Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»

О.М.Полещук

Основные понятия теории вероятностей и математической статистики

Учебное пособие

Москва

2020

УДК 519.21
ББК 22.171
П49 Полещук О.М. Основные понятия теории вероятностей и математической статистики: Учебное пособие. – М.: – Москва:, 2020. – 248 с.: ил.

Рецензенты: доктор технических наук, профессор Домрачев В.Г.,
доктор технических наук, профессор Ретинская И.В.

Авторы: Ольга Митрофановна Полещук, д.т.н., профессор

Оглавление

1. Классификация неопределенности

Любая наука занимается исследованием организованных объектов, состоящих из отдельных элементов, которые находятся друг с другом в определенной взаимосвязи. Такие организованные объекты получили название систем. Все системы делятся на два класса – детерминированные и недетерминированные. Детерминированные системы характеризуются тем, что в них никогда не возникает какой-либо неопределенности: если известно первоначальное состояние системы и известна поступающая на нее информация, то можно точно предсказать ее последующее состояние. Для недетерминированной системы невозможно сделать точное, однозначное предсказание ее последующего состояния. Неоднозначность состояния недетерминированной системы связана с возникающей в ней неопределенностью. Природа этой неопределенности может быть разной и чаще всего в одной системе заложено несколько видов неопределенности.

Результаты наблюдений за состоянием систем достаточно часто можно представить в виде последовательности действительных чисел

$$x_1, x_2, \ldots, x_n \;.$$

Чтобы из этой последовательности извлечь полезную информацию, необходимо иметь модель наблюдаемого явления. Для более полного понимания сущности происходящих в исследуемых системах процессов и их анализа строятся математические модели, которые широко применяются в технике, экономике, политике и ряде других областей.

Вероятностные модели, как известно, нашли широкое применение в тех случаях, когда наблюдаемые данные или исходы обладают, с одной стороны, некоторой степенью неопределенности, а, с другой стороны, свойством устойчивости частот. Это свойство означает, что частота появления явления при неограниченном наблюдении за ним колеблется около некоторого числа, принимаемого за его вероятность.

В обычном разговорном языке понятия неопределенности, нечеткости и случайности имеют тенденцию смешиваться в одно понятие, но в языке науки достаточно давно произошло их разграничение.

Исходя из этого, при рассмотрении такого философского понятия, как неопределенность, предлагается пользоваться следующей классификацией, изображенной на рисунке 1:

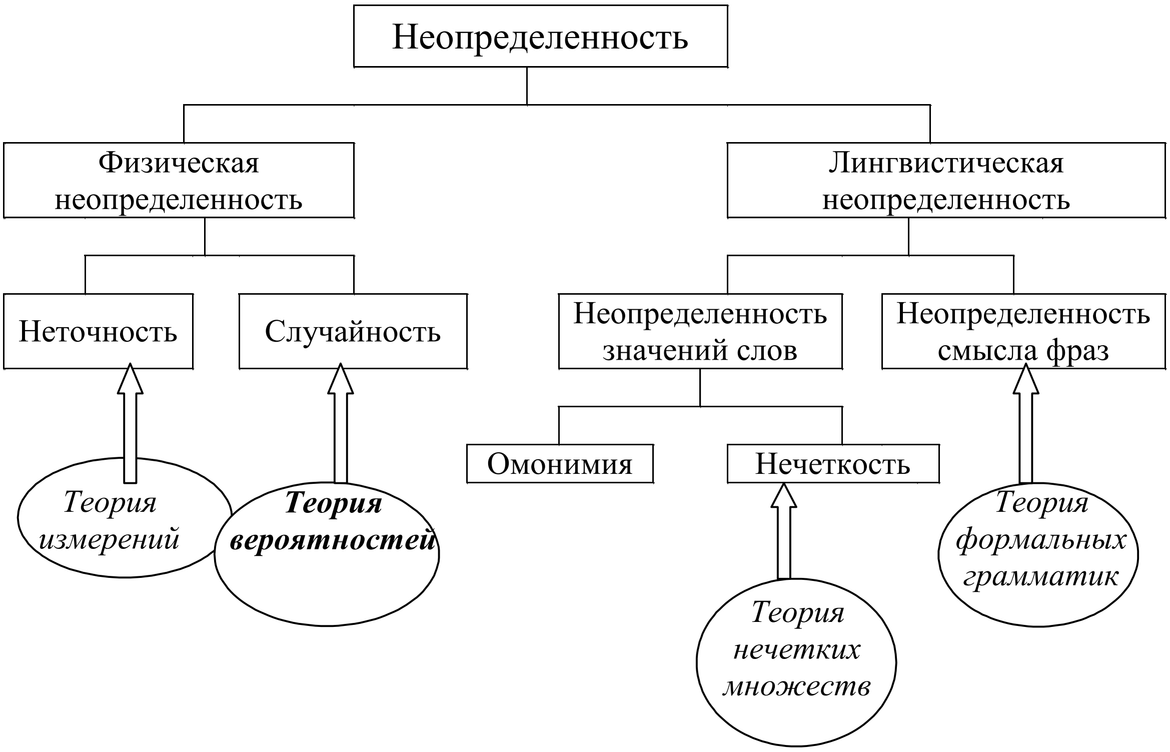

Рис. 1. Классификация неопределенности.

Предметом нашего рассмотрения является физическая неопределенность, которая носит название случайности и является предметом исследования теории вероятностей. Системы, неопределенность которых носит случайный характер, называются вероятностными системами.

Физическая неопределенность описывает неопределенность реального мира с точки зрения наблюдателя. Неточностью занимается теория измерений, а случайностью, как уже было сказано, теория вероятностей. Лингвистическая неопределенность включает в себя неопределенности понятий и конструкций естественного языка. Неопределенностью смысла фраз занимается теория формальных грамматик. Неопределенностью значений слов (нечеткостью) занимается теория нечетких множеств.

Нечеткость в отличие от случайности позволяет учитывать индивидуальные особенности лиц, производящих оценивание или описание рассматриваемых объектов и принимающих на основе полученной информации решения.

Понятие неопределенности в теории вероятностей формализуется путем введения распределения вероятностей на множестве всех возможных наблюдений.

Теория вероятностей как наука возникла в середине XVII века, а поэтому из всех теорий, в настоящее время занимающихся обработкой неопределенности, теория вероятностей наиболее развита.

2. Вероятностная модель эксперимента с конечным или счетным числом исходов

Рассмотрим некоторый эксперимент, все мыслимые исходы которого $\omega_1, \omega_2, ..., \omega_n$ конечны. Реальная природа этих исходов нас не интересует. Будем считать, что при появлении одного из исходов, появление остальных исключено (несовместность исходов), при проведении эксперимента обязательно произойдет один из $\omega_1, \omega_2, ..., \omega_n$ (полная группа исходов) и все исходы равновозможны (равновероятны). Исходы

$\omega_1, \omega_2, ..., \omega_n$ будем называть элементарными событиями, а их совокупность

$$\Omega = \{\omega_1, \omega_2, ..., \omega_n\}$$

(конечным) пространством элементарных событий или пространством исходов.

Например, при однократном подбрасывании монеты пространство элементарных событий состоит из двух элементов $\Omega = \{\Gamma, P\}$, Γ - герб, P - решетка.

Конечно, было бы совсем неинтересно рассматривать только элементарные события, да и в практических задачах трудно ограничиться только ими. Например, если рассмотреть выпадение четного числа очков при бросании игральной кости, то оказывается, что мы рассматриваем уже не одно элементарное событие, а совокупность целых трех элементарных событий, а именно – 2, 4, 6.

Определение элементарного события является тем кирпичиком, с которого начинается построение вероятностной модели эксперимента (с конечным числом исходов). И следующим необходимым понятием является понятие случайного события или просто события. Под этим термином понимается всякий факт, который может произойти или не произойти в результате поступления некой информации (испытания).

Событием (случайным событием) называется любое подмножество A пространства элементарных событий Ω, для которого по условиям эксперимента возможен ответ одного из двух типов: $\omega_i \in A$ или $\omega_i \notin A$.

Например, монета подбрасывается три раза. Пространство элементарных событий состоит из 8 элементов $\Omega = \{\Gamma\Gamma\Gamma, PPP, \Gamma PP, \Gamma P\Gamma, PP\Gamma, P\Gamma P, \Gamma\Gamma P, P\Gamma\Gamma\}$. Если мы рассмотрим множество $A = \{\Gamma\Gamma\Gamma, \Gamma\Gamma P, \Gamma P\Gamma, P\Gamma\Gamma\}$, то получим событие, состоящее в том, что выпало по крайней мере два герба. Если мы сможем зафиксировать результат только первого подбрасывания монеты, то рассматриваемое множество нельзя будет назвать событием, поскольку мы не сможем дать ответ, принадлежит ли конкретный исход множеству A, или не принадлежит.

Опираясь на некоторую систему множеств, являющихся событиями, можно образовывать новые события с логическими связками «или», «и» и «не», что на языке классической теории множеств означает «объединение», «пересечение», «дополнение».

Объединением событий A и B называется событие, состоящее в том, что произошло хотя бы одно из событий A, B, то есть

$$A \cup B = \{\omega \in \Omega : \omega \in A \, или \, \omega \in B\}.$$

Другими словами, объединением событий A и B называется событие, состоящее в том, произошло или событие A, или событие B.

Если события A и B не пересекаются, то их объединение называется суммой.

Пересечением событий A и B называется событие, состоящее в том, что произошло и событие A, и событие B, то есть

$$A \cap B = \{\omega \in \Omega : \omega \in A \, и \, \omega \in B\}.$$

События A и B называются **несовместными или непересекающимися**, если они не могут произойти одновременно. То есть событие $A \cap B$ (AB) не содержит ни одного элементарного события.

В теории вероятностей событие Ø, не содержащее ни одного элементарного события, называется **невозможным.**

Если бросается игральная кость (кубик с цифрами от 1 до 6 на его гранях), то событие «выпало 8» является невозможным.

Дополнением события A называется событие, состоящее из элементов Ω, не принадлежащих событию A, то есть

$$\overline{A} = \{\omega \in \Omega : \omega \notin A\}.$$

По определению события A и $\overline{A}$ не имеют общих элементов, а значит их пересечение является невозможным событием.

Разностью событий A и B называется событие, состоящее в том, что произошло событие A, но не произошло событие B, то есть

$$A \setminus B = \{\omega \in \Omega : \omega \in A, \omega \notin B\}.$$

Пространство Ω называется **достоверным** событием.

Если бросаются две игральные кости, то событие «в сумме выпало больше 1 и меньше 13» является достоверным.

Рассмотрим систему всех подмножеств Ω, включая невозможное и достоверное. Полученная система называется **алгеброй событий** – $\mathscr{A}$, то есть такой системой подмножеств множества Ω, что

1. $\Omega \in \mathscr{A}$.
2. $A \in \mathscr{A}$, $B \in \mathscr{A} \Rightarrow A \cup B \in \mathscr{A}, A \cap B \in \mathscr{A}$.
3. $A \in \mathscr{A} \Rightarrow \overline{A} \in \mathscr{A}$.

Таким образом, мы определили пространство элементарных событий Ω и алгебру событий $\mathscr{A}$.

Сделаем третий шаг в построении вероятностной модели эксперимента и определим вероятность события $A \in \mathscr{A}$.

Согласно **классическому определению вероятности события**

$$P(A) = \frac{m}{n},$$

где m - число элементарных событий, содержащихся в A, а n - число всех элементарных событий (элементов Ω).

Классическое определение вероятности события подразумевает, что всем элементарным событиям приписываются одинаковые вероятности, равные $\frac{1}{n}$. Вообще говоря, вопрос о том, как задавать вероятности элементарных событий, является достаточно трудным. Ответ на этот вопрос лежит вне рамок теории вероятностей.

Основной задачей нашего курса является не вопрос о том, как приписывать вероятности элементарным событиям, а как вычислять вероятности сложных событий по вероятностям элементарных событий.

Тройка $(\Omega, \mathscr{A}, P)$ называется **вероятностным пространством (вероятностной моделью)** эксперимента с конечным пространством исходов Ω и алгеброй событий $\mathscr{A}$.

Свойства вероятности события.

1. $0 \le P(A) \le 1$.

2. $P(\Omega) = 1, P(\varnothing) = 0$.

3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

4. $C = A + B, A \cap B = \varnothing$, тогда $P(C) = P(A) + P(B)$ - теорема сложения.

5. $P(\overline{A}) = 1 - P(A)$.

6. Если $A \subset B$, то $P(A) \le P(B)$.

Докажем свойства вероятности события.

Поскольку $0 \le m \le n$, то $0 \le \dfrac{m}{n} \le 1$ и соответственно $0 \le P(A) \le 1$.

Достоверное событие содержит все элементарные события, а невозможное событие не содержит ни одного элементарного события, поэтому $P(\Omega) = 1, P(\varnothing) = 0$.

Пусть событие A содержит m элементарных событий, а событие B содержит k элементарных событий. Поскольку события пересекаются, то будем считать, что пересечение $A \cap B$ содержит l элементарных событий. В объединении $A \cup B$ таким образом содержатся $m + k - l$ событий, а поэтому

$$P(A \cup B) = \frac{m}{n} + \frac{k}{n} - \frac{l}{n} = P(A) + P(B) - P(A \cap B).$$

Теорема сложения является следствием предыдущего свойства, так как по условию $l = 0$ и соответственно

$$P(C)=P(A)+P(B),\ C=A+B.$$

Согласно теореме сложения

$$P(\overline{A})+P(A)=P(\Omega)=1,$$

а поэтому

$$P(\overline{A})=1-P(A).$$

Если $A\subset B$, то событие B можно представить в виде

$$B=A+B\setminus A.$$

По теореме сложения $P(B)=P(A)+P(B\setminus A)$, а так как $P(B\setminus A)\geq 0$, то

$$P(A)\leq P(B).$$

Рассмотрим некоторый эксперимент, все мыслимые исходы которого $\omega_1,\omega_2,...,\omega_n,...$ счетны. Будем считать, что при появлении одного из исходов, появление остальных исключено (несовместность исходов), при проведении эксперимента обязательно произойдет один из $\omega_1,\omega_2,...,\omega_n,...$ (полная группа исходов) и все исходы не обязательно равновозможны (равновероятны). Исходы $\omega_1,\omega_2,...,\omega_n,...$ будем называть элементарными событиями, а их совокупность

$$\Omega=\{\omega_1,\omega_2,...,\omega_n,...\}$$

(счетным) пространством элементарных событий или пространством со счетным числом исходов.

Пример. Стрелок стреляет до тех пор, пока не попадет. Событие A состоит в том, что стрелок попал, событие $\overline{A}$ соответственно состоит в том, что стрелок не попал. Тогда

$$\Omega=\{A,\overline{A}A,\overline{A}\overline{A}A...,\overline{A}\overline{A}\overline{A}\overline{A}....A,...\}$$

Система $\mathscr{A}$ подмножеств Ω называется **σ-алгеброй** , если она является алгеброй и выполняется следующее свойство

2*) $\qquad A_n\in\mathscr{A},\ n=1,2,...\Rightarrow\bigcup_n A_n\in\mathscr{A},\ \bigcap_n A_n\in\mathscr{A}.$

Тогда для $A \in \mathscr{A}$

$$P(A) = \sum_{\omega_i \in A} p(\omega_i).$$

Тройка $(\Omega, \mathscr{A}, P)$ называется **вероятностным пространством** эксперимента со счетным числом исходов.

Вероятностные пространства с конечным или счетным числом исходов называются **дискретными вероятностными пространствами**.

Задачи.

Задача 1. Построить пространство элементарных событий при n - кратном подбрасывании монеты.

Задача 2. Построить пространство элементарных событий при однократном бросании игральной кости.

Задача 3. Построить пространство элементарных событий при проведении следующего эксперимента: бросается монета и если выпал герб, то бросается игральная кость, а если выпала решетка, то опять бросается монета.

Задача 4. В урне находятся N шаров, пронумерованных от 1 до N. Из урны извлекается шар и возвращается обратно в урну. Построить пространство элементарных событий при n-кратном извлечении шаров, $n \le N$.

Задача 5. В урне находятся N шаров, пронумерованных от 1 до N. Из урны извлекается шар. Построить пространство элементарных событий при n-кратном извлечении шаров, $n \le N$.

Задача 6. Студенты Иванов и Петров сдают экзамен. Событие A состоит в том, что Иванов успешно сдал экзамен, событие B состоит в том, что Петров успешно сдал экзамен. Найти событие, состоящее в том, что а) только один студент успешно сдал экзамен;

б) только Иванов успешно сдал экзамен;

в) только Петров успешно сдал экзамен;

г) оба студента не сдали экзамен.

Задача 7. Бросают монету и игральную кость. Событие A состоит в том, что выпал герб, событие B состоит в том, что выпало 5 очков. Найти событие, состоящее в том, что

а) выпала решетка и не выпало 5 очков;

б) выпал герб и не выпало 5 очков;

в) выпала решетка и выпало 5 очков.

Задача 8. Завод выпускает детали, среди которых есть окрашенные и неокрашенные, а также годные и бракованные. Событие A состоит в том, что деталь годная, событие B состоит в том, что деталь окрашенная. Найти событие, состоящее в том, что деталь бракованная и окрашенная.

Задача 9. Монета бросается два раза. Найти вероятность того, что выпадет по крайней мере один герб.

Задача 10. Монета бросается три раза. Найти вероятность того, что выпадет ровно две решетки.

Задача 11. Бросают игральную кость. Найти вероятность того, что выпадет больше четырех очков.

Задача 12. Бросают три игральные кости. Найти вероятность того, что в сумме выпадет 4 очка.

Задача 13. Вероятность, что студент Иванов сдаст экзамен на отлично равна 0.8, а вероятность, что студент Петров сдаст экзамен на отлично равна 0.4. Найти вероятность, что только один из студентов сдаст экзамен на отлично.

Задача 14. Бросаются две игральные кости. Найти вероятность, что на костях выпадет разное число очков, сумма которых не меньше 9.

Задача 15. Вероятность попадания в цель из первого орудия равна 0.8, из второго 0.7, из третьего 0.9. Найти вероятность, что

а) только третье орудие поразило цель;

б) все орудия поразили цель;

в) ни одно орудие не поразило цель;

г) хотя бы одно орудие поразило цель;

д) первое и второе орудия поразили цель;

е) только одно орудие поразило цель;

ж) только два орудия поразили цель.

Задача 16. Вероятность, что стрелок поразит цель, равна 0.7. Найти вероятность того, что

а) стрелок поразит цель со второго выстрела;

б) стрелок поразит цель меньше, чем с четвертого выстрела;

в) стрелок будет стрелять не больше четырех раз, пока не поразит цель.

Задача 17. В электрической цепи к сопротивлению R_1 последовательно подключено параллельное соединение сопротивлений R_2 и R_3. События $A_i, i=\overline{1,3}$ состоят в том, что в течение часа соответственно сопротивления $R_i, i=\overline{1,3}$ не выйдут из строя. Найти событие, состоящее в том, что в течение часа в цепи будет ток.

Задача 18. В условиях задачи 17 $P\left(\overline{A}_1\right)=0.3, P\left(\overline{A}_2\right)=0.2,\ P\left(\overline{A}_3\right)=0.4$. Найти вероятность, что в течение часа в цепи будет ток.

Задача 19. В результате проведения ремонтных работ (условие задачи 17) была увеличена вероятность безотказной работы сопротивления R_1 на a%. Найти вероятность, что в течение часа в цепи будет ток при a равном

а) 5%;

б) 10%;

в) 15%.

Задача 20. Брак при производстве изделия вследствие дефекта A составляет 12%, вследствие дефекта B 13%. Процент годной продукции составляет 85%. Найти вероятность, что наудачу взятое изделие имеет оба дефекта.

Задача 21. Брак при производстве изделия вследствие дефекта A составляет 11%, вследствие дефекта B 10%. Процент годной продукции составляет 87%. Найти вероятность, что наудачу взятое изделие имеет оба дефекта.

Задача 22. Брак при производстве изделия вследствие дефекта A составляет 14%, вследствие дефекта B 12%. Процент годной продукции составляет 82%. Найти вероятность, что наудачу взятое изделие с дефектом A имеет и дефект B.

Задача 23. Брак при производстве изделия вследствие дефекта A составляет 14%, вследствие дефекта B 12%. Процент годной продукции составляет 82%. Найти вероятность, что наудачу взятое изделие имеет

а) только дефект A;

б) только дефект B.

Задача 24. Брак при производстве изделия вследствие дефекта A составляет 14%, вследствие дефекта B 12%. Процент годной продукции составляет 82%. Найти вероятность, что наудачу взятое изделие с дефектом B имеет и дефект A.

3. Элементы комбинаторики

Описание пространства элементарных событий при решении практических задач существенным образом зависит от того, считаем ли мы выборки тождественного состава, например, такие, как, скажем, $\{1,2,4,5\}$ и $\{4,2,5,1\}$, различными или одинаковыми. Поэтому принято различать упорядоченные выборки и неупорядоченные. **Упорядоченные выборки** объявляются различными, если они состоят из одних и тех же элементов, но отличаются порядком следования этих элементов. **Неупорядоченные выборки** объявляются тождественными, если они состоят из одних и тех же элементов, но отличаются порядком следования этих элементов.

Рассмотрим n элементов и составим из них всевозможные упорядоченные выборки. На первом месте может стоять любой из n элементов, на втором месте любой из оставшихся $n-1$ элементов, на третьем месте любой из оставшихся $n-2$ элементов и т.д. Всего мы получим $n\cdot(n-1)\cdot(n-2)...2\cdot 1 = n!$ различных выборок, а их число

$$P_n = n!$$

называется числом перестановок из n элементов.

Рассмотрим n элементов и составим всевозможные упорядоченные выборки из k элементов. На первом месте может стоять любой из n элементов, на втором месте любой из оставшихся $n-1$ элементов, на третьем месте любой из оставшихся $n-2$ элементов и на k-ом месте любой из $n-k+1$ элементов. Всего мы получим $n\cdot(n-1)\cdot(n-2)\cdot\ldots\cdot(n-k+1) = \frac{n!}{(n-k)!}$ различных выборок, а их число

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

называется числом размещений из n элементов по k.

Рассмотрим n элементов и составим всевозможные неупорядоченные выборки из k элементов. Число этих выборок будет меньше, чем число размещений из n элементов по k ровно в $k!$ раз, поскольку именно $k!$ способами можно упорядочить k элементов.

Число неупорядоченных выборок из n элементов по k равно

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

и называется числом сочетаний из n элементов по k.

Мы рассматривали перестановки, сочетания и размещения при условии, что все элементы n различны. А теперь нас будут интересовать перестановки, размещения и сочетания с повторениями.

Рассмотрим n элементов и составим из них всевозможные упорядоченные выборки по k элементов, причем каждый из n элементов может повторяться до k раз. На первом месте может стоять любой из n элементов, на втором месте тоже любой из n элементов (поскольку элементы могут повторяться) и на каждом из k мест может стоять любой из n элементов. Всего мы получим n^k выборок, а их число

$$\bar{A}_n^k = n^k$$

называется числом размещений из n элементов по k с повторениями.

Рассмотрим n элементов, среди которых n_1 одинаковых элементов одного типа, n_2 одинаковых элементов второго типа, и т.д., n_k одинаковых элементов k-го типа, $\sum_{i=1}^{k} n_i = n$. Составим из n элементов всевозможные упорядоченные выборки, очевидно, что их будет меньше, чем число перестановок из n разных элементов в $n_1! \cdot n_2! \cdot \ldots \cdot n_k!$ раз. Число таких перестановок равно

$$P_n(n_1, n_2, \ldots, n_k) = \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_k!}$$

и называется числом перестановок с повторениями.

Рассмотрим n элементов и составим из них всевозможные неупорядоченные выборки по k элементов, причем каждый из n элементов может повторяться до k раз.

Будем условно записывать каждую выборку в виде совокупности нулей и единиц. Если выбран какой-то элемент, то ставим единицу, причем, если элемент повторяется, то ставим столько единиц, сколько раз повторяется этот элемент. Разные элементы отделяем друг от друга нулем. Таким образом, мы получим последовательность $n-1$ нулей и k единиц. Тогда число сочетаний с повторениями из n элементов по k - C_n^{-k} равно числу перестановок с повторениями - $P_{n+k-1}(n-1,k)$, то есть

$$C_n^{-k} = \frac{(n+k-1)!}{(n-1)!\,k!}.$$

Задачи.

Задача 25. Турист пришел в туристическое агентство, чтобы заказать тур по пяти европейским странам. В качестве бонуса при условии полной оплаты в день обращения туристу предложили выбрать индивидуальный маршрут следования. Сколько маршрутов может выбрать турист?

Задача 26. Турист пришел в туристическое агентство, чтобы заказать тур по пяти европейским странам, но у него не хватило денег на пять стран, а хватило только на три. В качестве бонуса при условии полной оплаты в день обращения туристу предложили выбрать индивидуальный маршрут следования. Сколько маршрутов может выбрать турист?

Задача 27. Турист пришел в туристическое агентство, чтобы заказать тур по пяти европейским странам, но у него не хватило денег на пять стран, а хватило только на три. Сколько различных вариантов может выбрать турист?

Задача 28. Сколько элементарных событий получится при проведении эксперимента, состоящем в бросании трех игральных костей?

Задача 29. Ребенок играет с картинками, на которых написаны буквы А, Р, Б, А, А, К,А, Д, Б, А, Р. Сколько различных буквосочетаний сложит ребенок?

Задача 30. Определить число размещений и сочетаний с повторениями из 3 элементов по 5.

Задача 31. Ребенок играет с четырьмя карточками, на которых написаны две буквы А и две буквы П. Найти вероятность, что ребенок сложит слово «папа».

Задача 32. В ящике содержится 10 деталей, помеченных номерами от 1 до 10. Из ящика наудачу извлекаются 6 деталей. Найти вероятность, что среди извлеченных окажутся а) деталь № 5, б) детали № 1 и № 4, в) детали № 2, № 3, № 6.

Задача 33. На полке случайным образом расставляют 8 книг, среди которых книга по кулинарии и справочник водителя. Найти вероятность, что эти две книги окажутся рядом.

Задача 34. Из колоды (36 карт) наудачу вынимают три карты. Найти вероятность, что среди них окажется ровно одна дама.

Задача 35. Из колоды (36 карт) наудачу вынимают три карты. Найти вероятность, что среди них окажется хотя бы один король.

Задача 36. Джо и Мо по очереди бросают кость. Найти вероятность того, что Джо выбросит больше очков, чем Мо.

Задача 37. В урне находятся 5 белых шаров и 6 черных. Из урны вынимаются 2 шара. Найти вероятность, что

а) оба шара белые;

б) оба шара черные;

в) один шар белый и один шар черный.

Задача 38. В урне находятся 2 красных шара, 3 черных и 3 белых шара. В урну добавляют 2 белых шара и после этого извлекают 3 шара. Найти вероятность, что

а) все шары белые;

б) все шары черные;

в) 2 шара белых и один красный;

г) 2 шара красных и один черный;

д) один шар белый, один шар черный и один шар красный.

4. Вероятностная модель эксперимента с бесконечным числом исходов. Геометрическое и статистическое определения вероятности события

В разделе 2 мы рассмотрели дискретные вероятностные пространства (вероятностные пространства с конечным или счетным числом исходов). Однако достаточно часто встречаются эксперименты, для которых число исходов несчетно.

Рассмотрим отрезок, на который случайным образом бросается точка. Отрезок содержит несчетное число точек, поэтому вероятность элементарного исхода (или попадания в определенную точку на отрезке) должна быть равна нулю. Если мы хотим определить вероятность попадания точки в правую половину отрезка, пользуясь дискретной моделью вероятностного пространства, то, складывая нулевые вероятности исходов, мы эту вероятность не определим. Однако интуитивно понятно, что эта вероятность должна быть равной 0.5.

Напомним определение понятие σ-алгебры подмножеств Ω.

Система $\mathscr{A}$ подмножеств Ω называется **σ-алгеброй** , если она является алгеброй , то есть

1. $\Omega \in \mathscr{A}$.

2. $A \in \mathscr{A}$, $B \in \mathscr{A} \Rightarrow A \cup B \in \mathscr{A}$, $A \cap B \in \mathscr{A}$.

3. $A \in \mathscr{A} \Rightarrow \overline{A} \in \mathscr{A}$.

и дополнительно выполняется следующее свойство

2^*) $\qquad A_n \in \mathscr{A}, n = 1,2,... \Rightarrow \bigcup_n A_n \in \mathscr{A}, \ \bigcap_n A_n \in \mathscr{A}.$

Пространство Ω вместе σ-алгеброй $\mathscr{A}$ называется **измеримым пространством** и обозначается $(\Omega,\mathscr{A})$.

Вероятностью или вероятностной мерой, заданной на $\mathcal{A}$, называется функция множеств $P(A)$, $A \in \mathcal{A}$, удовлетворяющая следующим условиям:

1. $P(A) \geq 0$,
2. $P(\Omega) = 1$;
3. $P\left(\bigcup_n A_n\right) = \sum_n P(A_n)$ для любых попарно непересекающихся $A_n \in \mathcal{A}$

(счетная аддитивность или **σ-аддитивность).**

Тройка $(\Omega, \mathcal{A}, P)$ называется **вероятностным пространством** и лежит в основе аксиоматики А.Н.Колмогорова.

Рассмотрим простейшие случаи вероятностных пространств и дадим геометрическое определение вероятности.

Будем предполагать, что область g с мерой $mesg$ (длина, площадь, объем) принадлежит области G с мерой $mesG$. Тогда $\Omega = G$, измеримые множества- отрезки, интервалы, полуинтервалы, подмножества плоской или объемной области G. В область G бросается точка. Вероятность события A, состоящего в том, что брошенная точка попадет в область g, полагается равной

$$P(A) = \frac{mesg}{mesG}$$

и называется геометрической вероятностью.

Как известно, взаимоотношение явления и его вероятностной модели имеет статистический характер, то есть обнаруживается при повторных наблюдениях за явлением.

Если какое-то испытание (эксперимент) повторяется n раз и при этом k раз происходит событие A, то число k называют частотой появления события A в n испытаниях, а отношение

$$h_n(A) = \frac{k}{n}$$

называют относительной частотой появления события A в n испытаниях.

Многочисленные наблюдения позволили сделать вывод, что частоты исходов в длинном ряду испытаний стабилизируются, их колебания с ростом числа испытаний становятся все менее значительными. Этот эмпирический факт, называемый законом устойчивости частот, наблюдается в самых различных ситуациях.

Выходя за пределы реального опыта, полагают, что при неограниченном повторении испытаний относительные частоты $h_n(A)$ стремятся к некоторому пределу $P(A)$, который принимается за статистическую вероятность события A:

$$P(A) = \lim_{n\to\infty} h_n(A) = \lim_{n\to\infty} \frac{k}{n}.$$

Задачи.

Задача 39. На отрезок AB длины l случайным образом ставится точка C. Найти вероятность события F, состоящего в том, что длина меньшего из отрезков AC и CB будет больше $\frac{l}{4}$.

Задача 40. В квадрат со стороной a вписан другой квадрат с вершинами, лежащими на серединах сторон первого квадрата. В первый квадрат случайным образом бросается точка. Найти вероятность, что она попадет во вписанный квадрат.

Задача 41. В круг радиуса R случайным образом бросается точка. Найти вероятность того, что эта точка попадет в кольцо шириной $R-r$.

Задача 42. На отрезок единичной длины случайным образом бросается точка. Найти вероятность того, что от точки до концов отрезка расстояние будет больше 0.1.

Задача 43. В круг радиуса 4 бросается точка. Найти вероятность того, что от точки до окружности (границы круга) расстояние будет не больше 1.

Задача 44. Молодой человек и девушка договорились о свидании у кинотеатра между 18 и 19 часами с условием, что первый пришедший ждет 15 минут и уходит. Найти вероятность, что они встретятся.

5. Наивероятнейшее число появлений события. Локальная и интегральная теоремы Муавра-Лапласа

Предположим, что проводится серия из n испытаний, которые состоят в наблюдении числа появлений события $A, P(A) = p$. Обозначим через $P_n(k)$ вероятность того, что событие A появилось k раз в n испытаниях. Легко найти, что

$$P_n(k) = C_n^k p^k q^{n-k}.$$

Пусть

$$P_n(k_0) = \max_{0 \le k \le n} P_n(k).$$

Тогда число k_0 называется наивероятнейшим числом появления события A в n испытаниях. Определим границы k_0.

Так как $P_n(k_0) = \max\limits_{0 \le k \le n} P_n(k)$, то $P_n(k_0) \ge P_n(k_0 + 1)$ и $P_n(k_0) \ge P_n(k_0 - 1)$.

В силу того, что $P_n(k) = C_n^k p^k q^{n-k}$, $q = 1 - p$, получаем два неравенства:

$$\frac{n!}{k_0!\,(n-k_0)!} p^{k_0} q^{n-k_0} \ge \frac{n!}{(k_0+1)!\,(n-k_0-1)!} p^{k_0+1} q^{n-k_0-1},$$

$$\frac{n!}{k_0!\,(n-k_0)!} p^{k_0} q^{n-k_0} \ge \frac{n!}{(k_0-1)!\,(n-k_0+1)!} p^{k_0-1} q^{n-k_0+1}.$$

После преобразования:

$$q(k_0+1)\geq p(n-k_0),\ p(n-k_0+1)\geq qk_0.$$

Таким образом

$$np-q\leq k_0\leq np+p.$$

При больших значениях n и k вычисление $P_n(k)$ (вероятности, что в n испытаниях событие A произошло k раз) и $P_n(k_1,k_2)=\sum\limits_{k=k_1}^{k_2}P_n(k)$ (вероятности, что в n испытаниях событие A произошло от k_1 до k_2 раз) превращается в сложную в техническом плане задачу. С этой задачей блестяще справился французский математик Муавр, который всю жизнь прожил в Англии, а потом ее обобщил тоже французский математик Лаплас, раскрыв тем самым более полно значение результата Муавра. Исторически теоремы Муавра-Лапласа (локальная и интегральная) являются одними из первых предельных теорем теории вероятностей.

Локальная теорема Муавра-Лапласа состоит в следующем:

$$P_n(k)=\frac{1}{\sqrt{2\pi npq}}e^{-\frac{x^2}{2}}(1+\alpha_n),$$

где $\alpha_n\to 0$ при $n\to\infty$ равномерно для всех k, для которых $x=\dfrac{k-np}{\sqrt{npq}}$ принадлежит какому-либо конечному интервалу.

Для практических задач используется приближенное равенство:

$$P_n(k)\approx\frac{1}{\sqrt{npq}}\varphi(x),$$

$$\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},\ x=\frac{k-np}{\sqrt{npq}},$$

Забегая вперед, отметим, что функция $\varphi(x)$ является плотностью распределения вероятностей стандартной нормальной случайной величины, о которой речь пойдет в разделе 7. Для значений $\varphi(x)$ составлена таблица (приложение 1). Поскольку $\varphi(x)$ является четной, то есть $\varphi(-x)=\varphi(x)$, то в таблице присутствуют только положительные значения аргумента.

Интегральная теорема Муавра-Лапласа состоит в следующем:

$$P_n(k_1,k_2)=(\Phi(x_2)-\Phi(x_1))(1+\alpha_n),$$

где $\alpha_n \to 0$ при $n \to \infty$ равномерно для всех $k \in (k_1,k_2)$, для которых $x=\frac{k-np}{\sqrt{npq}}$ принадлежит какому-либо конечному интервалу.

Для практических задач используется приближенное равенство:

$$P_n(k_1,k_2)\approx(\Phi(x_2)-\Phi(x_1)),$$

$$\Phi(x)=\int_0^x \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy,\ x_2=\frac{k_2-np}{\sqrt{npq}},x_1=\frac{k_1-np}{\sqrt{npq}}.$$

Функция $\Phi(x)=\int_0^x \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy$ называется функцией Лапласа, для ее значений составлена таблица (приложение 2). Значением функции Лапласа $\Phi(a)=\int_0^a \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}dy$ является площадь фигуры, ограниченная графиком функции $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ и двумя вертикальными прямыми $x=0, x=a$.

Функция Лапласа является нечетной, то есть $\Phi(-x)=-\Phi(x)$. В таблице даны значения функции Лапласа только при положительных значениях аргумента, чтобы найти значения функции Лапласа при отрицательных значениях аргумента, нужно воспользоваться ее свойством нечетности.

В таблице приложения 2 даны значения функции Лапласа при значениях аргумента $0 \le x \le 5$. Дело в том, что при $x > 5$ функция $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ резко стремится к нулю, а поэтому к площади фигуры, ограниченной ее графиком, практически уже ничего не добавляется. Это значит, что при всех значениях $x > 5$ функция Лапласа равняется $\Phi(5) = 0.5$.

Предположим, что p - вероятность появления события A в n испытаниях и $\lim\limits_{n\to\infty} p(n) = 0, \lim\limits_{n\to\infty} np = \lambda, \lambda > 0$, тогда справедлива **теорема Пуассона:**

$$\lim_{n\to\infty} P_n(k) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Задачи.

Задача 45. Вероятность обнаружения брака при проверке изделия на годность составляет 0.2. Найти наивероятнейшее число обнаруженных бракованных изделий при проверке его ста образцов.

Задача 46. Вероятность того, что студент забыл записать домашнее задание, равно 0.3. Найти наивероятнейшее число студентов, забывших записать домашнее задание на потоке из 240 человек.

Задача 47. Вероятность поражения мишени при одном выстреле равна 0.75. Найти вероятность, что при 10 выстрелах произойдет 8 поражений мишени.

Задача 48. Вероятность того, что сообщение проходит с помехами равно 0.2. Найти вероятность, что из 400 отправленных сообщений от 70 до 100 пройдут с помехами.

Задача 49. Известно, что в n испытаниях событие A произошло k раз, а p - вероятность появления события A в одном испытании. Используя интегральную теорему Муавра-Лапласа, найти $P\left(\left|\frac{k}{n} - p\right| < \varepsilon\right)$.

Задача 50. Вероятность стандартного изделия равна 0.8. Найти вероятность, что из пяти наудачу отобранных изделий два будут стандартными.

Задача 51. Вероятность стандартного изделия равна 0.8. Найти вероятность, что из пяти наудачу отобранных изделий не менее четырех будут стандартными.

Задача 52. Вероятность бракованного изделия равна 0.003. Найти вероятность, что из 1000 изделий бракованных будет не больше 3.

Задача 53. Вероятность сбоя системы в течение часа равна 0.002. Найти вероятность, что за 1000 часов функционирования системы будет меньше двух случаев сбоя.

