной генеральной совокупности с заданным МО $M(X) = m = \text{const}$ и неизвестной дисперсией σ^2 СКВ, вычисленная по формуле $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - m)^2$, определяет одновременно состоятельную несмещённую и эффективную оценки неизвестной дисперсии.

Примером задачи такого рода является статистический анализ нормального, или гауссовского, шума с нулевым МО.

10. 3. Распределение оценки математического ожидания для выборок из нормальной генеральной совокупности

Оценка МО по формуле выборочного среднего может рассматриваться как СВ в том смысле, что ее значение меняется случайным образом в зависимости от случайных значений выборочных данных $X_1, X_2, \ldots, X_n$.

Утверждение 10.6. Если СВ распределена по нормальному закону $X \sim N\left(m, \sigma^2\right)$ с параметрами m и σ^2, а $X_1, X_2, \ldots, X_n$- ряд независимых наблюдений над СВ X, то среднее выборочное значение $\overline{X}_n$ распределено по нормальному закону $\overline{X}_n \sim N\left(m,\ \sigma^2 / n\right)$.

Следствие. Нормированное среднее выборочное значение нормальной СВ, заданное выражением вида

$$\frac{\overline{X}_n - m}{\sqrt{\sigma^2/n}} = \frac{\overline{X}_n - m}{\sigma}\sqrt{n},$$

подчиняется нормальному распределению вероятности с МО, равным нулю, и дисперсией, равной единице.

Доказательство :

$$M\left(\frac{\overline{X}_n - m}{\sqrt{\sigma^2/n}}\right) = \frac{\sqrt{n}}{\sigma}\left(M\left(\overline{X}_n\right) - m\right) = 0;$$

$$D\left(\frac{\overline{X}_n - m}{\sqrt{\sigma^2/n}}\right) = \frac{n}{\sigma^2}\left(D\left(\overline{X}_n\right) + 0\right) = \frac{n}{\sigma^2}\frac{\sigma^2}{n} = 1.$$

Во многих прикладных задачах представляет интерес нормированная СВ вида

$$T \overset{\Delta}{=} \frac{\overline{X}_n - m}{\sqrt{S^2/n}},$$

определённая для случая неизвестной дисперсии $\sigma^2 = ?$, когда вместо неё используется несмещённая (исправленная) выборочная оценка $\hat{S}_n^2$.

Утверждение 10.7. Случайная величина T_n, определённая по выборке из нормальной генеральной совокупности с неизвестной дисперсией $\sigma^2 = ?$ и заданным МО, подчиняется закону t - распределения Стьюдента с n степенями свободы, который задается плотностью $f(t) = S(t,n),$ где n - параметр, определяемый объёмом выборки, причём $S(t,n) = S(-t,n)$ (функция чётная).

Значения плотности t - распределения подробно табулированы. Обычно используются таблицы t_n - статистики вида

$$t_n \overset{\Delta}{=} \mathrm{Arg}\{P(-t < T_n < t) = p\}, \quad p=\text{const}.$$

Учитывая зависимость t_n - статистики от заданной константы p, на практике используют двойную индексацию вида $t_{n,p}$.

Утверждение 10.8. Распределение Стьюдента асимптотически сходится к нормальному закону, т.е.

$$T_n \underset{n\to\infty}{\Rightarrow} N(0,1).$$

На практике достаточно, с точки зрения приемлемой аппроксимации распределения Стьюдента нормальным законом, иметь выборку объёмом $n \geq 50\ldots100$.

10. 4. Распределение оценки дисперсии для выборки из нормальной генеральной совокупности

При анализе распределения оценки неизвестной дисперсии следует иметь в виду два возможных случая:

А. Математическое ожидание СВ $M(X)=m=\text{const}$ априори известно.

Б. Математическое ожидание неизвестно.

Случай А. Пусть $X \sim N\left(m, \sigma^2\right)$, где m=const, а σ^2=? , и пусть имеется выборка объемом n наблюдений из заданной генеральной совокупности. Тогда оценка дисперсии может быть рассчитана по формуле

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - m\right)^2 .$$

Рассмотрим статистику вида

$$\frac{nS_n^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i - m}{\sigma}\right)^2 = \sum_{i=1}^{n} T_i ,$$

где T_i- случайная величина, распределённая по нормальному закону $T_i \sim N(0,1)$.

В силу независимости значений $X_1, X_2, \ldots, X_n$ значения статистик $T_1, T_2, \ldots, T_n$ также в совокупности независимы.

Утверждение 10.9. Случайная величина $\chi_n^2 \stackrel{\Delta}{=} \sum_{i=1}^{n} T_i^2$, вычисленная по n - выборке наблюдений из нормальной генеральной совокупности $N\left(m, \sigma^2\right)$, подчиняется χ^2 - распределению Пирсона $f\left(\chi^2, n\right)$ с n степенями свободы (n - параметр, χ^2 - аргумент). При этом МО χ^2- статистики равно

$$M\left(\chi^2\right) = \sum_{i=1}^{n} M\left(T_n^2\right) = \sum_{i=1}^{n} D\left(T_i\right) = n = \text{const},$$

а ее дисперсия $D\left(\chi^2\right) = 2n = \text{const}$.

Значения плотности χ^2- распределения Пирсона не вычисляются в элементарных функциях, а задаются соответствующими таблицами. Как правило, используются таблицы χ_n^2 – статистики (табл. 12.1) , отвечающей определению

$$\chi_n^2 = \mathrm{Arg}\left\{ P\left(\chi^2 > \alpha_n^2\right) = p \right\}.$$

Таблица 12.1

p / *n*	0,9	0,5	0,1	0,01
5	1,61	4,35	9,2	15,1
10	4,86	9,34	16,0	23,2
15	8,5	14,3	22,3	30,6

В соответствии с данными табл. 12.1 можно сделать следующие выводы:

для любого значения вероятности p = const статистика χ_n^2 возрастает при увеличении числа степеней свободы или объёма наблюдений n ;

для любого фиксированного объёма выборки n при увеличении значения порога α_n^2 значение вероятности

$$P\left(\chi^2 > \alpha_n^2\right) = \int_{\alpha_n^2}^{\infty} f\left(\chi^2, n\right) d\chi^2$$

монотонно убывает.

Случай Б. Пусть СВ X распределена нормально с параметрами m и σ^2, предполагаемыми неизвестными . Тогда оценка дисперсии может быть рассчитана по формуле исправленной выборочной дисперсии

$$\hat{S}_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}_n \right)^2.$$

Утверждение 10.10. Нормированная СВ

$$\frac{n\hat{S}^2}{\sigma^2} = \frac{n}{n-1} \sum \left(\frac{X_i - \overline{X}_n}{\sigma} \right)^2$$

подчиняется χ^2 - распределению с (n - *1*) степенями свободы.

Следствие. Вне зависимости от рассматриваемого случая при анализе качества оценки неизвестной дисперсии всегда используются таблицы χ^2-распределения. Различия между случаями сводятся лишь к различиям в числе степеней свободы распределения статистики χ^2_n.

11. СТАТИСТИЧЕСКАЯ ТЕОРИЯ ВЫБОРОЧНОГО МЕТОДА

11.1. Понятия доверительного интервала и доверительной вероятности

Рассмотренные выше численные оценки МО и дисперсии по конечной выборке наблюдений представляют собой примеры точечных оценок неизвестных параметров. Точечных в том смысле, что в качестве оценки параметра получают то или иное конкретное число, т.е. точки на числовой оси. Наряду с точечным оцениванием статистическая теория оптимальных оценок занимается и вопросами интервального оценивания неизвестных параметров. Задачу интервального оценивания в общем виде можно сформулировать следующим образом: по данным имеющейся выборки наблюдений $X_1, X_2, \ldots, X_n$ построить числовой интервал $\left[\Theta^{(1)}, \Theta^{(2)}\right]$, внутри которого с высокой вероятностью $p>0$ находится неизвестное значение оцениваемого параметра, т.е. $\Theta \in \left[\Theta^{(1)}, \Theta^{(2)}\right]$.

Интервальное оценивание особенно необходимо при малом числе наблюдений n, когда надёжность точечных оценок невелика.

Доверительным интервалом некоторого параметра Θ называют такой интервал $\left[\Theta_n^{(1)}, \Theta_n^{(2)}\right]$, для которого выполняется требование

$$P\left(\Theta_n^{(1)} \le \Theta \le \Theta_n^{(2)}\right) = p = \text{const} > 0.$$

При этом константа $p>0$ определяет *доверительную вероятность* попадания в доверительный интервал истинного значения неизвестного параметра. Чем больше вероятность p, тем надежнее интервальная оценка. На практике, как правило, выбирают $p \approx 1$, поэтому константу p часто задают

следующим образом: $p = 1 - \alpha$, где $\alpha << 1$ - *уровень значимости интервальной оценки.*

Доверительная вероятность p (или уровень значимости интервальной оценки α) тесно взаимосвязана с длиной доверительного интервала $\Delta_n = \Theta_n^{(2)} - \Theta_n^{(1)}$: чем больше p (меньше α), тем больше должен быть интервал Δ_n при фиксированном объёме n выборки. При увеличении n точность интервального оценивания возрастает, что проявляется в соответствующем уменьшении длины доверительного интервала Δ_n. В пределе при $n \to \infty$ длина доверительного интервала Δ_n при p=const стремится к нулю, что характерно для точечных оценок параметров. Т.е. в асимптотическом случае интервальная и точечная оценки сходятся к одной и той же величине (как правило, к истинному значению неизвестного параметра).

11. 2. Построение доверительного интервала для оценки математического ожидания по выборке из нормальной генеральной совокупности

Случай А. Дисперсия $\sigma^2 = \text{const}$ - известный параметр. В рассматриваемом случае имеем следующее равенство :

$$P\left\{\left|\frac{\overline{X}_n - m}{\sigma}\sqrt{n}\right| < z_p\right\} = \Phi\left(z_p\right),$$

где $\overline{X}_n -$ несмещённая оценка МО, $\Phi\left(z_p\right) -$ интеграл вероятности (функция Лапласа), вычисленный в точке z_p.

Указанное равенство основывается на том, что нормированная СВ

$$Z = \frac{\left(\overline{X}_n - m\right)}{\sigma}\sqrt{n}$$

подчинена нормальному закону распределения с параметрами m=0 и $\sigma^2 = 1$, т.е. $Z \sim N(0,1)$.

Приравнивая значения $\Phi\left(z_p\right) = p = \text{const}$, по таблицам нормального распределения находим соответствующее значение порогового уровня z_p, где индекс *p* указывает на очевидную взаимосвязь порогового уровня с константой *p*. После этого перепишем первоначальное равенство следующим образом:

$$P\left\{-z_p < \frac{\overline{X}_n - m}{\sigma}\sqrt{n} < z_p\right\} = p,$$

или в эквивалентном виде

$$P\left\{\overline{X}_n - \frac{z_p\sigma}{\sqrt{n}} < m < \overline{X}_n + \frac{z_p\sigma}{\sqrt{n}}\right\} = p.$$

Форма последнего выражения соответствует определению доверительного интервала для оценки неизвестного МО *m*.

Таким образом, с заданной доверительной вероятностью *p>0* можно утверждать, что интервал

$$\left[\overline{X}_n - \frac{z_p\sigma}{\sqrt{n}};\quad \overline{X}_n + \frac{z_p\sigma}{\sqrt{n}}\right]$$

является доверительным для искомой оценки МО. Его длина $\Delta_p = 2\frac{z_p\sigma}{\sqrt{n}}$ асимптотически стремится к нулю при $n \to \infty$.

Случай Б. Дисперсия $\sigma^2 = ?$- неизвестный параметр. В рассматриваемом случае имеем следующее равенство :

$$P\left\{|T_n| < t_{n,p}\right\} = p = \text{const},$$

где $T_n = \frac{\overline{X}_n - m}{\hat{S}}\sqrt{n}$ - нормированная СВ, распределённая по закону Стьюдента; *n* - объём выборки наблюдений (параметр распределения).

Задаваясь некоторым значением константы *p* как доверительной вероятностью и пользуясь таблицами *t*-распределения, находим соответствующее значение статистики $t_{n,p}$. После этого путем эквивалентных преобразований приходим к окончательному выражению

$$P\left\{\overline{X}_n - \frac{t_{n,p}\,\hat{S}}{\sqrt{n}} < m < \overline{X}_n + \frac{t_{n,p}\,\hat{S}}{\sqrt{n}}\right\} = p.$$

Таким образом, с доверительной вероятностью p можно утверждать, что неизвестное значение МО находится в пределах доверительного интервала

$$\left[\overline{X}_n - \frac{t_{n,p}\,\hat{S}}{\sqrt{n}};\quad \overline{X}_n + \frac{t_{n,p}\,\hat{S}}{\sqrt{n}}\right].$$

Его длина $\Delta_n = 2\dfrac{t_{n,p}\,\hat{S}}{\sqrt{n}}$ в асимптотике монотонно сходится к нулю. Кроме того, можно утверждать, что при больших n точность интервальной оценки в случае Б почти не отличается от точности интервальной оценки в случае А, так как при $n >> 1$ приближенно выполняется равенство

$$\frac{2t_{n,p}\,\hat{S}}{\sqrt{n}} \cong \frac{2z_p\sigma}{\sqrt{n}} \quad \left(\hat{S} \to \sigma;\quad t_{n,p} \to z_p\right)$$

11. 3. Построение доверительного интервала для оценки дисперсии по выборке из нормальной генеральной совокупности

Задаваясь доверительной вероятностью $p>0$, определим искомый доверительный интервал для оценки неизвестной дисперсии из следующего выражения :

$$P\left\{\chi_1^2 < \frac{n\hat{S}}{\sigma^2} < \chi_2^2\right\} = p = 1-\alpha,$$

где $\dfrac{n\hat{S}^2}{\sigma^2} = \chi^2 -$ нормированная СВ, распределённая по закону Пирсона с n или $n-1$ степенями свободы в зависимости от рассматриваемого случая: известно или неизвестно априори МО $M(X)$.

Прежде чем воспользоваться таблицами χ^2 – распределения, преобразуем выражение для доверительного интервала в эквивалентный вид

$$P\left\{\chi_1^2 < \chi^2 < \chi_2^2\right\} = P\left(\chi^2 > \chi_1^2\right) - P\left(\chi^2 > \chi_2^2\right).$$

Приравнивая $P\left(\chi^2 > \chi_1^2\right)$ значению $1 - \frac{\alpha}{2} = \text{const}$, а $P\left(\chi^2 > \chi_2^2\right)$ значению $\frac{\alpha}{2}$, по таблицам χ^2 – распределения определим соответствующие значения двух пороговых уровней χ_1^2 и χ_2^2 . Подставим найденные значения пороговых уровней в выражение для доверительного интервала и после этого запишем

$$\chi_1^2 < \frac{n\hat{S^2}}{\sigma^2} < \chi_2^2,$$

или в эквивалентном виде

$$\frac{n\hat{S^2}}{\chi_2^2} < \sigma^2 < \frac{n\hat{S^2}}{\chi_1^2}.$$

Таким образом , с доверительной вероятностью $p = 1 - \alpha$ значение неизвестной дисперсии не выходит за пределы доверительного интервала

$$\left[\frac{nS^2}{\chi_2^2};\quad \frac{nS^2}{\chi_1^2}\right].$$

Его длина $\Delta_n = nS^2\left(\chi_1^{-2} - \chi_2^{-2}\right)$ стремится к нулю при $n \to \infty$.

11. 4. Определение объёма выборки в задачах статистического оценивания

Чтобы по данным имеющейся выборки $X_1, X_2, \ldots, X_n$ конечного объёма $n < \infty$ можно было с уверенностью судить об интересующих нас признаках генеральной совокупности X , выборка должна быть

представительной (репрезентативной). Выборка является *представительной*, если, во-первых, она сформирована случайным образом, т.е. объективна, и, во-вторых, если ее объём n не ниже некоторого порогового значения $n \geq n^*$. Определение порогового значения n^* объёма используемой выборки - основной предмет исследования статистической теории выборочного метода.

Рассмотрим случай нормальной генеральной совокупности в задаче оценивания неизвестного МО . Имеем два варианта интервальной оценки :

длиной $\Delta_1 = 2z_p\sigma/\sqrt{n}$ при заданной дисперсии σ^2;

длиной $\Delta_2 = 2t_{n,p}\,\hat{S}/\sqrt{n}$ при неизвестной дисперсии.

С учётом этого получаем два варианта выражений для требуемого объёма выборочных данных

$$n_1 = \left(\frac{2z_p\sigma}{\Delta_1}\right)^2;$$

$$n_2 = \left(\frac{2t_{n,p}\,\hat{S}}{\Delta_2}\right)^2$$

при заданных длинах Δ_1 и Δ_2 .

Введём следующие обозначения : $\delta_1 = \Delta_1/\sigma -$ относительная длина доверительного интервала для первого варианта интервальной оценки; $\delta_2 = \Delta_2/\sigma -$ относительная длина доверительного интервала для второго варианта интервальной оценки.

Таким образом, окончательно получаем

$$n_1 \geq n_1^* = \left(\frac{2z_p}{\delta_1^*}\right)^2; \qquad n_2 \geq n_2^* = \left(\frac{2t_{n,p}}{\delta_2^*}\right).$$

Полученные выражения определяют минимальный объём выборки наблюдений, при котором гарантируется требуемое качество интервальной оценки неизвестного МО для каждого из рассмотренных случаев.

Пример. Пусть задана доверительная вероятность p=0,95, а требуемая точность интервальной оценки отвечает условию

$$\delta_1^2 = \delta_2^2 = 0{,}1 \quad (10\%).$$

Тогда по таблицам нормального распределения при доверительной вероятности p=0,95 находим значение порогового уровня $z_p = 1{,}96$. И, следовательно, получаем

$$n_1^* = \frac{(2 \cdot 1{,}96)^2}{0{,}1} = 154.$$

Для другого случая (дисперсия неизвестна) по таблицам распределения Стьюдента определяем значение статистики $t_{n,p} = 2{,}0 \ldots 2{,}2$ при $n \geq 10 \ldots 30$. В результате получаем

$$n_2^* \geq \frac{(2 \cdot 2{,}2)^2}{0{,}1} = 194.$$

В общем случае выполняется соотношение $n_2^* \geq n_1^*$ с равенством в асимптотическом случае $n \to \infty$.

12. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

12.1. Понятие статистической гипотезы

При исследовании статистических закономерностей наблюдатели (исследователи) выдвигают предварительные собственные гипотезы, т.е. предварительные суждения, которые в дальнейшем на основе выборочных данных либо доказываются, либо опровергаются. Под *статистической гипотезой* понимают любое высказывание или суждение о статистических свойствах или законе распределения анализируемой СВ .

Статистические гипотезы разделяются на гипотезы о параметрах распределения и о законах распределения (более общие гипотезы). Задача проверки статистических гипотез о параметрах распределения в общем случае формулируется следующим образом. Пусть $f(x,\Theta)$ - закон, зависящий от параметра Θ, значение которого заранее неизвестно. Предположим, что проверяется гипотеза о том, что $\Theta = \Theta_0 = \text{const}$. Назовём эту гипотезу исходной или нулевой и обозначим H_0. Гипотезу о том, что $\Theta = \Theta_1 \neq \Theta_0$, назовём конкурирующей и обозначим H_1. Конкурирующую гипотезу часто называют *альтернативной гипотезой* или просто *альтернативой*. Таким образом, возникает задача проверки нулевой гипотезы H_0 против альтернативы H_1 на основании имеющихся выборочных данных $X_1, X_2, \ldots, X_n$ конечного объёма $n < \infty$.Решение поставленной задачи, по сути, состоит в разделении всего пространства или множества значений выборочных данных $X_1, X_2, \ldots, X_n$ на два непересекающихся подмножества (областей пространства) *O* и *W*, таких, что принимается решение в пользу H_0, если выборка $\{X_1, X_2, \ldots, X_n\}$ принадлежит области *O*, и в пользу H_1, если выборка принадлежит подмножеству *W*, причём *O*+*W*=*X* (рис. 12.1).

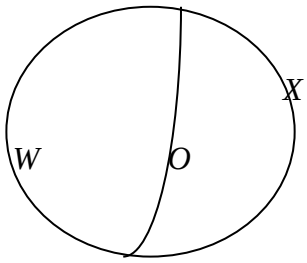


Рис. 12. 1

Возникает вопрос, какими принципами следует руководствоваться при определении каждого из подмножеств *O* или *W* соответственно. Впервые такие принципы были сформулированы в работах математиков Неймана и Пирсона. В соответствии с ними в задачах проверки статистических гипотез учитываются две ошибки: *ошибка первого рода* – принято решение в пользу H_1 при справедливости H_0; *ошибка второго рода* – принято решение в пользу H_0 при справедливости H_1. Указанные ошибки иллюстрируются в табл. 12.1.

Таблица 12.1

Гипотеза H_0	Верна	Неверна
Решение: отвергается	Ошибка первого рода	Нет ошибки
Решение: принимается	Нет ошибки	Ошибка второго рода

При этом область W называют критической областью выборочного пространства (гипотеза H_0 в этой области отвергается). Область O называют областью допустимых значений (гипотеза H_0 в этой области принимается). Соответствующие вероятности ошибок вычисляются согласно следующим выражениям:

- вероятность ошибки первого рода

$$\alpha = P\{X_1, X_2, ..., X_n \subset W | H_0\};$$

- вероятность ошибки второго рода при справедливости гипотезы H_1

$$\beta = P\{X_1, X_2, ..., X_n \subset O | H_1\}.$$

Теория проверки статистических гипотез исходит из того, что при любом уменьшении вероятности ошибки первого рода α возрастает вероятность ошибки второго рода β и наоборот, т.е. в принципе не удается свести к нулю или одновременно минимизировать вероятности ошибок первого и второго рода. Поэтому любой из возможных критериев принятия решения в данном случае основывается на принципиально противоположных требованиях к указанным вероятностям ошибочных решений.

12. 2. Проверка простой гипотезы против простой альтернативы

Статистическая гипотеза называется *простой*, если соответствующее суждение относится к вполне определённому распределению.

Пример 1. Гипотеза о равенстве МО нормальной СВ X некоторой константе: $M(X)=m$, если одновременно задана дисперсия $D(X) = \sigma^2$,-простая гипотеза. Напротив, $M(X) \neq m$ или $|M(X) - m| > 0$ – гипотеза сложная.

Задача проверки простой гипотезы H_0 против простой альтернативы H_1 формулируется следующим образом: пользуясь выборкой $X_1, X_2, ..., X_n$ объёмом n из генеральной совокупности X, проверить гипотезу о законе распределения вида $f(x_1, x_2, ..., x_n) = f(x_1, x_2, \ldots, x_n | H_0)$ против альтернативы $f(x_1, x_2, \ldots, x_n) = f(x_1, x_2, \ldots, x_n | H_1)$, где H_0 - нулевая гипотеза, а H_1 - альтернатива.

Ее решение состоит в проверке принадлежности имеющейся выборки наблюдений $X_1, X_2, ..., X_n$ критической области W выборочного пространства, которая в общем случае определяется из условия

$$W: \frac{f(X_1, X_2, \ldots, X_n | H_1)}{f(X_1, X_2, \ldots, X_n | H_0)} \stackrel{\Delta}{=} l(X_1, X_2, \ldots, X_n) \geq l_0 = \text{const}.$$

Здесь $l(X_1, X_2, \ldots, X_n)$ - отношение правдоподобия, используемое в роли решающей статистики, при $l(X_1, X_2, \ldots, X_n) \geq l_0$ гипотеза H_0 отвергается (или принимается гипотеза H_1); l_0 - относительный (безразмерный) пороговый уровень.

Следуя данному решающему правилу (алгоритму принятия решения), получим вероятность ошибки первого рода

$$\alpha = \int_W f(x_1, x_2, \ldots, x_n | H_0) dx_1 dx_2 \cdots dx_n$$

и вероятность ошибки второго рода

$$\beta = 1 - \int_W f(x_1, x_2, \ldots, x_n | H_1) dx_1 dx_2 \cdots dx_n ,$$

где $\int_W (\cdot)$ - многомерный интеграл в пределах заданной критической области W.

Очевидно, что границы критической области W в рассматриваемой задаче однозначно определяются пороговым уровнем l_0. В зависимости от выбора l_0 будем иметь различные критерии оптимального решения поставленной задачи. К числу критериев проверки простых гипотез теория вероятностей относит

следующие: критерий наименьшего среднего риска (критерий Байеса), критерий идеального наблюдателя, или максимума апостериорной вероятности, критерий максимального правдоподобия и критерий Неймана - Пирсона.

В критерии *наименьшего среднего риска* (критерий Байеса) полагают относительный пороговый уровень равным

$$l_0 = \frac{\Pi_{10}}{\Pi_{01}} \frac{q}{p},$$

где Π_{10} – коэффициент потерь, связанный с ошибкой первого рода (H_1/H_0); Π_{01} – коэффициент потерь, связанный с ошибкой второго рода (H_0/H_1); $q = P(H_0)$ – априорная вероятность нулевой гипотезы $H_0 (p + + q \equiv 1)$; $p = P(H_1)$ - априорная вероятность альтернативного распределения H_1.

Использование данного критерия гарантирует достижение минимальных средних потерь, т.е. потерь по множеству наблюдений, однозначно связанных с коэффициентами Π_{10} и Π_{01}. Средние потери - это средний риск принятия каждого конкретного решения. Недостатком данного критерия является наибольший объём требуемой априорной информации, которая во многих практических задачах отсутствует.

В критерии *максимума апостериорной вероятности* полагают $l_0 = q/p$. Это частный случай предыдущего критерия при равенстве $\Pi_{10} = \Pi_{01} = \text{const}$.

В критерии *максимального правдоподобия* $l_0 = 1$. Его можно считать частным случаем предыдущего критерия при $\Pi_{10} = \Pi_{01} = \text{const}$ и $q = p = 0{,}5$ (случай равновероятностных простых гипотез).

В критерии *Неймана - Пирсона* пороговый уровень l_0 рассчитывается заранее из учета требований к максимальной вероятности ошибки первого рода:

$$l_0 = \mathrm{Arg}\left\{ \int\limits_{W_0(l_0)} f\left(x_1, x_2, \ldots, x_n | H_0\right) dx_1 dx_2 \ldots dx_n = \alpha^* \right\}, \alpha^* \neq 0 = \mathrm{const}.$$

Среди данных критериев наибольшее распространение получили критерии максимального правдоподобия и Неймана - Пирсона как не требующие наличия большой априорной информации.

12. 3. Проверка гипотез о математическом ожидании нормальной генеральной совокупности

Пусть требуется проверить исходную гипотезу

$$H_0 : X \sim N\left(m_0, \sigma^2\right)$$

против альтернативы

$$H_1 : X \sim N\left(m_1, \sigma^2\right),$$

где $\sigma^2 = D(X) = \mathrm{const}$.

Нетрудно убедиться, что предложенная задача относится к случаю проверки простых гипотез. Поэтому воспользуемся для ее решения выводами предыдущего раздела.

Имеем:

функцию правдоподобия общего вида

$$f\left(X_1, X_2, \ldots, X_n | H_j\right) =$$

$$= \prod_{i=1}^{n} f(X_i | H_j) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(X_i - m_j\right)^2 \right] \quad ;$$

отношение правдоподобия общего вида

$$l\left(X_1, X_2, \ldots, X_n\right) = \exp\left[\frac{1}{2\sigma^2} \left\{ 2\left(m_1 - m_0\right) \sum_{i=1}^{n} X_i - n\left(m_1^2 - m_0^2\right) \right\} \right].$$

Воспользуемся критерием максимального правдоподобия. Из условия $l(\cdot) \geq 1$ (или $\ln l \geq 0$) при дополнительном предположении $m_1 > m_0$ получим в итоге следующее оптимальное правило:

$$W: \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X_n} \geq \frac{m_1 + m_0}{2}.$$

Здесь $\overline{X_n}$ - минимальная решающая статистика, которая называется достаточной статистикой задачи; $\frac{m_1 + m_0}{2}$ – пороговый уровень.

Полученное правило принятия решения в задаче проверки простых гипотез о МО по критерию максимального правдоподобия формулируется следующим образом. *Если средняя выборочная величина* $\overline{X_n}$ *превышает порог* $\frac{m_1 + m_0}{2}$, *то гипотеза* H_0 *отвергается (принимается* H_1*). В противном случае принимается гипотеза* H_0.

Дадим количественную характеристику эффективности полученного алгоритма оптимального решения. Вероятность ошибки первого рода равна

$$\alpha = P\left\{\overline{X_n} > \frac{m_1 + m_0}{2} \middle| H_0\right\}.$$

Поскольку $\overline{X_n}|H_0 \sim N\left(m_0, \sigma^2/n\right)$, где n - объём выборки, то вероятность случайного события

$$P\left\{\overline{X_n} > \frac{m_1 + m_0}{2} \middle| H_0\right\} = P\left\{\frac{\overline{X_n} - m_0}{\sigma}\sqrt{n} > \frac{m_1 - m_0}{2\sigma}\sqrt{n}\right\} =$$

$$= \frac{1}{2}\left[1 - P\left\{\left|\frac{\overline{X_n} - m_0}{\sigma}\sqrt{n}\right| < \frac{m_1 - m_0}{2\sigma}\sqrt{n}\right\}\right] = \frac{1}{2}\left[1 - \Phi\left(\frac{m_1 - m_0}{2\sigma}\sqrt{n}\right)\right] =$$

$$= \frac{1}{2}\left[1 - \Phi\left(z_p\right)\right],$$

где $\Phi\left(z_p\right)$ – значение интеграла вероятности или функции Лапласа в точке

$$z_p = \frac{m_1 - m_0}{2\sigma}\sqrt{n}\,.$$

Аналогично получаем выражение для вероятности ошибки второго рода

$$P\left\{\overline{X_n} < \frac{m_1 + m_0}{2}\middle| H_1\right\} = \frac{1}{2}\left[1 - \Phi\left(\frac{m_1 - m_0}{2\sigma}\sqrt{n}\right)\right] = \alpha\,.$$

Таким образом, чем больше разность $m_1 - m_0$, т.е. чем больше различия в гипотезах H_0 и H_1, тем выше эффективность оптимального алгоритма и меньше вероятности ошибок α и β.

Рассмотренная задача и алгоритм её оптимального решения имеют непосредственное отношение к проблеме оптимального обнаружения и распознавания сигналов на фоне случайных помех во многих областях радиотехники.

12. 4. Проверка гипотезы о равенстве математических ожиданий двух независимых нормальных генеральных совокупностей

Рассмотрим две СВ $X \sim N\left(m_x, \sigma_x^2\right)$ и $Y \sim N\left(m_y, \sigma_y^2\right)$, причём будем полагать, что их дисперсии одинаковы $\sigma_x^2 = \sigma_y^2 = \sigma$, но σ^2 - в общем случае неизвестна. Пусть имеются две независимые выборки из данных генеральных совокупностей $X_1, X_2, \ldots, X_{n_1}$ объёма $n_1 > 1$ и $Y_1, Y_2, \ldots, Y_{n_2}$ объёма $n_2 \neq n_1$. Поставим задачу проверки гипотезы $H_0 : m_x = m_y$ против альтернативы $H_1 : m_x \neq m_y$. Это классическая формулировка задачи проверки сложной гипотезы против сложной альтернативы. Принятие любой гипотезы в данном случае не позволяет однозначно определить законы распределения СВ *X* и *Y*.

В общем случае задачи такого рода не имеют строго оптимального решения. Поэтому преобразуем исходную формулировку задачи в

эквивалентный вид, вводя дополнительно СВ T=X-Y. При этом $T \sim N\left(m_T, \sigma_T^2\right)$, где $m_T = m_x - m_y$, а $\sigma_T^2 = \sigma_x^2 + \sigma_y^2 = 2\sigma^2$.

Тогда задача проверки рассматриваемых гипотез может быть сформулирована так :

$$H_0: m_T = 0;$$

$$H_1^{'}: |m_T| > 0.$$

Её решение по аналогии с решением задачи проверки простых гипотез принимает следующий вид:

$$W: \left|\overline{T}_{n_1+n_2}\right| > t_{n_1+n_2-2,p},$$

где $\overline{T}_{n_1+n_2}$ - среднее выборочное значение СВ T по выборке суммарного объёма $n_1 + n_2$:

$$\overline{T}_{n_1+n_2} = \frac{\overline{X}_{n_1} - \overline{Y}_{n_2}}{\sqrt{\left(\frac{1}{n_1} - \frac{1}{n_2}\right)\frac{(n_1-1)\hat{S}_x^2 + (n_2-1)\hat{S}_y^2}{n_1+n_2-2}}}$$

- нормированная СВ, имеющая распределение Стьюдента с k, n_1, n_2 степенями свободы; в знаменателе дроби дается выборочная оценка неизвестной дисперсии σ_T^2 с использованием двух выборочных дисперсий $\hat{S}_x^2$ и $\hat{S}_y^2$ СВ X и Y соответственно. При этом $t_{n_1+n_2-2,p}$ определяет p% точку распределения Стьюдента, заданную соответствующими таблицами распределения для уровня значимости решения $\alpha = 1 - p$

12. 5. Проверка гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей

Следуя логике и схеме предыдущих задач из области проверки статистических гипотез, сформулируем следующую актуальную для практики задачу.

Пусть

$$X \sim N\left(m_x, \sigma_x^2\right);$$

$$Y \sim N\left(m_y, \sigma_y^2\right)$$

- две нормальные генеральные совокупности, МО и дисперсии которых заранее неизвестны. Проверим гипотезу о равенстве дисперсий этих генеральных совокупностей в предположении, что $\sigma_x^2 \geq \sigma_y^2$. Ограничения данного типа нельзя считать существенными для практики, поскольку всегда имеется возможность искусственно обеспечить превышение одной дисперсии над другой (например, умножением на масштабный коэффициент).