Задача 54. Брак при производстве изделия вследствие дефекта A составляет 12%, вследствие дефекта B 13%. Процент годной продукции составляет 85%. Установите соответствие между объемом n произведенной продукции и наиболее вероятным (наивероятнейшим) количеством изделий только с дефектом A

а) $n = 100$;

б) $n = 200$;

в) $n = 400$.

Задача 55. Брак при производстве изделия вследствие дефекта A составляет 11%, вследствие дефекта B 10%. Процент годной продукции составляет 87%. Установите соответствие между объемом n произведенной продукции и наиболее вероятным (наивероятнейшим) количеством изделий только с дефектом B

а) $n = 300$;

б) $n = 400$;

в) $n = 500$.

Задача 56. Брак при производстве изделия вследствие дефекта A составляет 14%, вследствие дефекта B 12%. Процент годной продукции составляет 82%. Установите соответствие между объемом n произведенной продукции и наиболее вероятным (наивероятнейшим) количеством изделий с двумя дефектами

а) $n = 200$;

б) $n = 300$;

в) $n = 400$.

6. Условная вероятность, формула полной вероятности, формула Байеса

Условная вероятность события A при условии, что произошло событие B, по определению полагается равной

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(AB)}{P(B)}.$$

События A и B называются **независимыми**, если

$$P(AB) = P(A) \cdot P(B).$$

События $A_1, A_2, \dots, A_n$ называются **независимыми**, если

$$P(A_1A_2...A_n)=P(A_1)\cdot P(A_2)\cdot...\cdot P(A_n).$$

Пусть A - некоторое событие, $B_1,\ B_2,...,B_n$ - попарно несовместные события, такие что $A\subset\bigcup\limits_{i=1}^{n}B_i$. Тогда имеет место **формула полной вероятности:**

$$P(A)=\sum_{i=1}^{n}P(A|B_i)\cdot P(B_i).$$

Пусть A - некоторое событие, которое может наступить только при появлении одного из попарно несовместных событий $B_1,\ B_2,...,B_n$, таких что $A\subset\bigcup\limits_{i=1}^{n}B_i$. Заранее неизвестно, какое из событий $B_1,\ B_2,...,B_n$ наступит, поэтому задаются априорные вероятности $P(B_i), i=\overline{1,n}$ (вероятности, полученные до опыта). Сами события $B_1,\ B_2,...,B_n$ называются **гипотезами**. Когда событие A при одной из гипотез B_i наступает, то вычисляются апостериорные вероятности (вероятности, полученные после опыта) $P(B_i|A)$, позволяющие сделать вывод о правильности априорных вероятностей и переоценить их. Помогает в этом **формула Байеса**:

$$P(B_i|A)=\frac{P(B_iA)}{P(A)}=\frac{P(A|B_i)\cdot P(B_i)}{\sum\limits_{i=1}^{n}P(A|B_i)\cdot P(B_i)}.$$

В большинстве случаев исследование систем сопряжено с измерением каких-либо показателей или характеристик этих систем. При измерении значения какого-либо показателя вероятностной системы приходится иметь дело со случайными величинами, о которых речь пойдет в следующем разделе.

Задачи.

Задача 57. В колоде 36 карт. Последовательно извлекаются две карты. Найти вероятность, что вторая карта туз при условии, что первая карта тоже туз.

Задача 58. В колоде 36 карт. Последовательно извлекаются две карты. Найти вероятность, что вторая карта дама при условии, что первая карта туз.

Задача 59. В экспериментах на встречных электрон-позитронных пучках вероятность того, что за единицу времени произойдет j столкновений, сопровождающихся рождением новых элементарных частиц, равна

$$p_j = \frac{e^{-\lambda}\lambda^j}{j!},\ j = 0,1,...$$

где λ - положительный параметр.

При каждом столкновении в результате взаимодействия могут возникнуть различные группы элементарных частиц, при этом вероятность появления каждой группы фиксирована и не зависит от результатов взаимодействия при остальных столкновениях. В качестве одной из таких групп рассматривается пара μ-мезонов, а через p обозначается вероятность их появления при одном столкновении. Чему равна вероятность события A_k, при котором за единицу времени рождается k пар μ-мезонов?

Задача 60. Рассматриваются две выборки изделий, произведенных на разных заводах. Выборка изделий с первого завода содержит 24 изделия, из которых 8 бракованных. Выборка изделий со второго завода содержит 18 изделий, из которых 6 бракованных. Изделие, произведенное на втором заводе, перекладывается к изделиям, произведенным на первом заводе. Найти вероятность того, что изделие, извлеченное из выборки с первого завода, будет бракованным.

Задача 61. Два завода выпускают изделия. С первого завода в магазин отбирается 45% изделий, со второго завода 55%. На первом заводе 90% годных изделий, на втором заводе 80% годных изделий. Найти вероятность, что случайно выбранное в магазине изделие окажется годным.

Задача 62. В трех урнах находятся белые и черные шары. В первой урне находятся 2 белых и 3 черных шара, во второй урне находятся 1 белый и 1 черный шар, а в третьей урне находятся 7 белых и 8 черных шаров. Найти вероятность, что шар, взятый из случайно выбранной урны, окажется белым.

Задача 63. Изделие выпускается двумя заводами. Объем продукции второго завода в k раз превосходит объем продукции первого завода. Доля брака первого завода равна p_1, доля брака второго завода равна p_2. Изделия, выпущенные двумя заводами за один и тот же промежуток времени, смешали и пустили в продажу. Найти вероятность того, что приобретенное изделие выпущено вторым заводом, если оно оказалось бракованным.

Задача 64. Имеются пять урн следующего состава: две урны содержат 2 белых и 3 черных шара, две урны содержат 1 белый и 4 черных шара и одна урна содержит 4 белых и 1 черный шар. Из одной случайным образом выбранной урны взят шар. Он оказался белым. Найти вероятность, что этот шар вынут из урны:

а) третьего состава,

б) первого состава,

в) второго состава.

Задача 65. На столе преподавателя лежат два конверта с задачами, каждая из которых написана на отдельном листе. В первом конверте 24 задачи, из них 16 повышенной сложности. Во втором конверте 18 задач, из них 12 повышенной сложности. Преподаватель перекладывает из второго конверта в первый одну задачу, после этого студент берет из первого конверта одну задачу. Найти вероятность, что студент не возьмет задачу повышенной сложности.

Задача 66. В условиях задачи 65 найти вероятность, что студент возьмет задачу повышенной сложности.

Задача 67. На первую станцию поступило два сообщения без искажений и одно с искажениями, на вторую станцию поступило одно сообщение без искажений и четыре сообщения с искажениями. Для передачи сообщения случайным образом выбирается станция, после чего наудачу передается одно из поступивших на эту станцию сообщений. Найти вероятность, что будет передано сообщение без искажений.

Задача 68. На первую станцию поступило три сообщения без искажений и пять с искажениями, на вторую станцию поступило четыре сообщение без искажений и четыре сообщения с искажениями, на третью станцию поступило пять сообщений без искажений и три сообщения с искажениями. С первой станции на вторую случайным образом отправили два сообщения, после этого со второй станции на третью отправили одно сообщение. Найти вероятность, что сообщение, которое наудачу будет оправлено с третьей станции, окажется без искажений.

Задача 69. В урне содержится N шаров, из них n белых, а $N-n$ черных. Какова вероятность извлечь на втором шаге белый шар?

Задача 70. В урне находятся две монеты, одна из них симметричная (вероятности выпадения герба и решетки равны $\frac{1}{2}$), а вторая несимметричная (вероятность выпадения герба равна $\frac{1}{3}$, а вероятность выпадения решетки равна $\frac{2}{3}$). Из урны извлекается монета и подбрасывается. Результатом подбрасывания является герб. Найти вероятность, что из урны была извлечена симметричная монета.

Задача 71. В группе 25 студентов, преподаватель подготовил для экзамена 25 билетов. Студент Петров знает только билет № 5 и не может решить идти ему на экзамен первым или последним. В каком случае из двух у него больше вероятность взять билет № 5?

Задача 72. Пусть событие А не зависит от самого себя. Показать, что вероятность такого события равна 0 или 1.

Задача 73. Пусть вероятность события А равна 0 или 1. Показать, что событие А и любое событие В независимы.

7. Случайные величины и их числовые характеристики

Случайной величиной называют числовую функцию на пространстве элементарных событий.

Случайные величины бывают дискретные и непрерывные. Возможные значения дискретной случайной величины составляют конечное или счетное множество. Возможные значения непрерывной случайной величины составляют прямую, луч, отрезок. Случайные величины могут быть заданы своими распределениями вероятностей.

Если случайная величина X дискретна, то закон распределения вероятностей может быть задан в виде таблицы 1, в первой строке которой перечислены возможные значения случайной величины, а во второй строке – соответствующие этим значениям вероятности:

Таблица 1.

x_i	x_1	x_2	$\dots$	x_n
p_i	p_1	p_2	$\dots$	p_n

Если случайная величина X непрерывна, то ее закон распределения задается через функцию распределения вероятностей

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

или через плотность распределения вероятностей (если $F(x)$ дифференцируема)

$$f(x) = F'(x).$$

Функция распределения вероятностей случайной величины (дискретной и непрерывной) обладает следующими свойствами.

1. $P(x_1 < X \le x_2) = F(x_2) - F(x_1)$.

2. Для $x_2 \ge x_1$ $F(x_2) \ge F(x_1)$, то есть функция распределения вероятностей является неубывающей.

3. $0 \le F(x) \le 1$.

4. $F(-\infty) = 0, F(+\infty) = 1$.

5. $F(x)$ непрерывна справа.

Плотность распределения вероятностей непрерывной случайной величины обладает следующими свойствами.

1. $f(x) \ge 0$.

2. $P(x_1 < X \le x_2) = \int\limits_{x_1}^{x_2} f(x)dx$.

3. $\int\limits_{-\infty}^{+\infty} f(x)dx = 1$.

Случайные величины X, Y называются независимыми, если для любых $x, y \in R$

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

Это определение по аналогии может быть продолжено на группу случайных величин.

Знание закона распределения случайной величины дает полную информацию об этой величине и, исходя из этого, представляет ее самую информативную характеристику. В реальных исследованиях знание точных законов распределения случайных величин значительно упростили бы решения ряда задач, но, к сожалению, эта информация, чаще всего, бывает недоступна. Однако, бывают доступны и очень полезны, может быть, менее информативные, но гораздо простые характеристики случайных величин. Такими характеристиками являются математическое ожидание случайной величины X - $M(X)$, и ее дисперсия $D(X)$.

Для дискретной случайной величины X математическое ожидание равно

$$M(X) = \sum_k x_k p_k.$$

Для непрерывной случайной величины X с плотностью распределения вероятностей $f(x)$ математическое ожидание равно

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

Математическое ожидание случайной величины имеет следующие свойства.

1. Если $X \ge 0$, то $M(X) \ge 0$.

2. $M(aX + bY) = aM(X) + bM(Y), a, b - const$.

3. Если $X \ge Y$, то $M(X) \ge M(Y)$.

4. $|M(X)| \le M|X|$.

5. Если X, Y - независимы, то $M(XY) = M(X)M(Y)$.

Дисперсией случайной величины X называется величина

$$D(X)=M(X-M(X))^2.$$

Величина $\sigma=\sqrt{D(X)}$ называется **средним квадратическим отклонением** или стандартным отклонением.

$$D(X)=M\left(X^2-2XM(X)+(M(X))^2\right)=M\left(X^2\right)-(M(X))^2.$$

Для дискретной случайной величины X дисперсия равна

$$D(X)=\sum_k (x_k-M(X))^2 p_k=\sum_k x_k^2 p_k-\left(\sum_k x_k p_k\right)^2.$$

Для непрерывной случайной величины X с плотностью распределения вероятностей $f(x)$ дисперсия равна

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

Дисперсия случайной величины имеет следующие свойства.

1. $D(X)\geq 0$.
2. $D(a)=0, a=const$.
3. $D(aX)=a^2D(X)$.
4. Если случайные величины X,Y - независимы, то $D(X+Y)=D(X)+D(Y)$.

Свойство 4 будет выполняться и при меньшем предположении, нежели независимость X,Y, а именно при некоррелированности X,Y, то есть когда $\operatorname{cov}(X,Y)=M(X-M(X))(Y-M(Y))=0$.

Математическое ожидание является аналогом среднего (среднего арифметического) для значений случайной величины. Определение математического ожидания позволяет заменить случайную величину на ее среднее значение, которое является детерминированным и позволяет работать со случайной величиной (точнее с этим значением) детерминированными методами.

Дисперсия определяет меру разброса значений случайной величины вокруг ее среднего значения (математического ожидания). Чем больше дисперсия, тем больше разброс, тем менее информативно значение среднего (математического ожидания). Чем меньше дисперсия, тем более информативно среднее значение случайной величины и более прогнозируемы возможные значения случайной величины.

Существуют другие показатели меры расположения значений случайной величины – **мода** и **медиана**.

Для непрерывной случайной величины **мода** определяется как то то ее значение, при котором плотность распределения вероятностей имеет локальный максимум. Случайная величина может иметь одну моду и называться в этом случае одномодальной и несколько мод и называться в этом случае многомодальной. **Модой** дискретной случайной величины называется ее наиболее вероятное значение. Медиана m определяется как значение случайной величины, которое разделяет множество возможных значений этой величины на две равновероятные области

$$P(x \le m) = P(x > m) = \frac{1}{2}.$$

В случае непрерывного на $(-\infty;+\infty)$ распределения медиана m удовлетворяет соотношению

$$\int_{-\infty}^{m} f(x)dx = \int_{m}^{\infty} f(x)dx = \frac{1}{2},$$

где $f(x)$ - плотность случайной величины X. Геометрически, медиана - это абсцисса точки, в которой площадь, ограниченная кривой плотности распределения, делится пополам,

Мерой рассеяния возможных значений случайной величины X около ее математического ожидания $M(X)$ является не только дисперсия, но и центральные моменты l-го порядка ($1 \le l < \infty$)

$$\mu_l = M\left(X - M\left(X\right)\right)^l .$$

$\mu_1 = 0$, μ_3 характеризует асимметрию распределения случайной величины. Наряду с центральными моментами l-го порядка случайной величины широко используются начальные моменты l-го порядка ($1 \le l < \infty$)

$$\nu_l = M\left(X\right)^l .$$

Самым известным начальным моментом является начальный момент первого порядка или математическое ожидание.

Пусть X, Y - независимые случайные величины, имеющие распределение Бернулли:

$x_i\,(y_i)$	0	1
p_i	q	p

Рассмотрим случайную величину Z, равную сумме случайных величин X, Y: $Z = X + Y$. Значениями случайной величины Z являются 0, 1, 2.

$$P(Z=0) = P(X=0, Y=0) = q^2,$$

$$P(Z=1) = P(X=1, Y=0) + P(X=0, Y=1) = 2pq,$$

$$P(Z=2) = P(X=1, Y=1) = p^2 .$$

По индукции легко устанавливается, что если $X_1, X_2, \ldots, X_n$ - независимые случайные величины, имеющие распределение Бернулли, то случайная величина $X = X_1 + X_2 + \ldots + X_n$ имеет биномиальное распределение:

$$P(X=k) = C_n^k p^k q^{n-k}, k = \overline{0,n} .$$

Найдем математическое ожидание и дисперсию случайной величины, имеющей распределение Пуассона с параметром λ: $P(X=k)=\dfrac{\lambda^k e^{-\lambda}}{k!}, k=\overline{0,\infty}$.

$$M(X)=\sum_{k=0}^{\infty}k\frac{\lambda^k e^{-\lambda}}{k!}=\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}e^{-\lambda}}{(k-1)!}=\lambda\sum_{m=0}^{\infty}\frac{\lambda^m e^{-\lambda}}{m!}=\lambda .$$

$$M\left(X^2\right)=\sum_{k=0}^{\infty}k^2\frac{\lambda^k e^{-\lambda}}{k!}=\lambda\sum_{k=1}^{\infty}\frac{k\lambda^{k-1}e^{-\lambda}}{(k-1)!}=\lambda\left(\sum_{k=1}^{\infty}\frac{(k-1)\lambda^{k-1}e^{-\lambda}}{(k-1)!}+\sum_{k=1}^{\infty}\frac{\lambda^{k-1}e^{-\lambda}}{(k-1)!}\right)=$$
$$=\lambda^2+\lambda.$$

$$\left(\sum_{k=1}^{\infty}\frac{(k-1)\lambda^{k-1}e^{-\lambda}}{(k-1)!}=\sum_{m=0}^{\infty}\frac{m\lambda^m e^{-\lambda}}{m!}=\lambda,\ \sum_{k=1}^{\infty}\frac{\lambda^{k-1}e^{-\lambda}}{(k-1)!}=\sum_{m=0}^{\infty}\frac{\lambda^m e^{-\lambda}}{m!}=1\right)$$
$$D(X)=\lambda^2+\lambda-\lambda^2=\lambda .$$

Найдем математическое ожидание и дисперсию случайной величины, имеющей геометрическое распределение $P(X=k)=q^{k-1}p,\ q=1-p,\ k=\overline{1,\infty}$.

Чтобы понять, что это за распределение, предположим, что стрелок стреляет до тех пор, пока не попадет в мишень. Вероятность попадания при одном выстреле равна p. Результаты выстрелов не зависят друг от друга. Тогда случайная величина, равная числу выстрелов, имеет геометрическое распределение.

$$M(X)=\sum_{k=1}^{\infty}kpq^{k-1}=p\sum_{k=1}^{\infty}kq^{k-1}=p\frac{d}{dq}\left(\sum_{k=1}^{\infty}q^k\right)=p\frac{d}{dq}\left(\frac{q}{1-q}\right)=$$
$$=p\left(\frac{1-q+q}{(1-q)^2}\right)=\frac{p}{(1-q)^2}=\frac{p}{p^2}=\frac{1}{p}.$$

$$M\left(X^2\right)=\sum_{k=1}^{\infty}k^2pq^{k-1}=p\sum_{k=1}^{\infty}k^2q^{k-1}=p\frac{d}{dq}\left(\sum_{k=1}^{\infty}kq^k\right)=p\frac{d}{dq}\left(q\sum_{k=1}^{\infty}kq^{k-1}\right)=$$

$$=p\frac{d}{dq}\left(\frac{q}{(1-q)^2}\right)=p\left(\frac{(1-q)^2+2q(1-q)}{(1-q)^4}\right)=\frac{p\left(1-q^2\right)}{(1-q)^4}=\frac{1+q}{p^2}.$$

.

$$D(X)=\frac{1+q}{p^2}-\frac{1}{p^2}=\frac{q}{p^2}.$$

Найдем функцию распределения вероятностей, математическое ожидание и дисперсию случайной величины, равномерно распределенной на отрезке $[a,b]$.

Как известно, плотность распределения вероятностей случайной величины, равномерно распределенной на $[a,b]$ равна

$$f(x)=\begin{cases}0, x\notin[a,b]\\ \dfrac{1}{b-a}, x\in[a,b]\end{cases}.$$

$$F(x)=\begin{cases}0, x<a,\\ \int\limits_a^x f(y)dy, x\in[a,b],\\ 1, x>b.\end{cases}=\begin{cases}0, x<a,\\ \dfrac{x-a}{b-a}, x\in[a,b],\\ 1, x>b.\end{cases}$$

$$M(X)=\int\limits_a^b\frac{x}{b-a}dx=\frac{b^2-a^2}{2(b-a)}=\frac{a+b}{2}.$$

$$M\left(X^2\right)=\int\limits_a^b\frac{x^2}{b-a}dx=\frac{b^3-a^3}{3(b-a)}=\frac{a^2+ab+b^2}{3},$$

$$D(X)=\frac{a^2+ab+b^2}{3}-\left(\frac{a+b}{2}\right)^2=\frac{(b-a)^2}{12}.$$

Найдем функцию распределения вероятностей, математическое ожидание и дисперсию случайной величины, имеющей показательное распределение с плотностью распределения вероятностей $f(x)=\begin{cases} 0, x<0, \\ \lambda e^{-\lambda x}, x \ge 0, \end{cases} \quad \lambda > 0.$

$$F(x)=\begin{cases} 0, x<0, \\ \int_0^x \lambda e^{-\lambda y} dy, x \ge 0. \end{cases} = \begin{cases} 0, x<0, \\ 1-e^{-\lambda x}, x \ge 0. \end{cases}$$

$$M(X)=\int_0^\infty \lambda x e^{-\lambda x} dx = -\int_0^\infty x de^{-\lambda x} = -xe^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx = -\frac{1}{\lambda} e^{-\lambda x}\Big|_0^\infty = \frac{1}{\lambda},$$

$$M(X^2)=\int_0^\infty \lambda x^2 e^{-\lambda x} dx = -\int_0^\infty x^2 de^{-\lambda x} = -x^2 e^{-\lambda x}\Big|_0^\infty + 2\int_0^\infty xe^{-\lambda x} dx = \frac{2}{\lambda^2},$$

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

Найдем математическое ожидание и дисперсию случайной величины, имеющей н**ормальный закон распределения** или **распределение Гаусса** с плотностью распределения вероятностей

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

$M(X)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty} xe^{-\frac{(x-a)^2}{2\sigma^2}} dx$, сделаем замену $\frac{x-a}{\sigma}=y, dx=dy$, тогда

$$M(X)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{+\infty} xe^{-\frac{(x-a)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} (a+\sigma y) e^{-\frac{y^2}{2}} dy = a\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} dy +$$

$$+\frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} ye^{-\frac{y^2}{2}} dy = a - \frac{\sigma}{\sqrt{2\pi}}\int_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} d\left(-\frac{y^2}{2}\right) = a - \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}\Big|_{-\infty}^{+\infty} = a,$$

$\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^{-\frac{y^2}{2}}dy = 1$ по свойству плотности распределения вероятностей случайной величины, а функция $\varphi(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$ является плотностью распределения вероятностей нормальной случайной величины, которая называется стандартной нормальной случайной величиной и имеет $M(X) = 0, D(X) = 1$.

$$M\left(X^2\right) = \frac{1}{\sqrt{2\pi}\sigma}\int\limits_{-\infty}^{+\infty} x^2 e^{-\frac{(x-a)^2}{2\sigma^2}} dx \text{ , сделаем замену } \frac{x-a}{\sigma} = y, dx = dy,$$

тогда

$$M\left(X^2\right) = \frac{1}{\sqrt{2\pi}\sigma}\int\limits_{-\infty}^{+\infty} x^2 e^{-\frac{(x-a)^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} (a+\sigma y)^2 e^{-\frac{y^2}{2}} dy = a^2\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} dy +$$

$$+\frac{2a\sigma}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} ye^{-\frac{y^2}{2}} dy + \frac{\sigma^2}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} y^2 e^{-\frac{y^2}{2}} dy = a^2 -$$

$$-\frac{2a\sigma}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} d\left(-\frac{y^2}{2}\right) - \frac{\sigma^2}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} y\ d\left(e^{-\frac{y^2}{2}}\right) = a^2 -$$

$$-\frac{\sigma}{\sqrt{2\pi}}e^{-\frac{y^2}{2}}\Big|_{-\infty}^{+\infty} - \frac{\sigma^2}{\sqrt{2\pi}}ye^{-\frac{y^2}{2}}\Big|_{-\infty}^{+\infty} + \frac{\sigma^2}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^{-\frac{y^2}{2}} dy = a^2 + \sigma^2.$$

$$D(X) = a^2 + \sigma^2 - a^2 = \sigma^2.$$

Рассмотрим нормально распределенную случайную величину X, $M(X) = a, D(X) = \sigma^2$ и найдем вероятность

$$P(b < X \le c).$$

Согласно свойству плотности распределения вероятностей случайной величины, эта вероятность равна определенному интегралу от плотности $f(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-a)^2}{2\sigma^2}}$. Интеграл вычислить достаточно сложно, поскольку он относится к, так называемым, «неберущимся» интегралам, которые нельзя вычислить стандартными методами. Поэтому гораздо проще провести некоторые преобразования и воспользоваться таблицей для функции Лапласа $\Phi(x)$ (приложение 2). Эти преобразования состоят в том, чтобы нормальную случайную величину X ($M(X)=a, D(X)=\sigma^2$) преобразовать в стандартную нормальную случайную величину ($M(X)=0, D(X)=1$). Используя свойства математического ожидания и дисперсии, легко получить, что $\frac{X-a}{\sigma}$ является стандартной нормальной случайной величиной. Таким образом

$$P(b < X \le c) = P\left(\frac{b-a}{\sigma} < \frac{X-a}{\sigma} \le \frac{c-a}{\sigma}\right) = \Phi\left(\frac{c-a}{\sigma}\right) - \Phi\left(\frac{b-a}{\sigma}\right).$$

Правило трех сигм.

$$P(|X-a| \le \delta) = P\left(\left|\frac{X-a}{\sigma}\right| \le \frac{\delta}{\sigma}\right) = 2\Phi\left(\frac{\delta}{\sigma}\right).$$

Полагая в последней формуле $\delta = 3\sigma$, и находя по таблице значение функции $\Phi(3)$, получим

$$P(|X-a| \le 3\sigma) = P\left(\left|\frac{X-a}{\sigma}\right| \le 3\right) = 2\Phi(3) = 0.997\,.$$

То есть, правило трех сигм состоит в том, что если случайная величина распределена по нормальному закону, то модуль ее отклонения от математического ожидания не превосходит утроенного стандартного отклонения (с вероятностью близкой к единице).

Задачи.

Задача 74. Случайная величина задана распределением Бернулли

x_i	**0**	**1**
p_i	q	p

Найти функцию распределения вероятностей, математическое ожидание и дисперсию.

Задача 75. Случайная величина X задана распределением

x_i	**1**	**3**	**5**	**8**
p_i	**0.15**	**0.2**	**0.25**	**0.4**

Найти $P(1 \le X < 8)$.

Задача 76. В условиях задачи 75 выбрать правильную функцию распределения вероятностей:

а) $F(x)=\begin{cases} 1, x \le 1 \\ 0.6,\ 1 < x \le 3 \\ 0.35,\ 3 < x \le 5 \\ 0.15,\ 5 < x < 8 \\ 0, x \ge 8 \end{cases}$ б) $F(x)=\begin{cases} 0, x \le 1 \\ 0.6,\ 1 < x \le 3 \\ 0.15,\ 3 < x \le 5 \\ 0.35,\ 5 < x < 8 \\ 0, x \ge 8 \end{cases}$

в) $F(x)=\begin{cases} 0, x \le 1 \\ 0.35,\ 1 < x \le 3 \\ 0.15,\ 3 < x \le 5 \\ 0.6,\ 5 < x < 8 \\ 1, x \ge 8 \end{cases}$ г) $F(x)=\begin{cases} 0, x < 1 \\ 0.15,\ 1 \le x < 3 \\ 0.35,\ 3 \le x < 5 \\ 0.6,\ 5 \le x < 8 \\ 1, x \ge 8 \end{cases}$.

Задача 77. Функция распределения вероятностей случайной величины X имеет вид

$$F(x)=\begin{cases} 0, x<2 \\ 0.21,\ 2\le x<6 \\ p,\ 6\le x<7 \\ 0.77,\ 7\le x<11 \\ 1, x\ge 11 \end{cases}.$$

Выберите возможное значение p:

а) 0.15;

б) 0.39;

в) 0;

г) 0.88.

Задача 78. Известен прогноз (в %) выполнения плана рабочим

x_i	**90**	**100**	**110**	**120**
p_i	**0.2**	**0.5**	**0.2**	**0.1**

За каждый процент перевыполнения плана рабочий получает премию в размере 100 рублей, а за каждый процент невыполнения плана штраф в размере 90 рублей. Найти распределение размера премии и ее математическое ожидание.

Задача 79. Найти математическое ожидание и дисперсию биномиальной случайной величины, заданной распределением

x_i	0	…	k	…	n
p_i	$C_n^0 p^0 q^n$	…	$C_n^k p^k q^{n-k}$	…	$C_n^n p^n q^0$

Задача 80. Значения случайной величины X

а) увеличили в 3 раза;

б) уменьшили в 3 раза;

в) увеличили на 4 единицы;

г) уменьшили на 5 единиц.

Как изменится $M(X)$?

Задача 81. Значения случайной величины X

а) увеличили в 3 раза;

б) уменьшили в 3 раза;

в) увеличили на 4 единицы;

г) уменьшили на 5 единиц.

Как изменится $D(X)$?

Задача 82. Функция распределения вероятностей случайной величины X имеет вид

$$F(x)=\begin{cases} 0, x<0 \\ \dfrac{x^2}{25},\ 0\le x<5. \\ 1, x\ge 5 \end{cases}$$

Найти плотность распределения вероятностей.

Задача 83. В условиях задачи 82 найти $P(1<X\le 4)$, $P(2<X\le 6)$.

Задача 84. Плотность распределения вероятностей случайной величины X имеет вид

$$f(x)=\begin{cases} 0, x<0 \\ cx^2,\ 0\le x<4. \\ 0, x\ge 4 \end{cases}$$

Найти значение c и функцию распределения вероятностей.

Задача 85. Случайная величина X задана распределением

x_i	**2**	**6**
p_i	**0.4**	**0.6**

Найти $M(X)$, $D(X)$.

Задача 86. Плотность распределения вероятностей случайной величины X имеет вид

$$f(x)=\begin{cases} 0, x<0 \\ \dfrac{x}{8}, \ 0\le x<4. \\ 0, x\ge 4 \end{cases}$$

Найти $M(X), D(X)$.

Задача 87. Функция распределения вероятностей случайной величины X имеет вид

$$F(x)=\begin{cases} 0, x<0 \\ \dfrac{x}{5}, \ 0\le x<5. \\ 1, x\ge 5 \end{cases}$$

Найти плотность распределения вероятностей, $M(X), D(X)$.

Задача 88. Найти $M(X), D(X), \sigma$ случайной величины X, равномерно распределенной на отрезке $[2,8]$.

Задача 89. Диаметр круга x измерен приближенно, причем $3\le x\le 5$. Рассматривая диаметр круга, как случайную величину, распределенную равномерно на отрезке $[3,5]$, найти математическое ожидание площади круга.

Задача 90. Случайная величина X имеет плотность распределения вероятностей

$$f(x)=\begin{cases} a\,cos^2 x, |x|\le \dfrac{\pi}{2}, \\ 0, \ |x|>\dfrac{\pi}{2}. \end{cases}$$

Найти a и $M(X)$.

8. Неравенство Чебышева, закон больших чисел, центральная предельная теорема

Неравенство Чебышева.

Для любого $\varepsilon > 0$

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

или

$$P(|X - M(X)| \le \varepsilon) \ge 1 - \frac{D(X)}{\varepsilon^2} .$$

Доказательство. Предположим, что случайная величина X является непрерывной и имеет плотность распределения вероятностей $f(x)$.

$$D(X) = \int_{-\infty}^{\infty} (x - M(X))^2 f(x) dx = \int_{|x-M(X)| \le \varepsilon} (x - M(X))^2 f(x) dx + \int_{|x-M(X)| > \varepsilon} (x - M(X))^2 f(x) dx \ge$$

$$\ge \int_{|x-M(X)| > \varepsilon} (x - M(X))^2 f(x) dx \ge \varepsilon^2 \int_{|x-M(X)| > \varepsilon} f(x) dx = \varepsilon^2 P(|x - M(X)| > \varepsilon) \Rightarrow$$

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

Так как

$P(|X - M(X)| > \varepsilon) + P(|X - M(X)| \le \varepsilon) = 1$, то

$$P(|X - M(X)| \le \varepsilon) \ge 1 - \frac{D(X)}{\varepsilon^2} .$$

Доказательство для дискретной случайной величины проводится аналогично с заменой соответствующей формулы для дисперсии.

Большую роль в теории вероятностей играют ряд теорем под общим названием закон больших чисел. Оказывается, что при определенных условиях суммарное поведение большого числа случайных величин утрачивает случайный характер и становится закономерным. Например, поведение одной молекулы непредсказуемо, а вот вычислить долю молекул, которые будут двигаться с заданной скоростью или в заданном объеме уже представляется возможным.

Условия, при которых поведение достаточно большого числа случайных величин утрачивает случайный характер, указаны, например, в теореме Чебышева.

Закон больших чисел в форме Чебышева (теорема Чебышева).

Если случайные величины $X_1, X_2, ..., X_n$ - независимы, имеют математические ожидания и дисперсии, ограниченные одной и той же постоянной, то для любого $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} M(X_i)}{n}\right| \le \varepsilon\right) = 1$$

или $\frac{\sum_{i=1}^{n} X_i}{n}$ по вероятности сходится к $\frac{\sum_{i=1}^{n} M(X_i)}{n}$.

Доказательство. Рассмотрим случайную величину $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. Как легко определить,

$$M(\overline{X}) = \frac{\sum_{i=1}^{n} M(X_i)}{n},\ D(\overline{X}) = \frac{\sum_{i=1}^{n} D(X_i)}{n^2}.$$

Применим к случайной величине $\overline{X}$ неравенство Чебышева.

$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n} M(X_i)}{n}\right| \le \varepsilon\right) \ge 1 - \frac{\sum_{i=1}^{n} D(X_i)}{\varepsilon^2 n^2} \ge 1 - \frac{c}{\varepsilon^2 n}.$$

По условию теоремы $D(X_i)\le c, i=\overline{1,n}$ (равномерная ограниченность дисперсий), поэтому, переходя к пределам, получим

$$\lim_{n\to\infty} P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n}-\frac{\sum_{i=1}^{n} M(X_i)}{n}\right|\le\varepsilon\right)\ge 1.$$

А поскольку вероятность события по определению не может быть больше единицы, то получаем

$$\lim_{n\to\infty} P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n}-\frac{\sum_{i=1}^{n} M(X_i)}{n}\right|\le\varepsilon\right)=1.$$

Если все случайные величины имеют одно и то же математическое ожидание, то есть $M(X_i)=a, i=\overline{1,n}$, то $\lim_{n\to\infty} P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n}-a\right|\le\varepsilon\right)\ge 1.$

Закон больших чисел в форме Бернулли (теорема Бернулли).

Пусть k - число появлений события A в n испытаниях. Вероятность появления события A в одном испытании равна p. Тогда для любого $\varepsilon>0$

$$\lim_{n\to\infty} P\left(\left|\frac{k}{n}-p\right|\le\varepsilon\right)=1$$

или $\frac{k}{n}$ по вероятности сходится к p.

Доказательство. Число появлений события A в n испытаниях можно представить в виде суммы независимых бернуллиевских случайных величин

$$k = X_1 + X_2 + \ldots + X_n,\ X_i = \begin{cases} 0, q \\ 1, p \end{cases}, i = \overline{1,n}.$$

Тогда $M(X_i) = p,\ D(X_i) = pq, i = \overline{1,n}$. Поскольку $p + q = 1$, то $D(X_i) = pq \le \frac{1}{4}$, а значит можно применить теорему Чебышева. В результате получим

$$\lim_{n\to\infty} P\left(\left|\frac{k}{n} - p\right| \le \varepsilon\right) = 1.$$

Многие задачи теории вероятностей занимаются изучением поведения сумм случайных величин при определенных условиях. Существенную роль для решения этих задач играет центральная предельная теорема. Приведем эту теорему при условии, что все случайные величины имеют одинаковое распределение.

Центральная предельная теорема.

Пусть $X_1, X_2, \ldots, X_n$ - независимые, одинаково распределенные случайные величины, $M(X_i) = a, D(X_i) = \sigma^2$, $i = \overline{1,n}$, тогда

$$\lim_{n\to\infty} P\left(\frac{\sum_{i=1}^{n} X_i - na}{\sigma\sqrt{n}} \le x\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy.$$

Рассмотрим партию из n изделий, в которой оказалось m бракованных. Неизвестная вероятность p появления бракованного изделия оценивается величиной $\frac{m}{n}$. Докажем, что $\frac{m}{n}$ по вероятности сходится к p или, что

$$\lim_{n\to\infty} P\left(\left|\frac{m}{n} - p\right| \le \varepsilon\right) = 1.$$

Число бракованных изделий m можно представить в виде суммы независимых случайных величин:

$$m = X_1 + X_2 + \ldots + X_n,\ X_i = \begin{cases} 1, p \\ 0, q \end{cases},\ p + q = 1, i = \overline{1,n}.$$

$$M\left(\frac{m}{n}\right) = \frac{1}{n} \cdot \sum_{i=1}^{n} M(X_i) = M(X_1) = p,$$

Вычислим дисперсию $D\left(\frac{m}{n}\right)$ и воспользуемся неравенством Чебышева:

$$D\left(\frac{m}{n}\right) = \frac{1}{n^2} D(m) = \frac{1}{n^2} \sum_{i=1}^{n} D(X_i) = \frac{1}{n} D(X_1) = \frac{pq}{n},$$

$$P\left(\left|\frac{m}{n} - p\right| \geq \varepsilon\right) \leq \frac{D\left(\frac{m}{n}\right)}{\varepsilon^2} = \frac{pq}{n\varepsilon^2} \to 0, n \to \infty.$$

Переходя к пределам, получим $\lim\limits_{n\to\infty} P\left(\left|\frac{m}{n} - p\right| < \varepsilon\right) = 1.$

Задачи.

Задача 91. Вероятность появления события A в каждом из 100 испытаний равна 0.8. Оцените с помощью неравенства Чебышева вероятность появления события A в 100 испытаниях от 70 до 90 раз.

Задача 92. Вероятность появления события A в каждом из 50 испытаний равна 0.7. Оцените с помощью неравенства Чебышева вероятность появления события A в 50 испытаниях от 30 до 40 раз.

Задача 93. Вероятность появления события A в каждом из 200 испытаний равна 0.6. Оцените с помощью неравенства Чебышева вероятность появления события A в 200 испытаниях от 100 до 140 раз.

Задача 94. Пусть $X_1, X_2, ..., X_n$ - независимые случайные величины, имеющие одинаковые математические ожидания $M(X_1)=3$ и дисперсии $D(X_1)=5$, $n=100$. Оцените с помощью неравенства Чебышева вероятность нахождения суммы случайных величин в пределах отрезка $[250,350]$.

Задача 95. Пусть $X_1, X_2, ..., X_n$ - независимые случайные величины, имеющие одинаковые математические ожидания $M(X_1)=4$ и дисперсии $D(X_1)=8$, $n=200$. Оцените с помощью неравенства Чебышева вероятность нахождения суммы случайных величин в пределах отрезка $[700,900]$.

Задача 96. Пусть $X_1, X_2, ..., X_n$ - независимые случайные величины, равномерно распределенные на $[0,\theta]$. Доказать, что для любого $\varepsilon > 0$

$$\lim_{n\to\infty} P\left(\left|X_{(n)} - \theta\right| \le \varepsilon\right) = 1,$$

где $X_{(n)} = \max(X_1, X_2, ..., X_n)$.