Пусть имеются две выборки наблюдений:

$$X_1, X_2, ..., X_n \text{ объёма } n_1 < \infty;$$

$$Y_1, Y_2, \ldots, Y_n \text{ объёма } n_2 \neq n_1.$$

С учётом этого сформулируем задачу в терминах проверки двух статистических гипотез

$$H_0: \sigma_x^2 = \sigma_y^2 \quad ;$$

$$H_1: \sigma_x^2 > \sigma_y^2 \quad .$$

Это классический случай проверки сложной гипотезы против сложной альтернативы (аналогичный случай рассматривался в предыдущем разделе темы). Решение такой задачи основывается на двух независимых оценках неизвестных дисперсий. Обозначим их как

$$\hat{S}_x^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2;$$

$$\hat{S}_y^2 = \frac{1}{n_2 - 1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2$$

и после этого воспользуемся решающим правилом вида

$$W: \frac{\hat{S}_x^2}{\hat{S}_y^2} > F_{\alpha,k_1,k_2} = \text{const},$$

где $\hat{S}_x^2 / \hat{S}_y^2 \overset{\Delta}{=} F$ случайная решающая статистика, подчинённая закону F - распределения Фишера с k_1, k_2 степенями свободы. При этом $k_1 = n_1 - 1$, $k_2 = n_1 - 1$, $F_{\alpha,k_1,k_2} - \alpha\%$ точка F-распределения Фишера, заданная вероятностью ошибки первого рода : $P\left\{F > F_{\alpha,k_1,k_2}\right\} = \alpha$.

Как и в предыдущем разделе, здесь контролируется только вероятность ошибки первого рода, что можно рассматривать как своеобразную плату за сложность проверяемых гипотез.

Пример 1. Пусть заданы следующие характеристики:

$$\hat{S}_x^2 = 9{,}6; \quad n_1 = 10;$$
$$\hat{S}_y^2 = 5{,}7; \quad n_2 = 15.$$

С использованием их будем иметь решающую статистику $F = \dfrac{\hat{S}_x^2}{\hat{S}_y^2} = \dfrac{9{,}6}{5{,}7} = 1{,}68$

со степенями свободы $k_1 = 9, \quad k_2 = 14.$ При заданном уровне значимости $\alpha = 5 \cdot 10^{-2}$ по таблицам F - распределения находим $F_{0.05;9;14} = 2{,}65$. При этом $1{,}68 < 2{,}65$, следовательно, в имеющихся выборочных данных не находим

достаточных оснований для отклонения нулевой гипотезы о равенстве двух дисперсий. Поэтому гипотеза H_0 может быть принята как справедливая.

Примечание. Непрерывная СВ X имеет *F*-распределение Фишера с k_1, k_2 степенями свободы (натуральные числа), если соответствующая ей плотность распределения выражается формулой

$$f(x)=\begin{cases}\dfrac{\left(\dfrac{k_1}{k_2}\right)^{\frac{k_1}{2}} x^{\frac{k_1}{2}-1}}{B\left(\dfrac{k_1}{2},\dfrac{k_2}{2}\right)}\left(1+\dfrac{k_1}{k_2}x\right)^{-\frac{(k_1+k_2)}{2}}, x>0;\\ 0, x\le 0,\end{cases}$$

где $B(k_1/2, k_2/2)$ - бета-функция двух аргументов.

Видно, что данная плотность распределения не вычисляется в элементарных функциях, поскольку бета-функция относится к классу специальных функций. На практике пользуются приближёнными значениями данной плотности распределения, вычисленными с высокой степенью точности на ЭВМ по известным аппроксимациям бета-функции. Полученные приближённые значения представлены в виде таблицы *F*-распределения (имеются в специальной литературе).

12.6. Проверка гипотез о законе распределения

Рассмотрим СВ *X*, распределённую по некоторому закону с неизвестной интегральной функцией распределения *F(x)*. Предположим при этом, что нам известна область её допустимых значений $[x_{\min}, x_{\max}]$. И пусть имеется случайная выборка $X_1, X_2, \ldots, X_n$ конечного объёма $n<\infty$ из данной генеральной совокупности. Поставим следующую статистическую задачу: проверить гипотезу $H_0: F(x)=F^*(x)$ против альтернативы

$H_1: F(x) \neq F^*(x)$, где $F^*(x)$ определяет некоторый гипотетический закон распределения.

Таким образом, имеем задачу проверки простой гипотезы против сложной альтернативы. Задача решается в терминах вариационных рядов. По имеющимся выборочным данным $X_1, X_2, \ldots, X_n$ построим ИВР (табл.12.2).

Таблица 12. 2

№ варианта	Интервал	Эмпирическая частота	Теоретическая частота
1	$[x_{\min}, x_1]$	m_1	np_1
2	$[x_1, x_2]$	m_2	np_2
...	...	...	...
L	$[x_{l-1}, x_{\max}]$	m_l	np_l
Всего	$[x_{\min}, x_{\max}]$	$\sum m_i = n$	n

В правой графе таблицы представлены для сопоставления теоретические частоты каждого варианта, определённые по гипотетическому закону распределения согласно формуле

$$np_i = n \int_{x_{i-1}}^{x_i} f^*(x)dx = \left(F^*(x_i) - F^*(x_{i-1})\right)n .$$

В литературе по математической статистике строго доказано, что СВ $\sum_{i=1}^{l} \frac{(m_i - np_i)^2}{np_i} = \chi^2$ подчинена распределению χ^2 - Пирсона с $k = l - r - 1$ степенями свободы. Здесь *l* - число вариантов вариационного ряда, а *r* - число неизвестных параметров гипотетического распределения $F^*(x)$, которые требуют предварительного оценивания по имеющейся выборке наблюдений.

Описанное правило принятия решений записывают следующим образом:

$$W: \chi^2 > \chi^2_{k,\alpha},$$

где $\chi^2_{k,\alpha}$ - $\alpha\%$ точка распределения Пирсона с *k* степенями свободы.

Если решающая статистика χ^2 удовлетворяет данному соотношению, то принимается решение в пользу гипотезы H_1. В противном случае полагают, что нет оснований для отклонения исходной гипотезы. Но здесь, как и в предыдущих случаях, контролируется только вероятность ошибки первого рода.

Рассмотрённое правило принятия решения получило название критерия χ^2 - Пирсона. Наряду с ним на практике для проверки гипотез о законах распределения применяют критерии Колмогорова, Смирнова и др. Большинство из них имеют ту же структуру, что и рассмотренный критерий Пирсона. Их отличия связаны с тем, что вместо статистики χ^2 используются другие варианты решающей статистики.

13. ОСНОВЫ КОРРЕЛЯЦИОННОГО АНАЛИЗА

13.1. Предмет корреляционного анализа

При статистическом анализе случайных процессов и систем часто исследуют взаимное влияние одних СВ на другие. Указанное влияние может проявляться в самых разных формах. Наиболее общая форма отображается в многомерном законе распределения

$$F(x,y) \stackrel{\Delta}{=} P(X < x, Y < y).$$

В большинстве задач оценить или задать совместное распределение системы СВ не представляется возможным ввиду недостаточности априорных сведений. В таких случаях переходят к упрощенным формам характеристик взаимозависимости СВ. В числе наиболее распространённых из них - коэффициент взаимной, или парной, корреляции $\rho_{x,y}$, который определяется по выражению

$$\rho_{x,y} \stackrel{\Delta}{=} \frac{1}{\sqrt{\sigma_x^2 \sigma_y^2}} K_{xy}.$$

Он характеризует одну из важнейших форм зависимости двух СВ, а именно корреляционную зависимость. Исследование корреляционной зависимости двух генеральных совокупностей X и Y по имеющимся последовательным наблюдениям $X_1, X_2, \ldots, X_n$ и $Y_1, Y_2, \ldots, Y_n$ равного объёма $n < \infty$ является основным предметом исследования корреляционного анализа.

13. 2. Общие положения корреляционного анализа

По определению коэффициент взаимной корреляции равен

$$\rho_{x,y} \stackrel{\Delta}{=} \frac{1}{\sqrt{\sigma_x^2 \sigma_y^2}} M\left\{\left(X - m_x\right)\left(Y - m_y\right)\right\},$$

где $M\left\{\left(X - m_x\right)\left(Y - m_y\right)\right\}$ - второй центральный (корреляционный) момент.

Оценкой неизвестного коэффициента корреляции по выборкам $\{X_i\}$ и $\{Y_i\}$ из рассматриваемых генеральных совокупностей является выборочный коэффициент корреляции r_n, где n - объём выборок. Для получения выборочного коэффициента корреляции оценивают, во - первых, средние значения двух генеральных совокупностей

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \; ; \quad \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i \; ,$$

во - вторых, неизвестные дисперсии этих совокупностей

$$S_x^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2 ; \; S_y^2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \overline{Y}_n\right)^2 .$$

На основе полученных данных вычисляют искомый коэффициент корреляции

$$r_n = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)\left(Y_i - \overline{Y}_n\right)}{\sqrt{S_x^2 S_y^2}} \; .$$

Утверждение 13.1. Выборочный коэффициент корреляции принимает свои возможные значения строго в интервале $[-1,+1]$.

Доказательство следует автоматически из очевидного соотношения

$$\frac{1}{n}\sum\left(\frac{\left(X_i-\overline{X}_n\right)}{S_x}\pm\frac{\left(Y_i-\overline{Y}_n\right)}{S_y}\right)^2\geq 0.$$

Путём простых вычислений получаем $\pm 2r_n+2\geq 0$ или

$$|r_n|\leq 1.$$

13. 3. Проверка гипотез о значимости коэффициента корреляции

Утверждение 13.2. Статистика $t=r_n\sqrt{(n-2)\Big/\left(1-r_n^2\right)}$, вычисленная по двум имеющимся выборкам из некоррелированных нормальных генеральных совокупностей *X* и *Y*, имеет распределение Стьюдента с $k = n - 2$ степенями свободы.

Основные свойства статистики *t*:

1) в случае слабой или нулевой корреляции $|t|<<1$;

2) случаю сильной корреляции соответствует значение статистики $|t|>>1$.

Поставим задачу проверки нулевой гипотезы $H_0:\rho_{x,y}=0$ против её альтернативы $H_1:\rho_{x,y}>0$.

Используя в роли решающей статистики статистику *t* и учитывая при этом её свойства, сформулируем следующее правило решения поставленной задачи:

$$\text{W:}\quad |t|>t_{\alpha,k},$$

где $t_{\alpha,k}$ - $\alpha\%$ точка распределения Стьюдента с *k* степенями свободы, полученная по таблицам распределения для заданного уровня значимости α. Величина α здесь определяет допустимую вероятность ошибки первого рода. Если предложенное решающее правило выполняется, то гипотеза H_0 отвергается, так как нет достаточных оснований считать рассматриваемые генеральные совокупности некоррелированными. В таком случае возникает следующая статистическая задача: на основании выборочных данных оценить количественно существующую корреляционную связь. На практике эта задача

решается посредством определения доверительного интервала для неизвестного коэффициента корреляции $\rho_{x,y}$.

Утверждение 13.3. Решающая статистика $z_n = \frac{1}{2}\ln\left(\frac{1+r_n}{1-r_n}\right)$, вычисленная по выборкам из двух коррелированных нормальных генеральных совокупностей *X* и *Y*, подчиняется нормальному закону

$$z_n \sim N\left(m_z, \sigma_z^2\right)$$

с параметрами $m_z = \frac{1}{2}\ln\left(\frac{1+\rho_{x,y}}{1-\rho_{x,y}}\right) + \frac{\rho_{x,y}}{2(n-1)}$, $\sigma_z^2 = \frac{1}{n-3}$ (доказано Р. Фишером).

Используемое в данном утверждении логарифмическое преобразование соответствует функции гиперболического арктангенса $z_n = Ar\tanh(r_n)$. Напротив, $r_n = \tanh(z_n)$- функция гиперболического тангенса. Значения этих функций подробно табулированы.

Следуя обычной методике определения доверительного интервала значений неизвестного параметра по выборке (раздел 13.2), получаем интервальную оценку коэффициента корреляции следующего вида:

$$\tanh(z_1) < \rho_{x,y} < \tanh(z_2).$$

Здесь: $z_{1,2} = \frac{1}{2}\ln\left(\frac{1-r_n}{1+r_n}\right) \mp Z_p \frac{1}{\sqrt{n-3}}$ - нижний и верхний пороговые уровни; Z_p - аргумент функции Лапласа, обеспечивающий равенство $\Phi\left(Z_p\right) = P$ (*P*- доверительная вероятность).

Соответственно этому длина доверительного интервала определяется простым соотношением

$$\Delta_\rho = 2\frac{Z_p}{\sqrt{n-3}}.$$

При увеличении объёма наблюдений n величина доверительного интервала уменьшается по закону $1/\sqrt{n}$.

13. 4. Некоторые сведения из теории регрессионного анализа

Определим функцию регрессии (или просто регрессию) СВ Y относительно X по выражению условного МО вида

$$M(Y/X=x)\overset{\Delta}{=}\int_{-\infty}^{+\infty} yf(y/x)dy \quad ,$$

где условная плотность вероятности $f(y/x)$ находятся по формуле Байеса

$$f(y/x)=\frac{f(x,y)}{f(x)}=\frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)dy}.$$

Аналогично, регрессия X относительно Y равна

$$M(X/Y=y)\overset{\Delta}{=}\int_{-\infty}^{+\infty} xf(x/y)dx.$$

Утверждение 13.4. Для случая двух коррелированных нормальных генеральных совокупностей X и Y регрессия Y относительно X определяется линейной зависимостью

$$M(Y/X=x)=M(Y)+\rho_{x,y}\frac{\sigma_y}{\sigma_x}(x-M(X)).$$

Доказательство. По определению корреляционного момента можно записать

$$K_{x,y}\overset{\Delta}{=}\int_{-\infty-\infty}^{+\infty+\infty}(x-m_x)(y-m_y)f(x,y)dxdy=$$

$$=\int_{-\infty}^{+\infty}(x-m_x)\int_{-\infty}^{+\infty}(y-m_y)f(y/x)dyf(x)dx=$$

$$=\int_{-\infty}^{+\infty}(x-m_x)[M(Y/X)-m_y]f(x)dx.$$

Найдем выражение для корреляционного момента $K_{x,y}$ при условии равенства $x = x_0$, т.е. при фиксированном значении X на некотором постоянном уровне. Для этого случая нормальная плотность распределения преобразуется в дельта - функцию Дирака, так как

$$\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left[-\frac{1}{2\sigma^2}(x-x_0)^2\right]\Bigg|_{\sigma_x\to 0} = \delta(x-x_0) \quad .$$

Используя фильтрующие свойства дельта-функции, при $x = x_0$ окончательно получаем

$$K_{xy}\big|_{x=x_0} = \int_{-\infty}^{+\infty}(x-m_x)\left[M\left(\frac{Y}{X}\right)-m_y\right]\delta(x-x_0)dx =$$

$$=(x_0-m_x)\left(M\left(\frac{Y}{x_0}\right)-m_y\right).$$

Или

$$M(Y/X=x_0) = M(Y)+\rho_{xy}\frac{\sigma_x\sigma_y}{(x_0-M(x))^2}(x_0-M(x)) =$$

$$= M(Y)+\rho_{xy}\frac{\sigma_y}{\sigma_x}(x_0-M(x))$$

для любого x_0 из области возможных значений $\{x\}$. Это и требовалось доказать.

Величина $\beta_y \overset{\Delta}{=} \rho_{xy}\frac{\sigma_y}{\sigma_x}$ называется *коэффициентом линейной регрессии* СВ Y на X. Аналогично, β_x - коэффициент линейной регрессии X на Y. Таким образом, между коэффициентом корреляции ρ_{xy} и любым из коэффициентов регресии β_x, β_y существует однозначная пропорциональная связь. Такого же рода взаимосвязь существует и между предметами корреляционного и регрессионного анализа. Предметом исследования регрессионного анализа является уравнение регрессии, которое записывается

$$\overline{Y}(X_i) = \overline{Y}_n + r_n \frac{S_y}{S_x}\left(X_i - \overline{X}_n\right)$$

или

$$\overline{X}(Y_i) = \overline{X}_n + r_n \frac{S_x}{S_y}\left(Y_i - \overline{Y}_n\right).$$

Понятие уравнения регрессии можно рассматривать как выборочный эквивалент понятия функции регрессии. Поясним это понятие.