Задача 97. Пусть X - случайная величина, имеющая распределение Пуассона с параметром λ. Доказать, что $\lim\limits_{\lambda\to\infty} P\left(\left|X-\lambda\right| \le \lambda\varepsilon\right) = 1$.

Задача 98. Показать, что $D(X)=0$ тогда и только тогда, когда X является константой с вероятностью 1.

9. Системы случайных величин и их числовые характеристики

Пусть X и Y - две случайные величины, значения которых могут быть получены в результате одного эксперимента. Совместной функцией распределения вероятностей этих случайных величин называется функция двух переменных

$$F(x,y) = \mathrm{P}(X \le x, Y \le y).$$

Совместной плотностью распределения вероятностей этих случайных величин называется смешанная частная производная

$$f(x,y) = \frac{\partial^2 F(x,y)}{\partial x \partial y}.$$

В реальных ситуациях информация о законах распределения систем случайных величин, чаще всего, бывает недоступна, зато представляется возможным получить такие числовые характеристики поведения этой системы, как коэффициент ковариации и коэффициент корреляции.

Коэффициентом ковариации случайных величин X и Y называется величина

$$\operatorname{cov}(X,Y) = M[(X - M(X))(Y - M(Y))].$$

Преобразуем последнюю формулу

$$\begin{aligned}\operatorname{cov}(X,Y) &= M[(X - M(X))(Y - M(Y))] = M(XY) - M(X)M(Y) - \\ &- M(X)M(Y) + M(X)M(Y) = M(XY) - M(X)M(Y).\end{aligned}$$

Если $\operatorname{cov}(X,Y) = 0$, то случайные величины X, Y называются некоррелированными.

Поскольку $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$, а $\operatorname{cov}(X,X) = D(X)$, $\operatorname{cov}(Y,Y) = D(Y)$, то ковариационная матрица двух случайных величин представляет собой матрицу размера 2×2, симметричную относительно главной диагонали. На главной диагонали ковариационной матрицы расположены дисперсии случайных величин (которые по свойству всегда неотрицательны).

Для системы n случайных величин ковариационная матрица имеет размеры $n \times n$ и аналогичные свойства.

Покажем, что **из независимости случайных величин следует их некоррелированность.**

Если случайные величины X, Y - независимы, то по свойству математического ожидания

$$M(XY) = M(X)M(Y)..$$

В этом случае

$$\operatorname{cov}(X,Y) = M(XY) - M(X)M(Y) = 0,$$

а значит случайные величины X, Y являются некоррелированными.

Обратное, вообще говоря, неверно.

Покажем на примере, что **из некоррелированности случайных величин не следует их независимость.**

Рассмотрим случайную величину ξ, имеющую следующее распределение:

ξ_i	0	$\frac{\pi}{2}$	π
p_i	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

и две случайные величины - $X = \sin\xi$, $Y = \cos\xi$.

Тогда

X_i	0	1
p_i	$\frac{2}{3}$	$\frac{1}{3}$

Y_i	-1	0	1
p_i	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

Покажем, что случайные величины X, Y - некоррелированны и для этого вычислим $M(XY)=M\left(\frac{1}{2}sin\,2\xi\right)=0$ и $M(X)=\frac{1}{3}$, $M(Y)=0$.

Таким образом, $\operatorname{cov}(X,Y)=0$ и X, Y являются некоррелированными.

Однако они не являются независимыми, более того, они функционально зависимы - $X^2+Y^2=\sin^2\xi+\cos^2\xi=1$.

Коэффициентом корреляции случайных величин X и Y называется величина

$$r_{x,y}=\frac{cov(X,Y)}{\sqrt{D(X)D(Y)}}.$$

Поскольку $\operatorname{cov}(X,Y)=\operatorname{cov}(Y,X)$, а $\operatorname{cov}(X,X)=D(X)$, $\operatorname{cov}(Y,Y)=D(Y)$, то корреляционная матрица двух случайных величин представляет собой матрицу размера 2×2, симметричную относительно главной диагонали, на которой расположены единицы. Остальные ее элементы не превосходят по модулю единицы. Для системы n случайных величин корреляционная матрица имеет размеры $n\times n$ и аналогичные свойства.

Покажем, что $-1\le r_{x,y}\le 1$.

Обозначим $\sqrt{D(X)}=\sigma_x$, $\sqrt{D(Y)}=\sigma_y$ и рассмотрим две случайные величины:

$$Z_1=\sigma_y X-\sigma_x Y,\ Z_2=\sigma_y X+\sigma_x Y.$$

Найдем их математические ожидания:

$$M(Z_1)=\sigma_y M(X)-\sigma_x M(Y),\ M(Z_2)=\sigma_y M(X)+\sigma_x M(Y)$$

и математические ожидания Z_1^2, Z_2^2:

$$M\left(Z_1^2\right)=\sigma_y^2 M\left(X^2\right)-2\sigma_x\sigma_y M(XY)+\sigma_x^2 M\left(Y^2\right),$$
$$M\left(Z_2^2\right)=\sigma_y^2 M\left(X^2\right)+2\sigma_x\sigma_y M(XY)+\sigma_x^2 M\left(Y^2\right).$$

Вычислим дисперсии случайных величин Z_1, Z_2:

$$D(Z_1)=M\left(Z_1^2\right)-(M(Z_1))^2=\sigma_y^2 M\left(X^2\right)-2\sigma_x\sigma_y M(XY)+\sigma_x^2 M\left(Y^2\right)-$$
$$-\sigma_y^2 M(X)^2+2\sigma_x\sigma_y M(X)M(Y)-\sigma_x^2 M(Y)^2=2\sigma_y^2\sigma_x^2-2\sigma_x\sigma_y \operatorname{cov}(X,Y)\geq 0,$$

$$cov(X,Y)\leq\sigma_x\sigma_y,\ r_{x,y}=\frac{cov(X,Y)}{\sigma_x\sigma_y}\leq 1.$$

$$D(Z_2)=M\left(Z_2^2\right)-(M(Z_2))^2=\sigma_y^2 M\left(X^2\right)+2\sigma_x\sigma_y M(XY)+\sigma_x^2 M\left(Y^2\right)-$$
$$-\sigma_y^2 M(X)^2-2\sigma_x\sigma_y M(X)M(Y)-\sigma_x^2 M(Y)^2=2\sigma_y^2\sigma_x^2+2\sigma_x\sigma_y \operatorname{cov}(X,Y)\geq 0,$$
$$cov(X,Y)\geq -\sigma_x\sigma_y,\ r_{x,y}=\frac{cov(X,Y)}{\sigma_x\sigma_y}\geq -1.$$

Следовательно

$$-1\leq r_{x,y}\leq 1.$$

Покажем, что если $r_{x,y}=\pm 1$, то существуют такие a,b, что соответственно $Y=\pm aX+b$.

Если $r_{x,y}=1$, то отсюда следует, что $D(Z_1)=0$, то есть $Z_1=c=const$ или

$$Z_1=\sigma_y X-\sigma_x Y=c\Rightarrow Y=\frac{\sigma_y}{\sigma_x}X-\frac{c}{\sigma_x}=aX+b.$$

Если $r_{x,y}=-1$, то отсюда следует, что $D(Z_2)=0$, то есть

$Z_2=d=const$ или $Z_2=\sigma_y X+\sigma_x Y=d\Rightarrow Y=-\dfrac{\sigma_y}{\sigma_x}X-\dfrac{d}{\sigma_x}=-aX+b.$

Если случайный вектор (X, Y) имеет невырожденное $(|\rho| \neq 1)$ нормальное распределение, то его плотность $f(x, y)$ равна (см. рис.2)

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times$$

$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_1)^2}{\sigma_1^2} - 2\rho\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right]\right\}$$

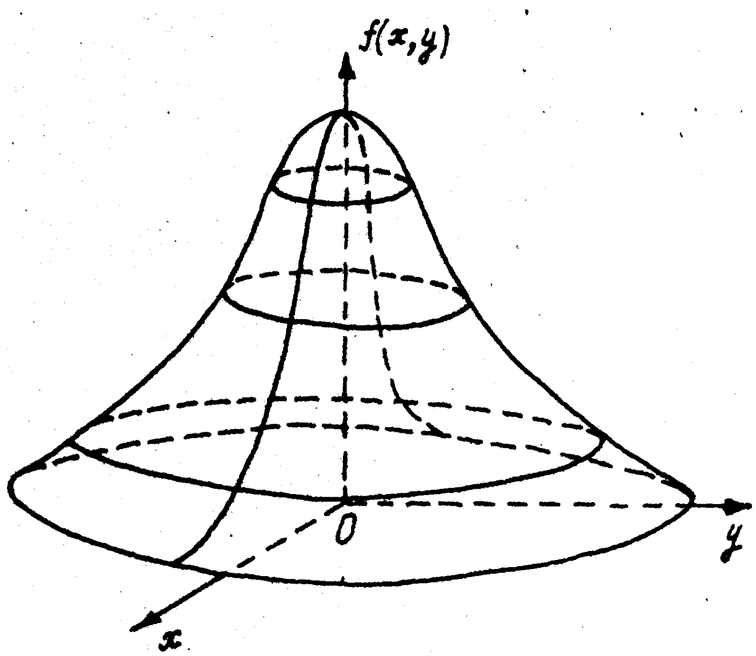

Рис. 2. График плотности двумерного нормального распределения.

Смысл величин $m_1, m_2, \sigma_1, \sigma_2$ и ρ следующий:

$$m_1 = MX\ , \quad \sigma_1^2 = DX\ , \quad m_2 = MY\ , \quad \sigma_2^2 = DY$$

$$M(XY) = M(YX) = \sigma_1\sigma_2\rho + m_1 m_2$$

$$\rho = \frac{M(X-m_1)(Y-m_2)}{\sqrt{DX \cdot DY}}$$ - коэффициент корреляции X и Y.

Для нормального закона справедливо следующее правило: если компоненты двумерного нормального вектора некоррелированы (т.е. $\rho = 0$), то они и независимы.

Задачи.

Задача 99а. Даны распределения дискретных случайных величин X и Y.

x_i	**1**	**2**
p_i	**0.2**	**0.8**

y_i	**3**	**5**
p_i	**0.4**	**0.6**

Найти распределение $X+Y$.

Задача 99б. Даны распределения дискретных случайных величин X и Y.

x_i	**2**	**3**
p_i	**0.3**	**0.7**

y_i	**1**	**4**
p_i	**0.9**	**0.1**

Найти распределение $X+Y$ и XY.

Задача 100. Даны распределения дискретных случайных величин X и Y.

x_i	**1**	**3**
p_i	**0.6**	**0.4**

y_i	**5**	**6**
p_i	**0.7**	**0.3**

Найти распределение $X+Y$ и XY.

Задача 101. Дано совместное распределение случайных величин X и Y.

$Y \backslash X$	**1**	**3**
1	**0.15**	**0.1**
3	**0.25**	**0.05**
6	**0.05**	**0.4**

Найти условное распределение $P(X|Y=6)$.

Задача 102. В условиях задачи 101 найти $P(1<X\le 3)$.

Задача 103. Дано совместное распределение случайных величин X и Y.

$Y \backslash X$	**2**	**5**
1	**0.05**	**0.15**
2	**0.2**	**0.15**
3	**0.25**	**0.2**

Найти $P(1<Y\le 3)$.

Задача 104. В условиях задачи 103 найти $P(Y|X=5), P(Y|X=2), P(X|Y=2), P(1\le Y<3)$.

Задача 105. Какая из матриц может быть ковариационной матрицей двух случайных величин?

а) $\begin{pmatrix} -2 & 3 \\ 3 & 4 \end{pmatrix}$ б) $\begin{pmatrix} 2 & -3 \\ 3 & 4 \end{pmatrix}$ в) $\begin{pmatrix} -1 & 3 \\ 3 & -3 \end{pmatrix}$ г) $\begin{pmatrix} 2 & -3 \\ -3 & 4 \end{pmatrix}$

Задача 106. Какая из матриц может быть ковариационной матрицей двух случайных величин?

а) $\begin{pmatrix} -1 & -3 \\ 3 & 5 \end{pmatrix}$ б) $\begin{pmatrix} 1 & -3 \\ 4 & 2 \end{pmatrix}$ в) $\begin{pmatrix} 1 & 3 \\ 3 & 3 \end{pmatrix}$ г) $\begin{pmatrix} 2 & 3 \\ -3 & 4 \end{pmatrix}$

Задача 107. Какая из матриц может быть корреляционной матрицей двух случайных величин?

а) $\begin{pmatrix} 1 & 07 \\ 0.8 & 1 \end{pmatrix}$ б) $\begin{pmatrix} 1 & -0.3 \\ -0.3 & 1 \end{pmatrix}$ в) $\begin{pmatrix} 1 & 0.9 \\ 0.9 & 3 \end{pmatrix}$ г) $\begin{pmatrix} -1 & 0 \\ -3 & 1 \end{pmatrix}$

Задача 108. Какая из матриц может быть корреляционной матрицей двух случайных величин?

а) $\begin{pmatrix} 2 & 07 \\ 0.7 & 1 \end{pmatrix}$ б) $\begin{pmatrix} 0 & -0.3 \\ -0.3 & 1 \end{pmatrix}$ в) $\begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}$ г) $\begin{pmatrix} 1 & 0 \\ 0.8 & 1 \end{pmatrix}$

Задача 109. Плотность распределения вероятностей случайной величины X равна $f(x)$, а плотность распределения вероятностей случайной величины Y равна $g(y)$. X и Y независимы. Найти функцию распределения случайной величины $Z = X + Y$.

Задача 110. Функция распределения вероятностей случайной величины X равна $F(x)$. Найти функцию распределения вероятностей случайной величины $Y = aX + b,\ a,b$ - const.

10. Выборка. Вероятностная модель эксперимента

Математическая статистика - это раздел математики, в котором изучаются методы сбора, систематизации и обработки результатов наблюдений массовых случайных явлений с целью выявления существующих закономерностей. Выводы о закономерностях, которым подчиняются явления, изучаемые методами математической статистики, всегда основываются на ограниченном, выборочном числе наблюдений. Поэтому естественно предположить, что эти выводы при большем числе наблюдений могут оказаться иными. Чтобы быть в состояния высказать более определенное суждение об изучаемом явлении, математическая статистика опирается на теорию вероятностей.

Теория вероятностей занимается формально-логическим изучением закономерностей случайных явлений и имеет дело с абстрактными описаниями действительности или с математическими моделями случайных явлений. Математическая статистика, в отличие от теории вероятностей, имеет дело с результатами наблюдений над случайными явлениями.

Оценив неизвестные величины или зависимости между ними по данным наблюдений, исследователь выдвигает ряд гипотез, предположений о том, что рассматриваемое явление можно описать той иди иной вероятностной теоретической моделью. Далее, используя методы математической статистики, можно дать ответ на вопрос, какую из гипотез или моделей следует принять. Именно эта модель и есть закономерность изучаемого явления. Таков типичный путь исследования на основе аппарата математической статистики.

Математическая статистика, опираясь на вероятностные модели при изучении закономерностей массовых, случайных явлений, в свою очередь, влияет на развитие теории вероятностей. Окружающий нас мир многообразен, и задачи, возникающие при изучении тех или иных случайных явлений, при обработке результатов требуют разработки новых вероятностных моделей. Таким образом, математическая статистика и теория вероятностей - две неразрывно связанные науки.

Вероятностная модель ставит в соответствие результатам наблюдений

$$x_1, x_2, \ldots, x_n \tag{1}$$

последовательность случайных величин

$$X_1, X_2, \ldots, X_n \ . \tag{2}$$

Предполагается, что случайные величины $X_1, X_2, \ldots, X_n$ независимы и имеют одно и то же распределение с функцией распределения $F(x)$. Полагают, что наблюдения (1) являются значениями величин (2) при осуществлении вероятностного эксперимента. Несмотря на различие объектов (1) и (2), в математической статистике принято называть и то и другое ***выборкой из генеральной совокупности.***

Количество наблюдений n называется объемом выборки.

Статистическим закономерностям числовых данных реального эксперимента (1) отвечают вероятностные утверждения для случайных величин (2).

Рассмотрим эксперимент, состоящий в n-кратном бросании игральной кости и построим вероятностную модель эксперимента.

Результаты наблюдений имеют вид

$$x_1, x_2, \ldots, x_n \ ,$$

где $x_i = \overline{1,6}, i = \overline{1,n}$.

Вероятностную модель представим следующим образом:

$$X_1, X_2, \ldots, X_n ,$$

где $X_i, i = \overline{1,n}$ - независимые случайные величины, имеющие дискретное распределение и принимающие значения 1,2,3,4,5,6 с равными вероятностями $\frac{1}{6}$.

Рассмотрим урну с N шарами, занумерованными числами от 0 до $N-1$. Опыт заключается в n - кратном извлечении шара «наугад» с последующим его возвращением в урну. Построим вероятностную модель опыта.

Все возможные исходы опыта могут быть представлены множеством из N^n последовательностей

$$x_1, x_2, ..., x_n \ ,$$

где $x_i = \overline{0, N-1}, i = \overline{1, n}$.

Вероятностная модель опыта будет иметь вид

$$X_1, X_2, ..., X_n \ ,$$

где $X_i, i = \overline{1, n}$ - независимые случайные величины, имеющие дискретное распределение и принимающие значения $0, 1, ..., N-1$ с равными вероятностями $\frac{1}{N}$.

Если число наблюдений велико, то необходима группировка данных, которая выполняется следующим образом. Выделяются все несовпадающие между собой значения выборки, для каждого из таких значений определяется общее число повторений. В результате получаются два ряда чисел. Первый ряд содержит все несовпадающие значения из выборок, расположенные порядке возрастания. Числа второго ряда показывают количество повторений каждого из этих значений в выборке.

Количество повторений соответствующего элемента x_i в выборке называют **частотой** этого элемента и обозначают n_i. **Относительной частотой** элемента x_i называют число $\frac{n_i}{n}$, где n - объем выборки.

$$\sum_i n_i = n, \ \sum_i \frac{n_i}{n} = 1.$$

Пример. Подсчет числа деревьев определенного вида на 60 делянках дал следующие результаты

15	13	15	14	16	14	12	14	14	15	13	11	13	15	14
13	15	14	12	14	14	15	15	12	12	13	15	16	14	13
14	13	15	14	13	14	15	14	15	14	14	13	14	15	13
14	13	13	11	12	14	13	12	11	15	15	13	13	13	14

Провести группировку данных.

Решение. Группировку осуществим следующим образом. В первую строку таблицы впишем все возможные значения из выборки в порядке возрастания. Просматривая подряд заданную выборку и, считая повторения каждого элемента, заполняем вторую строку. Отсюда сгруппированный ряд имеет следующий вид:

x_i	11	12	13	14	15	16
n_i	3	6	16	19	14	2

Такая форма записи сгруппированного ряда называется **статистическим рядом**.

Если выборка имеет большое число различных элементов, то группировка заключается в том, что диапазон выборки делится на определенное число частей (интервалов), а затем подсчитывается число значений выборки, попавшее в каждый из интервалов. При такой группировке возникает вопрос о длине интервалов Δx и расположении границ. Решение этого вопроса рассмотрим на примере.

Пример. Получены следующие данные измерения толщины кремниевой подложки в микронах

5,39	**5,43**	**5,49**	**5,42**	**5,45**	**5,44**	**5,64**

5,42	**5,52**	**5,35**	**5,45**	**5,37**	**5,54**	**5,18**
5,38	**5,45**	**5,48**	**5,32**	**5,48**	**5,66**	**5,46**
5,47	**5,26**	**5,26**	**5,44**	**5,46**	**5,43**	**5,61**
5,51	**5,33**	**5,55**	**5,58**	**5,51**	**5,42**	**5,47**
5,30	**5,43**	**5,46**	**5,50**	**5,29**	**5,44**	**5,36**
5,40	**5,50**	**5,41**	**5,56**	**5,42**	**5,34**	**5,57**
5,40	**5,44**	**5,55**	**5,44**	**5,69**	**5,33**	**5,39**
5,28	**5,47**	**5,37**	**5,50**	**5,60**	**5,41**	**5,58**
5,43	**5,52**	**5,45**	**5,37**	**5,45**	**5,54**	**5,44**
5,46	**5,48**	**5,54**	**5,57**	**5,38**	**5,40**	
5,53	**5,34**	**5,32**	**5,50**	**5,46**	**5,23**	
5,55	**5,36**	**5,52**	**5,44**	**5,52**	**5,45**	
5,47	**5,59**	**5,39**	**5,28**	**5,43**	**5,47**	
5,24	**5,45**	**5,62**	**5,31**	**5,41**	**5,40**	

Произвести группировку данных.

Решение. Самое простое предложение - разделить разность между наибольшим $(x_{\max} = x_{(n)} = 5{,}69)$ и наименьшим $(x_{\min} = x_{(1)} = 5{,}18)$ значениями на принятое число частей (например, $k = 11$), после чего границы интервалов находятся сразу. В нашем примере 5,69 - 5,18 = 0,51 и $\Delta x = \frac{0{,}51}{11} = 0{,}04636...$ Так как здесь $x_{\max} - x_{\min}$ не делится без остатка на k (в пределах принятой точности), то производится округление длины интервала в сторону увеличения. В противном случае общая длина интервалов уменьшилась бы так, что крайние значения выборки не попали бы в него.

Если какое-либо значение попадает на границу интервалов, то его относят к правому интервалу. Так, если границы интервалов: ...; 5,30; 5,35; 5,40; 5,45;…, то значение 5,35 относят к интервалу 5,35 - 5,40. Иными словами этот интервал содержит значения x удовлетворяющие условию $5{,}35 \le x < 5{,}40$.

Принимая $\Delta x = 0{,}05$ и $k = 11$, получим, что весь диапазон выборки имеет длину $0{,}05 \cdot 11 = 0{,}55$.

Следовательно, в качестве границ диапазона значений можно принять либо 5,175 - 5,725 (расширив диапазон в сторону больших значений, случай а))г либо 5,145 - 5,695 (расширив его в сторону меньших значений, случай б)), либо какие-нибудь промежуточные значения, например 5,165 - 5,715. Чтобы пояснить, как отразится на виде распределения тот или иной выбор границ диапазона выборки, приведем ниже результаты группировки для двух рассмотренных случаев.

Границы интервалов	Метки	Частота
Случай а)		
5,175 - 5,225	𝍷	1
5,225 - 5,275	𝍷𝍷𝍷𝍷	4
5,275 - 5,325	𝍸 𝍷𝍷	7
5,325 - 5,375	𝍸𝍸 𝍷	11
3,375 - 5,425	𝍸𝍸𝍸 𝍷	16
5,425 - 5,475	𝍸𝍸𝍸𝍸𝍸𝍸	30
5,475 - 5,525	𝍸𝍸 𝍷𝍷𝍷𝍷	14
5,525 - 5,575	𝍸 𝍷𝍷𝍷	8
5,575 - 5,625	𝍸 𝍷	6
5,625 - 5,675	𝍷𝍷	2
5,675 - 5,725	𝍷	1

<table>
<tr><th colspan="3">Случай б)</th></tr>
<tr><td>5,145 - 5,195</td><td>|</td><td>1</td></tr>
<tr><td>5,195 - 5,245</td><td>||</td><td>2</td></tr>
<tr><td>5,245 - 5,295</td><td>𝍸</td><td>5</td></tr>
<tr><td>5,295 - 5,345</td><td>𝍸 |||</td><td>8</td></tr>
<tr><td>5,345 - 5,395</td><td>𝍸𝍸 ||</td><td>12</td></tr>
<tr><td>5,395 - 5,445</td><td>𝍸𝍸𝍸𝍸 |||</td><td>23</td></tr>
<tr><td>5,445 - 5,495</td><td>𝍸𝍸𝍸𝍸 ||</td><td>22</td></tr>
<tr><td>5,495 - 5,545</td><td>𝍸𝍸 ||||</td><td>14</td></tr>
<tr><td>5,545 - 5,595</td><td>𝍸 ||</td><td>7</td></tr>
<tr><td>5,595 - 5,645</td><td>||||</td><td>4</td></tr>
<tr><td>5,645 - 5,695</td><td>||</td><td>2</td></tr>
</table>

Несмотря на видимую несхожесть, полученные ряды отражают одно и то же фактическое распределение.

Задачи.

Задача 111. Известно, что вероятность появления бракованного изделия равна p. В партии из n изделий обнаружено m бракованных. Наугад извлекается одно изделие. Построить вероятностную модель эксперимента.

Задача 112. Бросаются две монеты. Построить вероятностную модель эксперимента.

Задача 113. n раз бросается монета. Построить вероятностную модель эксперимента.

Задача 114. Объем выборки равен $n = 50$. Найти x.

x_i	11	12	13	14	15	16
n_i	3	6	x	19	14	2

Задача 115. Объем выборки равен $n = 100$. Найти x.

x_i	1	2	5	7	8	10
n_i	12	18	12	10	x	20

Задача 116. Объем выборки равен $n = 100$. Найти относительную частоту x и частоту n_i элемента выборки 4.

x_i	-1	0	3	4	9	11
$\frac{n_i}{n}$	0.05	0.15	0.3	x	0.2	0.1

Задача 117. Объем выборки равен $n = 50$. Найти относительную частоту x и частоту n_i элемента выборки 2.

x_i	1	2	3	4	5	6
$\frac{n_i}{n}$	0.1	x	0.05	0.15	0.2	0.2

11. Эмпирическая функция распределения, гистограмма

Произвольная случайная величина X характеризуется своей функцией распределения вероятностей $F(x)$. Если эта функция неизвестна, но известна выборка (1), числовые данные которой являются значениями случайной величины X, то возможно построить ***эмпирическую функцию распределения вероятностей*** $F_n(x)$, которая служит оценкой теоретической функции распределения вероятностей $F(x)$.

Если обозначить через $\mu_n(x)$ число тех значений $x_1, x_2, ..., x_n$, которые меньше или равны x, то

$$F_n(x) = \frac{\mu_n(x)}{n}. \qquad (3)$$

Таким образом, функция $F_n(x)$, так, как она определена, есть функция распределения дискретной случайной величины, принимающей значения $x_1, x_2, ..., x_n$ с равными вероятностями $\frac{1}{n}$. Если среди значений $x_1, x_2, ..., x_n$ k одинаковых, например, $x_1 = x_2 = ... = x_k$, то вероятность значения x_1 равна $\frac{k}{n}$.

Другими словами, эмпирическая функция распределения $F_n(x)$ есть частота события $X \le x$ в серии из n независимых измерений случайной величины X.

Следовательно, из теоремы Бернулли следует, что эмпирическая функция распределения сходится по вероятности к функции распределения генеральной совокупности, когда объем выборки неограниченно возрастает. Таким образом, $F_n(x)$ является приближенным значением функции распределения и обладает следующими свойствами:

1) значения $F_n(x)$ принадлежат отрезку [0,1];

2) $F_n(x)$ - неубывающая функция;

3) если $x_{\max}$ - наибольший элемент выборки, а $x_{\min}$ - наименьший, то

$$F_n(x) = \begin{cases} 0 & \text{при } x \le x_{\min} \\ 1 & \text{при } x > x_{\max} \end{cases}$$

График эмпирической функции распределения представляет собой ступенчатую линию со скачками в точках, определяемых элементами выборки.

Эмпирическая функция распределения $F_n(x)$ может быть записана следующим образом

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{x_i \le x\}},$$

где

$$I_{\{x_i \le x\}} = \begin{cases} 1, x_i \le x \\ 0, x_i > x \end{cases}.$$

Переходя к вероятностной модели, получим, что эмпирическая функция распределения $F_n(x)$ может быть записана следующим образом

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le x\}}, \qquad (4)$$

где

$$I_{\{X_i \le x\}} = \begin{cases} 1, P(X_i \le x) \\ 0, P(X_i > x) \end{cases},$$

а $X_1, X_2, ..., X_n$ - независимые одинаково распределенные случайные величины, принимающие значения $x_1, x_2, ..., x_n$ и имеющие теоретическую функцию распределения вероятностей $F(x)$.

Представление (4) позволяет видеть, что при каждом фиксированном x эмпирическая функция распределения $F_n(x)$ является случайной величиной.

Пусть $F(x)$ - теоретическая функция распределения случайной величины X. Найдем

$$P\left(F_n(x) = \frac{m}{n}\right),\ 0 \le m \le n.$$

Поскольку $\mu_n(x)$ представимо в виде суммы независимых одинаково распределенных случайных величин

$$\mu_n(x) = I_{\{X_1 \le x\}} + I_{\{X_2 \le x\}} + ... + I_{\{X_n \le x\}},$$

таких, что

$$I_{\{X_i \le x\}} = \begin{cases} 1, P(X_i \le x) \\ 0, P(X_i > x) \end{cases}, i = \overline{1,n},$$

то событие $F_n(x) = \frac{m}{n}$ состоит в том, что m случайных величин $I_{\{X_i \le x\}}, i = \overline{1,n}$ равны единице, а $n - m$ случайных величин $I_{\{X_i \le x\}}, i = \overline{1,n}$ равны нулю.

Поскольку

$$P\left(I_{\{X_i \le x\}} = 1\right) = P(X_i \le x), \text{ то}$$

$$P\left(I_{\{X_i \le x\}} = 1\right) = F(x), P\left(I_{\{X_i \le x\}} = 0\right) = P(X_i > x) = 1 - F(x),$$

то отсюда следует

$$P\left(F_n(x)=\frac{m}{n}\right)=C_n^m\left[F(x)\right]^m\left[1-F(x)\right]^{n-m}.$$

Вычислим математическое ожидание $F_n(x)$.

$$M\left(F_n(x)\right)=\frac{1}{n}M\left(\sum_{i=1}^{n}I_{\{X_i\le x\}}\right)=\frac{1}{n}\sum_{i=1}^{n}M\left(I_{\{X_i\le x\}}\right)=\frac{1}{n}\left(n\left[1\cdot F(x)+0\cdot\left(1-F(x)\right)\right]\right)=$$
$$=F(x).$$

Пример. Построить эмпирическую функцию распределения вероятностей по выборке, представленной в таблице 2.

Таблица 2. Сгруппированная выборка.

x_i	2	5	8	12
$\frac{n_i}{n}$	$\frac{2}{9}$	$\frac{1}{9}$	$\frac{2}{9}$	$\frac{4}{9}$

Решение. Эмпирическая функция распределения делает скачок в точке 2, величина скачка равна $\frac{2}{9}$, следующий скачок происходит в точке 5, его величина скачка равна $\frac{1}{9}$, очередные скачки происходят в точках 8 и 12 и имеют величины соответственно $\frac{2}{9}$ и $\frac{4}{9}$. То есть

$$F_n(x) = \begin{cases} 0, x < 2 \\ \frac{2}{9}, 2 \le x < 5 \\ \frac{1}{3}, 5 \le x < 8 \\ \frac{5}{9}, 8 \le x < 12 \\ 1, x \ge 12 \end{cases}.$$

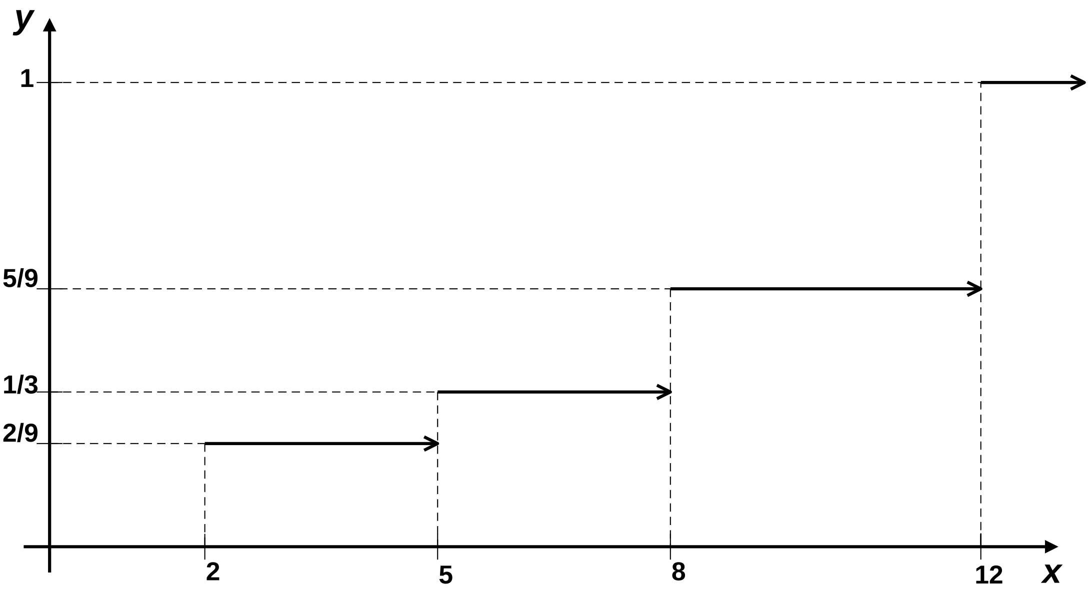

Рис. 2. График эмпирической функции распределения.

Пример. Найти эмпирическую функцию распределения вероятностей по данному распределению выборки, представленной в таблице 3.

Таблица 3. Сгруппированная выборка.

x_i	1	4	6
n_i	10	15	25

Решение. Найдем объём выборки: $n = 10 + 15 + 25 = 50$. Наименьший элемент выборки равен 1, поэтому $F_n(x) = 0$ при $x < 1$.

Значение $x < 4$, а именно $x_1 = 1$, наблюдалось 10 раз, следовательно,

$$F_n(x) = \frac{10}{50} = 0.2 \quad \text{при} \quad 1 \le x < 4$$

Значение $x < 6$, а именно $x_1 = 1$ и $x_2 = 4$, наблюдалось 10+15=25 раз, следовательно,

$$F_n(x) = \frac{25}{50} = 0{,}5 \ \text{при} \quad 4 \le x < 6$$

Так как $x = 6$ - наибольшее значение из выборки, то $F_n(x) = 1$ при $x \ge 6$. В результате эмпирическая функция распределения имеет вид:

$$F_n(x) = \begin{cases} 0 & \text{при} \quad x < 1 \\ 0{,}2 & \text{при} \quad 1 \le x < 4 \\ 0{,}5 & \text{при} \quad 4 \le x < 6 \\ 1 & \text{при} \quad x \ge 6 \end{cases}$$

Пусть $X_1, X_2, ..., X_n$ - последовательность независимых одинаково распределенных случайных величин с эмпирической функцией распределения $F_n(x)$ и теоретической функцией распределения $F(x)$. Найдем эмпирическую функцию распределения $G_n(y)$ для случайных величин $F(X_1), F(X_2), ..., F(X_n)$, если $F(x)$ - непрерывна и монотонна.

$$G_n(y) = \frac{1}{n}\sum_{i=1}^{n} I_{\{F(X_i) \le y\}} = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_i \le F^{-1}(y)\}} = F_n\left(F^{-1}(y)\right).$$

Оценим отклонение $F_n(x)$ от $F(x)$. Применяя неравенство Чебышева, получим

$$P\left(\sqrt{n}\left|F_n(x) - F(x)\right| > t\right) \le \frac{F(x)\left(1 - F(x)\right)}{t^2} \qquad (5)$$

Лучшую оценку для вероятности в левой части соотношения (5) при больших n можно получить применением интегральной теоремы Муавра-Лапласа:

$$\mathrm{P}\left(\sqrt{n}\left|F_n(x)-F(x)\right|>t\right)\cong 1-2\Phi\left(\frac{t}{\sqrt{F(x)(1-F(x))}}\right), \tag{6}$$

где

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_0^x e^{-\frac{y^2}{2}}dy.$$

Замечание. Недостаток оценок (5) и (6) состоит в том, что они зависят от функции $F(x)$, которую мы и пытаемся оценить по результатам наблюдений. Заменяя $F(x)(1-F(x))$ ее наибольшим значением $\frac{1}{4}$, эту зависимость можно устранить, но ценой огрубления оценки.

Пусть $F_n(x)$ - эмпирическая, а $F(x)$ - теоретическая функции распределения вероятностей для последовательности независимых одинаково распределенных случайных величин $X_1, X_2, ..., X_n$. Применяя закон больших чисел, получим

$$\lim_{n\to\infty} P\left(\left|F_n(x)-F(x)\right|<\varepsilon\right)=\lim_{n\to\infty} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} I_{\{X_i\le x\}}-F(x)\right|<\varepsilon\right)=1. \tag{7}$$

Мерой расхождения между теоретической и эмпирической функциями распределения вероятностей служит ***статистика Колмогорова***

$$D_n(x)=\sup_x\left|F_n(x)-F(x)\right|.$$

Статистикой принято называть любую функцию от результатов наблюдений и от параметров, значения которых известны.

Согласно теореме Гливенко - Кантелли

$$\lim_{n\to\infty} P\left(\sup_x \left|F_n(x)-F(x)\right|<\varepsilon\right)=1. \qquad (8)$$

Полученное соотношение (8) можно прокомментировать достаточно легко: полоса $F_n(x)\pm\varepsilon$, $-\infty<x<+\infty$, ширины 2ε и границами, строящимися по полученной выборке объема n, с вероятностью $1-\delta$ (чем больше объем выборки n, тем ближе δ к 1) заключает внутри себя неизвестную теоретическую функцию распределения вероятностей $F(x)$.

Конкретно для статистики Колмогорова было найдено предельное распределение, позволяющее обеспечить достаточную степень точности при определении расхождения между $F_n(x)$ и $F(x)$:

$$\lim_{n\to\infty} P\left(D_n(x)\le\frac{y}{\sqrt{n}}\right)=\sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2y^2}, y>0.$$

Если объем выборки n большой, то для представления о виде ее распределения строится ***гистограмма.*** Для этого интервал U, где наблюдается выборка, разбивается на непересекающиеся полуинтервалы $(y_k, y_{k+1}]$ длины h:

$$U=\bigcup_{k=0}^{m-1}(y_k, y_{k+1}].$$

Для каждого полуинтервала $(y_k, y_{k+1}]$ определяется частота n_k и относительная частота $m_k=\frac{n_k}{n}, k=\overline{0, m-1}$ попадания в этот полуинтервал элементов выборки. После этого строятся прямоугольники длиной $y_{k+1}-y_k=h$ и высотой $\frac{m_k}{h}$.