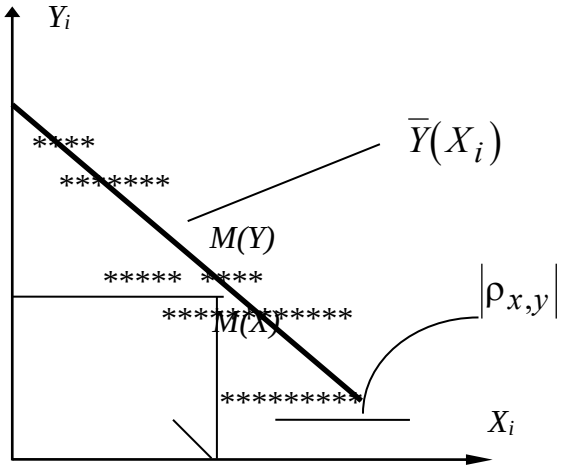


Рис 13.1

На рис. 13.1 звёздочками обозначены отдельные исходы наблюдений над двумя генеральными совокупностями Y и X. Множество звёздочек - это поле событий (элементарных исходов). Сплошная прямая построена по уравнению регрессии $\overline{Y}(X_i)$. Она определяет линейный характер зависимости условного среднего значения СВ Y от аргумента X.

Неравенство нулю соответствующего коэффициента корреляции $\rho_{x,y}$ предполагает наличие линейной связи между выборочными данными $Y_1, Y_2, \ldots, Y_n$ и $X_1, X_2, \ldots, X_n$ в статистическом смысле. Поэтому часто на практике применительно к нормальному распределению вероятностей коэффициент корреляции $\rho_{x,y}$ отождествляют с линейной зависимостью СВ Y и X.

Задача оценивания коэффициентов регрессии β_x и β_y решается по аналогии с основной задачей корреляционного анализа: вместо неизвестных коэффициентов регресии используются их выборочные оценки

$$b_y = r_n \frac{S_y}{S_x};$$

$$b_x = r_n \frac{S_x}{S_y}.$$

При анализе эффективности этих оценок также существует прямая аналогия с идеями корреляционного анализа. Без доказательства сформулируем следующее ключевое утверждение.

Утверждение 13.5. Доверительный интервал для оценки коэффициента регресии по выборкам $\{x_i\}$ и $\{y_i\}$ из двух коррелированных нормальных генеральных совокупностей равных объёмов *n* определяется выражением

$$b_y - t_{\alpha,k} \frac{S_y}{S_x} \sqrt{\frac{1-r_n^2}{n-2}} < \beta_y < b_y + t_{\alpha,k} \frac{S_y}{S_x} \sqrt{\frac{1-r_n^2}{n-2}},$$

где $t_{\alpha,k}$ - $\alpha\%$ точка (α - уровень значимости) распределения Стъюдента с $k = n - 2$ степенями свободы.

Величина установленного доверительного интервала возрастает при увеличении требований к надёжности интервальной оценки (α - мало) и зависит от соотношения выборочных дисперсий двух генеральных совокупностей. Чем больше коррелированность этих двух совокупностей (больше r_n), тем меньше длина доверительного интервала и, следовательно, точнее интервальная оценка неизвестного коэффициента регрессии *Y* на *X*.

14 ЛАБОРАТОРНЫЕ РАБОТЫ. МЕТОДИЧЕСКИЕ УКАЗАНИЯ И КОНТРОЛЬНЫЕ ВОПРОСЫ К НИМ

ЛАБОРАТОРНАЯ РАБОТА 1

Основы статистического описания

Порядок выполнения работы

1. Упорядочить выборку (таблица 1).

2. Построить эмпирическую функцию распределения и гистограмму.

3. Найти выборочные математическое ожидание, дисперсию, коэффициент асимметрии и эксцесс, выборочную медиану.

4. Составить отчет, в котором привести расчетную таблицу 2, график эмпирической функции распределения, гистограмму и вычисленные значения выборочных числовых характеристик.

5.Ответить устно на контрольные вопросы.

Упорядочение выборки

Упорядочение выборки удобно сделать следующим образом: выписать все выборочные значения x_i (на отдельную карточку); расположить карточки в порядке возрастания выборочных значений.

В дальнейшем под номером выборочного значения будет пониматься номер после упорядочения.

Построение эмпирической функции распределения и гистограммы

Полагая вероятность каждого значения x_i равной 1/n, получим распределение выборки. **Эмпирическая функция распределения** или функция распределения выборки $F^*(x)$ тогда будет определяться следующим образом:

$$F^*(x) = \frac{n_x}{n}, \qquad (1)$$

где n - объем выборки, n_x - число выборочных значений, меньших x.

При большом объеме $(n \geq 50)$ выборку целесообразно предварительно подвергнуть **группировке** следующим образом.

Построить интервал, содержащий все выборочные значения, левый и правый концы этого интервала – приближенные (с малым числом значащих

цифр) значения наименьшего и наибольшего элементов выборки, причем в первом случае приближение берется с недостатком, а во втором – с избытком.

Таблица Л1

2,551 016	1,182 147	0,784 914	0,543 969	0,370 497
0,123 475	0,028 970	1,514 199	0,938 832	0,644 410
0,294 172	0,172 712	0,071 059	2,269 003	1,143 489
0,529 767	0,359 708	0,226 162	0,116 188	0,022 701
0,912 381	0,627 765	0,432 929	0,284 614	0,644 831
0,108 985	0,016 493	1,391 414	0,887 003	0,611 552
2,075 379	1,107 083	0,744 048	0,515 881	0,349 101
0,275 199	0,157 049	0,057 722	1,927 765	1,072 682
0,502 296	0,338 672	0,209 119	0,101 864	0,010 347
0,862 617	0,595 748	0,409 327	0,265 924	0,149 362
1,808 440	1,040 075	0,705 673	0,489 001	0,328 414
0,094 823	0,004 260	1,288 835	0,839 146	0,580 335
0,256 785	0,141 768	0,044 663	1,708 286	1,099 087
0,475 982	0,318 322	0,193 528	0,087 860	3,666 642
0,816 525	0,565 292	0,386 585	0,247 778	0,144 266
1,621 985	0,979 563	0,669 527	0,463 229	0,308 390
0,080 974	2,740 878	1,200 742	0,794 363	0,550 603
0,238 898	0,126 852	0,031 872	1,546 169	0,951 370
0,450 731	0,298 613	0,176 366	0,074 163	2,383 084
0,773 600	0,536 252	0,364 641	0,230 143	0,119 526
1,478 563	0,924 394	0,635 357	0,438 479	0,288 987
0,067 429	2,156 978	1,123 547	0,753 194	0,522 222
0,221 509	0,112 284	0,019 357	1,417 563	0,898 534
0,426 461	0,279 507	0,160 612	0,060 761	1,991 291
0,733 434	0,508 501	0,343 442	0,212 993	0,105 125
1,361 986	0,837 702	0,802 959	0,414 671	0,270 168
0,054 165	1,860 452	1,054 846	0,714 200	0,495 074
0,204 591	0,098 048	0,007 049	1,310 953	0,849 819
0,403 098	0,620 967	0,145 245	0,047 639	1,752 317
0,695 695	0,481 930	0,322 938	0,196 300	0,091 049

1,263 774	0,826 816	0,572 158	0,391 737	0,251 900
0,041 180	1,660 161	0,992 954	0,677 647	0,469 056
0,188 118	0,084 128	3,014 916	1,219 908	0,804 629
0,380 578	0,242 962	0,234 861	0,644 572	0,764 142
1,449 796	0,065 355	0,217 582	0,421 017	0,724 561
1,338 022	0,051 159	0,200 768	0,389 853	0,687 338
1,243 237	0,038 235	0,184 395	0,375 517	0,662 208
1,160 953	0,025 573	0,168 440	0,353 952	0,618 948
1,088 252	0,013 163	0,152 881	0,333 106	0,587 369
1,023 134	0,000 994	0,137 701	0,312 933	0,557 309

Полученный интервал разделить на m равных частей (интервалов), заботясь при этом (для упрощения вычислений) о том, чтобы границы интервалов были числами с малым количеством значащих цифр.

Все расчеты удобно оформить в виде таблицы 2.

Таблица Л2

1	2	3	4	5	6
№ п/п	J (до объединения)	n_i (до объединения)	J (после объединения)	n_i (после объединения)	P_i^*
1	(0,000-0,185)	50	(0,000-0,185)	50	0,252525
2	(0,185-0,370)	38	(0,185-0,370)	38	0,191919
3	(0,370-0,555)	28	(0,370-0,555)	28	0,141414
4	(0,555-0,740)	22	(0,555-0,740)	22	0,111111
5	(0,740-0,925)	17	(0,740-0,925)	17	0,085859
6	(0,925-1,110)	11	(0,925-1,295)	20	0,101010
7	(1,110-1,295)	9			
8	(1,295-1,480)	7	(1,295-1,850)	14	0,070707
9	(1,480-1,665)	4			
10	(1,665-1,850)	3			
11	(1,850-2,035)	3	(1,850-2,775)	9	0,045454
12	(2,035-2,220)	2			
13	(2,220-2,405)	2			
14	(2,405-2,590)	1			
15	(2,590-2,775)	1			

Продолжение таблицы Л2.

16	(2,775-2,960)	0	(2,775-3,700)	2	
17	(2,960-3,145)	1			
18	(3,145-3,330)	0			
19	(3,330-3,515)	0			
20	(3,515-3,700)	1			

7	8	9	10	11	12
$\tilde{x}_i$	$F^*(\tilde{x}_i)$	$\tilde{x}_i P_i^*$	$\tilde{x}_i^2 P_i^*$	$\tilde{x}_i^3 P_i^*$	$\tilde{x}_i^4 P_i^*$
0,092 500	0	0,023358	0,002161	0,000200	0,000018
0,277 500	0,252525	0,653258	0,014779	0,004101	0,001138
0,462 500	0,444444	0,665404	0,030249	0,013990	0,006470
0,647 500	0,585858	0,071944	0,046584	0,030163	0,019530
0,832 500	0,696969	0,071478	0,059505	0,049538	0,041240
1,110 000	0,782228	0,112121	0,124454	0,138144	0,153340
1,572 500	0,883838	0,111187	0,174841	0,274937	0,432339
2,312 500	0,954545	0,105119	0,243072	0,562104	1,299866

Выбрав число интервалов m = 20 (оно в процессе обработки может быть изменено), обозначим границы $x^{(0)}, x^{(1)}, ..., x^{(m)}$. Поместим значения границ в графу 2. Подсчитаем число элементов n_i выборки, попавших в i-й интервал. Может случиться, что отдельные значения x_i совпадут со значениями границ некоторых интервалов.

В таких случаях поступают различными способами: а) уславливаются

все такие элементы совокупности относить либо к правому, либо к левому интервалу; б) эти элементы учитываются и в левом и в правом интервале, полагая, что в каждый интервал попало по 1/2 элемента. Значения n_i поместим в графу 3.

Интервалы, в которые попадает малое число элементов (концевые), рекомендуется объединить. Для построения гистограммы желательно, чтобы число элементов в каждом интервале было не меньше 10, а число интервалов не меньше 8. Уточненные границы интервалов поместим в графу 4, а число элементов, попавших в каждый интервал после уточнения границ, в графу 5.

Обозначим середины полученных интервалов через $\tilde{x}_i$ и поместим эти значения в графу 7.

Замечание. В случае малого объема выборки в качестве $\tilde{x}_i$ будет выступать сами выборочные значения, которые поместим в графу 7. При этом графы 2-5 не заполняются.

Определим частоту попадания выборочных значений в i-й интервал:

$$P_i^* = \frac{n_i}{n}, \qquad (2)$$

где n_i - число элементов, попавших в i-й интервал, n - объем выборки. Полученные значения P_i^* поместим в графу 6. Их можно проверить, вычислив сумму частот $\sum_{i=1}^{m_1} P_i^*$, где m_i - число уточненных интервалов. Эта сумма должна быть равна единице.

Замечание. При малом объеме выборки частота каждого выборочного значения постоянна и равна $1/n$.

Эмпирическая функция распределения строится как функция распределения выборки $\tilde{x}_1$, $\tilde{x}_2$, ..., $\tilde{x}_{m_1}$, причем значение $\tilde{x}_1$ повторяется n_1 раз, $\tilde{x}_2 - n_2$ раза и т.д. (все элементы исходной выборки, попавшие в i-й интервал, заменены на $\tilde{x}_i$).

Очевидно, что для всех $x \in (\tilde{x}_i, \tilde{x}_{i+1}]$

$$F^*(x) = \frac{n_x}{n} = \frac{n_1 + n_2 + \ldots + n_{i-1}}{n} = \frac{n_1}{n} + \frac{n_2}{n} + \ldots + \frac{n_{i-1}}{n}.$$

Таким образом,

$$F^*(x) = \sum_{s=1}^{i-1} P_i^*, \quad x \in (\tilde{x}_i, \tilde{x}_{i+1}] \tag{3}$$

Кроме того, $F^*(x) = 0$ при $x \le x_1$ и $F^*(x) = 1$ при $x > \tilde{x}_{m_1}$.

Значения $F^*(x_i)$ можно поместить в графу 8. Пример графика эмпирической функции распределения представлен на рис.1.

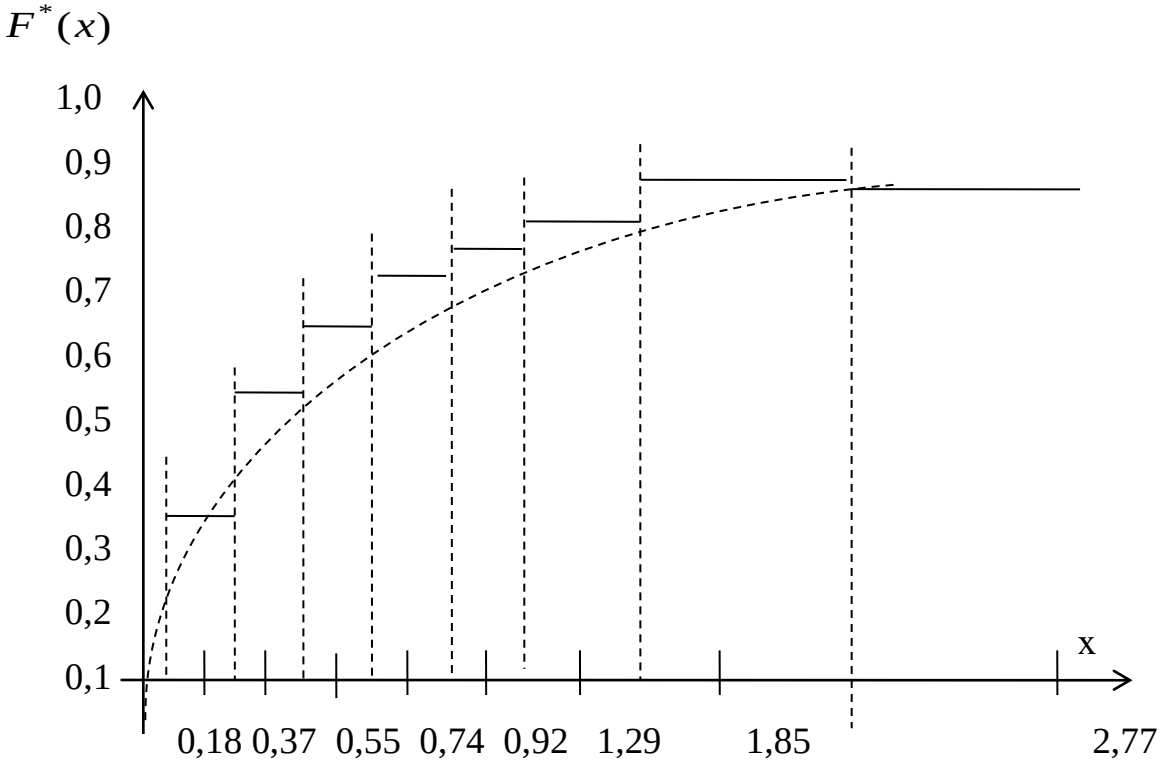


Рисунок 1

Доказано, что эмпирическая функция распределения, построенная по всей выборке, сходится по вероятности к функции распределения генеральной совокупности при $n \to \infty$. Но и эмпирическая функция распределения, построенная по группированной выборке, дает достаточно хорошее представление о функции распределения генеральной совокупности.

Если рассматривается выборка из генеральной совокупности значений

непрерывной случайной величины, то можно для более наглядного представления о функции распределения генеральной совокупности построить ломаную, соединив точки $\left(x^{(i)}, F^*(x^{(i)})\right)$ (рис.1, пунктирная линия).

Замечание. График эмпирической функции распределения можно построить непосредственно, взяв примерно двадцать идущих подряд элементов исходной (до упорядочения) выборки.

Гистограмма строится по группированной выборке следующим образом: над каждым интервалом (графа 4) строим прямоугольник, площадь которого равна частоте попадания в данный интеграл, т. е. высота прямоугольника равна частоте, деленной на длину соответствующего интервала (рис. 2).

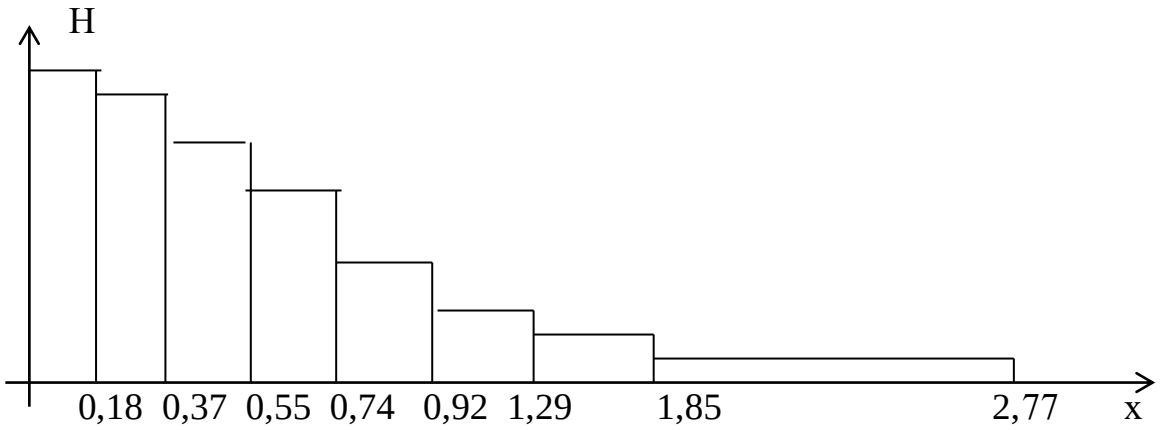


Рисунок 2

Плавная кривая, проведенная по средним точкам верхних оснований прямоугольников гистограммы для выборки из генеральной совокупности значений непрерывной случайной величины, дает представление о графике плотности распределения генеральной совокупности.

Замечание. При построении плавной кривой не следует стремиться проводить ее строго через средние точки оснований, так как число элементов, попадающих в каждый интервал, является случайным, и даже в том случае, когда исследуемая случайная величина строго следует закону распределения, характеризуемому плотностью вероятности $f(x)$, эти точки не совпадут точно с точками теоретической кривой.

Нахождение числовых характеристик выборки

Среднее значение выборки вычисляется по формуле

$$\overline{x} = \sum_{i=1}^{m_1} \widetilde{x}_i \cdot P_i^*, \qquad (4)$$

где $\widetilde{x}_i$ - середины уточненных интервалов (см. табл. 2, гр. 7), P_i^* - частоты попадания в i-й интервал (графа 6). Промежуточные результаты $\widetilde{x}_i \cdot P_i^*$ можно поместить в графу 9.

Выборочная дисперсия вычисляется по формуле

$$S^2 = \sum_{i=1}^{m_1} \left(\widetilde{x}_i - \overline{x}\right)^2 \cdot P_i^* \qquad (5)$$

или

$$S^2 = \sum_{i=1}^{m_1} \widetilde{x}_i^2 \cdot P_i^* - \overline{x}^2. \qquad (6)$$

Промежуточные значения $\widetilde{x}_i^2 \cdot P_i^*$ можно поместить в графу 10.

Выборочные коэффициент асимметрии и эксцесс вычисляются по формулам

$$A^* = \frac{m_3}{S^3}; \quad E^* = \frac{m_4}{S^4} - 3, \qquad (7)$$

где m_3, m_4 - выборочные центральные моменты 3 и 4-го порядка соответственно:

$$m_3 = \sum_{i=1}^{m_1} \left(\widetilde{x}_i - \overline{x}\right)^3, \quad m_4 = \sum_{i=1}^{m_1} \left(\widetilde{x}_i - \overline{x}\right)^4 \cdot P_i^* \qquad (8)$$

или

$$m_3 = \sum_{i=1}^{m_1} \widetilde{x}_i^3 \cdot P_i^* - 3\overline{x} \sum_{i=1}^{m_1} \widetilde{x}_i^2 \cdot P_i^* + 2\overline{x}^3, \qquad (9)$$

$$m_4 = \sum_{i=1}^{m_1} \widetilde{x}_i^4 \cdot P_i^* - 4\overline{x} \sum_{i=1}^{m_1} \widetilde{x}_i^3 \cdot P_i^* + 6\overline{x}^2 \sum_{i=1}^{m_1} \widetilde{x}_i^2 \cdot P_i^* - 3\overline{x}^4 \qquad (10)$$

Промежуточные значения $\widetilde{x}_i^3 \cdot P_i^*$, $\widetilde{x}_i^4 \cdot P_i^*$ можно поместить в графы 11

и 12.

Замечание. Более точными будут выборочные характеристики, полученные по всей выборке. Например,

$$\overline{x} = \sum^{n} x_i \frac{1}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad (11)$$

Аналогично для S^2, A^*, E^*.

Выборочная медиана M_e^* может быть определена по всей упорядоченной выборке следующим образом: если объем выборки n - нечетное число, то

$$M_e^* = x_{\left[\frac{n}{2}\right]+1} \qquad (12)$$

если n четное, то

$$M_e^* = \frac{1}{2}\left[x_{[n/2]} + x_{[n/2]+1}\right] \qquad (13)$$

где $[n/2]$ - целая часть $n/2$.

Для выборки из генеральной совокупности значений непрерывной случайной величины выборочная медиана может быть определена по графику эмпирической функции распределения как абсцисса точки с ординатой 1/2.

Найденные числовые характеристики выборки могут служить оценками соответствующих характеристик генеральной совокупности.

Пример выполнения и оформления лабораторной работы

Дана выборка, содержащая двести элементов (табл.1). Упорядочим выборку. Наименьшее число равно 0,000 994, наибольшее 3,666 642. Интервал (0,000; 3,700) разделим на 20 равных частей. Границы интервалов занесем в графу 2 табл. 2.

Число элементов, попавших в i-й интервал, занесем в графу № 3. Два числа – 3,014 916, 3,666 642, резко отличающиеся от других и полученные,

видимо, за счет грубых ошибок опыта можно отбросить. Таким образом, $n = 198$.

Объединим интервалы так, чтобы новые интервалы содержали не менее 8-10 элементов. Новые границы интервалов, а также число элементов, попавших в уточненные интервалы, поместим в графы 4 и 5. В графу 6 поместим частоты попаданий в каждый интервал. Далее таблица 2 заполняется в соответствии с описанием работ.

По полученным данным строится график эмпирической функции распределения (рис.1) и гистограмма (рис.2).

По формулам (4), (6), (7) вычисляются выборочные среднее, дисперсия, коэффициент асимметрии и эксцесс.

Предварительно удобно вычислить следующие суммы:

$$d_1 = \sum_{i=1}^{m_1} \tilde{x}_i \cdot P_i^* = 0{,}613\ 862; \qquad d_3 = \sum_{i=1}^{m_1} \tilde{x}_i^3 \cdot P_i^* = 1{,}073\ 177;$$

$$d_2 = \sum_{i=1}^{m_1} \tilde{x}_i^2 \cdot P_i^* = 0{,}695645; \qquad d_4 = \sum_{i=1}^{m_1} \tilde{x}_i^4 \cdot P_i^* = 1{,}953941.$$

Тогда $\bar{x} = d_1 = 0{,}613\,862$.

$$S^2 = d_2 - \bar{x}^2 = 0{,}318\ 818; \quad S = \sqrt{S^2} = 0{,}564\ 640;$$

$$m_3 = d_3 - 3\bar{x}d_2 + 2\bar{x}^3 = 1{,}020\ 170;$$

$$m_4 = d_4 - 4\bar{x}d_3 + 6\bar{x}^2 d_2 - 3\bar{x}^4 = 0{,}465\ 641;$$

$$A^* = \frac{m_3}{S^3} = 5{,}666\,493; \quad E^* = \frac{m_4}{S^4} - 3 = 1{,}581\ 046.$$

Выборочную медиану M_e^* определим по графику эмпирической функции распределения: $M_e^* = 0{,}45$.

Контрольные вопросы

1. Разъясните понятия: генеральная совокупность, случайная выборка, оценка.

2. Что такое эмпирическая функция распределения и как она вычисляется

по данным выборки ?

3. Что такое гистограмма распределения и как она строится по данным выборки ?

4. Какие оценки параметров называются точечными ?

5.Что такое состоятельность, эффективность и несмещенность точечных оценок?

6. В чем состоит метод моментов определения точечных оценок ?

ЛАБОРАТОРНАЯ РАБОТА 2

Построение доверительных интервалов математического ожидания и дисперсии в случае выборки из нормальной генеральной совокупности

Порядок выполнения работы

1. По данной выборке найти оценки математического ожидания и дисперсии.

2. Найти доверительный интервал для математического ожидания, соответствующий доверительной вероятности $\gamma = 0{,}9;\ \gamma = 0{,}95;\ \gamma = 0{,}99$.

3. Найти доверительный интервал для дисперсии, соответствующий доверительной вероятности $\gamma = 0{,}9;\ \gamma = 0{,}98$.

4. Составить отчет, в котором привести исходный статистический материал, использованные расчетные формулы, результаты счета.

5. Ответить устно на контрольные вопросы.

Нахождение оценок математического ожидания и дисперсии

По данной выборке находим оценки математического ожидания и дисперсии генеральной совокупности:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i , \qquad (1)$$

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

или

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^{\,2} - \bar{x}^2, \qquad (3)$$

где n-объем выборки, x_i - элементы выборки $(i = 1,2,...,n)$.

Построение доверительного интервала математического ожидания

Полученные на первом этапе оценки называются **точечными** и являются случайными величинами, изменяющимися от выборки к выборке. Использование точечных оценок, построенных по выборкам малого объема (n ~ 10), может привести к существенным ошибкам. Например, среднее арифметическое $\bar{x}$, как оценка математического ожидания α, имеет дисперсию σ^2/n. При больших n дисперсия оценки мала, и реализация оценки $\bar{x}$ весьма тесно концентрируется около своего математического ожидания, равного α. При малых объемах выборки дисперсия оценки $\bar{x}$ может быть большой.

В случае использования точечных оценок, построенных по выборкам малого объема, необходимо указать, с какой степенью уверенности можно говорить о том, что отклонение оценки a^* от оцениваемого параметра a не превзойдет определенную величину.

По заданной вероятности γ (как правило, 0,9; 0,95; 0,99) определим число ε_γ, такое, что $P\left(\left|a^* - a\right| < \varepsilon_\gamma\right) = \gamma$,

или, что то же самое,

$$P\left(a^* - \varepsilon_\gamma < a < a^* + \varepsilon_\gamma\right) = \gamma. \qquad (4)$$

Интервал $\left(a^* - \varepsilon_\gamma;\ a^* + \varepsilon_\gamma\right)$, с вероятностью γ содержащий истинное значение оцениваемого параметра a, называется **доверительным интервалом**; границы его - случайные величины. Вероятность γ называется **доверительной вероятностью.**

Доверительный интервал может быть несимметричным относительно оцениваемого параметра.

В случае выборки из нормальной генеральной совокупности оценка имеет

нормальное распределение с параметрами $\left(\alpha, \frac{\sigma}{\sqrt{n}}\right)$, где α, σ - параметры нормальной генеральной совокупности. Если параметр σ известен, то

$$P\left(\left|\bar{x} - \alpha\right| < \varepsilon_\gamma\right) = 2\Phi\left(\frac{\varepsilon_\gamma}{\sigma_{\bar{x}}}\right) = 2\Phi\left(\frac{\varepsilon_\gamma \sqrt{n}}{\sigma}\right), \quad (5)$$

где $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{x^2}{2}} dx$ - функция Лапласа. Из равенства

$$2\Phi\left(\frac{\varepsilon_\gamma \sqrt{n}}{\sigma}\right) = \gamma, \quad (6)$$

используя табл. А4, можно определить ε_γ. Интервал $\left(\bar{x} - \varepsilon_\gamma; \bar{x} + \varepsilon_\gamma\right)$ является доверительным интервалом для математического ожидания, соответствующим доверительной вероятности γ.

Если параметр σ не известен, то простая замена этого параметра в формуле (5) его оценкой $S = \sqrt{S^2}$ в случае малой выборки может привести к существенным ошибкам.

В этом случае можно воспользоваться случайной величиной $t = \sqrt{n-1} \cdot \frac{\bar{x} - \alpha}{3}$, где α - математическое ожидание генеральной совокупности, $\bar{x}$ и S - оценки параметров нормальной генеральной совокупности. В курсе математической статистики доказывается, что случайная величина t в выборке из нормальной генеральной совокупности имеет распределение Стьюдента (t - распределение) с (n-1) степенями свободы, распределение, не зависящее от параметров генеральной совокупности.

Пусть число $t_{\gamma, n-1}$ таково, что

$$P\left(|t| < t_{\gamma, n-1}\right) - \gamma, \quad (7)$$

где γ - заданная доверительная вероятность.

Равенство (7) означает, что $\left|\sqrt{n-1},\dfrac{\bar{x}-\alpha}{S}\right| < t_{\gamma,n-1}$ с вероятностью γ.

Последнее неравенство эквивалентно следующему:

$$\bar{x} - t_{\gamma,n-1}\frac{S}{\sqrt{n-1}} < \alpha < \bar{x} + t_{\gamma,n-1}\frac{S}{\sqrt{n-1}}. \qquad (8)$$

Следовательно, интервал $\left(\bar{x} - t_{\gamma,n-1}\dfrac{S}{\sqrt{n-1}};\ \bar{x} + t_{\gamma,n-1}\dfrac{S}{\sqrt{n-1}}\right)$ является доверительным интервалом математического ожидания, соответствующим доверительной вероятности γ.

Значения $t_{\gamma,n-1}$, зависящие от γ и числа степеней свободы $k = n-1$, могут быть определены по табл. А2.

Построение доверительного интервала дисперсии

В курсе математической статистики доказано, что в выборке из нормальной генеральной совокупности с параметрами (α,σ) случайная величина $\dfrac{nS^2}{\sigma^2}$, где S^2- оценка неизвестной дисперсии, равная $\dfrac{1}{n}\sum_{i=1}^{n}(x_i - \alpha)^2$, имеет распределение χ^2 с n степенями свободы. Если параметр α неизвестен, то в выражении S^2 можно заменить α на его оценку $\bar{x}$; в этом случае случайная величина $\dfrac{nS^2}{\sigma^2}$ также имеет распределение χ^2, но уже с $(n-1)$, а не с n степенями свободы.

Пусть числа χ_1^2 и χ_2^2 выбраны таким образом, что

$$P(\chi_1^2 < \chi^2 < \chi_2^2) = \gamma, \qquad (9)$$

где γ - заданная доверительная вероятность.

Равенство (9) означает, что $\chi_1^2 < \dfrac{nS^2}{\sigma^2} < \chi_2^2$ с вероятностью γ. Последнее двойное неравенство эквивалентно следующему:

$$\frac{nS^2}{\chi_2^2} < \sigma^2 < \frac{nS^2}{\chi_1^2}. \qquad (10)$$

Следовательно, $\left(\frac{nS^2}{\chi_2^2};\ \frac{nS^2}{\chi_1^2}\right)$ является доверительным интервалом дисперсии, соответствующим доверительной вероятности γ.

Однако по заданной вероятности γ можно построить множество доверительных интервалов для дисперсии. Принято χ_1^2 и χ_2^2 выбирать так, чтобы вероятности $P\left(\chi^2 < \chi_1^2\right)$ и $P\left(\chi^2 > \chi_2^2\right)$ были равны и равны $\frac{1-\gamma}{2}$ (рис. 1).

Соответствующие значения χ_1^2 и χ_2^2 могут быть определены по табл. А3.

Замечание. При больших объемах выборок можно воспользоваться тем, что рассмотренные оценки математического ожидания и дисперсии распределены асимптотически нормально.

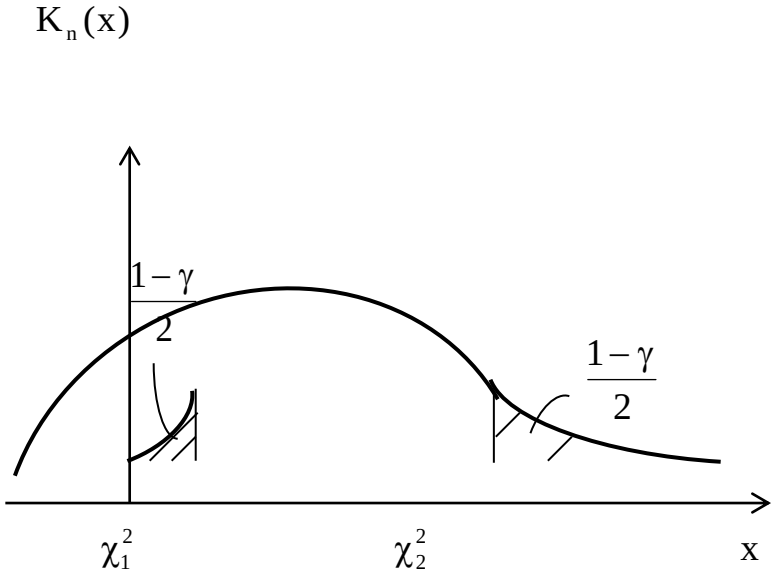


Рисунок 1. График плотности $K_n(x)$ распределения

Пример выполнения и оформления лабораторной работы

Дана выборка объемом $n = 20$ (табл. 1) из нормальной генеральной совокупности.