Площадь всех построенных прямоугольников равна

$$\sum_{k=0}^{m-1} h\cdot\frac{m_k}{h}=\sum_{k=0}^{m-1}\frac{n_k}{n}=\frac{n}{n}=1.$$

Как известно, если $f(x)$ - плотность распределения вероятностей некоторой случайной величины, то

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

Таким образом, если разбиение построено удачно, то гистограмма будет напоминать график плотности (если она существует) распределения вероятностей случайной величины (случайных величин), значениями которой являются элементы выборки. Если разбиение мелкое, то гистограмма не дает представления о плотности распределения вероятностей из-за случайных флуктуаций. Если разбиение крупное, то гистограмма тоже не дает представления о плотности распределения вероятностей из-за того, что теряется много информации.

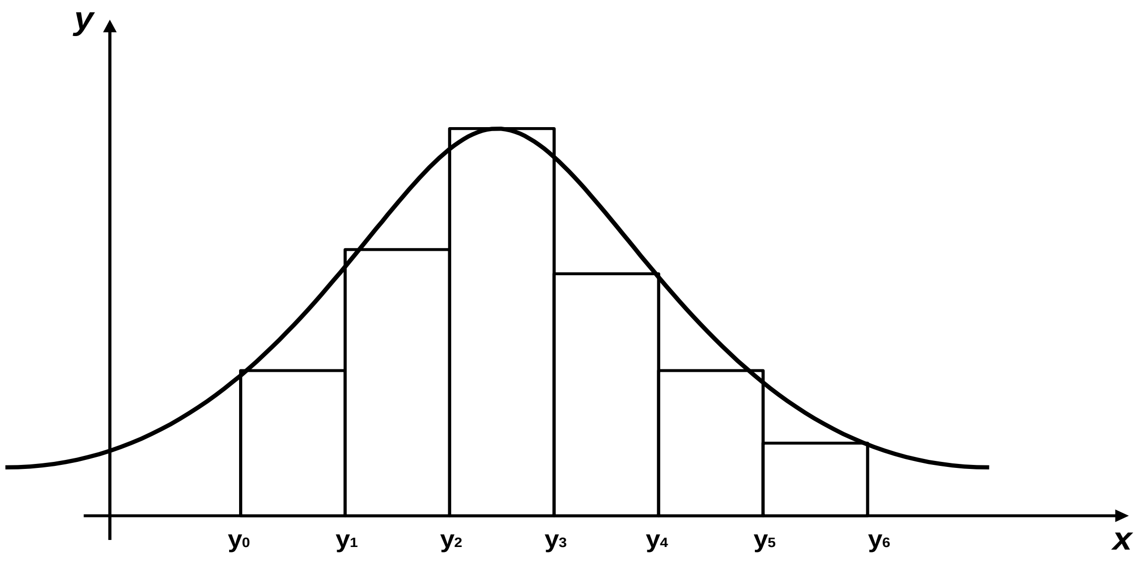

Рис. 3. Гистограмма

Пример. Построить гистограмму по сгруппированным данным, где n - объем выборки, равный 100.

Группы (интервал)	[-30; -10]	(-10,0]	(0,15]	(15,42]

Число элементов в	28	34	25	13
Частота $\frac{n_i}{n}$	0,28	0,34	0,25	0,13

Решение. Вычислим высоты прямоугольников (ординаты гистограммы) по формуле $y_i = \frac{n_i}{nh_i}$, где h_i - длина соответствующего интервала

$$y_1 = \frac{0,28}{20} = 0,014 \quad \text{для всех} \quad x \in [-30;-10]$$

$$y_2 = \frac{0,34}{10} = 0,034 \quad \text{для всех} \quad x \in (-10;0]$$

$$y_3 = \frac{0,25}{15} = 0,016 \quad \text{для всех} \quad x \in (0;15]$$

$$y_4 = \frac{0,13}{27} = 0,0048 \quad \text{для всех} \quad x \in (15;42]$$

Гистограмма приведена на рис.4.

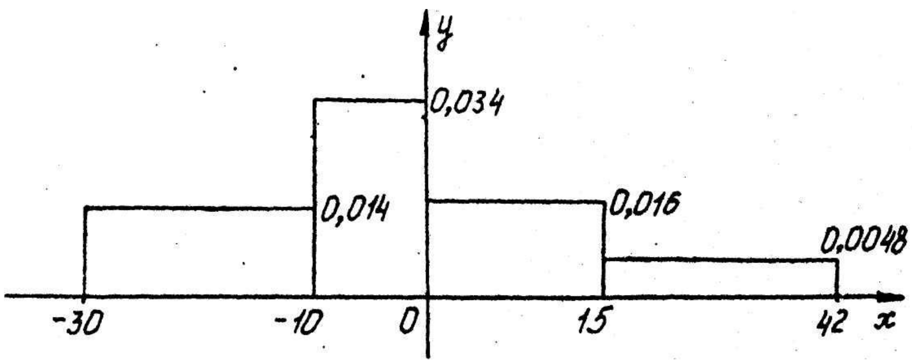

Рис.4. Гистограмма.

Задачи.

Задача 118. Вычислить дисперсию $F_n(x)$.

Задача 119. Построить эмпирическую функцию распределения по следующей выборке: -1; -2; 1; 3; 2.

Задача 120. Построить эмпирическую функцию распределения по данному распределению выборки:

x_i	1	4	6
n_i	10	15	25

Задача 121. Построить гистограмму и эмпирическую функцию распределения вероятностей по выборке:

x_i	-1	0	1	2
$\frac{n_i}{n}$	$\frac{2}{7}$	$\frac{1}{7}$	$\frac{1}{7}$	$\frac{3}{7}$

Задача 122. Построить гистограмму и эмпирическую функцию распределения вероятностей по выборке: -0.15, -0.17, -0.21, -0.33, -0.19, -0.04, -0.02, 0.13, 0.16, 0.24, 0.36, 0.56, 0.61, 0.77.

Задача 123. В течение 24 ч регистрирующее устройство контроля каждый час фиксирует напряжение сети. После первичной обработки данных получено распределение выборки в интервальной форме.

Интервалы, *В*	213-215	215-217	217-219	219-221	221-223
Число элементов	1	3	6	10	4

Построить гистограмму.

12. Выборочное среднее и выборочная дисперсия

Поскольку эмпирическая функция распределения вероятностей $F_n(x)$ служит оценкой для теоретической функции распределения вероятностей $F(x)$, то следует ожидать, что математическое ожидание и дисперсия, вычисленные по эмпирической функции распределения, будут служить оценками для теоретического математического ожидания и теоретической дисперсии.

Как уже было сказано, $F_n(x)$ отвечает дискретному распределению, представленному в таблице 4.

Таблица 4. Дискретная случайная величина.

X_i	x_1	x_2	…	x_n
p	$\frac{1}{n}$	$\frac{1}{n}$	…	$\frac{1}{n}$

Поэтому представляется логичным в качестве оценок математического ожидания и дисперсии независимых и одинаково распределенных случайных величин $X_1, X_2, ..., X_n$ использовать выборочное среднее $\bar{x}$ (математическое ожидание случайной величины из таблицы 4) и выборочную дисперсию s_0^2:

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

$$s_0^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 .$$

Если известно математическое ожидание $M(X_i)=a, i=\overline{1,n}$, то в качестве оценки дисперсии используется величина

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

Переходя к случайным величинам, получим

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

$$S_0^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2=\frac{1}{n}\sum_{i=1}^{n}X_i^2-\overline{X}^2,$$

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

Выборочное среднее и выборочная дисперсия являются во многих случаях хорошими, а в некоторых оптимальными (в определенном смысле) оценками теоретических значений математического ожидания и дисперсии.

Поскольку $X_1,X_2,...,X_n$ являются случайными величинами, то и $\overline{X}, S_0^2, S^2$ являются случайными величинами, а поэтому мы можем вычислить их математические ожидания.

$$M\left(\overline{X}\right)=M\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n}\sum_{i=1}^{n}M(X_i)=\frac{nM(X_i)}{n}=M(X_i),$$

$$M\left(S_0^2\right)=M\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right)=M\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)-M\left(\overline{X}^2\right)=\frac{nM\left(X_i^2\right)}{n}-$$

$$M\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)^2=M\left(X_i^2\right)-M\left(\frac{X_1^2+\ldots X_n^2+X_1X_2+X_1X_3\ldots.+X_{n-1}X_n}{n^2}\right)=$$

$$M\left(X_i^2\right)-M\left(\frac{X_1^2+\ldots X_n^2}{n^2}\right)-M\left(\frac{X_1X_2+X_1X_3\ldots.+X_{n-1}X_n}{n^2}\right)=M\left(X_i^2\right)-$$

$$\frac{nM\left(X_i^2\right)}{n^2}-\frac{n(n-1)M\left(X_iX_j\right)}{n^2}=\frac{n-1}{n}M\left(X_i^2\right)-\frac{n-1}{n}M\left(X_i\right)M\left(X_j\right)=$$

$$\frac{n-1}{n}M\left(X_i^2\right)-\frac{n-1}{n}\left(M\left(X_i\right)\right)=\frac{n-1}{n}D\left(X_i\right),$$

$$M\left(S^2\right)=M\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i-a\right)^2\right)=M\left(\frac{1}{n}\sum_{i=1}^{n}X_i^2\right)-2aM\left(\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)\right)+a^2=M\left(X_i^2\right)-$$

$$2a^2+a^2=D\left(X_i\right).$$

$\overline{X}$ и S^2 в среднем совпадают соответственно с теоретическим математическим ожиданием и теоретической дисперсией, оценками которых они и являются. Мы говорим в среднем, поскольку $M\left(\overline{X}\right)=M\left(X_i\right)$, а $M\left(S^2\right)=D\left(X_i\right)$. А вот S_0^2 в среднем не совпадает с теоретической дисперсией, оценкой которой является ($M\left(S_0^2\right)=\frac{n-1}{n}D\left(X_i\right)$). Эту ситуацию можно исправить, если рассмотреть величину

$$s_1^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2,\ (S_1^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2),$$

которая называется исправленной выборочной дисперсией и в среднем совпадает с $D\left(X_i\right)$. Исправленная выборочная дисперсия и выборочная дисперсия связаны отношением

$$s_1^2=\frac{n}{n-1}s_0^2$$

Необходимо отметить, что на практике возможны выборки, состоящие из большого числа многократно повторяющихся элементов. В этом случае при вычислении $\overline{x}$, s^2, s_0^2 и s_1^2 полезно пользоваться следующими формулами:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i \ ,$$

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

$$s_0^2 = \frac{1}{n}\sum_{i=1}^{k} \left(x_i - \overline{x}\right)^2 n_i \ ,$$

$$s_1^2 = \frac{1}{n-1}\sum_{i=1}^{k} \left(x_i - \overline{x}\right)^2 n_i \ ,$$

где n_i - число повторений значения x_i в выборке.

Пример. В таблице 5 приведены полученные данные и частоты, с которыми эти данные наблюдались. Вычислить выборочное среднее и выборочную дисперсию.

Таблица 5. Группированная выборка.

x_i	-5	-4	-1	0	1	2
n_i	30	20	10	30	20	10

Решение. Объем выборки равен $n = \sum_{i=1}^{6} n_i = 30 + 20 + 10 + 30 + 20 + 10 = 120$, получим новую таблицу 6, в которой $m_i = \frac{n_i}{n}, i = \overline{1,6}$.

Таблица 6.

x_i	-5	-4	-1	0	1	2
m_i	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{4}$	$\frac{1}{6}$	$\frac{1}{12}$

Получим

$$\bar{x} = \sum_{i=1}^{6} m_i x_i = -\frac{5}{4} - \frac{4}{6} - \frac{1}{12} + \frac{1}{6} + \frac{2}{12} = -\frac{5}{3},$$

$$s_0^2 = \sum_{i=1}^{6} m_i x_i^2 - \bar{x}^2 = \frac{25}{4} + \frac{16}{6} + \frac{1}{12} + \frac{1}{6} + \frac{4}{12} - \frac{25}{9} = 6\frac{13}{18}.$$

Ответ. $\bar{x} = -\frac{5}{3}$, $s_0^2 = 6\frac{13}{18}$.

Задачи.

Задача 124. В таблице 7 приведены полученные данные и частоты, с которыми эти данные наблюдались. Вычислить выборочное среднее и выборочную дисперсию.

Таблица 7. Группированная выборка.

x_i	1	3	6	26
n_i	8	40	10	2

Задача 125. В таблице 8 приведены полученные данные и частоты, с которыми эти данные наблюдались. Вычислить выборочное среднее и выборочную дисперсию.

Таблица 8. Группированная выборка.

x_i	-2	-1	0	6
n_i	20	25	25	30

Задача 126. В таблице 9 приведены полученные данные и частоты, с которыми эти данные наблюдались. Вычислить выборочное среднее, выборочную дисперсию и исправленную выборочную дисперсию.

Таблица 9. Группированная выборка.

x_i	-4	0	3	5
n_i	10	15	20	55

Задача 127. В таблице 10 приведены полученные данные и частоты, с которыми эти данные наблюдались. Вычислить выборочное среднее, выборочную дисперсию и исправленную выборочную дисперсию.

Таблица 10. Группированная выборка.

x_i	0	2	5	12
n_i	10	40	30	20

Задача 128. Вычислить $\overline{x}$ и s_1^2 по значениям x_i, равным 27, 22, 24 и 26.

Задача 129. Вычислить $\overline{x}$ и s_1^2 по приведенным в таблице 11 значениям.

Таблица 11. Группированная выборка.

x_i	7	8	9	10	11	12	13
n_i	11	10	21	25	23	11	12

Задача 130. В таблице 12 приведены результаты измерения роста (в сантиметрах) случайно отобранных 100 студентов.

Таблица 12. Группированная выборка.

Рост	Число студентов		Рост	Число студентов
154-158	8		170-174	12
158-162	14		174-178	8
162-166	26		178-182	2
166-170	28		182-186	2

Найти выборочные среднее и исправленную выборочную дисперсию роста обследованных студентов.

Указание. Середины интервалов принять в качестве x_i.

13. Основные законы распределения в математической статистике. Распределение среднего и дисперсии выборки из нормальной генеральной совокупности

Случайная величина X имеет нормальное распределение с параметрами m и σ^2, если плотность распределения имеет вид (рис.5):

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

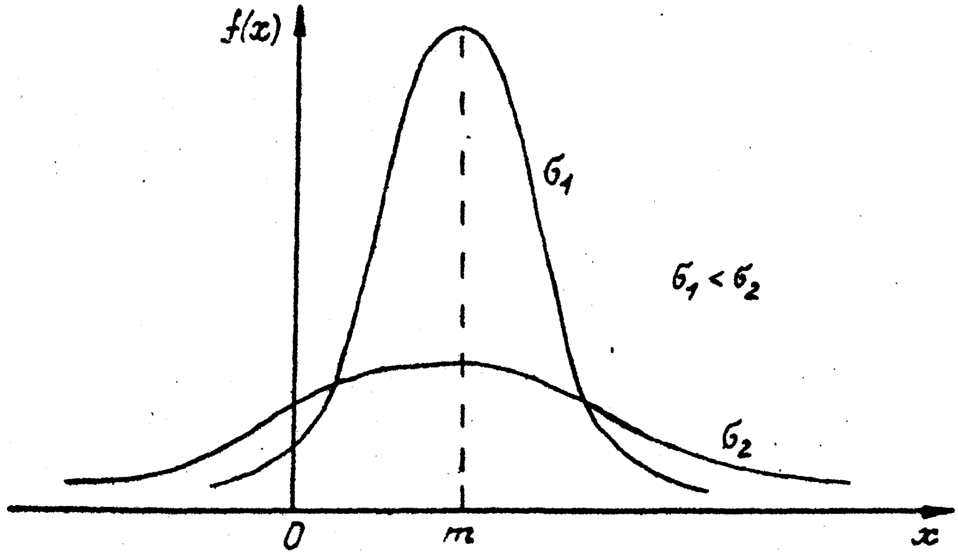

Рис. 5. Графики плотностей нормального распределения с одинаковым математическим ожиданием m и разными дисперсиями $\sigma_1^{\ 2}, \sigma_2^{\ 2}$.

Нормальное распределение с параметрами 0 и 1 называют стандартным нормальным распределением и обозначают $N(0,1)$. Значения функции распределения стандартной нормальной случайной величины

$$F(x) = \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt = \frac{1}{2} + \Phi(x)$$

легко находятся с помощью функции Лапласа $\Phi(x) = \int_{0}^{x} e^{-\frac{t^2}{2}} dt$ (приложение 2).

Пусть X имеет нормальное распределение о параметрами m и σ^2 . Тогда центральные моменты $\mu_n = M(X-m)^2$ равны $\mu_{2k+1} = 0$, $\mu_{2k} = 1 \cdot 3 \cdot .... \cdot (2k-1)\sigma^{2k}$,

математическое ожидание $MX = m$, дисперсия $DX = \sigma^2$; коэффициент асимметрии $a_x = \frac{\mu_3}{\sigma^3} = 0$ и коэффициент эксцесса $\varepsilon_x = \frac{\mu_4}{\sigma^4} - 3 = 0$. Более того, случайная величина $Y = \frac{X-m}{\sigma}$ имеет стандартное нормальное распределение.

Нормально распределенная случайная величина с большой вероятностью принимает значения, близкие к своему математическому ожиданию, что выражается правилом сигм:

$$P(|X-m|<k\sigma)=\begin{cases}0{,}6827, & k=1\\ 0{,}9545, & k=2\\ 0{,}9973, & k=3\end{cases}$$

Чаще всего используется правило трех сигм, т.е. $k=3$.

Пусть $X_1, X_2, ..., X_n$ - независимые случайные величины с общим распределением $N(0,1)$ (нормальное распределение с нулевым математическим ожиданием и единичной дисперсией). Тогда случайная величина

$$\chi^2=\sum_{i=1}^{n}X_i^2$$

имеет распределение χ^2 с n степенями свободы или распределение χ_n^2.

Плотность распределения вероятностей χ_n^2 имеет вид:

$$f_{\chi_n^2}(x)=\begin{cases}\dfrac{1}{2^{\frac{n}{2}}\cdot\Gamma\left(\dfrac{n}{2}\right)}\cdot x^{\frac{n}{2}-1}\cdot e^{-\frac{x}{2}}, x>0\\ 0, x\le 0\end{cases},$$

где

$$\Gamma(n)=\int_0^{\infty}x^{n-1}e^{-x}dx.$$

Графики плотности распределения χ^2 с n степенями свободы асимметричны и, начиная с $n=2$, имеют по одному максимуму в точке $x=n-2$ (рис. 6). Причем с ростом n кривая плотности приближается к симметричной функции.

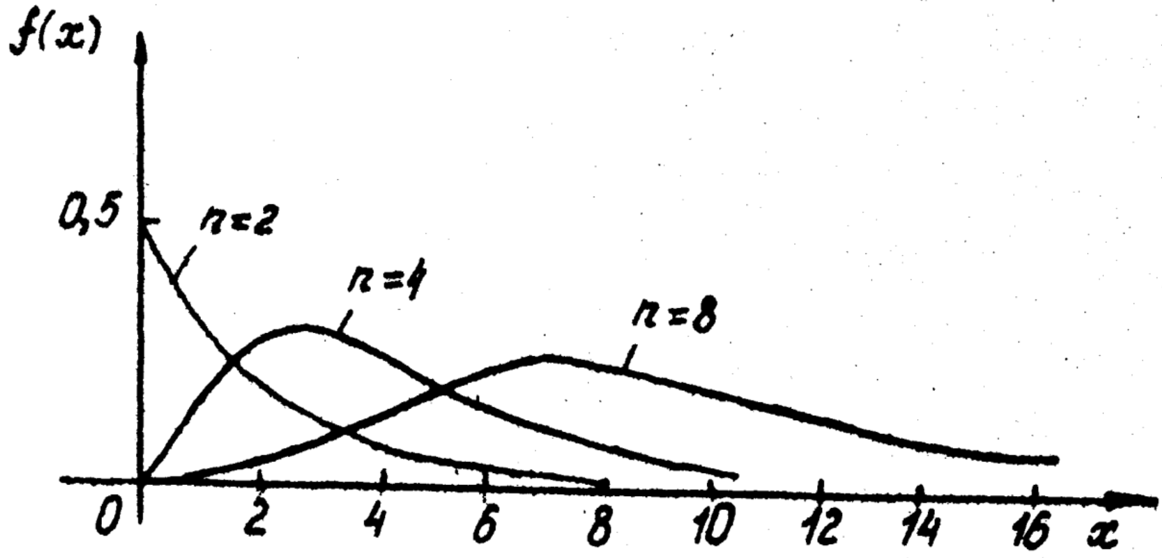

Рис.6. Графики плотности распределения χ^2 с n степенями свободы.

Пусть случайная величина Y имеет распределение $N(0,1)$, а независимая от Y случайная величина Z принадлежит χ_n^2 (имеет распределение χ^2 с n степенями свободы). Тогда случайная величина

$$X = \frac{Y}{\sqrt{\frac{Z}{n}}}$$

имеет распределение Стьюдента с n степенями свободы (t_n-распределение) и плотностью распределения вероятностей:

$$f_{t_n}(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\,\Gamma\left(\frac{n}{2}\right)} \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \; x \in (-\infty, +\infty).$$

Графики плотности случайной величины, имеющей распределение Стьюдента, при любом $n = 1,2,....$ симметричны относительно оси ординат (рис.7), поэтому при любом $n = 1,2,.....$ математическое ожидание равно нулю.

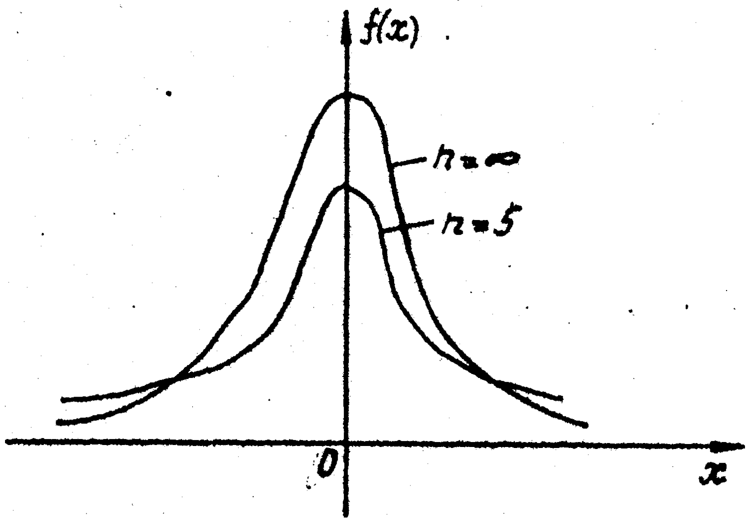

Рис.7. Графики плотности распределения Стьюдента.

С ростом n распределение Стьюдента приближается к $N(0,1)$.

Пусть U и V - независимые случайные величины, распределенные по закону χ^2 с n_1 и n_2 степенями свободы соответственно. Тогда случайная величина

$$X = \frac{\dfrac{U}{n_1}}{\dfrac{V}{n_2}}$$

имеет распределение Фишера с n_1 и n_2 степенями свободы (F_{n_1,n_2}-распределение) и плотностью распределения вероятностей

$$f(x) = \begin{cases} \dfrac{\Gamma\left(\dfrac{n_1+n_2}{2}\right)\cdot n_1^{\frac{n_1}{2}}\cdot n_2^{\frac{n_2}{2}}}{\Gamma\left(\dfrac{n_1}{2}\right)\cdot\Gamma\left(\dfrac{n_2}{2}\right)}\cdot x^{\frac{n_1}{2}-1}\cdot(n_2+n_1x)^{-\frac{n_1+n_2}{2}}, x>0 \\ 0, x\le 0 \end{cases}.$$

Графики плотности распределения случайной величины асимметричны, имеют длинные "хвосты" и достигают максимума вблизи точки $x=1$ (рис.8).

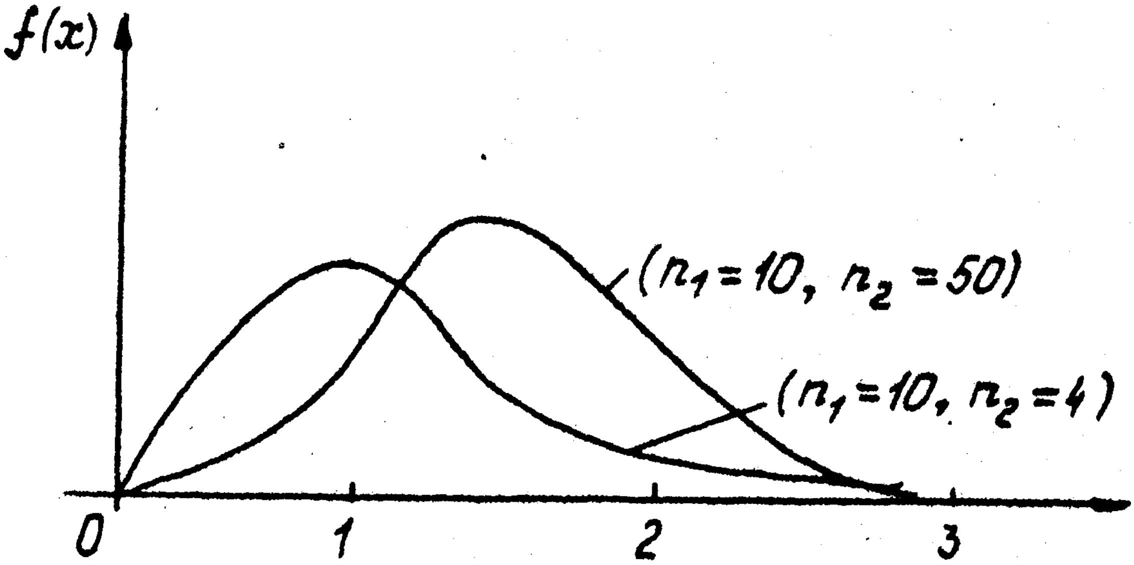

Рис.8. Графики плотности распределения Фишера.

Пусть $X_1, X_2, \ldots, X_n$ - независимые случайные величины с общим распределением $N\left(a, \sigma^2\right)$ (нормальное распределение с математическим ожиданием a и дисперсией σ^2). Рассмотрим выборочное среднее $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ и выборочную дисперсию $S_0^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}^2$.

Так как сумма независимых нормально распределенных случайных величин имеет нормальное распределение, то для нормально распределенной случайной величины $\overline{X}$ остается найти математическое ожидание и дисперсию:

$$M\left(\overline{X}\right)=\frac{1}{n}nM\left(X_1\right)=a, D\left(\overline{X}\right)=\frac{1}{n^2}nD\left(X_1\right)=\frac{\sigma^2}{n}.$$

Таким образом, $\overline{X}$ имеет распределение $N\left(a,\frac{\sigma^2}{n}\right)$.

Тогда $\frac{\left(\overline{X}-a\right)\sqrt{n}}{\sigma}$ тоже имеет нормальное распределение но с нулевым математическим ожиданием и единичной дисперсией (стандартное нормальное распределение $N(0,1)$):

$$M\left(\frac{\left(\overline{X}-a\right)\sqrt{n}}{\sigma}\right)=\frac{(a-a)\sqrt{n}}{\sigma}=0,\ D\left(\frac{\left(\overline{X}-a\right)\sqrt{n}}{\sigma}\right)=\frac{\sigma^2}{n}\times\frac{n}{\sigma^2}=1.$$

Такое нормальное распределение называется стандартным. Для работы с этим распределением, как уже было сказано) используется таблица для функции Лапласа $\Phi(x)$ (приложение 2).

Рассмотрим случайные величины $S^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-a\right)^2$, $S_0^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$, $S_1^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2$, где $a=M\left(X_1\right)$.

Случайная величина S^2 похожа на случайную величину, имеющую распределение χ_n^2, только без множителя $\frac{1}{n}$, поэтому рассмотрим $nS^2=\sum_{i=1}^{n}\left(X_i-a\right)^2$. В скобках стоят случайные величины, которые одинаково распределены и независимы. Так как $X_1, X_2,\ldots,X_n$ имеют распределение $N\left(a,\sigma^2\right)$, то $X_1-a, X_2-a,\ldots,X_n-a$ имеют распределение $N\left(0,\sigma^2\right)$. Поскольку в распределении χ_n^2 присутствует сумма квадратов стандартных нормальных случайных величин, то нормируем

$X_1-a, X_2-a, \ldots, X_n-a$ множителем $\dfrac{1}{\sigma}$, в результате чего получим, что случайные величины $\dfrac{X_i-a}{\sigma}, i=\overline{1,n}$ имеют стандартное нормальное распределение, а случайная величина $\dfrac{nS^2}{\sigma^2}=\sum_{i=1}^{n}\left(\dfrac{X_i-a}{\sigma}\right)^2$ имеет распределение χ_n^2.

Рассмотрим случайные величины $\dfrac{nS_0^2}{\sigma^2}$ и $\dfrac{(n-1)S_1^2}{\sigma^2}$. В отличие от $\dfrac{nS^2}{\sigma^2}$ эти случайные величины имеют распределение не χ_n^2, а χ_{n-1}^2, а случайные величины

$$\sqrt{n-1}\cdot\frac{\overline{X}-a}{S_0}=\frac{\sqrt{n}\cdot\dfrac{\overline{X}-a}{\sigma}}{\dfrac{\dfrac{\sqrt{n}S_0}{\sigma}}{\sqrt{n-1}}}$$

$$\sqrt{n}\cdot\frac{\overline{X}-a}{S_1}=\frac{\sqrt{n}\cdot\dfrac{\overline{X}-a}{\sigma}}{\dfrac{\dfrac{\sqrt{n-1}S_1}{\sigma}}{\sqrt{n-1}}}$$

имеют распределение t_{n-1}.

Если случайный вектор (X,Y) имеет невырожденное $(|\rho|\neq 1)$ нормальное распределение, то его плотность $f(x,y)$ (плотность двумерного нормального распределения) равна

$$f(x,y)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_1)^2}{\sigma_1^2}-2\rho\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2}+\frac{(y-m_2)^2}{\sigma_2^2}\right]\right\}$$

(рис. 9)

Смысл величин $m_1, m_2, \sigma_1, \sigma_2$ и ρ следующий: $m_1 = MX$, $\sigma_1^2 = DX$, $m_2 = MY$, $\sigma_2^2 = DY$, $M(XY) = M(YX) = \sigma_1\sigma_2\rho + m_1 m_2$,

$$\rho = \frac{M(X - m_1)(Y - m_2)}{\sqrt{DX \cdot DY}} -$$

коэффициент корреляции компонент X и Y.

Плотность невырожденного нормального распределения сохраняет постоянные значения на следующих кривых

$$\frac{1}{2(1-\rho^2)}\left[\frac{(x-m_1)}{\sigma_1^2} - 2\rho\frac{(x-m_1)(y-m_2)}{\sigma_1\sigma_2} + \frac{(y-m_2)^2}{\sigma_2^2}\right] = \lambda^2$$

называемых эллипсами равных вероятностей, причем вероятность попадания вектора (X, Y) внутрь такого эллипса равна $1 - e^{\lambda^2}$.

Для нормального закона справедливо следующее правило: если компоненты двумерного нормального вектора некоррелированы (т.е. $\rho = 0$), то они и независимы. При этом оси координат Ox называются главными осями рассеивания.

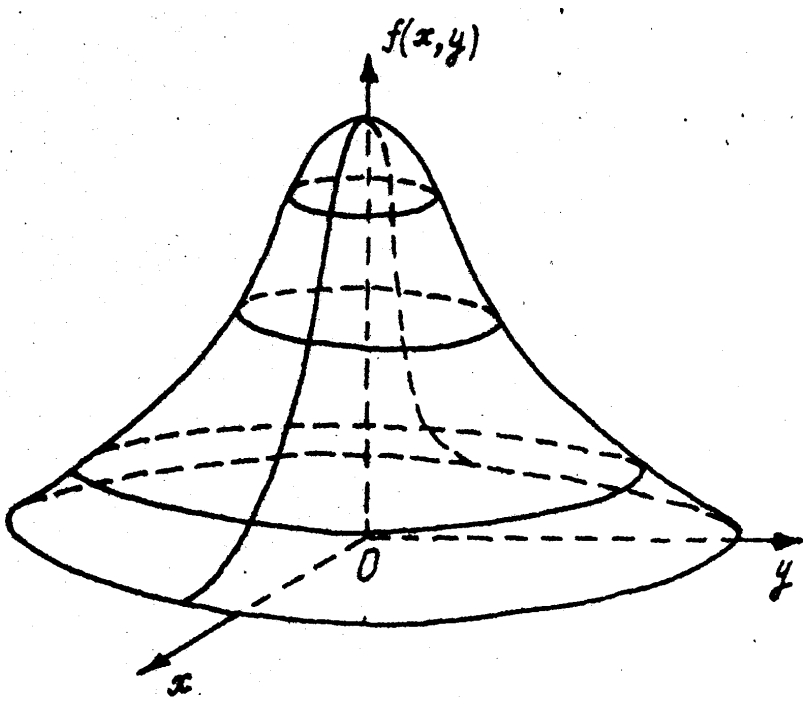

Рис. 9. График плотности двумерного нормального распределения.

Рассмотренные в этом разделе распределения имеют существенное значение для построения точечных и интервальных оценок (доверительных интервалов) для неизвестных параметров.

14. Порядковые статистики

Расположим элементы выборки

$$x_1, x_2, \ldots, x_n$$

в порядке возрастания:

$$x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)} \; . \tag{9}$$

Или, переходя к случайным величинам, запишем

$$X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)} \; . \tag{10}$$

Ряд (9), также как и ряд (10), называют **вариационным рядом**, а элементы вариационного ряда называют **порядковыми статистиками**.

Размахом вариационного ряда называется разность между его максимальным и минимальным значениями, то есть $x_{(n)} - x_{(1)}$.

Пример. Толщина кремниевой пластины измерена несколько раз. Получены следующие значения в микронах: 5,39; 5,42; 5,38; 5,40; 5,47; 5,51; 5,30; 5,40; 5,28; 5,43. Построить вариационный ряд.

Решение. Вариационный ряд имеет вид:

5,28; 5,30; 5,38; 5,39; 5,40; 5,40; 5,42; 5,43; 5,47; 5,51.

Порядковые статистики широко используются в статистических задачах. Например, в задаче оценивания содержимого урны. Предполагается, что урна содержит неизвестное число N пронумерованных шаров. С целью определения N осуществляется повторный выбор шаров с возвращением. Пусть номера выбранных шаров представлены в виде выборки

$$x_1, x_2, \ldots, x_n.$$

Вполне естественно в качестве оценки N предложить

$$\hat{N} = x_{(n)}.$$

Пример. Найти распределение $X_{(1)}$.

Решение.

$$P\left(X_{(1)} \le x\right) = 1 - P\left(X_{(1)} > x\right) = 1 - P\left(X_i > x, i = \overline{1,n}\right) =$$
$$= 1 - \left[P\left(X_1 > x\right)\right]^n = 1 - \left[1 - F(x)\right]^n.$$

Ответ. $P\left(X_{(1)} \le x\right) = 1 - \left[1 - F(x)\right]^n$.

Пример. Найти $P\left(X_{(k)} \le x, X_{(k+1)} > x\right)$.

Решение. Событие $\left(X_{(k)} \le x, X_{(k+1)} > x\right)$ состоит в том, что ровно k случайных величин из $X_1, X_2, \ldots, X_n$ меньше или равны x, а остальные $n-k$ случайных величин больше x, отсюда

$$P\left(X_{(k)} \le x, X_{(k+1)} > x\right) = C_n^k \left[F(x)\right]^k \left[1 - F(x)\right]^{n-k}.$$

Эту задачу можно решить по-другому. Пользуясь формулой полной вероятности, запишем:

$$P\left(X_{(k)} \le x\right) = P\left(X_{(k)} \le x, X_{(k+1)} \le x\right) + P\left(X_{(k)} \le x, X_{(k+1)} > x\right).$$

Так как

$P\left(X_{(k)} \le x, X_{(k+1)} \le x\right) = P\left(X_{(k+1)} \le x\right)$, то

$$P\left(X_{(k)} \le x, X_{(k+1)} > x\right) = P\left(X_{(k)} \le x\right) - P\left(X_{(k+1)} \le x\right) =$$

$$= \sum_{i=k}^{n} C_n^i \left[F(x)\right]^i \left[1 - F(x)\right]^{n-i} - \sum_{i=k+1}^{n} C_n^i \left[F(x)\right]^i \left[1 - F(x)\right]^{n-i} =$$

$$= C_n^k \left[F(x)\right]^k \left[1 - F(x)\right]^{n-k}.$$

Ответ. $P\left(X_{(k)} \le x, X_{(k+1)} > x\right) = C_n^k \left[F(x)\right]^k \left[1 - F(x)\right]^{n-k}$.

Пример. Урна содержит неизвестное число N шаров, занумерованных числами от 1 до N. С целью определения N совершается повторный выбор шаров с возвращением. Пусть $x_1, x_2, ..., x_n$ - номера извлеченных шаров. В качестве оценки предлагается максимальное значение из наблюдений

$$\hat{N} = x_{(n)}.$$

Вычислить математическое ожидание $\hat{N}$.

Решение. Переходя к вероятностной модели испытаний, введем независимые одинаково распределенные случайные величины $X_1, X_2, ..., X_n$, которые принимают значения $1,2,...,N$ с равными вероятностями $\frac{1}{N}$ и независимые одинаково распределенные случайные величины $Y_1, Y_2, ..., Y_n$, такие что $Y_i = \frac{X_i}{N}, i = \overline{1,n}$. Случайные величины $Y_1, Y_2, ..., Y_n$ с хорошей степенью приближения могут рассматриваться как равномерно распределенные на $[0,1]$, так что

$$P\left(X_{(n)} \le Ny\right) = P\left(Y_{(n)} \le y\right) = \left[P\left(Y_1 \le y\right)\right]^n \cong y^n, 0 \le y \le 1.$$

Совершенно ясно, оценка $\hat{N}$ всегда дает заниженный результат, то есть

$$\hat{N} \le N.$$

Вычисляя

$$M\left(Y_{(n)}\right) \cong \int_0^1 yny^{n-1} dy = \frac{n}{n+1},$$

найдем

$$M\left(\hat{N}\right) = NM\left(Y_{(n)}\right) \cong N \cdot \frac{n}{n+1}.$$

Отсюда видно, что оценка $\hat{N}$ в среднем почти равна оцениваемому параметру N. Умножая $\hat{N}$ на $\dfrac{n+1}{n}$, мы добьемся того, что в среднем $\hat{N}$ равна N, что только улучшает ее свойства.

Ответ. $M\left(\hat{N}\right) \cong N \cdot \dfrac{n}{n+1}$.

Напомним, что медианой случайной величины называется значение, делящее распределение на две равновероятные половины. В случае непрерывного на $(-\infty;+\infty)$ распределения медиана m удовлетворяет соотношению

$$\int\limits_{-\infty}^{m} f(x)dx = \int\limits_{m}^{\infty} f(x)dx = \frac{1}{2},$$

где $f(x)$ - плотность случайной величины X. Геометрически, медиана - это абсцисса точки, в которой площадь, ограниченная кривой плотности распределения, делится пополам,

Модой дискретной случайной величины называется ее наиболее вероятное значение. Мода непрерывной случайной величины - это значение, при котором функция плотности распределения случайной величины максимальна.

В качестве оценки медианы $\tilde{m}$ в выборке объема $2n+1$ можно взять значение $x_{(n+1)}$ в вариационном ряде. Если объем выборки равен $2n$, то в качестве оценки медианы $\tilde{m}$ можно взять $\frac{1}{2}(x_{(n)} + x_{(n+1)})$. Поэтому для группированной выборки

$$\tilde{m} = \alpha_k + \left[\frac{\frac{n}{2} - (n_1 + n_2 + ... + n_{k-1})}{n_k}\right] h,$$

где α_k - нижняя граница интервала, которому принадлежит медиана, h - длина этого интервала, n_k - число элементов выборки в этом интервале, $n_1 + n_2 + ... + n_{k-1}$ - число элементов выборки в интервалах, лежащих слева от интервала, содержащего медиану. Оценкой моды унимодального (с одним экстремумом) распределения служит элемент выборки, встречающийся с наибольшей частотой. Для группированной выборки моду можно оценить по следующей формуле:

$$\tilde{d} = \alpha_k + \frac{(n_k - n_{k-1})}{(n_k - n_{k-1}) + (n_k - n_{k+1})} h,$$

где α_k - нижняя граница интервала, содержащего наибольшее число элементов выборки, h - длина этого интервала, n_k - число элементов в этом интервале, n_{k-1} и n_{k+1} - число элементов выборки в соседних с ним интервалах (соответственно слева и справа).

Пример. Оценить моду и медиану для выборки из таблицы 13.

Таблица 13. Группированная выборка.

Рост	Число студентов	Рост	Число студентов
154-158	8	170-174	12
158-162	14	174-178	8
162-166	26	178-182	2
166-170	28	182-186	2

Решение. Оценкой медианы будет значение $\frac{1}{2}(x^{(50)} + x^{(51)})$.Определим в каком интервале она лежит. Поскольку 8 + 14 + 26 = 48 < 50, а 8 + 14 + 26 + 28 = 76 > 51,то оба значения $x^{(50)}$ и $x^{(51)}$ лежат в интервале (166 - 170). В предположении равновероятности любого значения данного интервала, очевидно, можно утверждать, что медиана лежит ближе к левому концу этого интервала. Оценим медиану и моду:

$$\tilde{m} = 166 + \frac{50 - (8 + 14 + 26)}{28} \cdot 4 \approx 166 + 0,286 = 166,286$$

$$\tilde{d} = 166 + \frac{(28 - 26)}{(28 - 26) + (28 - 12)} \cdot 4 \approx 166,44.$$

Задачи.

Задача 131. Оценить значение моды и медианы генеральной совокупности по выборке из таблицы 14.

Таблица 14. Группированная выборка.

x_i	7	8	9	10	11	12	13
n_i	11	10	21	25	23	11	12

Задача 132. Регистрировано число деревьев определенного вида на каждой из 10 выбранных делянок. Получены числа: 15; 13; 15; 14; 16; 14; 12; 14; 14; 15. Построить вариационный ряд. Найти моду и медиану.

Задача 133. Найти размах, моду и медиану вариационного ряда

1;1;1;2;3;3;5.

Задача 134. Найти размах, моду и медиану вариационного ряда

3;4;4;4;5;5;6;6;6;7;9;16.

Задача 135. Найти размах, моду и медиану вариационного ряда

10;11;11;11;11;13;15;16;16;16;16;35.

Задача 136. Найти распределение $X_{(n)}$.

Задача 137. Найти распределение $X_{(k)}$.

15. Точечные оценки неизвестных параметров. Состоятельность и несмещенность оценок

Как известно, статистикой называется функция от результатов наблюдений (функция от выборки). Статистики используются для построения точечных оценок неизвестных параметров различных законов распределений.

Оценка $\hat{\theta}_n$ неизвестного параметра θ называется ***несмещенной***, если

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

Свойство несмещенности оценки неизвестного параметра является вполне логичным, поскольку оно говорит о том, что оценка должна в среднем (в левой части стоит математическое ожидание параметра) соответствовать оцениваемому параметру θ.

Разность $b_n = M\left(\hat{\theta}_n\right) - \theta$ называется ***смещением оценки***.

Оценка $\hat{\theta}_n$ неизвестного параметра θ называется ***асимптотически несмещенной***, если

$$\lim_{n \to \infty} b_n = 0 .$$

Оценка $\hat{\theta}_n$ неизвестного параметра θ называется состоятельной, если

$$\lim_{n \to \infty} \mathrm{P}\left(\left|\hat{\theta}_n - \theta\right| < \varepsilon\right) = 1 .$$

Свойство состоятельности оценки неизвестного параметра говорит о том, что маленькая разность (меньше ε) между оценкой $\hat{\theta}_n$ и оцениваемым параметром θ является достоверным событием.

Несмещенная оценка $\hat{\theta}_n$ неизвестного параметра θ называется ***эффективной***, если при заданном объеме выборки n она имеет наименьшую дисперсию (меньшее отклонение от оцениваемого параметра θ).

Эффективность оценки не всегда удается проверить, зато это свойство дает ключ к сравнению, например, двух несмещенных (и состоятельных) оценок неизвестного параметра. Лучше будет та оценка, которая имеет меньшую дисперсию.

К сожалению, на практике при оценке параметров не всегда оказывается возможным одновременное выполнение требований несмещенности, эффективности и состоятельности.

Пример. В партии из n изделий оказалось m бракованных. Неизвестная вероятность p появления бракованного изделия оценивается величиной $\frac{m}{n}$. Охарактеризовать свойства этой оценки.

Решение. Представим m в следующем виде:

$$m = X_1 + X_2 + \ldots + X_n, \; X_i = \begin{cases} 1, p \\ 0, q \end{cases}, \; p + q = 1, \; i = \overline{1,n}.$$

Следовательно

$$M\left(\frac{m}{n}\right) = \frac{1}{n} \cdot \sum_{i=1}^{n} M(X_i) = M(X_1) = p,$$

и $\frac{m}{n}$ является несмещенной оценкой для параметра p.

Вычислим дисперсию $D\left(\frac{m}{n}\right)$ и воспользуемся неравенством Чебышева:

$$D\left(\frac{m}{n}\right) = \frac{1}{n^2} D(m) = \frac{1}{n^2} \sum_{i=1}^{n} D(X_i) = \frac{1}{n} D(X_1) = \frac{pq}{n},$$

$$P\left(\left|\frac{m}{n} - p\right| \geq \varepsilon\right) \leq \frac{D\left(\frac{m}{n}\right)}{\varepsilon^2} = \frac{pq}{n\varepsilon^2} \to 0, n \to \infty.$$

Получим $\lim\limits_{n\to\infty} P\left(\left|\frac{m}{n}-p\right|<\varepsilon\right)=1$ и состоятельность оценки $\frac{m}{n}$.

Ответ. Оценка $\frac{m}{n}$ является несмещенной и состоятельной оценкой для p.

Пример. Пусть $x_1, x_2, ..., x_n$ - выборка из равномерного на $[0,\theta]$ распределения с неизвестным параметром θ. В качестве оценки параметра θ предлагается $x_{(n)}$. Охарактеризовать свойства этой оценки.

Решение. Перейдем к вероятностной модели выборки

$$X_1, X_2, ..., X_n,$$

где $X_i, i=\overline{1,n}$ - независимые одинаково распределенные случайные величины, такие что

$$P(X_i \le x)=F_\theta(x)=\begin{cases} 0, x<0 \\ \frac{x}{\theta}, 0\le x\le\theta \\ 1, x>\theta \end{cases}, \; i=\overline{1,n},$$

тогда

$$M\left(X_{(n)}\right)=\int\limits_0^\theta x\cdot\frac{nx^{n-1}}{\theta^n}dx=\frac{\theta n}{n+1}\to\theta, n\to\infty$$

и следовательно оценка $X_{(n)}$ является асимптотически несмещенной.

$$P\left(\left|X_{(n)}-\theta\right|\ge\varepsilon\right)=P\left(X_{(n)}\ge\theta+\varepsilon\right)+P\left(X_{(n)}\le\theta-\varepsilon\right)=\left(\frac{\theta-\varepsilon}{\theta}\right)^n,$$

где $\mathrm{P}\left(X_{(n)}\ge\theta+\varepsilon\right)=0$, так как значения выборки (и в частности максимальное значение $X_{(n)}$) не могут быть больше θ.

Получаем

$$\lim_{n\to\infty} P\left(\left|X_{(n)}-\theta\right|<\varepsilon\right)=\lim_{n\to\infty}\left(1-\left(\frac{\theta-\varepsilon}{\theta}\right)^n\right)=1,$$

так как

$$\lim_{n\to\infty}\left(\frac{\theta-\varepsilon}{\theta}\right)^n = 0 \ \left(\frac{\theta-\varepsilon}{\theta}<1\right).$$

Таким образом, оценка $X_{(n)}$ является состоятельной.

Ответ. $X_{(n)}$ является асимптотически несмещенной и состоятельной оценкой для параметра θ.

Пример. Пусть $x_1, x_2, ..., x_n$ - выборка из равномерного на $[0,\theta]$ распределения с неизвестным параметром θ. Сравнить две оценки параметра θ - $\frac{n+1}{n}X_{(n)}$ и $2\overline{X}$.

Решение. Из предыдущих примеров следует, что обе оценки несмещенные и состоятельные. Лучшей из них будет та, у которой меньше дисперсия.

$$D\left(2\overline{X}\right) = \frac{4}{n^2}\sum_{i=1}^{n} D(X_1) = \frac{4}{n}\left(\int_0^\theta \frac{x^2}{\theta}dx - \frac{\theta^2}{4}\right) = \frac{\theta^2}{3n},$$

$$D\left(\frac{n+1}{n}X_{(n)}\right) = \left(\frac{n+1}{n}\right)^2 \cdot \left(D\left(X_{(n)}\right)\right) = \left(\frac{n+1}{n}\right)^2\left(n\int_0^\theta \frac{x^{n-1}}{\theta^n}x^2dx - \frac{n^2}{(n+1)^2}\theta^2\right) =$$

$$= \frac{(n+1)^{\ 2}}{n(n+2)}\theta^2 - \theta^2 = \frac{1}{n(n+2)}\theta^2.$$

При единичном объеме выборки дисперсии совпадают, но в этом случае и сами оценки совпадают. При объеме выборки больше единицы дисперсия оценки $\frac{n+1}{n}X_{(n)}$ в $\frac{n+2}{3}$ раза меньше дисперсии оценки $2\overline{X}$. С этой точки зрения оценка $\frac{n+1}{n}X_{(n)}$ лучше оценки $2\overline{X}$.

Ответ. Оценка $\frac{n+1}{n}X_{(n)}$ лучше оценки $2\overline{X}$.

Пример. Рассматривается выборка из некоторого распределения с неизвестными математическим ожиданием a и дисперсией σ^2. В качестве оценок a и σ^2 предлагаются соответственно выборочное среднее и выборочная дисперсия. Исследовать свойства этих оценок.

Решение. Вычислим $M(\overline{X})$ и $M(S_0^2)$.

$$M(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} M(X_i) = M(X_1) = a\,,$$

$$M(S_0^2) = \frac{1}{n}\sum_{i=1}^{n} M(X_i^2) - M\left((\overline{X})^2\right) = \frac{1}{n}\sum_{i=1}^{n} M(X_i^2) - \frac{1}{n^2}\left(\sum_{i=1}^{n} M(X_i^2) + n(n-1)(M(X_1))^2\right) =$$
$$= M(X_1^2) - \frac{1}{n}M(X_1^2) - \frac{n-1}{n}(M(X_1))^2 = \frac{n-1}{n}M(X_1^2) - \frac{n-1}{n}(M(X_1))^2 = \frac{n-1}{n}\sigma^2.$$

Как видно из вычислений, оценка $\overline{X}$ несмещенная, а оценка S_0^2 асимптотически несмещенная, которую можно сделать несмещенной, умножив на $\dfrac{n}{n-1}$ и получив исправленную выборочную дисперсию S_1^2 (становится понятным, почему она называется исправленной и что именно она исправляет):

$$\frac{n}{n-1}S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = S_1^2.$$

По закону больших чисел

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} M(X_i)\right| < \varepsilon\right) = P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - a\right| < \varepsilon\right) \to 1,\ n \to \infty,$$

следовательно, $\overline{X}$ является состоятельной оценкой для неизвестного параметра a (математического ожидания).

По закону больших чисел

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i^2-\frac{1}{n}\sum_{i=1}^{n}M\left(X_i^2\right)\right|<\varepsilon\right)=P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i^2-\left(\sigma^2+a^2\right)\right|<\varepsilon\right)\to 1,\, n\to\infty,$$

$$P\left(\left|\left(\overline{X}\right)^2-a^2\right|<\varepsilon\right)\to 1, n\to\infty.$$

Следовательно

$$P\left(\left|S_0^2-\sigma^2\right|<\varepsilon\right)=P\left(\left|\frac{1}{n}\sum_{i=1}^{n}X_i^2-\left(\overline{X}\right)^2-\sigma^2\right|<\varepsilon\right)\to 1,\, n\to\infty$$

и S_0^2 является состоятельной оценкой для неизвестного параметра σ^2 (дисперсии).

Ответ. Выборочное среднее $\overline{x}$ является несмещенной оценкой математического ожидания, выборочная дисперсия S_0^2 является асимптотически несмещенной оценкой дисперсии, исправленная выборочная дисперсия S_1^2 является несмещенной оценкой дисперсии, все три оценки являются состоятельными.

Для следующего примера следует напомнить, что распределение с плотностью

$$f(x)=\begin{cases} e^{-x}, x>0 \\ 0, x\le 0 \end{cases}$$

называется экспоненциальным. Это распределение часто используется в теории надежности как распределение времени до выхода из строя (до отказа) изделия. Сдвинутое распределение $f(x-\theta)$, $\theta>0$ используется для определения гарантийного срока θ, в течение которого отказ произойти не может.

Пример. Рассматривается выборка из сдвинутого экспоненциального распределения с плотностью $f(x-\theta)$, $\theta>0$. В качестве оценок неизвестного параметра θ предлагается $X_{(1)}$ и $\overline{X}-1$. Требуется выбрать лучшую оценку.

Решение.

$$P\left(X_{(1)} \le x\right) = 1 - P\left(X_{(1)} > x\right) = 1 - P\left(X_i > x, i = \overline{1,n}\right) =$$
$$= 1 - \left[P\left(X_1 > x\right)\right]^n = 1 - e^{-n(x-\theta)},\ x > \theta.$$

Следовательно

$$M\left(X_{(1)}\right) = -\int_{\theta}^{\infty} x d e^{-n(x-\theta)} = -\,x e^{-n(x-\theta)}\Big|_{\theta}^{\infty} + \int_{\theta}^{\infty} e^{-n(x-\theta)} dx = \theta + \frac{1}{n},$$

из чего заключаем, что оценка $X_{(1)}$ - асимптотически несмещенная и смещение можно легко устранить, если из нее вычесть $\frac{1}{n}$. Таким образом, в дальнейшем вместо оценки $X_{(1)}$ будем рассматривать оценку $X_{(1)} - \frac{1}{n}$.

$$P\left(\left|X_{(1)} - \frac{1}{n} - \theta\right| \ge \varepsilon\right) = P\left(X_{(1)} \ge \theta + \frac{1}{n} + \varepsilon\right) + P\left(X_{(1)} \le \theta + \frac{1}{n} - \varepsilon\right).$$

Вторая вероятность в правой части стремится к нулю при $n \to \infty$, поскольку $X_{(1)}$ не может быть меньше θ. Рассмотрим первую вероятность

$$P\left(X_{(1)} \ge \theta + \frac{1}{n} + \varepsilon\right) = e^{-n\left(\frac{1}{n} + \varepsilon\right)} \to 0,\ n \to \infty.$$

Получаем

$$P\left(\left|X_{(1)} - \frac{1}{n} - \theta\right| \ge \varepsilon\right) \to 0, n \to \infty,$$

откуда следует, что

$$P\left(\left|X_{(1)} - \frac{1}{n} - \theta\right| < \varepsilon\right) \to 1,\ n \to \infty.$$

Таким образом, оценка $X_{(1)}-\frac{1}{n}$ является состоятельной. Вычислим

$D\left(X_{(1)}-\frac{1}{n}\right)$:

$$D\left(X_{(1)}-\frac{1}{n}\right)=D\left(X_{(1)}\right)=-\int_{\theta}^{\infty}x^2de^{-n(x-\theta)}-\left(\theta+\frac{1}{n}\right)^2=-x^2e^{-n(x-\theta)}\Big|_{\theta}^{\infty}+$$

$$+2\int_{\theta}^{\infty}xe^{-n(x-\theta)}dx-\left(\theta+\frac{1}{n}\right)^2=\frac{1}{n^2}.$$

Займемся свойствами оценки $\overline{X}-1$.

$$M\left(\overline{X}-1\right)=M\left(X_1\right)-1=\int_{\theta}^{\infty}xe^{-(x-\theta)}dx-1=-xe^{-(x-\theta)}\Big|_{\theta}^{\infty}+\int_{\theta}^{\infty}e^{-(x-\theta)}dx-1=$$

$$=\theta-e^{-(x-\theta)}\Big|_{\theta}^{\infty}-1=\theta.$$

Таким образом, оценка $\overline{X}-1$ является несмещенной.

$$D\left(\overline{X}-1\right)=D\left(\overline{X}\right)=\frac{1}{n}D\left(X_1\right)=\frac{1}{n}\int_{\theta}^{\infty}x^2e^{-(x-\theta)}dx-\frac{1}{n}(\theta+1)^2=-\frac{1}{n}x^2e^{-(x-\theta)}\Big|_{\theta}^{\infty}+$$

$$+\frac{2}{n}\int_{\theta}^{\infty}xe^{-(x-\theta)}dx-\frac{1}{n}(\theta+1)^2=\frac{1}{n}\theta^2+\frac{2\theta}{n}+\frac{2}{n}-\frac{1}{n}(\theta+1)^2=\frac{1}{n}.$$

Применяя закон больших чисел, и учитывая, что $M\left(X_1\right)=1+\theta$, получим

$$P\left(\left|\overline{X}-1-\theta\right|<\varepsilon\right)=P\left(\frac{1}{n}\sum_{i=1}^{n}X_i-1-\theta\right)=P\left(\frac{1}{n}\sum_{i=1}^{n}X_i-M\left(X_1\right)\right)=$$

$$=P\left(\frac{1}{n}\sum_{i=1}^{n}X_i-\frac{1}{n}\sum_{i=1}^{n}M\left(X_i\right)\right)\to 1,\, n\to\infty.$$

Следовательно, оценка $\overline{X}-1$ является состоятельной.

Сравнивая дисперсии оценок $X_{(1)}-\frac{1}{n}$ и $\overline{X}-1$, видим, что дисперсия $X_{(1)}-\frac{1}{n}$ меньше дисперсии $\overline{X}-1$, а, учитывая, что они обе несмещенные и состоятельные, делаем вывод о предпочтительности оценки $X_{(1)}-\frac{1}{n}$.

Ответ. Выбирается оценка $X_{(1)}-\frac{1}{n}$, поскольку у нее меньше дисперсия, чем у оценки $\overline{X}-1$. Обе оценки несмещенные и состоятельные.

Задачи.

Задача 138. Дана выборка 8;10;12. Найти несмещенные оценки математического ожидания и дисперсии.

Задача 139. Дана выборка 14;18;22. Найти несмещенные оценки математического ожидания и дисперсии.

Задача 140. Дана выборка 13;16;19. Найти несмещенные оценки математического ожидания и дисперсии.

Задача 141. Пусть $x_1, x_2, ..., x_n$ - выборка из равномерного на $[0,\theta]$ распределения с неизвестным параметром θ. В качестве оценки параметра θ предлагается $2\overline{X}$. Охарактеризовать свойства этой оценки.

Задача 142. Рассматривается выборка из распределения с плотностью $\sigma^{-1} f\left(\sigma^{-1}(x-\mu)\right)$, $\mu>0, x>\mu$, где

$$f(x)=\begin{cases} e^{-x}, x>0 \\ 0, x\le 0 \end{cases}.$$

Требуется найти:

а) несмещенную оценку для параметра μ, если известен параметр σ;

б) несмещенную оценку для параметра σ, если известен параметр μ.

Задача 143. Наработки до отказа восьми элементов, выраженные в часах, равны:

10,8; 15,7; 12,0 ;9,2; 13,2; 11,6; 10,4; 12,7.

В предположении, что наработки до отказа всех элементов имеют одно и то же распределение, получить несмещенную оценку средней наработки элемента до отказа.

16. Методы получения точечных оценок неизвестных параметров

В предыдущем разделе мы рассмотрели свойства точечных оценок неизвестных параметров различных законов распределений. Однако прежде, чем изучать свойства точечных оценок, эти оценки необходимо получить.

16.1. Метод моментов

Пусть $x_1, x_2, \dots, x_n$ - выборка, принадлежащая закону распределения с функцией распределения вероятностей $F(x,\theta)$ или с плотностью распределения вероятностей $f(x,\theta)$. Как уже было выяснено выше, эмпирическая функция распределения хорошо оценивает теоретическую функцию распределения. Поэтому следует ожидать, что эмпирические моменты распределения будут служить хорошими оценками для теоретических моментов.

Исходя из этого, суть метода моментов можно сформулировать следующим образом: теоретические моменты приравниваются к эмпирическим моментам, а из полученных уравнений находятся неизвестные параметры распределений.

Будем считать, что выборка $x_1, x_2, ..., x_n$ принадлежит распределению с плотностью $f(x,\theta)$. Тогда для нахождения оценок для неизвестного параметра θ могут быть составлены следующие уравнения, в которые входят начальные эмпирические и теоретические моменты от первого до k-го порядка:

$$M(X_1) = \int_{-\infty}^{+\infty} x f(x,\theta) dx = \overline{X},$$

$$M(X_1^2) = \int_{-\infty}^{+\infty} x^2 f(x,\theta) dx = \overline{X^2},$$

$$\ldots$$

$$M(X_1^k) = \int_{-\infty}^{+\infty} x^k f(x,\theta) dx = \overline{X^k}.$$

Естественный вопрос, который здесь возникает, состоит в том, что поскольку для одного или нескольких неизвестных параметров распределений можно получить множество оценок, то какую из них следует использовать. Ответ на этот вопрос дает предыдущий раздел, в котором рассматриваются свойства оценок. То есть, получая несколько оценок неизвестного параметра по методу моментов, необходимо изучить свойства этих оценок и выбрать из них лучшую.

Следует сразу отметить, что метод моментов не всегда позволяет получить оптимальную (по свойствам) оценку неизвестного параметра распределения. Существуют другие методы получения точечных оценок, о которых речь пойдет ниже. Оценки, полученные разными методами, могут не совпадать и соответственно иметь различные свойства.

Пример. Выборка $x_1, x_2, ..., x_n$ принадлежит равномерному на $[0, \theta]$ распределению. Построить оценку для неизвестного параметра θ.

Решение.

$$M(X_1) = \int_0^\theta x \frac{1}{\theta} dx = \frac{\theta}{2} = \overline{X} \Rightarrow \hat{\theta}_1 = 2\overline{x},$$

$$M(X_1^2) = \int_0^\theta x^2 \frac{1}{\theta} dx = \frac{\theta^2}{3} = \overline{X^2} \Rightarrow \hat{\theta}_2 = \sqrt{3\overline{x^2}}.$$

Получено две оценки для θ (оценок может быть получено и больше, но с ростом порядка момента вид оценки только усложняется, а свойства ее не улучшаются). Ранее мы показали, что оценка $2\overline{X}$ является несмещенной и состоятельной, а оценка $\sqrt{3\overline{X^2}}$ (желающие могут проверить) несмещенной не является. Поэтому по методу моментов для параметра θ выбирается оценка $2\overline{x}$.

Ответ. $\hat{\theta} = 2\overline{x}$.

Вместе с тем, из примера предыдущего раздела следует, что для параметра θ существует оценка $\frac{n+1}{n} X_{(n)}$, которая лучше оценки $2\overline{X}$, поскольку тоже является несмещенной и состоятельной, но при этом имеет меньшую дисперсию.

Таким образом, для равномерного распределения метод моментов позволяет получить несмещенную и состоятельную оценку, но не позволяет получить оценку оптимальную в плане дисперсии.

Пример. Выборка принадлежит показательному закону распределения с плотностью

$$f(x,\alpha)=\begin{cases}\alpha e^{-\alpha x},\ x>0\\ 0,\ x\le 0\end{cases},\ \ \alpha>0.$$

Построить оценки для параметра α, приравнивая начальные эмпирический и теоретический моменты первого и второго порядка.

Решение.

$$M(X_1)=\int_0^\infty x\alpha e^{-\alpha x}dx=-\int_0^\infty x de^{-\alpha x}=-xe^{-\alpha x}\Big|_0^\infty+\int_0^\infty e^{-\alpha x}dx=\frac{1}{\alpha}=\overline{X}\Rightarrow$$

$$\Rightarrow\hat{\alpha}_1=\frac{1}{\overline{x}},$$

$$M\left(X_1^2\right)=\int_0^\infty x^2\alpha e^{-\alpha x}dx=-\int_0^\infty x^2 de^{-\alpha x}=-x^2e^{-\alpha x}\Big|_0^\infty+2\int_0^\infty xe^{-\alpha x}dx=\frac{2}{\alpha^2}=$$

$$=\frac{1}{n}\sum_{i=1}^n X_i^2=\overline{X^2}\Rightarrow\hat{\alpha}_2=\sqrt{\frac{2}{\overline{x^2}}}.$$

Ответ. $\hat{\alpha}_1=\frac{1}{\overline{x}},\ \hat{\alpha}_2=\sqrt{\frac{2}{\overline{x^2}}}.$

Пример. Выборка принадлежит двойному показательному закону распределения с плотностью

$$f(x,\lambda)=\frac{\lambda}{2}e^{-\lambda|x|},\ \ \lambda>0.$$

Построить оценку для параметра λ.

Решение. Вычислим $M(X_1)$:

$$M(X_1)=\int_{-\infty}^{+\infty} x\frac{\lambda}{2}e^{-\lambda|x|}dx=\frac{\lambda}{2}\int_{-\infty}^{0} xe^{\lambda x}dx+\frac{\lambda}{2}\int_{0}^{+\infty} xe^{-\lambda x}dx=\frac{1}{2}xe^{\lambda x}\Big|_{-\infty}^{0}-\frac{1}{2}\int_{-\infty}^{0} e^{\lambda x}dx-$$
$$-\frac{1}{2}xe^{-\lambda x}\Big|_{0}^{+\infty}+\frac{1}{2}\int_{0}^{+\infty} e^{-\lambda x}dx=0.$$

Вычислим $M\left(X_1^2\right)$:

$$M\left(X_1^2\right)=\int_{-\infty}^{+\infty} x^2\frac{\lambda}{2}e^{-\lambda|x|}dx=\frac{\lambda}{2}\int_{-\infty}^{0} x^2e^{\lambda x}dx+\frac{\lambda}{2}\int_{0}^{+\infty} x^2e^{-\lambda x}dx=\frac{1}{2}x^2e^{\lambda x}\Big|_{-\infty}^{0}-\int_{-\infty}^{0} xe^{\lambda x}dx-$$
$$-\frac{1}{2}x^2e^{-\lambda x}\Big|_{0}^{+\infty}+\int_{0}^{+\infty} xe^{-\lambda x}dx=\frac{2}{\lambda^2}=\overline{X^2}\ \Rightarrow\hat{\lambda}=\sqrt{\frac{2}{\overline{\overline{X^2}}}}.$$

Ответ. $\hat{\lambda}=\sqrt{\frac{2}{\overline{\overline{x^2}}}}$.

Пример. Выборка принадлежит равномерному на $[a,b]$ распределению. Найти оценки для параметров a и b.

Решение. Вычислим $M(X_1)$ и $M\left(X_1^2\right)$:

$$M(X_1)=\int_a^b x\frac{1}{b-a}dx=\frac{a+b}{2}=\overline{X},$$

$$M\left(X_1^2\right)=\int_a^b x^2\frac{1}{b-a}dx=\frac{x^3}{3(b-a)}\Big|_a^b=\frac{a^2+ab+b^2}{3}=\overline{X^2}\quad .$$

Получаем систему:

$$\begin{cases} a+b=2\overline{X} \\ a^2+ab+b^2=3\overline{X^2} \end{cases}.$$

Из первого уравнения $b=2\overline{X}-a$, подставляя во второе, получим:

$$a^2+a\left(2\overline{X}-a\right)+\left(2\overline{X}-a\right)^2=3\overline{X^2}, a^2-2a\overline{X}+4\overline{X}^2-3\overline{X^2}=0,$$
$$a_{1,2}=\overline{X}\pm\sqrt{3S_0^2},\ b_{1,2}=\overline{X}\pm\sqrt{3S_0^2}\Rightarrow\hat{a}=\overline{x}-\sqrt{3s_0^2},\ \hat{b}=\overline{x}+\sqrt{3s_0^2}.$$

Ответ. $\hat{a}=\overline{x}-\sqrt{3s_0^2}$, $\hat{b}=\overline{x}+\sqrt{3s_0^2}$.

Пример. Выборка принадлежит нормальному распределению с параметрами $\left(a, \sigma^2\right)$. Построить оценки для a и σ^2.

Решение. $M\left(X_1\right) = a = \overline{X} \Rightarrow \hat{a} = \overline{x}$, $M\left(X_1^2\right) = a^2 + \sigma^2 = \overline{x^2} \Rightarrow$

$\Rightarrow \hat{\sigma}^2 = \overline{x^2} - \overline{x}^2 = s_0^2$.

Как мы уже знаем, оценка $\overline{X}$ является несмещенной и состоятельной, а оценка S_0^2 является асимптотически несмещенной и состоятельной.

Ответ. $\hat{a} = \overline{x}$, $\hat{\sigma}^2 = s_0^2$.

Пример. Выборка принадлежит гамма – распределению с плотностью

$$f(x,p) = \begin{cases} [\Gamma(p)]^{-1} x^{p-1} e^{-x}, x > 0 \\ 0, x \le 0 \end{cases},$$

где $\Gamma(p) = \int\limits_0^\infty x^{p-1} e^{-x} dx$ - гамма – функция Эйлера. Найти оценку для параметра p.

Решение. Вычислим начальные моменты k-го порядка:

$$M\left(X_1^k\right) = \int\limits_0^\infty x^k f(x,p) dx = \frac{1}{\Gamma(p)} \int\limits_0^\infty x^k x^{p-1} e^{-x} dx = \frac{\Gamma(p+k)}{\Gamma(p)}.$$

Из определения $\Gamma(p)$, интегрированием по частям, нетрудно показать, что

$$\Gamma(p+1) = \Gamma(p),$$

откуда

$$M(X_1) = \frac{\Gamma(p+1)}{\Gamma(p)} = p = \overline{X} \Rightarrow \hat{p} = \overline{x}.$$

Ответ. $\hat{p} = \overline{x}$.

Задачи.

Задача 144. Выборка - 3, - 2, - 1, - 1, - 2, - 6, - 4, - 5, - 3 принадлежит равномерному на $[-\theta, 0]$ распределению. Найти методом моментов оценку для неизвестного параметра θ.

Задача 145. Выборка 9, 8, 3, 9, 6, 4, 5, 12 принадлежит равномерному на $[0, \theta]$ распределению. Найти методом моментов оценку для неизвестного параметра θ.

Задача 146. Выборка - 3, - 6, - 13, - 10, - 8, - 6, - 4, - 1, - 13 принадлежит равномерному на $[-\theta, 0]$ распределению. Найти методом моментов оценку для неизвестного параметра θ.

Задача 147. Выборка 3, 8, 2, 1, 8, 6, 8, 5, 7 принадлежит равномерному на $[0, \theta]$ распределению. Найти методом моментов оценку для неизвестного параметра θ.

Задача 148. Выборка принадлежит равномерному на $[-\theta, 0]$ распределению. Построить методом моментов оценку для параметра θ.

Задача 149. Выборка 0.15, 1.26, 2.34, 1.10, 2.56, 6.23, 4.78, 5.94, 3.54 принадлежит равномерному на $[0, \theta]$ распределению. Найти методом моментов оценку для неизвестного параметра θ.

Задача 150. Выборка 1.35, 1.68, 2.84, 1.19, 2.26, 6.78, 4.08, 5.62, 3.88 принадлежит равномерному на $[1, \theta]$ распределению. Найти методом моментов оценку для неизвестного параметра θ.

Задача 151. Выборка -0.15, -1.26, -2.34, -1.10, -2.56, -6.23, -4.78, -5.94, -3.54 принадлежит равномерному на $[-\theta, 0]$ распределению. Найти методом моментов оценку для неизвестного параметра θ.

16.2. Метод максимального правдоподобия

Предположим, что в результате эксперимента случайная величина X приняла значения

$$(x_1, x_2, \ldots, x_n) = x.$$

Обозначим через $F(y,\theta)$ функцию распределения вероятностей случайной величины X, параметр θ неизвестен.

Если X непрерывна, то через $f(y,\theta)$ обозначим плотность распределения ее вероятностей, то есть

$$F(y,\theta) = P(X \le y), F(y,\theta) = \int_{-\infty}^{y} f(z,\theta)dz.$$

Если случайная величина X дискретна, то обозначим через $\mathrm{P}(x_i,\theta)$ ее распределение вероятностей.

Функцией правдоподобия называется

$$L(x,\theta) = \prod_{i=1}^{n} \mathrm{P}(x_i,\theta)$$

в случае дискретной случайной величины и

$$L(x,\theta) = \prod_{i=1}^{n} f(x_i,\theta)$$

в случае непрерывной случайной величины.

Значение функции $L(x,\theta)$ при данном наблюдении x называется **правдоподобием** θ при этом x. Значительное практическое и теоретическое удобство достигается переходом к логарифмам.

Логарифмической функцией правдоподобия называют

$$l(x,\theta) = \ln L(x,\theta).$$

В схеме повторной независимой выборки

$$l(x,\theta) = \sum_{i=1}^{n} ln\, f(x_i,\theta) \quad (l(x,\theta) = \sum_{i=1}^{n} ln\, P(x_i,\theta)).$$

По методу максимума правдоподобия (максимального правдоподобия) за оценку неизвестного параметра θ берется то его значение, при котором $l(x,\theta)$ достигает максимума (функции $L(x,\theta)$ и $l(x,\theta)$ достигают максимума при одном и том же значении θ).

Следует отметить, что оценки неизвестных параметров, полученные разными методами, не обязательно совпадают. В частности оценки, полученные методом моментов, совсем не обязательно совпадают с оценками, полученными методом максимального правдоподобия.

Если у неизвестных параметров существуют эффективные оценки, то они получаются методом максимального правдоподобия.

Оценки неизвестных параметров, полученные методом максимального правдоподобия, состоятельны, хотя могут быть и смещенными.

Пример. Выборка $x_1, x_2, ..., x_n$ принадлежит распределению Бернулли с неизвестной вероятностью θ положительного исхода. Найти оценку максимального правдоподобия для параметра θ.

Решение. Случайная величина X имеет распределение Бернулли, если $P(X=1)=\theta, P(X=0)=1-\theta$. Таким образом, выборка $x_1, x_2, ..., x_n$ представляет собой набор из 1 и 0. Число единиц в этой выборке равно $\sum_{i=1}^{n} x_i$, а число нулей равно $n-\sum_{i=1}^{n} x_i$, $P(X_i=1)=\theta^{x_i}, P(X_i=0)=(1-\theta)^{1-x_i}, i=\overline{1,n}$.

Отсюда следует, что функция правдоподобия имеет вид:

$$L(x,\theta)=\theta^{\sum\limits_{i=1}^{n} x_i}\cdot(1-\theta)^{n-\sum\limits_{i=1}^{n} x_i}.$$

Логарифмическая функция правдоподобия выглядит следующим образом:

$$l(x,\theta)=\left(\sum_{i=1}^{n} x_i\right)\cdot\ln\theta+\left(n-\sum_{i=1}^{n} x_i\right)\cdot\ln(1-\theta).$$

Для нахождения точки максимума логарифмической функции правдоподобия найдем ее производную и приравняем к нулю.

$$\frac{\partial l(x,\theta)}{\partial\theta}=\left(\sum_{i=1}^{n} x_i\right)\cdot\theta^{-1}-\left(n-\sum_{i=1}^{n} x_i\right)\cdot(1-\theta)^{-1}=\left(\sum_{i=1}^{n} x_i\right)(\theta(1-\theta))^{-1}-n(1-\theta)^{-1}=.$$

$$= \left(\sum_{i=1}^{n} x_i \cdot \theta^{-1} - n \right) \cdot (1-\theta)^{-1} = 0.$$

Из этого уравнения находим

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

Полученное решение действительно является точкой максимума, поскольку производная $\frac{\partial l(x,\theta)}{\partial \theta}$ меняет знак с плюса на минус при переходе через эту точку. Поэтому оценкой максимального правдоподобия неизвестного параметра θ является $\hat{\theta} = \overline{x}$.

Ответ. $\hat{\theta} = \overline{x}$.

Пример. Выборка $x_1, x_2, ..., x_n$ принадлежит распределению Пуассона с неизвестным параметром λ. Найти оценку максимального правдоподобия для λ.