Таблица 1

	Элементы выборки	№ п/п	Элементы выборки
1	0,047	11	-1,7888
2	-0,451	12	-0,855
3	1,661	13	0,095
4	1,290	14	1,192
5	0,380	15	-0,059
6	0,496	16	0,118
7	-0,748	17	0,242
8	-0,083	18	1,739
9	-0,312	19	-0,412
10	-1,372	20	-0,426

Найдем по формулам (1) и (3) оценки математического ожидания и дисперсии.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{20}\left(0{,}047 - 0{,}451 + \ldots + 0{,}426\right) = 0{,}037;$$

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 = \frac{1}{20}\left(0{,}047^2 + 0{,}451^2 + \ldots + 0{,}426^2\right) - 0{,}037^2 = 0{,}819;$$

$$S = \sqrt{S^2} = 0{,}905.$$

Так как объем выборки невелик, для построения доверительного интервала для математического ожидания воспользуемся формулой (8):

$$\bar{x} - t_{\gamma,n-1} \cdot \frac{S}{\sqrt{n-1}} < \alpha < \bar{x} + t_{\gamma,n-1} \cdot \frac{S}{\sqrt{n-1}}.$$

Доверительную вероятность γ положим равной 0,95, $n = 20$. По табл. А2 по заданным γ и $k = n - 1$ определим $t_{\gamma,n-1} = 2{,}093$.

Доверительный интервал для математического ожидания, соответствующий доверительной вероятности $\gamma = 0{,}95$:

$$\left(0{,}037 - 2{,}093 \cdot \frac{0{,}905}{\sqrt{19}}; 0{,}037 + 2{,}093 \cdot \frac{0{,}905}{\sqrt{19}}\right) \text{ или } (-0{,}398;\ 0{,}472).$$

При построении доверительного интервала дисперсии положим $\gamma = 0{,}98$. Тогда $\frac{1-\gamma}{2} = 0{,}01$. χ_2^2 определим из условия $P(\chi^2 > \chi_2^2) = 0{,}01$; χ_1^2 определим из условия $P(\chi^2 > \chi_1^2) = 1 - \frac{1-\gamma}{2} = 0{,}99$ (рис. 1).

По табл. А3 по заданным вероятностям Р (0,01 и 0,99) и заданному числу степеней свободы $k = 19$ находим $\chi_1^2 = 7{,}633$, $\chi_2^2 = 36{,}191$.

Доверительный интервал дисперсии, соответствующий доверительной вероятности $\gamma = 0{,}98$, определяется по формуле (10):

$$\left(\frac{20 \cdot 0{,}819}{36{,}191};\ \frac{20 \cdot 0{,}819}{7{,}633}\right) \quad \text{или} \quad (0{,}450;\ \ 2{,}146\ ).$$

Контрольные вопросы

1. Какие оценки параметров называются точечными?

2. Что такое состоятельность, эффективность и несмещенность точечных оценок?

3. В чем состоит метод моментов определения точечных оценок?

4. Что такое доверительный интервал, доверительная вероятность?

5. Что такое χ^2 - распределение?

6. Чему равны параметры χ^2 - распределения?

7. Охарактеризуйте распределение Стьюдента.

8. Почему рассмотренный способ построения доверительных интервалов применим только в при выборке из нормальной генеральной совокупности?

ЛАБОРАТОРНАЯ РАБОТА 3

Проверка статистической гипотезы о законе распределения генеральной совокупности по критерию

Порядок выполнения работы

1. Для заданного статистического материала построить гистограмму и выдвинуть гипотезу о законе распределения генеральной совокупности.

2. Найти оценки неизвестных параметров распределения.

3.Проверить выдвинутую гипотезу по критерию χ^2 на уровнях значимости $\alpha = 0{,}05;\ \alpha = 0{,}01$.

4. Составить отчет, в котором привести графическое изображение исходной выборки в виде гистограммы или эмпирической функции распределения, расчетную таблицу, результаты проверки гипотезы.

5. Ответить устно на контрольные вопросы.

Построение гистограммы и выдвижение гипотезы о распределении генеральной совокупности

Данная выборка подвергается группировке и по группированной выборке строится гистограмма. Процесс группировки и построения гистограммы описан в лаб. работе 1.

По виду гистограммы выдвигается гипотеза о распределении генеральной совокупности.

Определение оценок параметров распределения

После выдвижения гипотезы о виде закона распределения определяем по группированной выборке оценки параметров распределения.

Например, для нормального закона оценки математического ожидания и среднего квадратичного отклонения; для показательного закона оценку параметра λ. Для большинства законов параметры либо являются

математическим ожиданием и дисперсией либо являются функциями этих числовых характеристик. Поэтому в подавляющем числе случаев для определения оценок параметров распределения достаточно определить оценки математического ожидания и дисперсии.

Оценки математического ожидания и дисперсии определяются по формулам

$$M^*[X] = \overline{x} = \sum_{i=1}^{m_1} P_i^* \tilde{x}_i ; \quad (1)$$

$$D^*[X] = S^2 = \sum_{i=1}^{m_1} \tilde{x}_i^2 P_i^* - \overline{x}^2 , \quad (2)$$

где m_1 - число уточненных интервалов, $\tilde{x}_i$ - середины уточненных интервалов; P_i^* - частоты попадания в i-й интервал.

Замечание. Оценки параметров выборки, определенные по негруппированной выборке, будут более точными, но громоздкость вычислений не всегда оправдывается улучшением результата. При использовании ЭВМ оценки параметров рекомендуется определять по всей выборке.

Проверка согласия теоретического и статистического распределений

При построении гистограммы была выдвинута гипотеза о законе распределения генеральной совокупности. Назовем этот закон распределения теоретическим. Проверим его согласие с распределением выборки.

Сущность проверки статистической гипотезы заключается в том, чтобы установить, согласуются или нет данные наблюдения и выдвинутая гипотеза, можно ли расхождения между гипотезой и результатом выборочных наблюдений отнести за счет случайной погрешности, обусловленной механизмом случайного отбора. При этом критерии в задачах проверки гипотез о параметрах распределения называют критериями значимости, а в задачах проверки гипотез о законах распределения – критериями согласия.

Идея проверки статистической гипотезы состоит в следующем. Пусть H_0

- выдвинутая гипотеза, которую назовем **основной**, противопоставляя ее множеству H_a **альтернативных** гипотез. Для проверки основной гипотезы H_0 проводится опыт, результатом которого является величина Z,скалярная мера близости между гипотетическим и эмпирическим распределениями или между гипотетической и эмпирической характеристиками распределения. Z представляет собой одномерную величину, значения которой изменяются от опыта к опыту. Предполагается закон распределения Z известным. По заданному $\alpha(\alpha << 1)$ область значений Z можно разбить на две области R_1 и R_2. Область R_2 определяется из условия $P(Z \in R_2 / H_0) = \alpha$ и называется **критической** областью гипотезы H_0 на **уровне значимости** α. Таким образом, при условии справедливости гипотезы H_0, попадание величины Z в критическую область R_2 есть событие маловероятное, практически невозможное.

Процедура проверки гипотезы заключается в следующем: по заданному уровню значимости α определяются R_1 и R_2, затем проводится опыт. Если его результат $Z \in R_2$, т. е. произошло событие практически невозможное при условии справедливости гипотезы H_0, то **гипотеза** H_0 **отвергается на уровне значимости** α. Если $Z \notin R_2$, т.е. $Z \notin R_1$, то **гипотеза** H_0 **не отвергается на уровне значимости** α.

Стандартным значением для уровня значимости является одно из следующих значений: 0,05; 0,01; 0,001. Величина Z называется **критерием** проверки гипотезы H_0.

Очевидно, что при такой проверке гипотезы правильная гипотеза может быть отвергнута. Ошибка, заключающаяся в том, что **отвергается верная гипотеза**, называется **ошибкой первого рода**. Вероятность такой ошибки равна $P(Z \in R_2 / H_0) = \alpha$. Выбор малого α гарантирует , что ошибка первого рода будет совершаться редко.

Возможна еще **ошибка второго рода**, состоящая в том, что гипотеза H_0,

будучи **неверной, не отвергается**. Вероятность ошибки второго рода равна $P(Z \in R_1 / H_a) = \beta$.

Величина

$$1 - \beta = 1 - P(Z \in R1 / H_a) = P(Z \notin R_1 / H_a) = P(Z \in R_2 / H_a)$$

называется **мощностью** критерия Z при заданном α. Для уменьшения вероятности ошибки второго рода или, что то же самое для увеличения мощности критерия, вероятность $P(z \in R_2 / H_a)$ должна быть возможно большей.

Таким образом, при выборе критической области R_2 будем руководствоваться следующими соображениями:

$$P(z \in R_2 / H_0) = \alpha,$$

$$P(z \in R_2 / H_a) = \max. \qquad (3)$$

Процесс проверки статистической гипотезы сводится к следующему:

- выдвигается основная гипотеза H_0 и множество альтернативных гипотез H_a;

- выбирается критерий, представляющий собой некоторую меру близости между гипотетическим и эмпирическим распределениями или между гипотетической и эмпирической характеристиками распределения;

- критерий выбирается так, чтобы его распределение было известно;

- назначается уровень значимости α и определяется критическая область R_2;

- производится опыт и по данным опыта (выборочным наблюдениям) вычисляется значение критерия $Z_{выб}$;

- если $z_{выб} \in R_2$, то гипотеза H_0 отвергается, если $z_{выб} \in R_1$, то гипотеза H_0 не отвергается на уровне значимости α.

Из большого числа различных критериев наибольшей распространенностью пользуется **критерий согласия** χ^2, предложенный К.

Пирсоном. В этом критерии в качестве меры расхождения теоретического и статистического распределений выбирается величина χ^2, определяемая равенством

$$\chi^2 = \sum_{i=1}^{m_1} \frac{(n_i - np_i)^2}{np_i}, \qquad (4)$$

где n – объем выборки; m_1 - число интервалов, на которые разбита выборка; n_i - число элементов выборки, попавших в i-й интервал; P_i - теоретическая вероятность попадания значений случайной величины в i-й интервал.

Вероятность P_i определяется в согласии с теоретическим законом распределения по формулам

$$P_i = F_т\left(x^{(i)}\right) - F_т\left(^{(i-1)}\right) \qquad (5)$$

или

$$P_i = \int_{x^{(i-1)}}^{x^{(i)}} f_т(x)dx, \qquad (6)$$

где $x^{(i)}$, $x^{(i-1)}$ - границы i-го интервала.

Примеры

Пусть выдвинуты гипотезы о распределении генеральной совокупности:

1) по показательному закону $P_i = \int_{x^{(i-1)}}^{x^{(i)}} f_т(x)dx = \int_{x^{(i-1)}}^{x^{(i)}} \lambda^* e^{-\lambda^* x}dx = e^{-\lambda^* x^{(i-1)}} - e^{-\lambda^* x^i}$, где $\lambda^{(*)}$- оценка параметра показательного закона распределения по выборке: $\lambda^{(*)} = \frac{1}{\overline{x}}$. Здесь $\overline{x}$ - оценка математического ожидания;

2) по нормальному закону $P_i = \Phi\left(\frac{x^{(i)} - \overline{x}}{S}\right) - \Phi\left(\frac{x^{(i)} - \overline{x}}{S}\right)$,

где $\overline{x}$ - оценка математического ожидания. А S^2 - оценка дисперсии по выборке.

$S = \sqrt{S^2}$ - оценка среднего квадратичного; $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{x^2}{2}} dx$ - функция Лапласа (таблица А4);

3) по закону Релея $P_i = \int_{x^{(i-1)}}^{x^{(i)}} f_r(x)dx = \int_{x^{(i-1)}}^{x^{(i)}} \frac{x}{\bar{a}^2} e^{-\frac{x^2}{2\bar{a}^2}} dx = e^{-\frac{\left(x^{(i-1)}\right)^2}{2\bar{a}^2}} - e^{-\frac{\left(x^{(i)}\right)^2}{2\bar{a}^2}}$,

где $\bar{a}$ -оценка параметра закона Релея по выборке : $\bar{a} = \sqrt{\frac{2}{\pi}} \cdot \bar{x}$;

4) по равномерному закону $P_i = \int_{x^{(i-1)}}^{x^{(i)}} f_r(x)dx = \frac{x^{(i)} - x^{(i-1)}}{\bar{b} - \bar{a}}$,

где $\bar{b}$ и $\bar{a}$ - оценки крайних значений выборки, которые находятся из этой системы $\bar{x} = \frac{\bar{a} + \bar{b}}{2}$; $S^2 = \frac{\left(\bar{b} - \bar{a}\right)^2}{12}$.

Случайная величина χ^2, независимо от вида закона распределения генеральной совокупности, при достаточно больших $n (n \geq 50)$ имеет распределение χ^2 с числом степеней свободы $k = m_1 - r - 1$, где m_1 - число интервалов, r – число параметров распределения, определенных по выборке.

Задаваясь **уровнем значимости** $(\alpha = 0{,}01; 0{,}05; 0{,}1)$, по табл. А3 определим критическое значение χ^2_α, при котором $P\left(\chi^2 > \chi^2_\alpha\right) = \alpha$. При больших $m_1 (m_1 > 30)$ χ^2 распределено асимптотически нормально и можно пользоваться таблицами нормального закона. Если $\chi^2 < \chi^2_\alpha$, то выдвинутая гипотеза о виде закона распределения генеральной совокупности не отвергается на уровне значимости α (гипотеза не противоречит опытным данным), если же $\chi^2 > \chi^2_\alpha$, то гипотеза отвергается на уровне значимости α.

Замечание. Критерий Пирсона обладает большей мощностью, если интервалы содержат примерно равное число элементов, при этом длины интервалов не обязательно должны быть равными. Поэтому при использовании

критерия Пирсона нужно произвести новое разбиение данной выборки на интервалы, содержащие примерно равное число элементов.

Замечание. Все расчеты вести с тем количеством знаков, с каким даны значения случайной величины (можно добавить один дополнительный знак).

Пример выполнения и оформления лабораторной работы

Дана выборка, содержащая двести элементов (см. лаб. раб. 1, табл. 1). Упорядочим выборку. Наименьшее число равно 0,009 94, наибольшее число равно 3,666 642. Интервал (0,0001; 3,700) разделим на 20 равных частей. Границы интервалов занесем в графу 2 табл. 1. Число элементов, попавших в i-й интервал, занесены в графу 3. Два числа – 3,014 916, 3,666 642 - резко отличающиеся от других и полученные, видимо, за счет грубых ошибок опыта, можно отбросить. Таким образом, $n = 198$. Объединим интервалы таким образом, чтобы новые интервалы содержали не менее 8-10 элементов. Новые границы интервалов, а также число элементов, попавших в уточненные интервалы, поместим в графы 4 и 5, в графу 6 поместим частоты попаданий в каждый интервал. По полученным данным построим гистограмму (рис.1, см. лаб. раб. 1). Вид гистограммы дает право выдвинуть гипотезу о показательном распределении генеральной совокупности.

Оценку параметра показательного закона можно определить следующим образом: $\lambda^* = \frac{1}{\bar{x}}$, где $\bar{x} = \sum_{i=1}^{m_1} \bar{x}_i P_i^*$. $m_1 = 8$ - число уточненных интервалов.

Для удобства значения $\tilde{x}_i P_i^*$ поместим в графу 8, значения $\tilde{x}_i = \frac{x^{(i)} + x^{(i-1)}}{2}$ помещены были предварительно в графу 6. Оценка математического ожидания $\bar{x} = 0{,}613\,862$, оценка параметра показательного закона $\lambda = 1{,}629\,030\,6$. Для вычисления величины χ^2- меры расхождения теоретического и статистического распределений - вычислим теоретические вероятности попаданий значений случайной величины в i-й интервал по

формуле (6). Значения $\frac{(n_i - np_i)^2}{np_i}$ для каждого i занесем в графу 11. Вычисленное значение $\chi^2 = 0{,}519\,65$.

В данном примере по выборке определен один параметр λ. Следовательно, $r = 1$ и число степеней свободы распределения $k = 8 - 1 - 1 = 6$. Зададимся уровнем значимости $\alpha = 0{,}1$. По табл. А3 находим $\chi^2_\alpha = 10{,}64$. Вычисленное значение $\chi^2 = 0{,}519\,65$ меньше $\chi^2_\alpha = 10{,}64$, следовательно, гипотеза не отвергается с уровнем значимости $\alpha = 0{,}1$.