Решение. Как известно,

$$P(X = x_i) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}.$$

Поэтому функция правдоподобия имеет вид:

$$L(x, \lambda) = \frac{\lambda^{x_1} e^{-\lambda}}{x_1!} \cdot \frac{\lambda^{x_2} e^{-\lambda}}{x_2!} \cdot \ldots \cdot \frac{\lambda^{x_n} e^{-\lambda}}{x_n!} = \frac{\lambda^{\sum_{i=1}^{n} x_i} e^{-n\lambda}}{x_1! x_2! \ldots x_n!}.$$

Логарифмическая функция правдоподобия имеет вид:

$$l(x, \lambda) = \left(\sum_{i=1}^{n} x_i \right) \cdot \ln \lambda - n\lambda - \ln(x_1! x_2! \ldots x_n!).$$

Для нахождения точки максимума логарифмической функции правдоподобия найдем ее производную и приравняем к нулю.

$$\frac{\partial l(x,\lambda)}{\partial \lambda} = \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0.$$

Отсюда критическая точка:

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

Найдем вторую производную функции $l(x,\lambda)$ по λ и убедимся, что она отрицательна в этой точке:

$$\frac{\partial^2 l(x,\lambda)}{\partial \lambda^2}\Big|_{\lambda=\overline{x}} = -\frac{\sum_{i=1}^{n} x_i}{\lambda^2}\Big|_{\lambda=\overline{x}} = -\frac{n^2}{\sum_{i=1}^{n} x_i} < 0.$$

Следовательно, $\overline{x}$ действительно является точкой максимума функции $l(x,\lambda)$ и оценкой максимального правдоподобия для неизвестного параметра λ.

Ответ. $\hat{\lambda} = \overline{x}$.

Пример. Выборка $x_1, x_2, ..., x_n$ принадлежит нормальному распределению с неизвестными параметрами (a,σ^2). Найти оценки максимального правдоподобия для этих параметров.

Решение. Плотность распределения вероятностей нормальной случайной величины с параметрами (a,σ^2) имеет вид:

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

Найдем функцию правдоподобия

$$L(x,a,\sigma^2) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-a)^2}$$

и логарифмическую функцию правдоподобия

$$l\left(x,a,\sigma^2\right)=-n\ln\sqrt{2\pi\sigma^2}-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-a\right)^2=-n\ln\sqrt{2\pi\sigma^2}-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2-$$

$$-\frac{n\left(\overline{x}-a\right)^2}{2\sigma^2}.$$

Найдем $\dfrac{\partial l\left(x,a,\sigma^2\right)}{\partial a}$ и $\dfrac{\partial l\left(x,a,\sigma^2\right)}{\partial \sigma^2}$.

$$\frac{\partial l\left(x,a,\sigma^2\right)}{\partial a}=\frac{n\left(\overline{x}-a\right)}{\sigma^2},$$

$$\frac{\partial l\left(x,a,\sigma^2\right)}{\partial \sigma^2}=-\frac{n}{2\sigma^2}+\frac{\left(\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2+n\left(\overline{x}-a\right)^2\right)}{2\sigma^4}.$$

Приравнивая к нулю найденные производные, найдем:

$$a=\overline{x},\ \sigma^2=\frac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{n}=s_0^2.$$

Для того, чтобы убедиться, что найденная точка является точкой максимума, покажем, что в этой точке

$$\frac{\partial^2 l\left(x,a,\sigma^2\right)}{\partial a^2}\cdot\frac{\partial^2 l\left(x,a,\sigma^2\right)}{\partial\left(\sigma^2\right)^2}-\frac{\partial^2 l\left(x,a,\sigma^2\right)}{\partial a\partial\sigma^2}>0.$$

Исключая случай $x_i=\overline{x}, i=\overline{1,n}$, при $a=\overline{x},\ \sigma^2=\dfrac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}{n}$

получим:

$$\frac{\partial^2 l\left(x,a,\sigma^2\right)}{\partial a^2}=-\left(\sigma^2\right)^{-1}=-\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2\right)^{-1}<0,$$

$$\frac{\partial^2 l\left(x,a,\sigma^2\right)}{\partial\left(\sigma^2\right)^2}=\left(\frac{n}{2\sigma^4}-\frac{1}{\sigma^6}\left(\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2+n\left(\overline{x}-a\right)^2\right)\right)=\frac{n}{2\sigma^4}-\frac{n}{\sigma^4}<0,$$

$$\frac{\partial^2 l\left(x,a,\sigma^2\right)}{\partial a\partial\sigma^2}=-\frac{n\left(\overline{x}-a\right)}{\sigma^4}=0.$$

Таким образом

$$\frac{\partial^2 l(x,a,\sigma^2)}{\partial a^2}\cdot\frac{\partial^2 l(x,a,\sigma^2)}{\partial(\sigma^2)^2}-\frac{\partial^2 l(x,a,\sigma^2)}{\partial a\partial\sigma^2}>0$$

и, следовательно, в точке $a=\overline{x},\ \sigma^2=\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}$ функция правдоподобия достигает максимума.

Ответ. $\hat{a}=\overline{x},\ \hat{\sigma}^2=s_0^2$.

Пример. Выборка $x_1,x_2,...,x_n$ принадлежит экспоненциальному распределению с плотностью $\sigma^{-1}f\left(\frac{x-\mu}{\sigma}\right)$, где

$$f(x)=\begin{cases}e^{-x}, x>0\\ 0, x\le 0\end{cases},\ x\ge\mu,\ -\infty<\mu<+\infty,\ \sigma>0.$$

Построить оценки максимального правдоподобия для неизвестных параметров μ и σ.

Решение. $L(x,\mu,\sigma)=\frac{1}{\sigma^n}\cdot e^{-\frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma}}$, $l(x,\mu,\sigma)=-n\ln\sigma-\frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma}$.

Очевидно, что при любом σ максимум функции $l(x,\mu,\sigma)$ по μ достигается при $\mu=\min\limits_{1\le i\le n}x_i=x_{(1)}$. Далее, из уравнения

$$\frac{\partial l(x,\mu,\sigma)}{\partial\sigma}=-\frac{n}{\sigma}+\frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2}=0$$

находим $\sigma=\frac{\sum_{i=1}^{n}(x_i-x_{(1)})}{n}$.

Таким образом, оценками максимального правдоподобия являются

$$\hat{\mu} = x_{(1)},\ \hat{\sigma} = \frac{\sum_{i=1}^{n}\left(x_i - x_{(1)}\right)}{n}.$$

Ответ. $\hat{\mu} = x_{(1)},\ \hat{\sigma} = \dfrac{\sum_{i=1}^{n}\left(x_i - x_{(1)}\right)}{n}$.

Задачи.

Задача 152. Выборка $x_1, x_2, ..., x_n$ принадлежит показательному распределению с плотностью распределения вероятностей

$$f(x, \lambda) = \lambda e^{-\lambda x}, x > 0, \lambda > 0.$$

Найти оценку максимального правдоподобия для неизвестного параметра λ.

Задача 153. Выборка $x_1, x_2, ..., x_n$ принадлежит равномерному на $[0, \theta]$ распределению. Построить методом максимального правдоподобия оценку для неизвестного параметра θ.

Задача 154. Дана выборка из равномерного на $[0, \theta]$ распределения: 0.10, 0.12, 0.11, 0.15, 0.20, 0.27, 0.23, 0.34, 0.60, 0.54, 0.48, 0.19, 0.38, 0.41. Найти методом максимального правдоподобия несмещенную и состоятельную оценку для неизвестного параметра θ. Решение обосновать.

Задача 155. Выборка 4. 11.45, 4.64, 2.46, 2.78, 3.66, 7.22, 6.44, 5.11, 9.47, 2.33, 1.55, 8.88 принадлежит распределению Пуассона с неизвестным параметром λ. Найти оценку максимального правдоподобия для λ.

Задача 156. Выборка 1, 1, 1, 0, 1, 0, 0, 0, 1, 1, принадлежит распределению Бернулли с неизвестной вероятностью θ положительного исхода. Найти оценку максимального правдоподобия для параметра θ.

17. Доверительное оценивание

В предыдущем разделе по результатам наблюдений $x_1, x_2, ..., x_n$ предлагались точечные оценки неизвестных параметров распределений. В настоящем разделе эти оценки будут использоваться при построении доверительных интервалов для неизвестных параметров различных распределений.

В ряде задач требуется не только найти для параметра θ подходящую оценку $\hat{\theta}$, но и указать, к каким ошибкам может привести замена параметра θ его оценкой $\hat{\theta}$. Другими словами, требуется оценить точность и надежность оценки. Такого рода задачи особенно актуальны при малом числе наблюдений, когда точечная оценка $\hat{\theta}$ в значительной мере случайна и приближенная замена θ на $\hat{\theta}$ может привести к серьезным ошибкам.

Для определения **точности** оценки $\hat{\theta}$ в математической статистике пользуются **доверительными интервалами**, а для определения **надежности - доверительными вероятностями.**

Интервал (θ_1, θ_2) называется α**-доверительным интервалом** или **доверительным интервалом с доверительной вероятностью** $1-\alpha$, если

$$P(\theta_1 \le x \le \theta_2) = 1-\alpha .$$

Чем меньше выбранная доверительная вероятность, тем меньше (уже) интервал для неизвестного параметра θ.

Чем больше доверительная вероятность, тем больше (шире) доверительный интервал параметра θ.

Если доверительная вероятность близка к 1, то интервальная оценка (доверительный интервал) неизвестного параметра мало пригодна для практики. Например, при оценивании дисперсии в генеральной совокупности можно было бы выбрать границы $\theta_H(x_1, x_2, ..., x_n) = 0$, $\theta_B(x_1, x_2, ..., x_n) = \infty$ (независимо от конкретных значений) и получить доверительный интервал $0 < \sigma^2 < \infty$, имеющий доверительную вероятность 1. Хотя в этом случае никакой пользы от такого доверительного оценивания мы не извлечем, так как наши знания о значении σ^2 не возрастут, поскольку мы и так знаем, что для любой случайной величины $0 < \sigma^2 < \infty$ и только для константы дисперсия равна нулю.

Назовем p - **квантилью** непрерывной функции распределения $F(x)$ решение $x_p = x_p(F)$ уравнения

$$F(x_p) = p, \; 0 < p < 1. \tag{11}$$

Для $p = \frac{1}{2}$ x_p называется медианой распределения, для $p = \frac{1}{4}$ и $p = \frac{3}{4}$ употребляется название квартиль. Если функция $F(x)$ строго монотонна, то $x_p = F^{-1}(p)$ определяется однозначно. В противном случае для некоторых p решением уравнения (11) является целый отрезок $[\underline{x}, \overline{x}]$ значений x_p. Так как $F(\underline{x}) = F(\overline{x}) = p$, то

$$\mathrm{P}(X \in [\underline{x}, \overline{x}]) = F(\overline{x}) - F(\underline{x}) = 0.$$

С точки зрения теории вероятностей значения x из отрезка $[\underline{x}, \overline{x}]$ вообще можно не принимать во внимание. Это рассуждение проведено для того, чтобы показать, что неоднозначность уравнения (11) несущественна. Чтобы устранить связанные с этой неоднозначностью формальные неудобства, можно принять за x_p при $p \neq \frac{1}{2}$ $x_p = \underline{x}$, а за $x_{\frac{1}{2}}$ середину отрезка $[\underline{x}, \overline{x}]$.

Замечания.

- Утверждение, что параметр θ о вероятностью $(1-\alpha)$ лежит в интервале $[\theta_H, \theta_B]$, не равнозначно утверждению, что вероятность попадания значения θ внутрь интервала $[\theta_H, \theta_B]$ равна $(1-\alpha)$, так как θ не является случайной величиной и дос-товерно лежит либо внутри, либо вне интервала.

- Границы доверительного интервала расширяется при уменьшении объема выборки я увеличении доверительной вероятности.

- Выбор доверительной вероятности не является математической задачей, а определяется конкретно решаемой задачей.

- Иногда на практике представляет интерес лишь один из двух доверительных пределов. В этом случае определяются односторонние доверительные интервалы:

$$P[\theta < \theta_B] = 1 - \alpha\ ,$$

$$P[\theta > \theta_H] = 1 - \alpha\ .$$

Пример. Можно ли в рамках математической теорииf не интересуясь природой выпускаемых изделий, решить вопрос о том, мала или велика вероятность α?

Решение. Допустим, что на двух предприятиях вероятность выпуска годных изделий $p = 1 - \alpha = 0{,}99$, т.е. вероятность выпуска бракованных изделий $\alpha = 0{,}01$.

Пусть одно предприятие выпускает электролампы, а другое - парашюты. Если на 100 ламп встретится одна бракованная, то с этим можно мириться при условии, что выбросить 1 % ламп дешевле, чем перестроить технологический процесс. Если же на 100 парашютов встретится один бракованный, что может повлечь за собой серьезные последствия, то такое положение неприемлемо. Следовательно, в первом случае вероятность брака приемлема, а во втором - нет, поэтому выбор доверительной вероятности $p = 1-\alpha$ следует производить, исходя из конкретных условий задачи.

Напомним (раздел 13), что случайная величина $\dfrac{\sqrt{n}\left(\overline{X}-a\right)}{\sigma}$ имеет распределение $N(0,1)$, случайная величина $\dfrac{nS^2}{\sigma^2} = \sum\limits_{i=1}^{n}\left(\dfrac{X_i - a}{\sigma}\right)^2$ имеет распределение χ_n^2, случайные величины $\dfrac{nS_0^2}{\sigma^2}$ и $\dfrac{(n-1)S_1^2}{\sigma^2}$ имеют распределение χ_{n-1}^2, а случайные величины $\sqrt{n-1}\cdot\dfrac{\overline{X}-a}{S_0} = \dfrac{\sqrt{n}\cdot\dfrac{\overline{X}-a}{\sigma}}{\dfrac{\dfrac{\sqrt{n}S_0}{\sigma}}{\sqrt{n-1}}}$ и

$\sqrt{n}\cdot\dfrac{\overline{X}-a}{S_1} = \dfrac{\sqrt{n}\cdot\dfrac{\overline{X}-a}{\sigma}}{\dfrac{\dfrac{\sqrt{n-1}S_1}{\sigma}}{\sqrt{n-1}}}$ имеют распределение t_{n-1}, где

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2,\ S_0^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,\ S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,\ a = M(X_1).$$

Пример. Выборка $x_1, x_2, \ldots, x_n$ принадлежит нормальному распределению с параметрами $\left(a, \sigma^2\right)$, σ^2 - известно, a - неизвестный параметр. Построить α - доверительный интервал для a.

Решение. Хорошо известно, что точечной оценкой параметра a (математического ожидания), является выборочное среднее $\overline{x}$. Поэтому рассмотрим случайную величину $\dfrac{\left(\overline{X}-a\right)\sqrt{n}}{\sigma} \in N(0,1)$.

$$\mathrm{P}\left(\left|\frac{\left(\overline{X}-a\right)\sqrt{n}}{\sigma}\right| \le x_{\frac{1-\alpha}{2}}\right) = \mathrm{P}\left(-x_{\frac{1-\alpha}{2}} \le \frac{\left(\overline{X}-a\right)\sqrt{n}}{\sigma} \le x_{\frac{1-\alpha}{2}}\right) = 2\Phi\left(x_{\frac{1-\alpha}{2}}\right) = 1-\alpha.$$

$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{y^2}{2}} dy$ - функция Лапласа.

Значение $x_{\frac{1-\alpha}{2}}$ определяется из таблицы для функции Лапласа (приложение 2) и соответствует значению $\Phi\left(x_{\frac{1-\alpha}{2}}\right) = \frac{1-\alpha}{2}$.

Таблицы могут быть составлены как для функции Лапласа $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{y^2}{2}} dy$, так и для функции распределения вероятностей стандартного нормального распределения $F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy$. Значения этих функций связаны следующим образом:

$$\Phi(x_p) = F(x_p) - \frac{1}{2} = p - \frac{1}{2}.$$

Возвращаясь к неизвестному параметру a для нормального распределения, получим

$$-x_{\frac{1-\alpha}{2}} \le \frac{(\overline{X} - a)\sqrt{n}}{\sigma} \le x_{\frac{1-\alpha}{2}},$$

$$-\overline{X}\sqrt{n} - x_{\frac{1-\alpha}{2}}\sigma \le -a\sqrt{n} \le -\overline{X}\sqrt{n} + x_{\frac{1-\alpha}{2}}\sigma,$$

$$\overline{X} - \frac{x_{\frac{1-\alpha}{2}}\sigma}{\sqrt{n}} \le a \le \overline{X} + \frac{x_{\frac{1-\alpha}{2}}\sigma}{\sqrt{n}}.$$

Получаем доверительный интервал для математического ожидания a с уровнем доверия (с надежностью) $1-\alpha$ (с известной дисперсией)

$$\overline{x} - \frac{x_{\frac{1-\alpha}{2}}\sigma}{\sqrt{n}} \le a \le \overline{x} + \frac{x_{\frac{1-\alpha}{2}}\sigma}{\sqrt{n}}. \qquad (12)$$

Значение $\dfrac{x_{\frac{1-\alpha}{2}}\sigma}{\sqrt{n}}$ **называется точностью оценки математического ожидания**.

Доверительный интервал для математического ожидания всегда симметричен относительно $\bar{x}$. Если доверительная вероятность увеличивается, то доверительный интервал становится шире, оставаясь при этом симметричным относительно $\bar{x}$. Если доверительная вероятность уменьшается, то доверительный интервал становится уже, оставаясь при этом симметричным относительно $\bar{x}$. Другими словами, $\bar{x}$ является серединой доверительного интервала. Чтобы проверить это, нужно сложить концы интервала и поделить пополам.

Чтобы у читателей не было путаницы в употреблении $\overline{X}$ и $\bar{x}$, то поясним еще раз, что с точки зрения математической статистики $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ является случайной величиной и используется при построении вероятностной модели $X_1, X_2, ..., X_n$ числовой выборки $x_1, x_2, ..., x_n$, а $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ есть число. При построении доверительных интервалов допустимо употребление $\overline{X}$ и $\bar{x}$, но с точки зрения автора более понятна окончательная запись доверительного интервала с $\bar{x}$ в форме (12).

Ответ. $\bar{x} - \dfrac{x_{\frac{1-\alpha}{2}}\sigma}{\sqrt{n}} \le a \le \bar{x} + \dfrac{x_{\frac{1-\alpha}{2}}\sigma}{\sqrt{n}}$.

Пример. Выборка $x_1, x_2, ..., x_n$ принадлежит нормальному распределению с параметрами (a, σ^2), a - известно, σ^2 - неизвестный параметр. Построить α - доверительный интервал для σ^2.

Решение. Известно, что точечной оценкой параметра σ^2 (дисперсии), является выборочная дисперсия s_0^2, но поскольку известно a, то рассмотрим статистику $\sum_{i=1}^{n}\left(\frac{X_i-a}{\sigma}\right)^2 \in \chi_n^2$.

Таким образом:

$$P\left(\delta_{n,1-\frac{\alpha}{2}} \le \frac{\sum_{i=1}^{n}(x_i-a)}{\sigma^2} \le \delta_{n,\frac{\alpha}{2}}\right)=1-\alpha,$$

$\delta_{n,1-\frac{\alpha}{2}}, \delta_{n,\frac{\alpha}{2}}$ определяются из таблицы для вероятностей $P\left(\chi_n{}^2 > \delta_{n,\gamma}\right)=\gamma$ (приложение 4).

$$\delta_{n,1-\frac{\alpha}{2}} \le \frac{\sum_{i=1}^{n}(x_i-a)}{\sigma^2} \le \delta_{n,\frac{\alpha}{2}},$$

$$\delta^{-1}_{n,\frac{\alpha}{2}}\sum_{i=1}^{n}(x_i-a)^2 \le \sigma^2 \le \delta^{-1}_{n,1-\frac{\alpha}{2}}\sum_{i=1}^{n}(x_i-a)^2.$$

Ответ. $\delta^{-1}_{n,\frac{\alpha}{2}}\sum_{i=1}^{n}(x_i-a)^2 \le \sigma^2 \le \delta^{-1}_{n,1-\frac{\alpha}{2}}\sum_{i=1}^{n}(x_i-a)^2$.

Пример. Выборка $x_1, x_2, ..., x_n$ принадлежит нормальному распределению с параметрами (a,σ^2). Построить α- доверительные интервалы для неизвестных параметров a и σ^2.

Решение. Начнем построение с параметра a. Как известно, точечными оценками для математического ожидания и дисперсии являются соответственно $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ и $S_0^2=\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)$ ($S_1^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)$.

Хорошо известно, что $\frac{(\overline{X}-a)\sqrt{n}}{\sigma} \in N(0{,}1)$, но σ^2 неизвестно, поэтому пользоваться статистикой $\frac{(\overline{X}-a)\sqrt{n}}{\sigma}$ с целью построения доверительного интервала для параметра a не представляется возможным. Поэтому будем использовать статистику

$$\sqrt{n}\cdot\frac{\overline{X}-a}{S_1} = \frac{\sqrt{n}\cdot\frac{\overline{X}-a}{\sigma}}{\frac{\frac{\sqrt{n-1}S_1}{\sigma}}{\sqrt{n-1}}}$$

(или $\sqrt{n}\cdot\frac{\overline{X}-a}{S_1} = \frac{\sqrt{n}\cdot\frac{\overline{X}-a}{\sigma}}{\frac{\frac{\sqrt{n-1}S_1}{\sigma}}{\sqrt{n-1}}}$), которая имеют распределение t_{n-1}

(распределение Стьюдента с $n-1$ степенью свободы).

Таким образом

$$P\left(\left|\frac{\sqrt{n}(\overline{X}-a)}{S_1}\right| \le \Delta_{n-1,\alpha}\right) = 1-\alpha,$$

где $\Delta_{n-1,\alpha}$ находится из таблицы для вероятностей $P(|t_{n-1}| > \Delta_{n-1,\alpha}) = \alpha$ распределения t_{n-1} (приложение 3).

Таким образом, α-доверительный интервал для a:

$$\bar{x} - \frac{\Delta_{n-1,\alpha}s_1}{\sqrt{n}} \le a \le \bar{x} + \frac{\Delta_{n-1,\alpha}s_1}{\sqrt{n}} \quad \left(\frac{\Delta_{n-1,\alpha}s_1}{\sqrt{n}}\right.$$ -**точность оценки**),

а α-доверительный интервал для σ^2:

$$\delta_{n-1,1-\frac{\alpha}{2}} \le \frac{(n-1)s_1^2}{\sigma^2} \le \delta_{n-1,\frac{\alpha}{2}},$$

$$\delta^{-1}_{n-1,\frac{\alpha}{2}}(n-1)s_1^{\ 2} \le \sigma^2 \le \delta^{-1}_{n-1,1-\frac{\alpha}{2}}(n-1)s_1^{\ 2}$$

или

$$\delta^{-1}_{n-1,1-\frac{\alpha}{2}}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2 \le \sigma^2 \le \delta^{-1}_{n-1,\frac{\alpha}{2}}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2$$

Ответ. $$\overline{x}-\frac{\Delta_{n-1,\alpha}s_1}{\sqrt{n}} \le a \le \overline{x}+\frac{\Delta_{n-1,\alpha}s_1}{\sqrt{n}},$$

$$\delta^{-1}_{n-1,\frac{\alpha}{2}}(n-1)s_1^{\ 2} \le \sigma^2 \le \delta^{-1}_{n-1,1-\frac{\alpha}{2}}(n-1)s_1^{\ 2}.$$

Пример. Из нормально распределенной генеральной совокупности извлечена выборка объема $n=10$, представленная в таблице 15.

Таблица 15. Группированная выборка.

x_i	-2	1	2	3	4	5
n_i	2	1	2	2	2	1

Оценить с надежностью 0,95 математическое ожидание m нормально распределенного признака генеральной совокупности.

Решение. Найдем $\bar{x}=2$; $s_1=\sqrt{s_1^2}=2{,}4$. Кроме того, $1-\alpha=0{,}95$. Следовательно, $\alpha=0{,}05$. Из таблицы приложения 4 для распределения Стьюдента $\Delta_{9;0,05}=2{,}262$. Найдем искомый доверительный интервал:

$$2-\frac{2{,}4}{\sqrt{10}}\cdot 2{,}262 < m < 2+\frac{2{,}4}{\sqrt{10}}\cdot 2{,}262$$

Откуда

$$0{,}3 < m < 3{,}7.$$

Задачи.

Задача 157. $\bar{x} = 21.5$. Какой из интервалов может быть доверительным интервалом для математического ожидания

а) $(-10;30)$, б) $(19;24)$, в) $(20;24)$, г) $(-22;22)$?

Задача 158. $\bar{x} = 6$. Какой из интервалов может быть доверительным интервалом для математического ожидания

а) $(0;10)$, б) $(-6;6)$, в) $(2;14)$, г) $(3;9)$?

Задача 159. Интервал $(15;25)$ является доверительным интервалом для неизвестного математического ожидания. Найти точность полученной интервальной оценки?

Задача 160. Интервал $(8;10)$ является доверительным интервалом для неизвестного математического ожидания. Найти точность полученной интервальной оценки?

Задача 161. Интервал $(12;14)$ является доверительным интервалом для неизвестного математического ожидания при определенном уровне надежности (доверительной вероятности) γ. Каким может быть доверительный интервал для неизвестного математического ожидания при увеличении γ

а) $(10;13)$, б) $(11;14)$, в) $(12;13)$, г) $(11;15)$?

Задача 162. Интервал $(4;8)$ является доверительным интервалом для неизвестного математического ожидания при определенном уровне надежности (доверительной вероятности) γ. Каким может быть доверительный интервал для неизвестного математического ожидания при увеличении γ

а) $(5;7)$, б) $(2;10)$, в) $(1;9)$, г) $(3;15)$?

Задача 163. Интервал $(2;4)$ является доверительным интервалом для неизвестного математического ожидания при определенном уровне надежности (доверительной вероятности) γ. Каким может быть доверительный интервал для неизвестного математического ожидания при уменьшении γ

а) $(1;5)$, б) $(1;3)$, в) $(2;6)$, г) $(2.5;3.5)$?

Задача 164. Интервал $(-3;3)$ является доверительным интервалом для неизвестного математического ожидания при определенном уровне надежности (доверительной вероятности) γ. Каким может быть доверительный интервал для неизвестного математического ожидания при уменьшении γ

а) $(1;5)$, б) $(-1;1)$, в) $(-2;3)$, г) $(-2;2)$?

Задача 165. Дан объем выборки $n=25$ из нормального распределения с неизвестными математическим ожиданием и дисперсией, $\Delta_{n-1,\alpha}=2{,}064$, $s_1^2=9$. Найти точность интервальной оценки для неизвестного математического ожидания.

Задача 166. Дан объем выборки $n=9$ из нормального распределения с неизвестными математическим ожиданием и дисперсией, $\Delta_{n-1,\alpha}=2{,}064$, $s_1^2=4$. Найти точность интервальной оценки для неизвестного математического ожидания.

Задача 167. $s_1^2=25$, $n=5$ Какой из интервалов может быть доверительным интервалом для дисперсии

а) $(-10;30)$, б) $(0;24)$, в) $(10;84)$, г) $(-10;42)$?

Задача 168. $s_1^2=9$, $n=5$ Какой из интервалов может быть доверительным интервалом для дисперсии

а) $(-1;10)$, б) $(3;32)$, в) $(0;8)$, г) $(-10;12)$?

Задача 169. Произведена выборка объемом $n=100$ из большой партии радиоламп. Средний срок службы радиоламп выборки оказался равным 500 часов. Найти с доверительной вероятностью $p=1-\alpha=0{,}95$ доверительный интервал для среднего срока службы радиолампы во всей партии, если среднее квадратическое отклонение срока службы составляет 40 часов.

Задача 170. На контрольных испытаниях 16 радиоламп были определены выборочные характеристики их срока службы, которые оказалась равными $\bar{x} = 300$ ч и $s_1^2 = 400$ ч^2. Считая срок службы каждой лампы нормально распределенной случайной величиной» определить доверительные интервалы для математического ожидания m и дисперсии σ^2 при доверительной вероятности $1-\alpha = 0{,}9$.

Задача 171. Выборка 4.6, -2.2, -1.6, 1.8, -1.9, 2.1, 4.1, 1.7, -2.4 принадлежит нормальному распределению, $\sigma^2 = 9$. Построить доверительный интервал для математического ожидания, $1-\alpha = 0.9$.

Задача 172. Выборка 1.6, 2.9, -1.1, 1.6, -1.3, 2.8, 4.9, 1.4, -2.6 принадлежит нормальному распределению, $\sigma^2 = 4$. Построить доверительный интервал для математического ожидания, $1-\alpha = 0.95$.

Задача 173. Выборка 3.1, 8.3, 4.9, -2.5, -3.9, 7.5 принадлежит нормальному распределению с математическим ожиданием 7. Построить доверительный интервал для σ^2, $1-\alpha = 0.95$.

Задача 174. Выборка -1.1, 5.3, 2.9, -2.7, 3.4, 7.2 принадлежит нормальному распределению с математическим ожиданием 3. Построить доверительный интервал для σ^2, $1-\alpha = 0.9$.

Задача 175. Выборка 1.6, 8.2, -1.7, 1.9, -1.3, 2.5, 4.8, -1.5, 2.7 принадлежит нормальному распределению с неизвестными математическим ожиданием и дисперсий. Построить доверительный интервал для математического ожидания и дисперсии, $1-\alpha = 0.9$.

Задача 176. Выборка -0.5, -0.3, 0.9, 1.2, 1.1, 1.3, 1.5, 2.8, 3.7, 1.6, 6.5 принадлежит нормальному распределению с неизвестными математическим ожиданием и дисперсий. Построить доверительный интервал для математического ожидания и дисперсии, $1-\alpha = 0.95$.

18. Процедуры проверки гипотез

Пусть $F(x,\theta)$ - функция распределения вероятностей случайной величины, значения которой являются элементами выборки

$$x_1, x_2, \ldots, x_n .$$

Любое предположение, выделяющее некоторое подмножество, которому может принадлежать θ ($F(x,\theta)$), в математической статистике называют ***гипотезой***.

Так, например, статистическая гипотеза утверждает: неизвестный параметр θ распределения вероятностей $F(x,\theta)$ принадлежит заданному подмножеству $H \subset \Theta$ множества возможных значений параметра θ. Подмножество, которое является дополнением к подмножеству H, называется ***альтернативой к гипотезе*** H.

Гипотеза называется простой, если подмножество H состоит из одного единственного значения θ, в противном случае гипотеза называется сложной.

Практическое применение математической статистики состоит в проверке фактического соответствия реальных результатов экспериментов предполагаемой гипотезе. С этой целью строится процедура проверки гипотезы (критерий согласия), который позволяет по результатам наблюдений принимать или отвергать гипотезу.

Построение начинается с определения двух непересекающихся подмножеств A_0 и A_1. Правило проверки гипотезы формулируется следующим образом: если значения соответствующей статистики K (функции от полученных наблюдений $K = f(x_1, x_2, \ldots, x_n)$) принадлежат A_0, то гипотеза H принимается; если значения соответствующей статистики K принадлежат A_1, то гипотеза H отвергается.

Подмножество A_1 (область непринятия гипотезы H) называется **критической областью**.

Подмножество A_0 (область принятия гипотезы H) называется **областью дополнительных значений или областью принятия гипотезы**.

Критические области бывают двусторонними или односторонними, которые в свою очередь разделяются на левосторонние и правосторонние.

Пусть Δ_1, Δ_2 - критические точки, которые используются для проверки основной гипотезы. Тогда **двусторонняя критическая область** определяется из условия

$$P(K < \Delta_1) + P(K > \Delta_2) = \alpha .$$

Правосторонняя критическая область определяется из условия

$$P(K > \Delta_2) = \alpha .$$

Левосторонняя критическая область определяется из условия

$$P(K < \Delta_1) = \alpha .$$

Поскольку события, состоящие в том, что значения функции от наблюдений принадлежат к A_0 или A_1, являются случайными, то применение процедуры проверки гипотезы сопряжено с ошибками двух родов: отвергнуть гипотезу, когда она верна (**ошибка первого рода**); принять гипотезу, когда она неверна (**ошибка второго рода**).

Если

$$\mathrm{P}(K \in A_0) = 1 - \alpha ,$$

то вероятность ошибки первого рода равна α. Значение вероятности ошибки первого рода называют уровнем значимости критерия.

Если вероятность ошибки второго рода равна β, то величину $1-\beta$ называют **мощностью критерия**.

При заданном уровне значимости критическую область строят так, чтобы мощность критерия была максимальной, тем самым уменьшают ошибку второго рода.

Существенную роль при проверке гипотез играют доверительные интервалы для неизвестных параметров, построенные с помощью их точечных оценок.

На рис. 10 показан способ интерпретации вероятностей ошибок первого и второго рода. Предположим, что статистика K, которая выбрана для проверки гипотез, имеет плотность распределения $f_\theta(t/\theta_0)$, если гипотеза H_0 справедлива и плотность распределения $f_\theta(t_1/\theta_1)$, если справедлива гипотеза H_1. Будем также считать, что критическая область, задаваемая соотношением $K > A$ ($(t > A)$, такая, как показано на рисунке. Вероятность ошибки первого рода α равна $\int_A^\infty f_\theta(t/\theta_0)dt$, а соответствующая ей область отмечена на рисунке наклонной штриховкой. Вероятность ошибки второго рода (состоящей в принятии H_0) равна $\beta = \int_{-\infty}^A f_\theta(t/\theta_1)dt$, а соответствующая ей область отмечена на рис.10 вертикальной штриховкой.

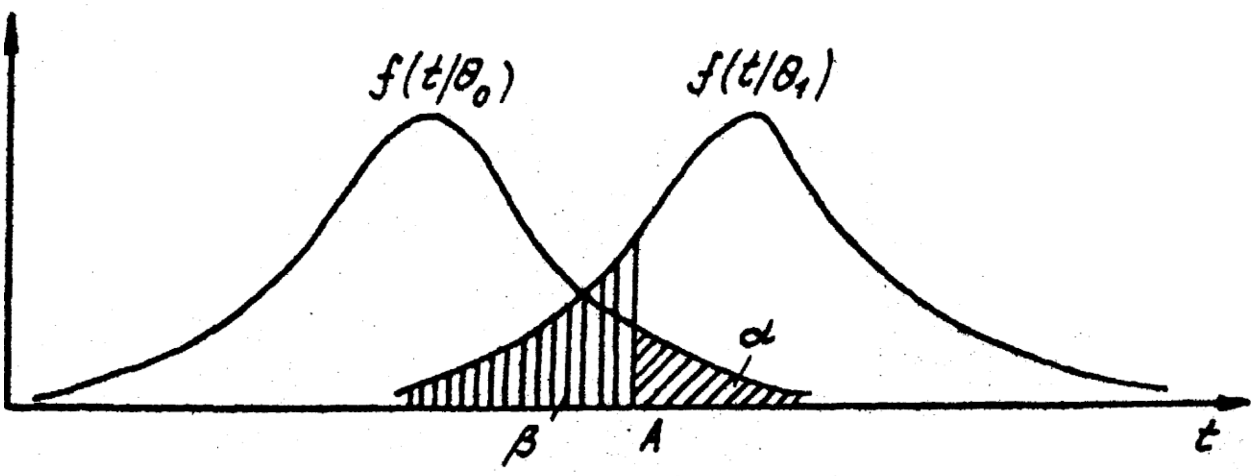

Рис. 10. Интерпретация вероятностей ошибок первого и второго рода.

Пример. Пусть $x_1, x_2, ..., x_n$ выборка объема $n = 9$ из нормального распределения с математическим ожиданием m и дисперсией $\sigma^2 = 16$. Требуется проверить простую гипотезу H_0: $m = 20$ против простой альтернативной гипотезы H_1: $m = 24$ при $\alpha = 0{,}05$. Вычислить вероятности ошибки второго рода β.

Решение. Рассмотрим статистику $Z = \bar{x} = \frac{1}{9}\sum_{i=1}^{9} x_i$. Если гипотеза H_0 верна, то Z имеет нормальное распределение с параметрами 20 и 16/9 с плотностью распределения (рис.11)

$$f_Z(Z/H_0) = \frac{1}{\sqrt{2\pi}\,\frac{4}{3}} e^{-\frac{9(z-20)^2}{32}}$$

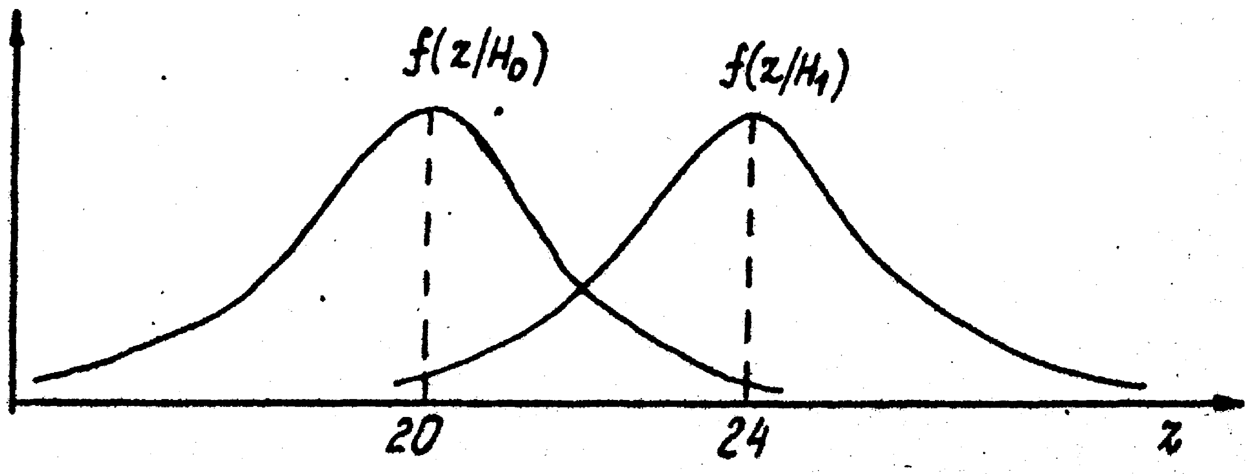

Рис.11. Плотности распределения вероятностей двух выборок.

Если справедлива гипотеза H_1, то статистика Z нормально распределена с параметрами 24 и 16/9 и плотностью распределения (рис.11).

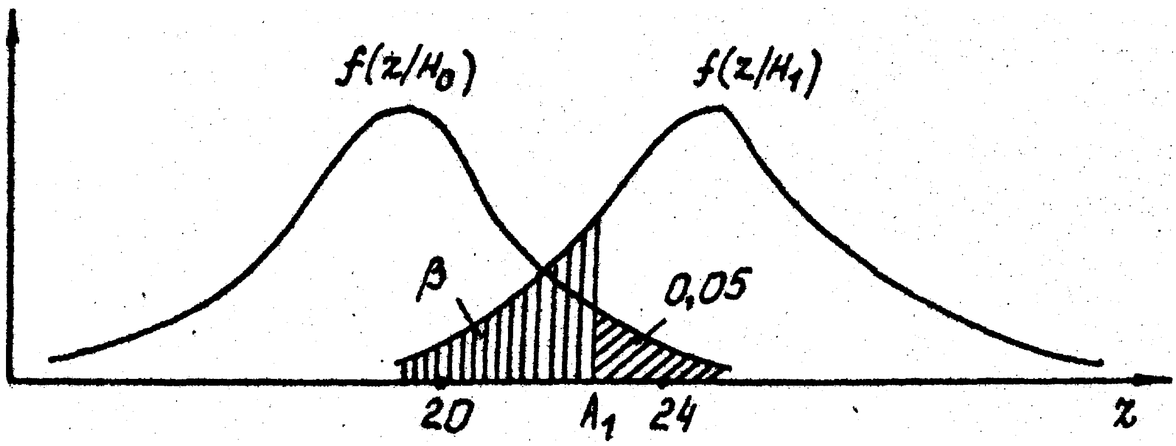

Рис. 12. Критические области.

$$f_Z(Z/H_1) = \frac{1}{\sqrt{2\pi}\,\frac{4}{3}} e^{-\frac{9(z-24)^2}{32}}$$

Рассмотрим различные способы выбора критической области V.

1) Условимся отклонять гипотезу H_0, когда $Z = \bar{x} > A_1$ (рис.12).

Тогда значение A_1 находим из следующего уравнения:

$$P\left(Z > A_1 / H_0\right) = \int\limits_{A_1}^{\infty} f_Z\,(Z / H_0)dz = 0{,}05\,.$$

или

$$P\left(Z > A_1 / H_0\right) = \frac{1}{2} - \Phi\left(\frac{A_1 - 20}{4/3}\right) = 0.05$$

где $\Phi(x) \sim$ функция Лапласа (приложение 2). Откуда

$$\Phi\left(\frac{A_1 - 20}{4/3}\right) = 0{,}5 - 0{,}05 = 0{,}45$$

или

$$\Phi\left(\frac{A_1 - 20}{4/3}\right) = \Phi(1{,}65)$$

Следовательно,

$$A_1 = \frac{4 \cdot 1{,}65}{3} + 20 = 22{,}2$$

и критическая область определяется неравенством: $Z = \bar{x} > 22{,}2$

При этом ошибки второго рода

$$\beta_1 = \int\limits_{-\infty}^{22,2} f_2(z / H_1)dz = \Phi_0\left(\frac{22{,}2 - 24}{4/3}\right) - \Phi_0(-\infty) = \Phi_0(-1{,}35) + 0{,}5 =$$
$$-0{,}41149 + 0{,}5 = 0{,}08851$$

2) Условимся отклонять гипотезу H_0 при $A_2 < Z < A_3$, где $P\left(A_2 < Z < A_3 / H_0\right) = 0{,}05$ (рис.13).

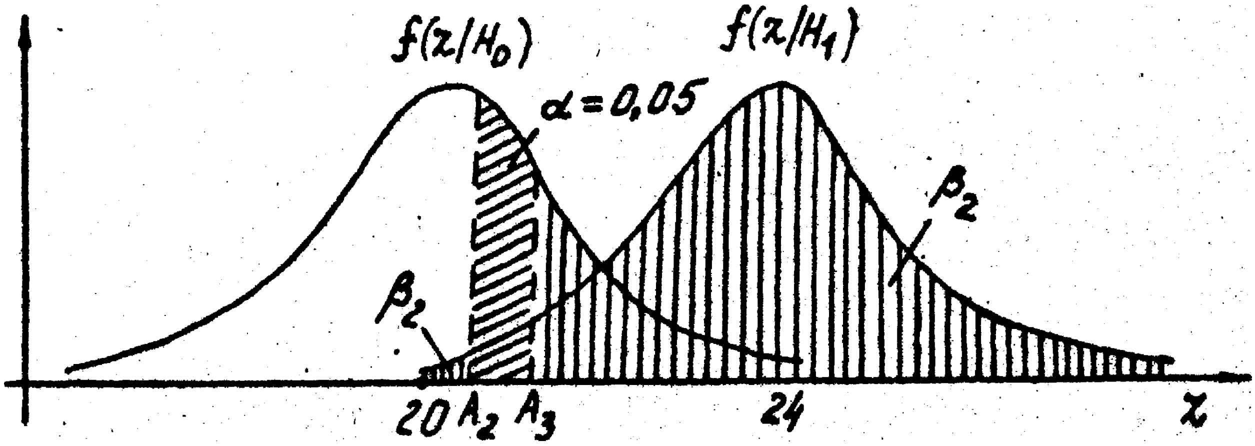

Рис. 13. Критические области.

Поскольку таких A_2 и A_3 можно подобрать бесконечно много, выберем для определенности $A_2 = 21$ и найдем A_3. Имеем

$$P(21 < Z < A_3 / H_0) = \Phi\left(\frac{A_3 - 20}{4/3}\right) - \Phi\left(\frac{21-20}{4/3}\right) = \Phi\left(\frac{A_3 - 20}{4/3}\right) - \Phi(0{,}75) = \alpha = 0{,}05$$

Отсюда

$$\Phi\left(\frac{A_3 - 20}{4/3}\right) = \Phi(0{,}75) + 0{,}05 = 0{,}27337 + 0{,}05 = 0{,}32337 = \Phi(0{,}93)$$

Следовательно, $\frac{A_3 - 20}{4/3} = 0{,}93$ и $A_3 = 21{,}24$. В этом случае критическая область V определяется неравенством:

$$21 < Z < 21{,}24,$$

а вероятность ошибки второго рода

$$\beta_2 = \int_{-\infty}^{21} f_Z(z/H_1)dz + \int_{21{,}4}^{\infty} f_Z(z/H_1)dz = \Phi\left(\frac{21-24}{4/3}\right) + \frac{1}{2} + \frac{1}{2} - \Phi\left(\frac{21{,}24-24}{4/3}\right) =$$
$$= 1 - \Phi(2{,}25) + \Phi(2{,}07) = 1 - 0{,}48778 + 0{,}48077 = 0{,}99299$$

3) Условимся отклонять гипотезу H_0 при $Z > A_5$ и $Z < A_4$, где

$P(Z > A_5 / H_0) = \dfrac{\alpha}{2} = 0{,}025$ и $P(Z < A_4 / H_0) = \dfrac{\alpha}{2} = 0{,}025$ (рис. 14).

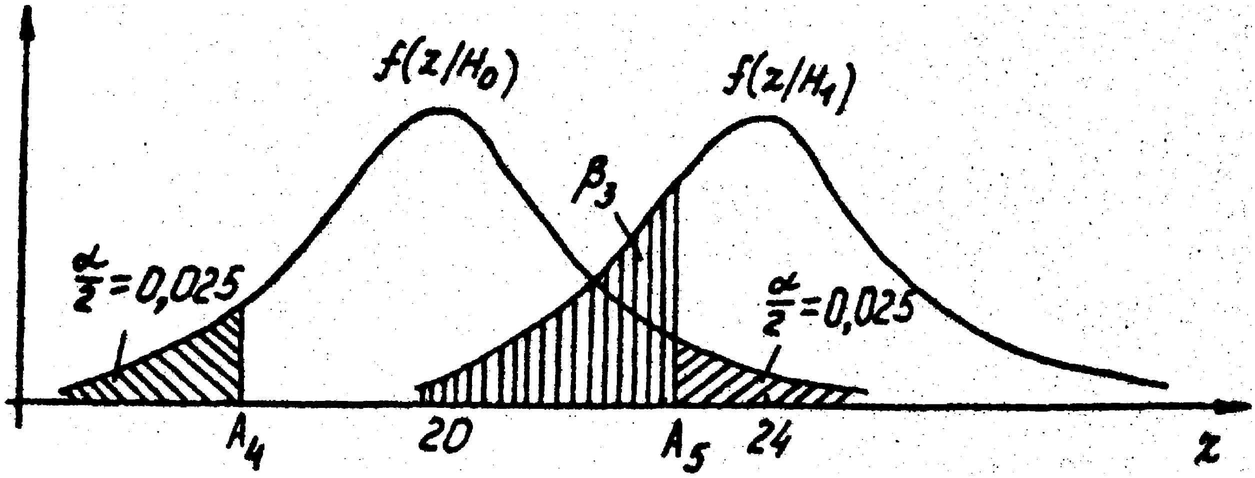

Рис.14. Критические области.

Откуда получим, что $A_4 = 17{,}4$; $A_5 = 22{,}6$ При этом вероятность ошибки второго рода

$$\beta_3 = P(A_4 < Z < A_5 / H_1) = 0{,}147$$

Из примера следует, что наилучшим критерием является тот, у которого при заданном уровне значимости имеет место наименьшая вероятность ошибки второго рода β, или что то же самое, наибольшая мощность критерия $1-\beta$.

В теории статистической проверки гипотез доказывается, что при фиксированном объеме выборки требование снижения вероятностей ошибок первого и второго рода противоречивы, так как с уменьшением α возрастает β, и наоборот, т.е. между α и β необходим компромисс. Компромиссные значения α и β выбирается в соответствии с важностью последствий ошибок первого и второго рода. Обычно это делается так: при выбранном уровне значимости α максимизируют мощность критерия по отношению к альтернативной гипотезе H_1.

Пример. Из нормальной генеральной совокупности с известной дисперсией $\sigma^2 = 5{,}2$ извлечена выборка $n = 100$ и по ней найдена выборочная средняя $\bar{x} = 27{,}56$. Требуется при уровне значимости $\alpha = 0{,}05$ проверить гипотезу H_0 относительно значения математического ожидания: $m_0 = 26$ против альтернативной гипотезы $H_1 : m_0 \neq 26$.

Решение. Различие между выборочным средним и предполагаемым значением математического ожидания m_0 можно исследовать, используя статистику критерия:

$$U = \frac{(\bar{x} - m_0)\sqrt{n}}{\sigma}.$$

$$P\left(\left| \frac{(\bar{x} - m_0)\sqrt{n}}{\sigma} \right| \leq \Delta_{\frac{1-\alpha}{2}} \right) = 1 - \alpha$$

Критическая область определяется соотношением:

$$\left| \frac{(\bar{x} - m_0)\sqrt{n}}{\sigma} \right| > \Delta_{\frac{1-\alpha}{2}},$$

где $\Delta_{\frac{1-\alpha}{2}}$ определяется из таблицы для функции Лапласа (приложение 2)

при

значении $\Phi\left(\Delta_{\frac{1-\alpha}{2}} \right) = \frac{1-\alpha}{2} = \frac{1-0{,}05}{2} = 0{,}475$.

$$\frac{(\bar{x} - m_0)\sqrt{n}}{\sigma} = \frac{27{,}56 - 26}{5{,}2/\sqrt{100}} = 3, \quad \Delta_{\frac{1-\alpha}{2}} = 1{,}96.$$

Так как $3 > 1{,}96$ то нулевую гипотезу отвергаем.

Замечание. При справедливости неравенства $\left|\frac{(\bar{x}-m_0)\sqrt{n}}{\sigma}\right| \le \Delta_{\frac{1-\alpha}{2}}$ еще утверждать нельзя, что основная гипотеза подтвердилась, можно только признать допустимость гипотезы для рассмотренных выборочных наблюдений до тех пор, пока более обстоятельные исследования не позволяют сделать противоположное заключение. Другими словами, с помощью проверки статистических гипотез можно лишь отвергнуть проверяемую гипотезу, но никогда нельзя доказать ее справедливость.

Задачи

Задача 177. Из нормальной генеральной совокупности с известной дисперсией σ^2 извлечена выборка объема n и по ней найдена выборочная средняя $\bar{x}$. Тогда для проверки гипотезы $H_0: M(X)=6$ против альтернативной гипотезы $H_1: M(X)\neq 6$ используется статистика критерия

а) $\frac{(\bar{x}+6)\sqrt{n}}{\sigma}$, б) $\frac{(\bar{x}-6)}{\sigma\sqrt{n}}$, в) $\frac{(\bar{x}-6)\sqrt{n}}{\sigma}$, г) $\frac{(\bar{x}-6)\sigma}{\sqrt{n}}$.

Задача 178. Из нормальной генеральной совокупности с известной дисперсией σ^2 извлечена выборка объема n и по ней найдена выборочная средняя $\bar{x}$. Тогда для проверки гипотезу $H_0: M(X)=0$ против альтернативной гипотезы $H_1: M(X)\neq 0$ используется статистика критерия

а) $\frac{\bar{x}\sqrt{n}}{\sigma}$, б) $\frac{\bar{x}}{\sigma\sqrt{n}}$, в) $\frac{\sqrt{n}}{\sigma\bar{x}}$, г) $\frac{\bar{x}\sigma}{\sqrt{n}}$.

Задача 179. По выборке $n=16$ из нормально распределенной генеральной совокупности получено $\bar{x}=0{,}154$. С уровнем значимости $\alpha=0{,}01$ проверить гипотезу $H_0: M(X)=0$ против альтернативной гипотезы $H_1: M(X)\neq 0$, если $\sigma^2=1$.

Задача 180. Из нормальной генеральной совокупности с неизвестной дисперсией σ^2 извлечена выборка объема $n = 36$ и по ней найдена исправленная выборочная дисперсия $s_1^2 = 4$. Чему равно наблюдаемое значение статистики критерия для проверки гипотезы $H_0 : D(X) = \sigma^2 = 5$ против альтернативной гипотезы $H_1 : D(X) \neq 5$?

Задача 181. Из нормальной генеральной совокупности с неизвестной дисперсией σ^2 извлечена выборка объема $n = 10$ и по ней найдена исправленная выборочная дисперсия $s_1^2 = 8$. Чему равно наблюдаемое значение статистики критерия для проверки гипотезы $H_0 : D(X) = \sigma^2 = 9$ против альтернативной гипотезы $H_1 : D(X) \neq 9$?

Задача 182. Пусть $x_1, x_2, ..., x_n$ выборка объема $n = 25$ из нормального распределения с математическим ожиданием m и дисперсией $\sigma^2 = 4$. Требуется проверить простую гипотезу H_0: $m = 10$ против простой альтернативной гипотезы H_1: $m = 8$ при $\alpha = 0{,}05$. Вычислить вероятность ошибки второго рода β.

Задача 183. Пусть $x_1, x_2, ..., x_n$ выборка объема $n = 13$ из нормального распределения с математическим ожиданием m и дисперсией $\sigma^2 = 6$. Требуется проверить простую гипотезу H_0: $m = 4$ против простой альтернативной гипотезы H_1: $m = 7$ при $\alpha = 0{,}05$. Вычислить вероятность ошибки второго рода β.

19. Сравнение математического ожидания и дисперсии в двух нормальных выборках. Критерии Фишера и Стьюдента

Имеются две независимые выборки, принадлежащие нормальным законам распределения:

$$X_{11}, X_{12}, \ldots, X_{1n_1} \in N\left(a_1, \sigma_1^2\right),$$
$$X_{21}, X_{22}, \ldots, X_{2n_2} \in N\left(a_2, \sigma_2^2\right).$$

Через a_1, σ_1^2 обозначены соответственно математическое ожидание и дисперсия случайных величин из первой выборки, а через a_2, σ_2^2 - математическое ожидание и дисперсия случайных величин из второй выборки.

В экспериментальных исследованиях часто возникает вопрос о сравнении a_1 и a_2, σ_1^2 и σ_2^2. Очевидно, что это сравнение будет происходить на основе соответствующих выборочных характеристик

$$\overline{X_k} = \frac{1}{n_k}\sum_{i=1}^{n_k} X_{ki}\,,\ S_k^2 = \frac{1}{n_k - 1}\sum_{i=1}^{n}\left(X_{ki} - \overline{X_k}\right)^2, k = \overline{1,2}\,.$$

Начнем с задачи сравнения σ_1^2 и σ_2^2. Хорошо известно, что

$\dfrac{(n_1 - 1)S_1^2}{\sigma_1^2} \in \chi^2_{n_1 - 1}$ (χ^2-распределению с $n_1 - 1$ степенями свободы),

$\dfrac{(n_2 - 1)S_2^2}{\sigma_2^2} \in \chi^2_{n_2 - 1}$ (χ^2-распределению с $n_2 - 1$ степенями свободы.

Как известно, отношение $\dfrac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \dfrac{S_1^2}{S_2^2 \lambda}$ имеет распределение Фишера с

$n_1 - 1$ и $n_2 - 1$ степенями свободы, $\lambda = \dfrac{\sigma_1^2}{\sigma_2^2}$. Тогда

$$\mathrm{P}\left(\Delta_{1-\frac{\alpha}{2}} \le \frac{S_1^2}{S_2^2 \lambda} \le \Delta_{\frac{\alpha}{2}}\right) = 1 - \alpha\,,$$

$\Delta_{\frac{\alpha}{2}}, \Delta_{1-\frac{\alpha}{2}}$ находятся из таблицы критических точек распределения Фишера с $n_1 - 1$ и $n_2 - 1$ степенями свободы (приложение 5).

Таким образом, для λ получаем доверительный интервал

$$\frac{s_1^2}{s_2^2 \Delta_{\frac{\alpha}{2}}} \le \lambda \le \frac{s_1^2}{s_2^2 \Delta_{1-\frac{\alpha}{2}}}.$$

В качестве основной гипотезы будем рассматривать $H_0 : \sigma_1^2 = \sigma_2^2$, в качестве конкурирующей гипотезы будем рассматривать $H_1 : \sigma_1^2 \ne \sigma_2^2$.

Критическая область строится на основе доверительного интервала для параметра λ. Подставляя вместо параметра λ единицу, получим:

$$\frac{s_1^2}{s_2^2 \Delta_{\frac{\alpha}{2}}} \le 1,\ 1 \le \frac{s_1^2}{s_2^2 \Delta_{1-\frac{\alpha}{2}}} ,$$

$$\Delta_{\frac{\alpha}{2}} \ge \frac{s_1^2}{s_2^2},\ \Delta_{1-\frac{\alpha}{2}} \le \frac{s_1^2}{s_2^2}$$

или

$$\Delta_{1-\frac{\alpha}{2}} \le \frac{s_1^2}{s_2^2} \le \Delta_{\frac{\alpha}{2}}.$$

Если это неравенство выполняется, то считается, что полученные наблюдения не противоречат гипотезе H_0, и она принимается. В противном случае гипотеза H_0 отвергается и принимается конкурирующая гипотеза H_1. То есть, если значение статистики $\frac{s_1^2}{s_2^2}$, вычисленное в соответствии с полученными данными, принадлежит отрезку $\left[\Delta_{1-\frac{\alpha}{2}}, \Delta_{\frac{\alpha}{2}}\right]$ (области принятия гипотезы), то основная гипотеза H_0 принимается.

Критической областью в этом случае является **двусторонняя область**

$$\left(-\infty, \Delta_{1-\frac{\alpha}{2}}\right) \cup \left(\Delta_{\frac{\alpha}{2}}, +\infty\right).$$

В качестве основной гипотезы рассмотрим $H_0 : \sigma_1^2 = \sigma_2^2$, в качестве конкурирующей гипотезы рассмотрим $H_1 : \sigma_1^2 > \sigma_2^2$.

Тогда областью принятия гипотезы H_0 является отрезок $\left[\Delta_{1-\frac{\alpha}{2}}, \Delta_{\frac{\alpha}{2}}\right]$, а критической областью является **односторонняя (правосторонняя) область** $\left(\Delta_{\frac{\alpha}{2}}, +\infty\right)$.

Рассмотрим основную гипотезу $H_0 : \sigma_1^2 = \sigma_2^2$ и конкурирующую гипотезу $H_1 : \sigma_1^2 < \sigma_2^2$.

Тогда областью принятия гипотезы H_0 по-прежнему является отрезок $\left[\Delta_{1-\frac{\alpha}{2}}, \Delta_{\frac{\alpha}{2}}\right]$, а критической областью является **односторонняя (левосторонняя область)** $\left(-\infty, \Delta_{1-\frac{\alpha}{2}}\right)$.

Построенные критерии называются соответственно двусторонним и односторонними **критериями Фишера**.

Решим задачу сравнения средних a_1 и a_2 в двух выборках

$$X_{11}, X_{12}, \ldots, X_{1n_1} \in N\left(a_1, \sigma_1^2\right),$$

$$X_{21}, X_{22}, \ldots, X_{2n_2} \in N\left(a_2, \sigma_2^2\right).$$

при условии равенства дисперсий $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Рассмотрим статистику

$$\overline{X_1} - \overline{X_2} = \frac{1}{n_1}\sum_{i=1}^{n_1} X_{1i} - \frac{1}{n_2}\sum_{i=1}^{n_2} X_{2i}\,.$$

Поскольку обе выборки принадлежат нормальным распределениям, то статистика $\overline{X_1} - \overline{X_2}$ тоже принадлежит нормальному закону распределения с параметрами $\left(a_1 - a_2, \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2}\right)$.

Соответственно статистика

$$\frac{\overline{X_1}-\overline{X_2}-(a_1-a_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

имеет стандартное (с параметрами $(0,1)$) нормальное распределение.

Известно, что

$$\frac{(n_1-1)S_1^{\,2}+(n_2-1)S_2^{\,2}}{\sigma^2}\in\chi^2_{n_1+n_2-2},$$

$$S_k^2=\frac{1}{n_k-1}\sum_{i=1}^{n}\left(X_{ki}-\overline{X_k}\right)^2, k=\overline{1,2}.$$

Поскольку

$$\frac{\overline{X_1}-\overline{X_2}-(a_1-a_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\in N(0,1),$$

то отношение

$$\frac{\dfrac{\overline{X_1}-\overline{X_2}-(a_1-a_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\dfrac{(n_1-1)S_1^2+(n_2-1)S_2^2}{\sigma^2(n_1+n_2-2)}}}=\frac{\overline{X_1}-\overline{X_2}-(a_1-a_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

имеет распределение Стьюдента ($t_{n_1+n_2-2}$-распределение) с n_1+n_2-2 степенями свободы. Построим доверительный интервал для разности a_1-a_2. С вероятностью $1-\alpha$ выполняется

$$-\Delta_{n_1+n_2-2,\alpha}\le\frac{\overline{X_1}-\overline{X_2}-(a_1-a_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\le\Delta_{n_1+n_2-2,\alpha},$$

где $\Delta_{n,\alpha}$ находится из таблицы для вероятностей $P(|t_n|>\Delta_{n,\alpha})=\alpha$ распределения Стьюдента (приложение 4).

$$S^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1+n_2-2)}.$$

После соответствующих преобразований получаем:

$$\overline{x_1} - \overline{x_2} - \Delta_{n_1+n_2-2,\alpha} s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le a_1 - a_2 \le \overline{x_1} - \overline{x_2} + \Delta_{n_1+n_2-2,\alpha} s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

В качестве основной гипотезы будем рассматривать $H_0 : a_1 = a_2$, в качестве конкурирующей гипотезы будем рассматривать $H_1 : a_1 \ne a_2$.

Подставляя в доверительный интервал ноль вместо $a_1 - a_2$, получаем:

$$-\Delta_{n_1+n_2-2,\alpha} \le \frac{\overline{x_2} - \overline{x_1}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \le \Delta_{n_1+n_2-2,\alpha}.$$

Таким образом, если значение статистики

$$\frac{\overline{x_2} - \overline{x_1}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

принадлежит отрезку $\left[-\Delta_{n_1+n_2-2,\alpha}, \Delta_{n_1+n_2-2,\alpha}\right]$, то гипотеза $H_0 : a_1 = a_2$ принимается.

Если значение статистики

$$\frac{\overline{x_2} - \overline{x_1}}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

попадают в двустороннюю критическую область $\left(-\infty, -\Delta_{n_1+n_2-2,\alpha}\right) \cup \left(\Delta_{n_1+n_2-2,\alpha}, +\infty\right)$, то принимается конкурирующая гипотеза $H_1 : a_1 \ne a_2$.

В качестве основной гипотезы рассмотрим $H_0 : a_1 = a_2$, в качестве конкурирующей гипотезы рассмотрим $H_1 : a_1 > a_2$.

Тогда областью принятия гипотезы H_0 является отрезок $\left[-\Delta_{n_1+n_2-2,\alpha},\Delta_{n_1+n_2-2,\alpha}\right]$, а критической областью является односторонняя область $\left(\Delta_{n_1+n_2-2,\alpha},+\infty\right)$.

Рассмотрим основную гипотезу $H_0 : a_1 = a_2$ и конкурирующую гипотезу $H_1 : a_1 < a_2$.

Тогда областью принятия гипотезы H_0 по-прежнему является отрезок $\left[-\Delta_{n_1+n_2-2,\alpha},\Delta_{n_1+n_2-2,\alpha}\right]$, а критической областью является односторонняя область $\left(-\infty,-\Delta_{n_1+n_2-2,\alpha}\right)$.

Построенные критерии называются соответственно двусторонним и односторонними **критериями Стьюдента**.

Задачи.

Задача 184. По двум выборкам случайных величин X и Y объемов соответственно $n=10, m=13$ вычислены исправленные выборочные дисперсии $s_{1x}^2 = 7.6,\ s_{1y}^2 = 4.7$. Как определяется критическая точка распределения Фишера для проверки гипотезы $H_0 : D(X) = D(Y)$ против альтернативной гипотезы $H_1 : D(X) > D(Y),\ \alpha = 0.01$?

а) $F_{кр}(0.01;10;13)$, б) $F_{кр}(0.01;9;12)$, в) $F_{кр}(0.01;12;9)$, г) $F_{кр}(0.01;13;10)$

Задача 185. По двум выборкам случайных величин X и Y объемов соответственно $n=24, m=33$ вычислены исправленные выборочные дисперсии $s_{1x}^2 = 1.6,\ s_{1y}^2 = 3.8$. Как определяется критическая точка распределения Фишера для проверки гипотезы $H_0 : D(X) = D(Y)$ против альтернативной гипотезы $H_1 : D(X) < D(Y),\ \alpha = 0.01$?

а) $F_{кр}(0.01;24;33)$, б) $F_{кр}(0.01;33;24)$, в) $F_{кр}(0.01;23;32)$, г) $F_{кр}(0.01;32;23)$

Задача 186. Для проверки точности двух станков проведены измерения некоторого признака выпускаемых ими однотипных изделий. По результатам $n_1 = 11$ измерений первого станка получено $s_x^2 = 81{,}68$, а по результатам $n_2 = 9$ второго станка найдено $s_y^2 = 22{,}60$.Можно ли на основании этих результатов сделать вывод, что точность второго станка выше точности первого (на основании 5%-ного уровня значимости)

Задача 187. Двумя приборами в одном и том же порядке измерены шесть деталей и получены следующие результаты измерений:

x_i	**2**	**3**	**5**	**6**	**8**	**10**
y_i	**10**	**3**	**6**	**1**	**7**	**4**

При уровне значимости 0,05 установить, значимо или незначимо различаются результаты измерений, в предложении, что $D_i = x_i - y_i$ распределены нормально.

Задача 188. Химическая лаборатория произвела в одном и том же порядке 8 проб двумя методами. Подучены следующие результаты (в первой строке указано содержание вещества в процентах в каждой пробе, определенное первым методом, во второй строке - вторым методом):

x_i	15	20	16	22	24	14	18	20
y_i	15	22	14	25	29	16	20	24

При уровне значимости $\alpha = 0{,}05$ установить, значимо или незначимо различаются средние результаты анализов в предположении, что они распределены нормально.

Задача 189. Две лаборатории одним и тем же методом, в одном и том же порядке определяли содержание углерода в 13 пробах нелегированной стали. Получены следующие результаты анализов (в первой отроке указано содержание углерода в процентах в каждой пробе, полученное первой лабораторией; во второй строке -второй лабораторией):

x_i	**0,18**	**0,12**	**0,12**	**0,08**	**0,08**	**0,12**	**0,19**	**0,32**	**0,27**	**0,22**	**0,34**	**0,14**	**0,46**
y_i	**0,16**	**0,09**	**0,08**	**0,05**	**0,13**	**0,10**	**0,14**	**0,30**	**0,31**	**0,24**	**0,28**	**0,11**	**0,42**

При уровне значимости 0,05 установить, значимо или незначимо различаются средние результаты анализа в предположении, что они нормально распределены.

20. Критерий χ^2

Рассмотрим выборку

$$x_1, x_2, ..., x_n,$$

которую будем использовать для проверки простой гипотезы H_0, состоящей в том, что исследуемая случайная величина X имеет функцию распределения $F(x)$ (или в частном случае дискретное распределение $(p_1, p_2, ..., p_m), \sum_{i=1}^{m} p_i = 1$).

Разбиваем числовую ось на непересекающиеся полуинтервалы

$$R = (y_0, y_1] \cup (y_1, y_2] \cup ... \cup (y_k, y_{k+1}), \; y_0 = -\infty, y_{k+1} = +\infty.$$

Определим $n_i, i=\overline{1,k+1}$ - числа попаданий элементов выборки $x_1, x_2, ..., x_n$ в полуинтервалы $\left(y_{i-1}, y_i\right]$, $i=\overline{1,k+1}$.

Очевидно, что $n_i, i=\overline{1,k+1}$ можно представить в виде следующей суммы (аналогичное представление мы уже использовали при определении эмпирической функции распределения):

$$n_i = \sum_{j=1}^{n} I_{\left\{x_j \in \left(y_{i-1}, y_i\right]\right\}},$$

где

$$I_{\left\{x_j \in \left(y_{i-1}, y_i\right]\right\}} = \begin{cases} 1, x_j \in \left(y_{i-1}, y_i\right] \\ 0, x_j \notin \left(y_{i-1}, y_i\right] \end{cases}, i=\overline{1,k+1},\ j=\overline{1,n}.$$

Обозначим через $p_i, i=\overline{1,k+1}$ вероятности $P\left(X \in \left(y_{i-1}, y_i\right]\right) = F\left(y_i\right) - F\left(y_{i-1}\right), i=\overline{1,k+1}$.

Вычислим $M\left(n_i\right)$:

$$M\left(n_i\right) = nM\left(I_{\left\{x_1 \in \left(y_{i-1}, y_i\right]\right\}}\right) = np_i.$$

Таким образом, если наша гипотеза H_0 верна, то $n_i, i=\overline{1,k+1}$ должны быть близки к $M\left(n_i\right) = np_i$. Мер этой близости можно предложить много, но К.Пирсоном была предложена следующая статистика:

$$\hat{\chi}^2 = \sum_{i=1}^{k+1} \frac{\left(n_i - np_i\right)^2}{np_i}.$$

При $n \to \infty$ распределение статистики $\hat{\chi}^2$ стремится к распределению χ_k^2.

Пусть $\chi_{k,\alpha}^2$ - решение уравнения

$$P\left(\chi_k^2 > \chi_{k,\alpha}^2\right) = \alpha.$$

Поскольку $\hat{\chi}^2 \to \chi_k^2$, то $P\left(\hat{\chi}^2 > \chi_{k,\alpha}^2\right) \to P\left(\chi_k^2 > \chi_{k,\alpha}^2\right) = \alpha$, то есть можно считать, что $P\left(\hat{\chi}^2 > \chi_{k,\alpha}^2\right) \approx \alpha$.

Если α мало, то событие $\hat{\chi}^2 > \chi_{k,\alpha}^2$ является практически невозможным. Если $\hat{\chi}^2 \le \chi_{k,\alpha}^2$, то считается, что полученные данные $x_1, x_2, \ldots, x_n$ не противоречат гипотезе H_0. Ошибка первого рода равна α.

Таким образом, область принятия гипотезы H_0:

$$\hat{\chi}^2 \le \chi_{k,\alpha}^2.$$

Критическая область:

$$\hat{\chi}^2 > \chi_{k,\alpha}^2.$$

Если известен тип распределения, но неизвестны его l параметров, то предельным распределением для

$$\chi^2 = \sum_{i=1}^{k+1} \frac{(n_i - np_i)^2}{np_i}$$

является χ_{k-l}^2.

Замечания.

1) Необходимым условием применения критерия χ^2 является наличие в каждом из интервалов по крайней мере 5 - 10 наблюдений. Поэтому все частоты $n_i < 5$ следует объединить путем объединения соседних интервалов, в этом случае соответствующие им теоретические частоты (np_i) также надо сложить.

2) При проверке гипотезы о законе распределения контролируется лишь ошибка первого рода.

Пример. Опыт состоит в бросании монеты, $n = 100$ и определении числа выпавших гербов $n_1 = 22$. Применяя χ^2- критерий, проверить симметричность монеты при уровне значимости 0.1.

Решение. Так как $n_1 = 22$, то $n_2 = 100 - 22 = 78$. Предполагая симметричность монеты, получаем $p_1 = p_2 = \frac{1}{2}, np_1 = 50,\ np_2 = 50$,

$$\chi^2 = \sum_{i=1}^{2} \frac{(n_i - np_i)^2}{np_i} = \frac{(22-50)^2}{50} + \frac{(78-50)^2}{50} = 3.14.$$

Находим по таблице χ^2-распределения значение $\chi^2_{1,0.1}$=2.71. Поскольку $3.14 > 2.71$, то данные не согласуются с гипотезой о симметричности монеты.

Ответ. Данные не согласуются с гипотезой о симметричности монеты.

Пример. Для управления качеством выполняемой стали необходимо знание законов распределения ее механических свойств. Для исследования взято $n = 374$ наблюдений над сталью марки 35ГС. Необходимо проверить гипотезу о том, что механическое свойство (предел текучести) этой стали имеет нормальный закон распределений, По исходным наблюдениям рассчитаны $\bar{x} = 42{,}37$ кг/мм^2, $s_1 = 0{,}94$ кг/мм^2 и данные сгруппированы в таблице ниже:

(x_i, x_{i+1})	**(40,41)**	**(41,42)**	**(42,43)**	**(43,44)**	**(44,45)**	**(45,46)**
n_i	**20**	**112**	**154**	**73**	**13**	**2**

Уровень значимости принять равным $\alpha = 0{,}01$.

Решение. В силу замечания необходимо объединить последние два столбца. В результате объединения получим следующий ряд распределения:

(x_i, x_{i+1})	**($-\infty$,41)**	**(41,42)**	**(42,43)**	**(43,44)**	**(44,$+\infty$)**

n_i	**20**	**112**	**154**	**73**	**15**

Найдем теперь p_i, $i=\overline{1,5}$, в предположении, что X нормально распределена с параметрами $\overline{x}$ и s_1^2. Получаем

$$p_1 = P(X < 41) = \frac{1}{2} + \Phi\left(\frac{41-42,37}{0,94}\right) = \frac{1}{2} - \Phi(1,46) = 0,5 - 0,4279 = 0,0721;$$

$$p_2 = P(41 \le X \le 42) = \Phi\left(\frac{42-42,37}{0,94}\right) - \Phi\left(\frac{41-42,37}{0,94}\right) = -\Phi(0,39) + \Phi(1,46) =$$
$$= -0,1517 + 0,4279 = 0,2762.$$

Аналогично вычисляем

$p_3 = 0,3971$; $p_4 = 0,2128$ и

$$p_5 = P(X \ge 44) = 1 - P(X < 44) = 1 - \left[\frac{1}{2} + \Phi\left(\frac{44-42,37}{0,94}\right)\right] = 0,5 - \Phi(1,73) = 0,5 -$$
$$- 0,4582 = 0,0418$$

Вычислим χ^2:

$$\chi^2 = \frac{(20 - 374\cdot 0,0721)^2}{374\cdot 0,0721} + \ldots + \frac{(15 - 374\cdot 0,0418)^2}{374\cdot 0,0418} = 2,231$$

и $\chi^2_{5-2-1,0.01}$:

$$\chi^2_{5-2-1,0.01} = 9.21$$

имеем 2,231<9.21, следовательно, нет оснований отвергать гипотезу о нормальности распределения предела текучести.

Пример. На телефонной станции проводились наблюдения за числом X неправильных соединений в минуту. Наблюдения в течение часа дали следующие результаты:

x_i	**0**	**I**	**2**	**3**	**4**	**5**	**6**	**7**
n_i	**8**	**17**	**16**	**10**	**6**	**2**	**0**	**I**

Определить выборочную среднюю и исправленную выборочную дисперсию неправильных соединений в минуту и проверить выполнение основного условия для распределения Пуассона ($M(X) = D(X)$). Найти теоретическое распределение Пуассона и проверить степень согласия теоретического и эмпирического распределений по критерию χ^2 при уровне значимости $\alpha = 0{,}05$.

Решение. Вычислим $s_1^2 \approx 2{,}1$ и $\bar{x} = 2$. Необходимое условие для распределения Пуассона $M(X) = D(X) = 2$ практически выполняется. Запишем теоретический закон Пуассона в виде:

$$p(m) = P(X = M) = \frac{\lambda^m}{m!} e^{-\lambda}$$

Так как математическое ожидание λ неизвестно, то подставим вместо него оценку $\tilde{\lambda} = \bar{x} = 2$. Имеем

$$P(X = 0) = \frac{2^m}{0!} e^{-2} = 0{,}1353;$$

$$P(X = 1) = \frac{2!}{1!} e^{-2} = 0{,}2707;$$

$$P(X = 2) = 0{,}2707;$$

$$P(X = 3) = 0{,}1804;$$

$$P(X = 4) = 0{,}0902$$

$$P(x = 5) = 0{,}0361;$$

$$P(X = 6) = 0{,}0120;$$

$$P(X \geq 7) = 1 - P(X < 7) = 1 - \sum_{i=0}^{6} P(X = i) = 0{,}0046$$

Запишем полученные результаты в следующую таблицу:

x_i	**0**	**I**	**2**	**3**	**4**	**5**	**6**	**7**
n_i	**8**	**17**	**16**	**10**	**6**	**2**	**0**	**I**
пр$_i$	**8,1**	**16,24**	**16,24**	**10,82**	**5,41**	**2,166**	**0,72**	**0,276**

Объединим столбцы 5, 6 и 7, так как в каждом из них мало наблюдений. Тогда таблица преобразуется к виду:

x_i	**0**	**I**	**2**	**3**	**4**	$\geq$**5**
n_i	**8**	**17**	**16**	**10**	**6**	**3**
пр$_i$	**8,1**	**16,24**	**16,24**	**10,82**	**5,41**	**3,16**

Вычислим значение χ^2

$$\chi^2_{л} = \frac{(8-8{,}1)^2}{8{,}1} + \frac{(17-16{,}24)^2}{16{,}24} + \frac{(16-16{,}24)^2}{16{,}24} + \frac{(10-10{,}82)^2}{10{,}82} + \frac{(6-5{,}41)^2}{5{,}41}$$

$$+\frac{(3-3{,}16)^2}{3{,}16} \approx 0{,}2$$

и $\chi^2_{6-1-1,0.05} = 9{,}488$

Поскольку 0,2<9,488, следовательно, нет оснований отвергать проверяемую гипотезу.

Задачи.