Таблица 1

№ п/п	J (до объединения)	n_i (до объединения)	J (после объединения)	n_i (после объединения)	P_i^*	$\tilde{x}_i$	$\tilde{x}_i P_i^*$		P_i	$\frac{(n_i - np_i)^2}{np_i}$
1	2	3	4	5	6	7	8	9	10	11
1	(0,000-0,185)	50	(0,000-0,185)	50	0,252 525	0,092 5	0,233 585		0,244 485	0,052 3541
2	(0,185-0,370)	38	(0,185-0,370)	38	0,191 919	0,277 5	0,053 2575		0,184 713	0,055 6801
3	(0,370-0,555)	28	(0,370-0,555)	28	0,141 414	0,462 5	0,065 4039		0,139 554	0,004 9092
4	(0,555-0,740)	22	(0,555-0,740)	22	0,111 111	0,647 5	0,071 9444		0,105 434	0,060 525
5	(0,740-0,925)	17	(0,740-0,925)	17	0,085 859	0,832 5	0,071 4778		0,079 657	0,095 5973
6	(0,925-1,110)	11	(0,925-1,295)	11	0,101 010	1,11	0,112 1211		0,105 650	0,040 3464
7	(1,110-1,295)	9								
8	(1,295-1,480)	7								
9	(1,480-1,665)	4	(1,295-1,850)	14	0,070707	1,572 5	0,111 1867		0,079 914	0,210 0245
10	(1,665-	3								

	1,850)									
11	(1,850-2,035)	3								
12	(2,035-2,220)	2								
13	(2,220-2,405)	2	(1,850-2,775)	9	0,045454	2,3125	0,1051123		0,045678	0,000 2145
14	(2,405-2,590)	1								
15	(2,590-2,775)	1								
16	(2,775-2,960)	0								
17	(2,960-3,145)	1								
18	(3,145-3,330)	0								
19	(3,330-3,515)	0								
20	(3,515-3,700)	1								

$\bar{x} = 0{,}613\,862\,0$ $k = 8-1-1 = 6$

$\lambda^* = 1{,}629\,030\,6$ $\chi^2_\alpha = 10{,}64$

$\chi^2 = 0{,}519\,65$

Контрольные вопросы

1. Что такое эмпирическая функция распределения и как она вычисляется по данным выборки?

2. Что такое гистограмма распределения и как она строится по данным выборки?

3. Объяснить содержательный смысл критерия χ^2 как меры близости эмпирического и теоретического распределений.

4. Как учитывается при пользовании критерием согласия χ^2 факт определения параметров теоретического распределения по данным выборки?

5. Почему при $\chi^2_{\text{выб}} > \chi^2_{\alpha}$ гипотезу следует отбросить?

ЛАБОРАТОРНАЯ РАБОТА 4

Моделирование случайных величин

При решении многих математических и прикладных задач возникает необходимость в моделировании случайных величин, т.е. в формировании последовательно-сти чисел, элементы которой можно рассматривать как результаты независимых измерений случайной величины с определенным законом распределения. Процесс получения указанной выше последовательности чисел в математической статистике принято называть построением (случайной) **выборки из генеральной совокуп-ности** с заданным законом распределения. Число элементов в выборке называется **объемом** выборки.

Целью настоящей работы является ознакомление с простейшими методами построения случайных выборок.

На практике построение выборок из генеральных совокупностей с заданными законами распределения производится путем преобразования выборки из какой-либо одной, так сказать «стандартной», генеральной совокупности. Обычно роль «стандартной» играет генеральная совокупность с равномерным распределением на интервале (0, 1). Плотность $f_0(x)$ этого распределения задается, как известно, следующим образом:

$$f_0(x) = \begin{cases} 1, & 0 < x < 1, \\ 0 & \text{в остальных случаях,} \end{cases} \qquad (1)$$

а функция распределения

$$f_0(x)=\begin{cases}0, & x\le 0,\\ x, & 0<x<1,\\ 1, & x>1.\end{cases} \qquad (2)$$

Случайную величину с плотностью $f_c(x)$ и функцией распределения $F_0(x)$ будем обозначать буквой X. а ее значения (элементы выборки) - $x_1, x_2,...$

Процесс построения выборки $y_1, y_2,...$ значений случайной величины Y путем преобразования выборки $x_1, x_2,...$ значений случайной величины X принято называть **разыгрыванием** случайной величины У.

Рассмотрим два метода разыгрывания случайных величин. При этом будем предполагать, что случайная величина У с функцией распределения $F(y)$, которая строго больше нуля в некотором интервале (a,b) и равна нулю вне этого интервала.

Первый метод разыгрывания заключается в решении уравнения

$$F(Y)=X \qquad (3)$$

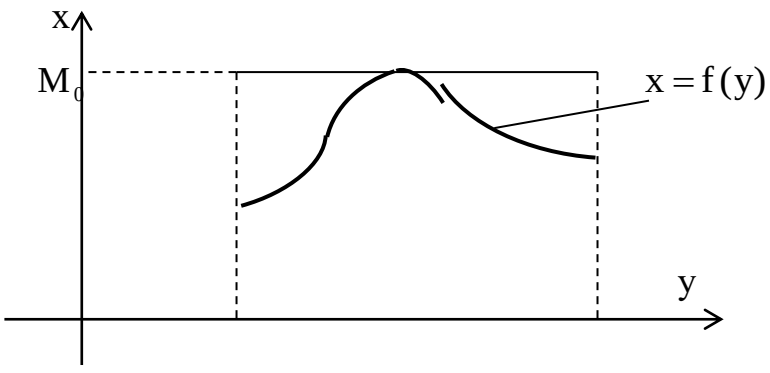
или эквивалентного ему уравнения

$$\int_a^y f(y)dy=X \qquad (4)$$

Подставляя в правую часть (3) или (4) вместо X последовательно элементы $x_1, x_2,...$ и решая уравнение относительно Y, получаем элементы $y_1, y_2,...$ из генеральной совокупности с распределением, задаваемым плотностью $f(y)$.

Второй метод разыгрывания (метод Неймана)

Пусть интервал (a,b) конечен и плотность $f(y)$ ограничена (рис.1):



а в

Рис. 1

Тогда выборку $y_1, y_2 \ldots$ из генеральной совокупности с законом распределения, задаваемым плотностью $f(y)$, можно получить из выборки $x_1, x_2 \ldots$ следующим образом:

1) образуем из элементов выборки $x_1, x_2 \ldots$ пары $(x_1, x_2), (x_3, x_4)$, …, $(x_{2i-1}, x_{2i}) \ldots$;

2) для каждой пары (x_{2i-1}, x_{2i}) строим точку Γ_i с координатами $\left(y_1^{(i)}, y_2^{(i)}\right)$, где $y_1^{(i)} = a + x_{2i-1}(b-a)$,

$$y_2^{(i)} = x_{2i} M_0 \qquad (5)$$

3) если $y_2^{(i)}, < y_1^{(i)}$ точка Γ_i лежит под кривой $x = f(y)$, то включаем в формируемую выборку очередной элемент, равный $y_1^{(i)}$, если $\left(y_2^{(i)} > f\left(y_1^{(i)}\right)\right)$ (точка Γ_i лежит над кривой $x = f(y)$), то не включаем в формируемую выборку никакого элемента.

Замечание 1. Первый метод разыгрывания применим непосредственно и к случаю $a = -\infty, b = \infty$. Если имеется таблица значений функции $F(y)$, то значения y_i можно находить приближенно из уравнения $F(y_i) = x_i$ с помощью этой таблицы.

Замечание 2. Для применения метода Неймана при $a = -\infty$, $b = \infty$ можно поступить следующим образом. Зададимся малым числом ε (например, $\varepsilon = 0{,}01$ или $0{,}0001$) и выберем числа a_ε и b_ε так, чтобы $p(a_\varepsilon < Y < b_\varepsilon) = 1 - \varepsilon$.

Затем применим метод Неймана, считая приближенно, что $f(y) = 0$ при $y \notin (a_\varepsilon, b_\varepsilon)$. Например, когда Y распределена нормально с параметрами m и σ, можно взять $\varepsilon = 0{,}003$, $a_\varepsilon = m - 3\sigma$, $b_\varepsilon = m + 3\sigma$.

Задание. Используя выборку из генеральной совокупности с равномерным распределением на интервале (0,1) (см. табл. А1), построить

выборку из генеральной совокупности с заданной плотностью $f(y)$.

Варианты заданий содержатся в табл. 3. Оформление работы провести по образцам, содержащимся в приведенных ниже примерах.

Пример 1. Построить выборку объемом 16 из генеральной совокупности с плотностью

$$f(y)=\begin{cases}0 & \text{при } y<0,\\ (1+y)^{-2} & \text{при } y>0.\end{cases}$$

В этом случае $F(y)=0$ при $y<0$, а при $y>0$

$$F(y)=\int_0^y \frac{dt}{(1+t)^2}=-\frac{1}{1+t}\Big|_0^y=-\frac{1}{1+y}+1.$$

Уравнение (3) будет иметь вид $1-\frac{1}{1+y}=x$ и решается в явном виде:

$$y=\frac{1}{1-x}-1 \qquad (6)$$

Используя выборку объемом 16 из генеральной совокупности с равномерным распределением (элементы $x_1, x_2, \ldots$ этой выборки в данном случае взяты из таблицы, отличной от табл. 1, см. лаб. раб. 1), вычисляем по формуле (6) соответствующие элементы $y_1, y_2, \ldots$ искомой выборки. Результаты вычислений оформляем в виде следующей таблицы (табл. 1).

Таблица 1

№ п/п	X	У	№ п/п	X	У
1.	0,865 15	6,415 65	9.	0,691 86	2,245 28
2.	0,907 95	9,863 66	10.	0,033 93	0,035 12
3.	0,661 55	1,954 65	11.	0,425 02	0,739 19
4.	0,664 34	1,979 20	12.	0,992 24	127,865 97
5.	0,565 58	1,301	13.	0,889	8,053 87

		92		55	
6.	0,123 32	0,140 67	14.	0,537 58	1,162 54
7.	0,943 77	16,784 10	15.	0,916 41	10,96 315
8.	0,578 02	1,369 78	16.	0,188 67	0,232 54

Пример 2. Построить выборку из генеральной совокупности с плотностью

$$f(y)=\begin{cases}\frac{1}{2}\left(1+3y^2\right), & 0<y<1,\\ 0 \text{ в остальных случаях.}\end{cases}$$

Очевидно, что для данной генеральной совокупности $F(y)=0$ при $y<0$, $F(y)=1$ при $y>1$, а при $0<y<1$

$$F(y)=\int_0^y \frac{1}{2}\left(1+3t^2\right)dt=\frac{1}{2}\left(t+t^3\right)\Big|_0^y=\frac{1}{2}\left(y+y^3\right)$$

Уравнение (3) имеет вид $\frac{1}{2}\left(y+y^3\right)=x$.

При применении первого способа разыгрывания случайной величины Y пришлось бы n раз (при различных значениях x, n – объем формируемой выборки) решить данное кубическое относительно уравнение, что представляет довольно громоздкую вычислительную задачу. Поэтому здесь целесообразно применить метод Неймана.

Используем ту же выборку $x_1, x_2, \ldots$, что и в примере 1. Образуем из этой выборки пары $(x_1, x_2), (x_3, x_4), \ldots$ В примере $a=0, b=1$. Кроме того, $M_0=\max\limits_{0\le y\le 1} f(y)=f(1)=2$, так как $f(y)$ возрастает при $0\le y\le 1$ и, следовательно, достигает максимума при $y=1$. Отсюда по формуле (5) получаем

$$y_1^{(i)} = x_{2i-1},$$
$$y_2^{(i)} = 2x_{2i}. \qquad (7)$$

Далее построение выборки $y_1, y_2, \ldots$ проводим следующим образом. Для каждой пары (x_{2i-1}, x_{2i}) вычисляем по формулам (7) числа $y_1^{(i)}$ и $y_2^{(i)}$, затем число $f\left(y_1^{(i)}\right) = \frac{1}{2}\left[1 + 3\left(y_1^{(i)}\right)^2\right]$ и проверяем выполнение неравенства $y_2^{(i)} < f\left(y_1^{(i)}\right)$.

Если это неравенство выполнено, то $y_1^{(i)}$ - очередной элемент формируемой выборки. В противном случае данная пара (x_{2i-1}, x_{2i}) не порождает элемента выборки $y_1, y_2, \ldots$ и следует перейти к рассмотрению следующей пары (x_{2i+1}, x_{2i+2}). Результаты вычислений представлены в табл. 2.

Таблица 2

i	x_{2i-1}	x_{2i}	$y_1^{(i)}$	$y_2^{(i)}$	><	$y_1^{(i)}$	y_i
1.	0,865 15	0,907 95	0,865 15	1,815 90		1,622 73	
2.	0,661 55	0,664 34	0,661 55	1,328 64		1,156 47	
3.	0,565 58	0,123 32	0,565 58	0,246 64		0,979 82	0,5655
4.	0,943 77	0,578 02	0,943 77	1,156 04		1,836 05	0,943 77
5.	0,691 86	0,033 93	0,691 86	0,067 86		1,218 00	0,691 8
6.	0,425 02	0,992 24	0,425 02	1,984 48		0,770 96	
7.	0,889 55	0,537 58	0,889 55	1,075 16		1,686 95	0,889 55
8.	0,916 41	0,188 67	0,916 41	0,377 34		1,759 71	0,916 41

Элементы формируемой выборки содержатся в последнем столбце табл. 2. Прочерки в последнем столбце соответствуют парам (x_{2i-1}, x_{2i}), не порождающим элементов выборки $y_1 y_2 \ldots$ В рассматриваемом примере сформирована выборка объемом 5. Для увеличения объема формируемой выборки следует, очевидно, увеличить объем исходной выборки из генеральной совокупности с равномерным распределением (см. лаб. раб. 1, нахождение числовых характеристик выборки).

Контрольные вопросы

1. Дать теоретико-вероятностное обоснование первого способа

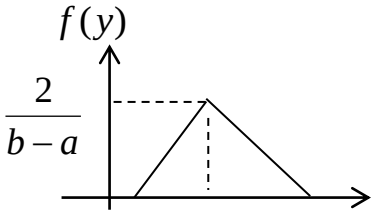
разыгрывания случайных величин.

2. Дать теоретико-вероятностное обоснование метода Неймана (Указание: убедиться предварительно, что пары $\left(y_1^{(i)}, y_2^{(i)}\right)$ можно рассматривать как выборку значений пары независимых случайных величин (y_1, y_2), имеющей равномерное распределение в прямоугольнике $(a < y < b,\ 0 < y_2 < M_0)$).

3. Доказать, что для получения по методу Неймана выборки достаточно большого объема N потребуется выборка из генеральной совокупности с равномерным распределением объема примерно $2M_0(b-a)\cdot N$.

Таблица 3

Плотность	№ варианта	Значения параметров	Плотность	№ варианта	Значения параметров
1	2	3	4	5	6
Закон Коши $\frac{1}{\pi}\frac{h}{h^2+(y-y_0)^2}$, $-\infty<y<\infty$.	1	$h=1; y_0=1$.	Нормальный закон $\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(y-m)^2}{2\sigma^2}\right\}$, $-\infty<y<\infty$.	1	$m=0;\sigma=0{,}5$.
	2	$h=1; y_0=2$.		2	$m=0;\sigma=0{,}2$.
	3	$h=2; y_0=1$.		3	$m=1;\sigma=0{,}4$.
	4	$h=2; y_0=2$.		4	$m=-1;\sigma=0{,}7$.
Закон Релея $\frac{y}{\sigma^2}e^{-\frac{y^2}{2\sigma^2}}$, $y>0$.	1	$\sigma=0{,}5$.	Усеченный нормальный $\frac{c}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{(y-m)^2}{2\sigma^2}\right\}$, $a<y<b$.	1	$m=0;\sigma=0{,}5$.
	2	$\sigma=1$.		2	$m=0;\sigma=0{,}2$.
	3	$\sigma=1{,}5$.		3	$m=1;\sigma=0{,}4$.
	4	$\sigma=2$.		4	$m=-1;\sigma=0{,}7$.
Закон Вейбулла; при $\alpha=1$-показательный закон $c\alpha y^{\alpha-1}e^{-cy^{\alpha}}$, $y>1$.	1	$\alpha=1;c=1$.	χ^2- распределение $\frac{y^{\frac{n}{2}-1}e^{-\frac{y}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}$, $y>0$.	1	$n=8$.
	2	$\alpha=1;c=2$.		2	$n=9$.
	3	$\alpha=0{,}5;c=1$.		3	$n=10$.
	4	$\alpha=3;c=1$.		4	$n=12$
1	2	3	4	5	6

Закон Лапласа (двусторонний экспоненциальный) $\frac{\lambda}{2}e^{-\lambda\lvert y-\mu\rvert}$, $-\infty < y < \infty$.	1	$\mu = 1; \lambda = 2$.	χ - распределение $\frac{y^{n-1}e^{-\frac{y^2}{2}}}{2^{\frac{m}{2}-1}\Gamma\left(\frac{n}{2}\right)}$, $y > 0$	1	$n = 1$.
	2	$\mu = 1; \lambda = 2{,}5$.		2	$n = 2$.
	3	$\mu = 2; \lambda = 3$.		3	$n = 3$.
	4	$\mu = 2; \lambda = 3{,}5$.		4	$n = 4$
Двойное показательное распределение $c\alpha \exp\{-\alpha y - ce^{-\alpha y}\}$, $-\infty < y < \infty$.	1	$c = 1; \alpha = 2$.	Бета-распределение $\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} y^{a-1}(1-y)^{b-1}$. $0 < y < 1$.	1	$a = 3; b = 3$.
	2	$c = 1; \alpha = 1{,}5$.		2	$a = 3; b = 2$.
	3	$c = 2; \alpha = 1{,}5$.		3	$a = 2; b = 3$.
	4	$c = 2; \alpha = 2$.		4	$a = 2; b = 4$.
				5	$a = 4; b = 2$
Симпсона (треугольный) закон 	1	$a = -1; b = 2$	Стьюдента (t- распределение) $\frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)} \cdot \left(1+\frac{y^2}{k}\right)^{-\frac{k+1}{2}}$, $-\infty < y < \infty$.	1	$k = 3$.
	2	$a = 0; b = 2$.		2	$k = 5$.
	3	$a = 0{,}5; b = 1{,}5$.		3	$k = 7$.
	4	$a = 0{,}5; b = 1$		4	$k = 15$.

Замечание. В некоторых вариантах заданий из табл. 3 выражение плотности $f(y)$ моделируемой случайной величины содержит значения гамма-функции $\Gamma(\alpha)$. Напомним, что $\Gamma(\alpha)$ определяется соотношением $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$.

При этом для всех натуральных m справедливы формулы

$$\Gamma(m) = (m-1)!,\ \Gamma\left(m + \frac{1}{2}\right) = \left(m - \frac{1}{2}\right) \cdot \left(m - \frac{3}{2}\right) \cdot \left(m - \frac{5}{2}\right) \ldots \frac{3}{2} \cdot \frac{1}{2} \cdot \sqrt{\frac{\pi}{2}}.$$

ПРИЛОЖЕНИЕ А

Таблица А1

Выборка из равномерной на (0,1) генеральной совокупности

0,573 132	0,439 157	0,305 182	0,141 206	0,037 231	0,903 255
0,769 281	0,635 306	0,501 331	0,307 356	0,233 381	0,099 406
0,965 31	0,831 456	0,697 481	0,563 506	0,429 531	0,295 556
0,161 581	0,027 606	0,893 631	0,769 656	0,625 681	0,491 706
0,357 781	0,223 756	0,089 781	0,955 806	0,821 831	0,687 856
0,553 881	0,419 906	0,285 931	0,151 956	0,017 981	0,884 006
0,750 031	0,616 055	0,482 080	0,348 105	0,214 130	0,080 155
0,746 180	0,812 205	0,678 230	0,544 255	0,410 280	0,276 305
0,142 330	0,008 355	0,874 380	0,740 405	0,606 430	0,472 455
0,338 480	0,204 505	0,070 530	0,936 555	0,802 580	0,668 605
0,534 630	0,400 655	0,266 680	0,132 705	0,998 730	0,864 755
0,730 780	0,596 805	0,462 830	0,328 855	0,194 880	0,060 905
0,926 929	0,792 954	0,658 979	0,525 004	0,391 029	0,257 054
0,123 099	0,989 104	0,855 129	0,721 154	0,587 179	0,453 204
0,319 229	0,185 254	0,051 279	0,917 304	0,783 329	0,649 354
0,515 379	0,381 404	0,247 429	0,113 454	0,979 479	0,845 504
0,711 529	0,577 554	0,443 579	0,209 604	0,175 629	0,041 654
0,907 679	0,773 704	0,639 729	0,505 754	0,371 778	0,237 803
0,103 828	0,969 853	0,835 878	0,701 903	0,567 928	0,433 953
0,299 978	0,166 003	0,032 028	0,898 053	0,764 078	0,630 103
0,496 128	0,362 153	0,228 178	0,094 293	0,960 228	0,826 253
0,692 278	0,558 303	0,424 328	0,290 353	0,156 378	0,022 403
0,888 428	0,754 453	0,620 478	0,486 503	0,352 528	0,218 553
0,084 578	0,950 603	0,816 628	0,682 652	0,548 677	0,414 702
0,280 727	0,146 752	0,021 777	0,878 802	0,744 827	0,610 852
0,476 877	0,342 902	0,208 927	0,074 952	0,940 977	0,807 002
0,673 027	0,539 052	0,405 077	0,271 102	0,136 127	0,003 152
0,869 177	0,735 202	0,601 227	0,467 252	0,333 277	0,199 302
0,065 327	0,231 352	0,797 377	0,663 402	0,529 427	0,195 452
0,291 477	0,127 501	0,993 529	0,859 851	0,725 573	0,591 601
0,457 626	0,323 651	0,189 979	0,055 701	0,921 226	0,787 751
0,653 776	0,519 801	0,385 826	0,251 951	0,117 886	0,983 901
0,849 926	0,715 951	0,581 976	0,448 001	0,314 026	0,180 051
0,046 076	0,912 101				

Таблица А2

Двусторонние границы t- распределения: значения t_γ, для которых

$$P(|t|<t_\gamma)=\tau .$$

	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	0,95	0,98	0,99	0,999
1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0,158	0,325	0,510	0,727	1,000	1,376	1,936	3,078	6,314	12,706	31,821	63,657	636,519
2	0,142	0,289	0,445	0,617	0,816	1,061	1,336	1,886	2,920	1,303	6,965	9,925	31,598
3	0,137	0,277	0,424	0,584	0,765	0,978	1,250	1,638	2,353	3,182	4,541	5,841	12,941
4	0,134	0,271	0,414	0,569	0,741	0,941	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,267	0,408	0,559	0,727	0,920	1,156	1,476	2,015	2,571	3,365	4,032	6,859
6	0,131	0,265	0,404	0,553	0,718	0,906	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,263	0,402	0,549	0,711	0,890	1,119	1,415	1,895	2,365	2,998	3,499	5,405
8	0,130	0,262	0,399	0,546	0,706	0,889	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,261	0,398	0,543	0,703	0,883	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,260	0,397	0,542	0,700	0,879	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,260	0,396	0,540	0,697	0,876	1,088	1,363	1,796	2,201	2,718	3,106	4,487
12	0,128	0,259	0,395	0,539	0,695	0,873	1,083	1,356	1,782	2,179	2,681	3,055	4,313
13	0,128	0,259	0,394	0,538	0,694	0,870	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,258	0,393	0,537	0,692	0,868	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,258	0,393	0,536	0,691	0,866	1,074	1,341	1,753	2,131	2,602	2,947	4,073

Окончание табл. А2

1	2	3	4	5	6	7	8	9	10	11	12	13	14
16	0,128	0,258	0,392	0,535	0,690	0,865	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,257	0,392	0,534	0,689	0,863	1,069	1,333	1,640	2,110	2,567	2,898	3,965
18	0,127	0,257	0,392	0,534	0,688	0,962	1,067	1,330	1,734	2,103	2,552	2,878	3,922
19	0,127	0,257	0,391	0,533	0,688	0,861	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,257	0,391	0,533	0,687	0,860	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,257	0,391	0,532	0,686	0,859	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,256	0,390	0,532	0,686	0,858	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,256	0,390	0,532	0,685	0,858	1,060	1,319	1,714	2,069	2,500	2,797	3,745
24	0,127	0,256	0,390	0,531	0,685	0,857	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,256	0,390	0,531	0,684	0,856	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,256	0,389	0,531	0,684	0,855	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,256	0,389	0,530	0,683	0,855	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,256	0,389	0,530	0,683	0,854	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,255	0,388	0,529	0,681	0,851	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,126	0,254	0,387	0,527	0,679	0,848	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,235	0,385	0,524	0,674	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,291

Таблица А3

χ^2 - распределение

Число степеней свободы k	χ^2_α как функции k и α. $\alpha = P(\chi^2 > \chi^2_\alpha)$															
	α =0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01	0,005	0,002	0,001
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	0,00016	0,0006	0,0039	0,016	0,064	0,148	0,455	1,07	1,64	2,7	3,8	5,4	6,6	7,9	9,5	10,83
2	0,020	0,040	0,103	0,211	0,446	0,713	1,386	2,41	3,22	4,6	5,9	7,9	9,2	10,6	12,4	13,8
3	0,115	0,185	0,352	0,584	1,005	1,424	2,366	3,67	4,64	6,3	7,8	9,8	11,3	13,8	14,8	16,3
4	0,30	0,43	0,71	1,06	0,65	2,19	3,36	4,9	6,0	7,8	9,5	11,7	13,3	14,9	16,9	16,5
5	0,55	0,75	1,14	1,61	2,34	3,00	4,35	6,1	7,3	9,2	11,1	13,4	15,1	16,8	18,9	20,5
6	0,87	1,13	1,63	2,20	3,07	3,83	5,35	7,2	8,6	10,6	12,6	15,0	16,8	18,5	20,7	22,5
7	1,24	1,56	2,17	2,83	3,82	4,67	6,35	3,4	9,8	12,0	14,1	16,6	18,5	20,3	22,6	24,3
8	1,65	2,03	2,73	3,49	4,59	5,53	7,34	9,5	11,0	13,4	15,5	18,2	20,1	22,0	24,3	26,1
9	2,09	2,53	3,32	4,17	5,38	6,39	8,34	10,7	12,2	14,7	16,9	19,7	21,7	23,6	26,1	27,9
10	2,56	3,06	3,94	4,86	6,18	7,27	9,34	11,8	13,4	16,0	18,3	21,2	23,2	25,2	27,7	29,6

Окончание табл. А3

11	3,1	3,6	4,6	5,6	7,0	8,1	10,3	12,9	14,6	17,3	19,7	22,6	24,7	26,8	29,4	31,3
12	3,6	4,2	5,2	6,3	7,8	9,0	11,3	14,0	15,8	18,5	21,0	24,1	26,2	28,3	30,9	32,9
13	4,1	4,8	5,9	7,0	8,6	9,9	12,3	15,1	17,0	19,8	22,4	25,5	27,7	29,8	32,5	34,5
14	4,7	5,4	6,6	7,8	9,5	10,8	13,3	16,2	18,2	21,1	23,7	26,9	29,1	31,3	34,0	36,1
15	5,2	6,0	7,3	8,5	10,3	11,7	14,3	17,3	19,3	22,3	25,0	28,8	30,6	32,8	35,6	37,1
16	5,8	6,6	8,0	9,3	11,2	12,6	15,3	18,4	20,5	23,5	26,3	29,6	32,0	34,3	37,1	39,3
17	6,4	7,3	8,7	10,1	12,0	13,5	16,3	19,6	21,6	24,8	27,6	31,0	33,4	35,7	38,6	40,3
18	7,0	7,9	9,4	10,9	12,9	14,4	17,3	20,6	22,8	26,0	28,9	32,3	34,8	37,2	40,1	42,3
19	7,6	8,6	10,1	11,7	13,7	15,4	18,3	21,7	23,9	27,2	30,1	33,7	36,2	38,6	41,6	43,8
20	8,3	9,2	10,9	12,4	14,6	16,3	19,3	22,8	25,0	28,4	31,4	35,0	37,6	40,0	43,0	45,3
21	8,9	9,9	11,6	13,2	15,4	17,2	20,3	23,9	26,2	29,6	32,7	36,3	38,9	41,4	44,5	46,8
22	9,5	10,6	12,3	14,0	16,3	18,1	21,3	24,9	27,3	30,8	33,9	37,7	40,3	42,8	45,9	48,3
23	10,2	11,3	13,1	14,8	17,2	19,0	22,3	26,0	28,4	32,0	35,2	39,0	41,6	44,2	47,8	49,7
24	10,9	12,0	13,8	15,7	18,1	19,9	23,3	27,1	29,6	33,2	36,4	40,3	43,0	45,6	48,7	51,2
25	11,5	12,7	14,6	16,5	18,9	20,9	24,3	28,2	30,7	34,4	37,7	41,6	44,3	46,9	50,1	52,6
26	12,2	13,4	15,4	17,3	19,8	21,8	25,3	29,2	31,8	35,6	38,9	42,9	45,6	48,3	51,6	54,1
27	12,9	14,1	16,2	18,1	20,7	22,7	26,3	30,3	32,9	36,7	40,1	44,1	47,0	49,6	52,9	55,5
28	13,6	14,8	16,9	18,9	21,6	23,6	27,3	31,4	34,0	37,9	41,3	45,4	48,3	51,0	54,4	56,9
29	14,3	15,6	17,7	19,8	22,5	24,6	28,3	32,5	35,1	39,1	42,6	46,7	49,6	52,3	55,7	58,3
30	15,0	16,8	18,5	20,6	23,4	25,5	29,3	33,5	36,3	40,3	43,8	48,0	50,9	53,7	57,1	59,7

Таблица А4

Функция Лапласа $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{x^2}{2}} dx$

	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0040	0080	0120	0160	0199	0239	0279	0319	0359
0,1	0,398	0438	0178	0517	0557	0596	0636	0675	0714	0753
0,2	0,0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0,3	0,1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0,4	0,1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0,5	0,1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0,6	0,2257	2291	2324	2357	2389	2422	2454	2485	2517	2549
0,7	0,2580	2611	2642	2673	2703	2734	2764	2794	2823	2852
0,8	0,2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0,9	0,3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1,0	0,3413	3437	3461	3485	3508	3531	3554	3577	3599	3621
1,1	0,3643	3665	3686	3708	3729	3749	3770	3790	3810	3830
1,2	0,3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1,3	0,4032	4049	4049	4082	4099	4115	4131	4147	4162	4177
1,4	0,4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1,5	0,4332	4345	4357	4370	4382	4394	4406	4418	4429	4441
1,6	4452	4463	4474	4484	4495	4505	4515	4525	4535	4545
1,7	4554	4564	4573	4582	4591	4599	4608	4616	4525	4633
1,8	0,4641	4649	4656	4664	4571	4678	4686	4693	4699	4706
1,9	0,4713	4719	4726	4732	4738	4744	4750	4756	4761	4767
2,0	0,47725	47778	47831	47882	47932	47981	48230	48077	48124	48169
2,1	0,48214	48257	48300	48341	48382	48422	48461	48500	48537	48574
2,2	0,48610	48645	48679	48713	48746	48778	48809	48840	48870	48899
2,3	0,48928	48956	48983	49010	49036	49061	49086	49111	49135	49158
2,4	0,49180	49202	49224	49245	49266	49286	49305	49324	49343	49361
2,5	0,49379	49396	49413	49430	49446	49461	49477	49492	49506	49520
2,6	0,49534	49547	49560	49573	49585	49597	49609	49621	49632	49643
2,7	0,49653	49664	49674	49683	49693	49702	49711	49720	49728	49736

2,8	0,49744	49752	49760	49767	49774	49781	49788	49795	49801	49807
2,9	0,49813	49819	49825	49831	49836	49841	49846	49851	49856	49861
3,0	0,49865	49903	49931	49952	49966	49977	49984	49989	49993	49995
4,0	0,999 68									
4,5	0,499 997									
5,0	0,499 99997									

ЛИТЕРАТУРА

1. **Вентцель Е.С**. Теория вероятностей, М., «Наука», 1973.

2. **Коваленко И. Н., Филиппов А. А.** Теория вероятностей и математическая статистика. М.: Высш. школа, 1973.

3. **Климов Г.П.,Кузьмин А.Д.** Вероятность, процессы, статистика: Задачи с решениями.-М.: Изд-во МГУ, 1985.

4. **Гмурман В. Е.** Теория вероятностей и математическая статистика. М.: Высш. школа, 1972.

5. **Гурский Е. Н.** Теория вероятностей с элементами математической статистики. М.: Высш. школа, 1971.

6. **Иванов-Мусатов О. С.** Теория вероятностей и математическая статистика. М.: Наука, 1979.

7. **Савченко В.В.** Теория вероятностей: конспект лекций.-Н.Новгород: НГТУ,1997.

8. Справочник по теории вероятностей и математической статистике // **В. С. Королюк, Н. Н. Портенко и др.** М.: Наука, 1985.

9. **Пугачев В. С.** Теория вероятностей и математическая статистика. М.: Наука, 1979.

10. Свешников А.А. Основы теории ошибок. Л.: Изд-во ЛГУ, 1972.

11. Математическая статистика: Метод. указания/ Сост.: **Н.С.Кац, М.Н. Соколовский, В.Д. Цветкова** .Омск, 1992.

12. Математическая статистика: Метод. указания к лабораторным работам. Сост.: Н.С. Кац. Омск, 1991.

Алданов Ербол Сатыбаевич

Математическая статистика
Учебное пособие

Подписано в печать с готового оригинал-макета 20.05. 2019г.
Печать офсетная. Формат бумаги 60х84/16.
Объем 8,5 усл. печ. л. Тираж 500 экз. Заказ № 50

Отпечатано в типографии «Университет «Туран-Астана»»
с готового набора:

Адрес: 010000, г. Астана, ул. Дукенулы, 29