Задача 190. Произведено $n = 200$ испытаний, в результате каждого из которых событие А появлялось в различные моменты времени, в итоге было получено эмпирическое распределение, приведенное в следующей таблице:

Интервал (x_{i-1}, x_i)	Частота n_i	Интервал (x_{i-1}, x_i)	Частота n_i
(2,4)	**21**	**(12-14)**	**14**
(4,6)	**16**	**(14-16)**	**21**
(6-8)	**15**	**(16,18)**	**22**
(8-10)	**26**	**(18,20)**	**18**
(10-12)	**22**	**(20,22)**	**25**

(в первой строке указаны интервалы времени в минутах, во второй - число появлений события А в интервале). Требуется при уровне значимости $\alpha = 0{,}05$ проверить гипотезу о том, что время появления событий распределено равномерно.

Задача 191. Проведено $n = 100$ опытов. Каждый опыт состоит из $N = 10$ испытаний, в каждом из которых вероятность p появления события А равна 0,3. В итоге получено следующее эмпирическое распределение (в первой строке указано число x_i - появлений события А в одном опыте; во второй - число опытов, в которых наблюдается x_i появлений события А):

x_i	**0**	**1**	**2**	**3**	**4**	**5**
n_i	**2**	**10**	**27**	**32**	**23**	**6**

Требуется при уровне значимости $\alpha = 0{,}05$ проверить гипотезу о том, что дискретная случайная величина X (число появлений события А) распределена по биномиальному закону.

Задача 192. Выборка разбита на 9 интервалов, $\alpha = 0.05$. Тогда критическая точка для проверки гипотезы о нормальном распределении выборки определяется следующим образом

а) $\chi^2_{9,0.05}$, б) $\chi^2_{8,0.05}$, в) $\chi^2_{6,0.05}$, г) $\chi^2_{5,0.05}$.

Задача 193. Выборка разбита на 12 интервалов, $\alpha = 0.05$. Тогда критическая точка для проверки гипотезы о распределении Пуассона определяется следующим образом

а) $\chi^2_{10,0.05}$, б) $\chi^2_{12,0.05}$, в) $\chi^2_{11,0.05}$, г) $\chi^2_{9,0.05}$.

21. Двумерная регрессионная модель

Общая задача исследования зависимостей, осуществляемая в рамках классического корреляционного и регрессионного анализа, может быть сформулирована следующим образом: - по результатам N измерений

$$x_i, y_i, i = \overline{1,n}$$

переменных X, Y построить такую функцию $f(X)$, которая бы наилучшим (в определенном смысле) образом восстанавливала значения переменной Y по значениям переменной X. Частными случаями этой общей задачи являются задачи построения:

а) линейной регрессионной модели

$$f(X) = a + bX;$$

б) показательной модели

$$f(X) = ae^{bX};$$

в) степенной модели

$$f(X) = bx^a.$$

В качестве меры отклонения функции $f(X)$ от Y можно предложить наиболее распространенные функционалы:

1) сумма квадратов отклонений $F = \sum_{i=1}^{n} (y_i - f(x_i))^2$;

2) сумма модулей отклонений $F = \sum_{i=1}^{n} |y_i - f(x_i)|$.

Оба функционала имеют свои достоинства и недостатки.

Плюсы суммы квадратов отклонений: легкость вычислительной процедуры; хорошие статистические свойства; простота математических выводов; возможность построения развитой теории, позволяющей провести проверку различных статистических гипотез.

Минусы суммы квадратов отклонений: чувствительность к выбросам.

Плюсы суммы модулей: робастность, то есть нечувствительность к выбросам.

Минусы суммы модулей отклонений: сложность вычислительной процедуры; необходимость большим отклонениям придавать больший вес; неоднозначность, то есть разным значениям параметра b могут соответствовать одинаковые суммы модулей отклонений.

21. 1. Метод наименьших квадратов (МНК)

Рассмотрим задачу наилучшей аппроксимации набора наблюдений $x_i, y_i, i = \overline{1,n}$ линейной функцией $f(X) = a + bX$ в смысле минимизации функционала

$$F = \sum_{i=1}^{n} \left(y_i - \left(a + bx_i\right)\right)^2 .$$

Запишем необходимые условия экстремума:

$$\frac{\partial F}{\partial a} = -2\sum_{i=1}^{n} \left(y_i - a - bx_i\right) = 0$$

$$\frac{\partial F}{\partial b} = -2\sum_{i=1}^{n} x_i\left(y_i - a - bx_i\right) = 0$$

или

$$\sum_{i=1}^{n} \left(y_i - a - bx_i\right) = 0$$

$$\sum_{i=1}^{n} x_i\left(y_i - a - bx_i\right) = 0 .$$

Раскрыв скобки, получим:

$$an + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2} = \sum_{i=1}^{n} x_i y_i .$$

Решая систему уравнений, находим неизвестные a и b:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{b}{n}\sum_{i=1}^{n} x_i .$$

Полученные результаты можно выразить, используя коэффициент корреляции X и Y - r_{xy}:

$$b = r_{xy}\frac{\sigma_y}{\sigma_x},$$

$$a = \overline{y} - b\overline{x}.$$

Коэффициент b называется **выборочным коэффициентом регрессии**.

Рассмотрим задачу наилучшей аппроксимации набора наблюдений $x_i, y_i, i = \overline{1,n}$ показательной функцией $f(X) = ae^{bX}$.

Логарифмируя обе части уравнения $Y = ae^{bX}$, получим $ln Y = ln\, a + bX = A + bX,\ A = ln\, a$. Для функционала

$$F = \sum_{i=1}^{n}\left(ln\, y_i - \left(A + bx_i\right)\right)^2$$

запишем необходимые условия экстремума:

$$\frac{\partial F}{\partial A} = -2\sum_{i=1}^{n}\left(ln\, y_i - A - bx_i\right) = 0$$

$$\frac{\partial F}{\partial b} = -2\sum_{i=1}^{n}x_i\left(ln\, y_i - a - bx_i\right) = 0$$

или

$$\sum_{i=1}^{n}\left(ln\, y_i - A - bx_i\right) = 0$$

$$\sum_{i=1}^{n}x_i\left(ln\, y_i - A - bx_i\right) = 0.$$

После преобразования получим:

$$An + b\sum_{i=1}^{n}x_i = \sum_{i=1}^{n}ln\, y_i$$

$$A\sum_{i=1}^{n}x_i + b\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}x_i\, ln\, y_i.$$

Находим неизвестные A и b:

$$b=\frac{n\sum_{i=1}^{n}x_i\,ln\,y_i-\left(\sum_{i=1}^{n}x_i\right)\left(\sum_{i=1}^{n}ln\,y_i\right)}{n\sum_{i=1}^{n}x_i^{\,2}-\left(\sum_{i=1}^{n}x_i\right)^2},$$

$$A=\frac{1}{n}\sum_{i=1}^{n}ln\,y_i-\frac{b}{n}\sum_{i=1}^{n}x_i\,.$$

Соответственно $a=e^A=e^{\frac{1}{n}\sum_{i=1}^{n}ln\,y_i-\frac{b}{n}\sum_{i=1}^{n}x_i}$.

Рассмотрим задачу наилучшей аппроксимации набора наблюдений $x_i,y_i,i=\overline{1,n}$ степенной функцией $f(X)=bx^a$.

Логарифмируя обе части уравнения $Y=bx^a$, получим $\ln Y=\ln b+a\ln X=B+a\ln X$, $B=\ln b$. Для функционала

$$F=\sum_{i=1}^{n}\left(ln\,y_i-\left(B+a\,ln\,x_i\right)\right)^2$$

запишем необходимые условия экстремума:

$$\frac{\partial F}{\partial B}=-2\sum_{i=1}^{n}\left(ln\,y_i-B-a\,ln\,x_i\right)=0$$

$$\frac{\partial F}{\partial a}=-2\sum_{i=1}^{n}ln\,x_i\left(ln\,y_i-B-a\,ln\,x_i\right)=0$$

или

$$\sum_{i=1}^{n}\left(ln\,y_i-B-a\,ln\,x_i\right)=0$$

$$\sum_{i=1}^{n}ln\,x_i\left(ln\,y_i-B-a\,ln\,x_i\right)=0\,.$$

После преобразования получим:

$$Bn + a\sum_{i=1}^{n} ln\, x_i = \sum_{i=1}^{n} ln\, y_i$$

$$B\sum_{i=1}^{n} ln\, x_i + a\sum_{i=1}^{n} (ln\, x_i)^2 = \sum_{i=1}^{n} ln\, x_i\, ln\, y_i \, .$$

Находим неизвестные B и a:

$$a = \frac{n\sum_{i=1}^{n} ln\, x_i\, ln\, y_i - \left(\sum_{i=1}^{n} ln\, x_i\right)\left(\sum_{i=1}^{n} ln\, y_i\right)}{n\sum_{i=1}^{n} (ln\, x_i)^2 - \left(\sum_{i=1}^{n} ln\, x_i\right)^2},$$

$$B = \frac{1}{n}\sum_{i=1}^{n} ln\, y_i - \frac{a}{n}\sum_{i=1}^{n} ln\, x_i \, .$$

Соответственно $b = e^B = e^{\frac{1}{n}\sum_{i=1}^{n}\ln y_i - \frac{a}{n}\sum_{i=1}^{n}\ln x_i}$.

21.2. Спецификация модели и основные гипотезы

Добавим к постановке задачи некоторые статистические данные и запишем линейное регрессионное уравнение в виде:

$$Y_i = a + bX_i + \varepsilon_i , i = \overline{1,n} ,$$

где X_i - неслучайная (детерминированная) величина, Y_i, ε_i - случайные величины, ε_i - ошибки регрессии.

Основные гипотезы:

1. $Y_i = a + bX_i + \varepsilon_i$, $i = \overline{1,n}$ - спецификация модели.

2. X_i - детерминированная величина; вектор $(X_1, X_2, ..., X_n)$ не коллинеарен вектору $(1,1,...,1)$.

3. $M(\varepsilon_i) = 0, D(\varepsilon_i) = \sigma^2$, $i = \overline{1,n}$, $M(\varepsilon_i, \varepsilon_j) = 0, i \neq j$.

Часто добавляется условие

4. ε_i- нормально распределенная случайная величина, $M(\varepsilon_i) = 0, D(\varepsilon_i) = \sigma^2$.

Как утверждает **теорема Гаусса – Маркова**, в этих предположениях оценки неизвестных параметров модели

$$\hat{b} = \frac{n\sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}$$

и

$$\hat{a} = \frac{1}{n}\sum_{i=1}^{n} Y_i - \frac{\hat{b}}{n}\sum_{i=1}^{n} X_i,$$

полученные по МНК, имеют наименьшую дисперсию в классе всех линейных несмещенных оценок.

Непосредственно из 1) - 4) следует, что Y_i- нормально распределенная случайная величина, $M(Y_i) = a + bX_i, D(Y_i) = \sigma^2$.

Нетрудно проверить, что

$$M(\hat{b}) = b,\ M(\hat{a}) = a, D(\hat{b}) = S_b^2 = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}, D(\hat{a}) = S_a^2 = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)},$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{, поэтому } \hat{a} \in N\left(a, \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)}\right),\ \hat{b} \in N\left(b, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right).$$

Обозначим через $e_i, i = \overline{1,n}$ разницу между действительным значением переменной Y и модельным значением этой переменной, то есть

$$e_i = Y_i - a - bX_i, i = \overline{1,n}.$$

Несмещенной оценкой дисперсии ошибок σ^2 является

$$S^2 = \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2 .$$

Нетрудно показать, что S^2 независима с $\hat{a}$ и $\hat{b}$, а $\frac{(n-2)S^2}{\sigma^2} \in \chi^2_{n-2}$.

Построим статистику для проверки гипотезы $H_0 : b = b_0$ против альтернативной гипотезы $H_1 : b \neq b_0$.

Поскольку $\hat{b} - b \in N\left(0, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\right) \in N(0, \sigma_b^2)$, то $\frac{\hat{b} - b}{\sigma_b} \in N(0,1)$. Из условия $\frac{(n-2)S^2}{\sigma^2} \in \chi^2_{n-2}$ следует, что

$$\frac{\dfrac{\hat{b}-b}{\sigma_b}}{\sqrt{\dfrac{(n-2)S^2}{\sigma^2(n-2)}}}=\frac{\hat{b}-b}{\dfrac{\sigma_b S}{\sigma}}=\frac{\hat{b}-b}{S_b}\in t_{n-2}$$

(распределению Стьюдента с $n-2$ степенями свободы).

Таким образом, для проверки гипотезы $H_0: b=b_0$ против альтернативной гипотезы $H_1: b\neq b_0$ будет использоваться статистика $\frac{\hat{b}-b}{S_b}$.

Построим доверительный интервал для b, используя распределение t_{n-2} и его двусторонние квантили $\hat{t}_{n-2,\alpha}$ ($P\left(\left|t_{n-2}\right|\leq \hat{t}_{n-2,1-\alpha}\right)=1-\alpha$):

$$P\left(-\hat{t}_{n-2,1-\alpha}\leq\frac{\hat{b}-b}{S_b}\leq\hat{t}_{n-2,1-\alpha}\right)=1-\alpha .$$

Если b_0 принадлежит отрезку $\left[\hat{b}-\hat{t}_{n-2,1-\alpha}S_b, \hat{b}+\hat{t}_{n-2,1-\alpha}S_b\right]$, то принимается гипотеза H_0, в противном случае принимается гипотеза H_1.

Если требуется проверить наличие связи между переменными X и Y, то используется статистика $\frac{\hat{b}}{S_b}$, тем самым проверяется равенство нулю коэффициента b.

При $\left|\frac{\hat{b}}{S_b}\right|>\hat{t}_{n-2,1-\alpha}$ делается вывод о достоверной связи между переменными X и Y, при $-\hat{t}_{n-2,1-\alpha}\leq\frac{\hat{b}-b}{S_b}\leq\hat{t}_{n-2,1-\alpha}$ делается вывод об ее отсутствии.

Аналогично можно показать, что $\frac{\hat{a}-a}{S_a}\in t_{n-2}$ и использовать эту статистику для проверки гипотез относительно коэффициента a.

Рассмотрим статистику $\frac{Y-(a+bX)}{\sigma}$, которая принадлежит стандартному нормальному распределению - $N(0{,}1)$. При известной σ^2 (дисперсии ошибок) можно было бы использовать $N(0{,}1)$ для прогнозирования значений Y в виде доверительных интервалов.

Поскольку σ^2 неизвестно, то будем использовать ее оценку S^2, для которой известно, что $\frac{(n-2)S^2}{\sigma^2} \in \chi^2_{n-2}$.

Таким образом,

$$\frac{\dfrac{Y-(a+bX)}{\sigma}}{\dfrac{S}{\sigma}} = \frac{Y-(a+bX)}{S} \in t_{n-2}$$

и используется для построения доверительных интервалов с целью прогнозирования значений Y:

$$a+bX-t_{n-2,\alpha}S \le Y \le a+bX+t_{n-2,\alpha}S\,.$$

Пример. По экспериментальным данным исследовать зависимость между переменными X, Y и Z, Y.

Y	X	Z
33	20	1.6
35	17	3.1
37	15	4.5
39	15	6.3
32	21	1.2
30	25	0.7
35	14	3.5
41	10	7.8

38	14	4.9
43	9	10.1

Решение. МНК дает следующие зависимости между исследуемыми переменными:

$$Y = -0.789X + 48.920,$$

$$Y = 1.331Z + 30.480.$$

Зададимся доверительной вероятностью 0.95, определим $t_{0.95}$ и обозначим статистику $\frac{\hat{b}}{S_b}$, вычисляемую по данным X, Y через $t(X)$, а вычисляемую по данным Z, Y через $t(Z)$.

Тогда $t(X) = 8.630 > t_{0.95} = 2.310,\ t(Z) = 16.330 > t_{0.95} = 2.310$, что говорит о достоверной статистической связи между X, Y и Z, Y.

Построим прогноз значения Y при $X = 54$. Используя ранее построенный доверительный интервал для Y и задавшись уровнем значимости 0.05 (доверительной вероятностью 0.95), получим:
$Y \in [-12.561, 25.189]$.

Задачи.

Задача 194. Выборочный коэффициент линейной регрессии равен 2, $\bar{x} = 10, \bar{y} = 35$. Определить уравнение регрессии.

Задача 195. Выборочный коэффициент линейной регрессии равен -3, $\bar{x} = 12, \bar{y} = 40$. Определить уравнение регрессии.

Задача 196. Выборочный коэффициент корреляции равен -0.8, $\sigma_x = 2,\ \sigma_y = 3,\ \bar{x} = 12, \bar{y} = 40$. Определить уравнение регрессии.

Задача 197. Выборочный коэффициент корреляции равен 0.9, $\sigma_x = 4,\ \sigma_y = 8,\ \bar{x} = 10, \bar{y} = 20$. Определить уравнение регрессии.

22. Оценка точного значения измеряемой величины

Рассмотрим результаты измерения некоторой величины

$$x_1, x_2, \ldots, x_n.$$

Будем предполагать, что эти результаты свободны от грубых и систематических ошибок.

Требуется оценить точное значение a измеряемой величины.

Если все измерения произведены с одинаковой точностью, то в качестве оценки a принимают среднее арифметическое значение результатов измерения, то есть

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

Переходя к вероятностной модели эксперимента, можно утверждать, что эта оценка является несмещенной и состоятельной.

Введем дополнительное условие, состоящее в том, что случайные ошибки измерений принадлежат стандартному нормальному закону распределения.

Используя доверительные интервалы, запишем

$$P\left(\left|\frac{\sqrt{n}\left(a-\bar{x}\right)}{\sigma}\right| \le x_{\frac{1-\alpha}{2}}\right) = 1-\alpha ,$$

$$\left|\frac{\sqrt{n}\left(a-\bar{x}\right)}{\sigma}\right| \le x_{\frac{1-\alpha}{2}} ,$$

$$\bar{x} - \frac{\sigma x_{\frac{1-\alpha}{2}}}{\sqrt{n}} \le a \le \bar{x} + \frac{\sigma x_{\frac{1-\alpha}{2}}}{\sqrt{n}} ,$$

где $\Phi\left(x_{\frac{1-\alpha}{2}}\right) = \frac{1}{\sqrt{2\pi}} \int\limits_0^{x_{\frac{1-\alpha}{2}}} e^{-\frac{y^2}{2}} dy = \frac{1-\alpha}{2}$.

Точность оценки выражается формулой

$$\Delta = \frac{\sigma x_{\frac{1-\alpha}{2}}}{\sqrt{n}}.$$

Если средняя квадратическая погрешность σ неизвестна, то вместо нее применяется $s_1 = \sqrt{\frac{1}{n-1}\sum\limits_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$, а доверительные интервалы строятся с помощью распределения Стьюдента.

Вопросы по теории вероятностей и математической статистике

1. Вероятностное пространство.
2. Свойства вероятности события, теорема сложения.
3. Число перестановок, сочетаний и размещений без повторений и с повторениями. Примеры применения.
4. Геометрические вероятности.
5. Наивероятнейшее число появлений события в ряде испытаний.
6. Локальная теорема Муавра-Лапласа.
7. Интегральная теорема Муавра-Лапласа.
8. Условная вероятность, теорема умножения.
9. Формула полной вероятности.
10. Формула Байеса.
11. Правило трех сигм.
12. Случайные величины: дискретные и непрерывные.
13. Функция распределения вероятностей случайной величины, ее свойства.
14. Плотность распределения вероятностей случайной величины, ее свойства.
15. Математическое ожидание случайной величины и ее свойства.
16. Дисперсия случайной величины и ее свойства.
17. Равномерно распределенная случайная величина, плотность и функция распределения ее вероятностей.
18. Равномерно распределенная случайная величина, ее математическое ожидание и дисперсия.
19. Показательно распределенная случайная величина, плотность и функция распределения ее вероятностей.

20. Показательно распределенная случайная величина, ее математическое ожидание и дисперсия.
21. Нормально распределенная случайная величина, плотность и функция распределения ее вероятностей.
22. Нормально распределенная случайная величина, ее математическое ожидание и дисперсия.
23. Неравенство Чебышева, закон больших чисел.
24. Центральная предельная теорема.
25. Ковариация и коэффициент корреляции двух случайных величин.
26. Выборка, гистограмма, полигон частот.
27. Эмпирическая функция распределения.
28. Выборочное среднее и выборочная дисперсия.
29. Вариационный ряд. Порядковые статистики, их распределения.
30. Свойство несмещенности точечных оценок неизвестных параметров. Примеры.
31. Свойство состоятельности точечных оценок неизвестных параметров. Примеры.
32. Свойства выборочного среднего и выборочной дисперсии.
33. Метод моментов в построении точечных оценок неизвестных параметров. Примеры.
34. Метод максимального правдоподобия в построении точечных оценок неизвестных параметров. Примеры.
35. Доверительное оценивание.
36. Доверительные интервалы для неизвестного математического ожидания в нормальной выборке.
37. Доверительные интервалы для неизвестной дисперсии в нормальной выборке.
38. Распределения хи-квадрат, Стьюдента, Фишера.

39. Критерий, критическая область, ошибки первого и второго рода, мощность критерия.

40.Критерий хи-квадрат Пирсона.

Ответы

№ задачи	Ответ
1	$\Omega = \{a_1, a_2, ..., a_n,\ a_i = Г\ или\ P\}$,общее число исходов 2^n
2	$\Omega = \{1,2,3,4,5,6\}$
3	$\Omega = \{Г1, Г2, Г3, Г4, Г5, Г6, РГ, РР\}$
4	$\Omega = \{a_1, a_2, ..., a_n\}, a_i = \overline{1, N}$
5	$\Omega = \{a_1, a_2, ..., a_n\}, a_i = \overline{1, N}$, $a_i \neq a_j, i \neq j$
6	а) $A\overline{B} + \overline{A}B$; б) $A\overline{B}$; в) $\overline{A}B$; г) $\overline{A}\overline{B}$.
7	а) $\overline{A}\overline{B}$; б) $A\overline{B}$; в) $\overline{A}B$.
8	$\overline{A}B$
9	**0.75**
10	**0.375**

11	$\frac{1}{3}$
12	$\frac{1}{72}$
13	**0.56**
14	$\frac{2}{9}$
15	а) 0.054; б) 0.504; в) 0.006; г) 0.994; д) 0.056; е) 0.092; ж) 0.398.
16	а) 0.21; б) 0.973; в) 0.9919.
17	$A_1(A_2 + A_3)$
18	0.644
19	а) 0.6762; б) 0.7084; в) 0.7406.
20	$P(A \cup B) = P(A) + P(B) - P(AB) \Rightarrow P(AB) = 0.12 + 0.13 - 0.15 = 0.1$
21	0.08

22	$\frac{4}{7}$
23	а) 0.06; б) 0.04
24	$\frac{2}{3}$
25	$P_5 = 5! = 5 \cdot 4 \cdot 3 \cdot 2 = 120$
26	$A_5^3 = \frac{5!}{2!} = 3 \cdot 4 \cdot 5 = 60$
27	$C_5^3 = \frac{5!}{2! \cdot 3!} = 10$
28	$6^3 = 216$
29	$P_{11}(5,2,2) = \frac{11!}{5!2!2!} = 83160$
30	$A_3^{-5} = 3^5 = 243, C_3^{-5} = \frac{7!}{5!2!} = 21$
31	$\frac{1}{6}$
32	**а)** $\frac{C_9^5}{C_{10}^6}$**, б)** $\frac{C_8^4}{C_{10}^6}$**, в)** $\frac{C_7^3}{C_{10}^6}$
33	**0.25**
34	$\frac{C_4^1 C_{32}^2}{C_{36}^3}$
35	$1 - \frac{C_{32}^3}{C_{36}^3}$
36	$\frac{5}{12}$
37	а) $\frac{C_5^2}{C_{11}^2} = \frac{4 \cdot 5}{10 \cdot 11} = \frac{2}{11}$;

	б) $\frac{C_6^2}{C_{11}^2}=\frac{5\cdot 6}{10\cdot 11}=\frac{3}{11}$; в) $\frac{C_5^1C_6^1}{C_{11}^2}=\frac{2\cdot 5\cdot 6}{10\cdot 11}=\frac{6}{11}$.
38	а) $\frac{C_5^3}{C_{10}^3}=\frac{3\cdot 4\cdot 5}{8\cdot 9\cdot 10}=\frac{1}{12}$; б) $\frac{C_3^3}{C_{10}^3}=\frac{2\cdot 3}{8\cdot 9\cdot 10}=\frac{1}{120}$; в) $\frac{C_5^2C_2^1}{C_{10}^3}=\frac{2\cdot 3\cdot 4\cdot 5}{8\cdot 9\cdot 10}=\frac{1}{6}$; г) $\frac{C_2^2C_3^1}{C_{10}^3}=\frac{2\cdot 3\cdot 3}{8\cdot 9\cdot 10}=\frac{1}{40}$; д) $\frac{C_5^1C_2^1C_3^1}{C_{10}^3}=\frac{2\cdot 3\cdot 2\cdot 3\cdot 5}{8\cdot 9\cdot 10}=\frac{1}{4}$.
39	$P(F)=\frac{\frac{l}{2}}{l}=\frac{1}{2}$
40	**0.5**
41	$\frac{R^2-r^2}{R^2}$
42	0.8
43	$\frac{\pi 4^2-\pi 3^2}{\pi 4^2}=\frac{7}{16}$
44	$\frac{60^2-45^2}{60^2}=0.4375$
45	**20**

46	**72**
47	**0.275** (локальная теорема Муавра-Лапласа) **0.282** (прямое определение вероятности)
48	**0.887**
49	$2\Phi\left(\varepsilon\sqrt{\frac{n}{pq}}\right)$
50	$C_5^2(0.2)^3(0.8)^2=0.0512$
51	$C_5^4(0.2)^1(0.8)^4+C_5^5(0.2)^0(0.8)^5=0.4096+0.32768=0.73728$
52	Теорема Пуассона $\lambda=np=1000\cdot 0.003=3,\ P_{1000}(\le 3)=\frac{3^0e^{-3}}{0!}+\frac{3^1e^{-3}}{1!}+$ $+\frac{3^2e^{-3}}{2!}+\frac{3^3e^{-3}}{3!}=e^{-3}+3e^{-3}+4.5e^{-3}+4.5e^{-3}=13e^{-3}$
53	Теорема Пуассона $\lambda=np=1000\cdot 0.002=2,\ P_{1000}(<2)=\frac{2^0e^{-2}}{0!}+\frac{2^1e^{-2}}{1!}=$ $=e^{-2}+2e^{-2}=3e^{-2}$
54	а) $n=2$; б) $n=4$; в) $n=8$.
55	а) $n=6$; б) $n=8$; в) $n=10$.
56	а) $n=16$; б) $n=24$; в) $n=32$.
57	$\frac{3}{35}$

58	$\frac{4}{35}$
59	$P(A_k)=\sum_{j=k}^{\infty}P(A_k\|B_j)\cdot P(B_j)=\sum_{j=k}^{\infty}\frac{e^{-\lambda}\lambda^j}{j!}\cdot\frac{j!}{k!\,(j-k)!}\cdot p^k(1-p)^{j-k}=$ $=\frac{e^{-\lambda}p^k\lambda^k}{k!}\cdot\sum_{j=0}^{\infty}\frac{(\lambda(1-p))^j}{j!}=\frac{e^{-\lambda p}(\lambda p)^k}{k!}.$
60	$\frac{1}{3}$
61	**0.845**
62	$\frac{1}{3}\left(\frac{2}{5}+\frac{1}{2}+\frac{7}{15}\right)=\frac{41}{90}$
63	B_1 -событие, что купленное изделие выпущено первым заводом, B_2 -событие, что купленное изделие выпущено вторым заводом. Если Y - объем продукции, выпущенной первым заводом, $P(B_1)=\frac{Y}{Y+kY}=\frac{1}{1+k}$, $P(B_2)=\frac{kY}{Y+kY}=\frac{k}{1+k}$, $P(A\|B_1)=p_1$, $P(A\|B_2)=p_2$, $P(B_2\|A)=\frac{\frac{k}{1+k}\cdot p_2}{\frac{1}{1+k}\cdot p_1+\frac{k}{1+k}\cdot p_2}=\frac{kp_2}{p_1+kp_2}.$
64	а) $P(B_3\|A)=\frac{2}{5}$, б) $P(B_1\|A)=\frac{2}{5}$, в) $P(B_2\|A)=\frac{1}{5}$.
65	$\frac{1}{3}$
66	$\frac{2}{3}$
67	$\frac{13}{30}$

68	**0.55**
69	$\frac{n}{N}$
70	$\frac{2}{5}$
71	**Одинаковая, равная** $\frac{1}{25}$
74	$F(x)=\begin{cases} 0, x<0 \\ q, 0 \le x < 1 \\ 1, x \ge 1 \end{cases}$, $M(X)=p$, $D(X)=pq$
75	**0.6**
76	**г)**
77	**б)**
78	<table><tr><td>x_i</td><td>**-900**</td><td>**0**</td><td>**1000**</td><td>**2000**</td></tr><tr><td>p_i</td><td>**0.2**</td><td>**0.5**</td><td>**0.2**</td><td>**0.1**</td></tr></table>
79	$M(X)=np$, $D(X)=npq$ (представить биномиальную случайную величину в виде суммы независимых случайных величин, имеющих распределение Бернулли)
80	а) увеличится в 3 раза; б) уменьшится в 3 раза; в) увеличится на 4 единицы; г) уменьшится на 5 единиц.
81	а) увеличится в 9 раз; б) уменьшится в 9 раз; в) не изменится; г) не изменится.

82	$f(x)=\begin{cases}0, x<0\\ \frac{2x}{25}, 0\le x<5\\ 0, x\ge 5\end{cases}$
83	$P(1<X\le 4)=F(4)-F(1)=\frac{16}{25}-\frac{1}{25}=\frac{3}{3},$ $P(2<X\le 6)=F(6)-F(2)=1-\frac{4}{25}=\frac{21}{25}$
84	$\int_0^4 cx^2dx=1,\ \frac{c4^3}{3}=1\Rightarrow c=\frac{3}{64},\ F(x)=\begin{cases}0, x<0\\ \frac{x^3}{64}, 0\le x<4\\ 1, x\ge 4\end{cases}.$
85	$M(X)=4.4,\ D(X)=3.84$.
86	$M(X)=\frac{8}{3},\ D(X)=\frac{8}{9}$
87	$f(x)=\begin{cases}0, x<0\\ \frac{1}{5}, 0\le x<5\\ 0, x\ge 5\end{cases},\ M(X)=2.5,\ D(X)=\frac{25}{12}$
88	$M(X)=5, D(X)=3, \sigma=\sqrt{3}$
89	$\frac{49\pi}{12}$
90	$a=\frac{2}{\pi}, M(X)=0$
91	$D(X)=npq=100\cdot 0.8\cdot 0.2=16,\ \varepsilon=10,\ \frac{D(X)}{\varepsilon^2}=0.16\Rightarrow P\ge 1-0.16=0.84$ $\varepsilon=10$, так как $M(X)=np=100\cdot 0.8=80\Rightarrow \varepsilon=90-80=10$
92	$D(X)=npq=50\cdot 0.7\cdot 0.3=10.5,\ \varepsilon=5,\ \frac{D(X)}{\varepsilon^2}=0.42\Rightarrow P\ge 1-0.42=0.58$

<table>
<tr><td>93</td><td>$D(X) = npq = 200 \cdot 0.6 \cdot 0.4 = 48,\ \varepsilon = 20,\ \frac{D(X)}{\varepsilon^2} = 0.12 \Rightarrow P \geq 1 - 0.12 = 0.88$</td></tr>
<tr><td>94</td><td>$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - a\right| \leq \varepsilon\right) \geq 1 - \frac{\sum_{i=1}^{n} D(X_i)}{\varepsilon^2 n^2} = 1 - \frac{D(X_1)}{\varepsilon^2 n}$$
$$\varepsilon n = 50, \varepsilon = \frac{1}{2} \Rightarrow P \geq 1 - \frac{5}{25} = 0.8$$</td></tr>
<tr><td>95</td><td>$P \geq 0.8$</td></tr>
<tr><td>99a</td><td><table><tr><td>4</td><td>5</td><td>6</td><td>7</td></tr><tr><td>0.08</td><td>0.32</td><td>0.12</td><td>0.48</td></tr></table></td></tr>
<tr><td>101</td><td><table><tr><td>1</td><td>3</td></tr><tr><td>$\frac{1}{9}$</td><td>$\frac{8}{9}$</td></tr></table></td></tr>
<tr><td>102</td><td>0.55</td></tr>
<tr><td>103</td><td>0.8</td></tr>
<tr><td>105</td><td>г)</td></tr>
<tr><td>107</td><td>б)</td></tr>
<tr><td>109</td><td>$$F(z) = \int_{-\infty}^{z} \int_{-\infty}^{+\infty} f(x) g(u - x) dx du$$</td></tr>
<tr><td>110</td><td>$$F\left(\frac{x - b}{a}\right)$$</td></tr>
<tr><td>114</td><td>6</td></tr>
<tr><td>115</td><td>28</td></tr>
<tr><td>116</td><td>0.2 и 20</td></tr>
<tr><td>117</td><td>0.3 и 15</td></tr>
<tr><td>118</td><td>$$\frac{F(x)(1 - F(x))}{n}$$</td></tr>
</table>

124	$\overline{x}=4,\ s_0^2=18.7$
128	$\overline{x}=24{,}75;\ s_1^2=4{,}917$.
129	$\overline{x}=10{,}06,\ s_1^2=2{,}99$
130	$\overline{x}=166{,}6,\ s_1^2=39{,}47$
131	Оценкой моды является значение 10, так как у него наибольшая частота. Оценкой медианы является значение 10, поскольку $x^{(57)}=10$.
132	**12; 13; 14; 14; 14; 14; 15; 15; 15; 16.** Мода и медиана равны14
133	Размах равен 4, мода равна 1, медиана равна 2
134	Размах равен 13, две моды равные 4 и 6, медиана равна 5.5
135	Размах равен 25, две моды равные 11 и 16, медиана равна 14
136	$P\left(X_{(n)}\le x\right)=\left[F(x)\right]^n$.
137	$P\left(X_{(k)}\le x\right)=\sum\limits_{i=k}^{n}C_n^i\left[F(x)\right]^i\left[1-F(x)\right]^{n-i}$.
138	$\overline{x}=10,s_1^2=4$
139	$\overline{x}=18,s_1^2=16$
140	$\overline{x}=16,s_1^2=9$

141	$2\overline{X}$ является несмещенной и состоятельной оценкой для параметра θ
142	а) $X_{(1)} - \frac{\sigma}{n}$; б) $\overline{X} - \mu$
143	$\tilde{m} = 11{,}95$
144	**6**
145	**14**
148	$\hat{\theta} = -2\overline{x}$
152	$\hat{\lambda} = \left(\overline{x}\right)^{-1}$
157	**б)**
158	**г)**
159	**5**
160	**1**
161	**г)**
162	**б)**
163	**г)**
164	**б), г)**

165	**1.2384**
166	**1.376**
167	**в)**
168	**б)**
169	$499{,}88 < m < 501{,}12$
170	$291{,}235$ ч $< m < 308{,}765$ ч; $240{,}038$ ч$^2 < \sigma^2 < 826{,}446$ ч^2
177	**в)**
178	**а)**
179	Гипотезу H_0 при уровне значимости 0,01 можно принять
180	$\frac{(n-1)s_1^2}{\sigma^2} = \frac{35 \cdot 4}{5} = 28$
181	8
184	**б)** Число степеней свободы на единицу меньше объема выборки, первым стоит число степеней свободы с большей дисперсией
185	**г)**

186	Гипотеза о равенстве дисперсий отвергается, поэтому точность второго станка выше точности первого
187	Средние результаты измерений различаются незначимо
188	Поскольку статистика, вычисленная на основе наблюдений $>\Delta_{14;0,975}$, то результаты анализов различаются значимо
189	Результаты анализа различаются незначимо
190	Данные наблюдений согласуются с основной гипотезой
191	Гипотеза о биномиальном распределении X отвергается
192	**в)**
193	**а)**
194	$y = 2x + 15$
195	$y = -3x + 76$
196	$y = -1.2x + 54.4$
197	$y = 1.8x + 2$

Приложение 1.

Таблица значений плотности стандартного нормального распределения

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

x	0	1	2	3	4	5	6	7	8	9
0,0	0,398942	0,398922	0,398862	0,398763	0,398623	0,398444	0,398225	0,397966	0,397668	0,397330
0,1	0,396953	0,396536	0,396080	0,395585	0,395052	0,394479	0,393868	0,393219	0,392531	0,391806
0,2	0,391043	0,390242	0,389404	0,388529	0,387617	0,386668	0,385683	0,384663	0,383606	0,382515
0,3	0,381388	0,380226	0,379031	0,377801	0,376537	0,375240	0,373911	0,372548	0,371154	0,369728
0,4	0,36827	0,366782	0,365263	0,363714	0,362135	0,360527	0,358890	0,357225	0,355533	0,353812
0,5	0,352065	0,350292	0,348493	0,346668	0,344818	0,342944	0,341046	0,339124	0,337180	0,335213
0,6	0,333225	0,331215	0,329184	0,327133	0,325062	0,322972	0,320864	0,318737	0,316593	0,314432
0,7	0,312254	0,310060	0,307851	0,305627	0,303389	0,301137	0,298872	0,296595	0,294305	0,292004
0,8	0,289692	0,287369	0,285036	0,282694	0,280344	0,277985	0,275618	0,273244	0,270864	0,268477
0,9	0,266085	0,263688	0,261286	0,258881	0,256471	0,254059	0,251644	0,249228	0,246809	0,24439
1,0	0,241971	0,239551	0,237132	0,234714	0,232297	0,229882	0,227470	0,22506	0,222653	0,220251
1,1	0,217852	0,215458	0,213069	0,210686	0,208308	0,205936	0,203571	0,201214	0,198863	0,196520
1,2	0,194186	0,19186	0,189543	0,187235	0,184937	0,182649	0,180371	0,178104	0,175847	0,173602
1,3	0,171369	0,169147	0,166937	0,164740	0,162555	0,160383	0,158225	0,15608	0,153948	0,151831
1,4	0,149727	0,147639	0,145564	0,143505	0,14146	0,139431	0,137417	0,135418	0,133435	0,131468
1,5	0,129518	0,127583	0,125665	0,123763	0,121878	0,120009	0,118157	0,116323	0,114505	0,112704
1,6	0,110921	0,109155	0,107406	0,105675	0,103961	0,102265	0,100586	0,098925	0,097282	0,095657
1,7	0,094049	0,092459	0,090887	0,089333	0,087796	0,086277	0,084776	0,083293	0,081828	0,08038
1,8	0,07895	0,077538	0,076143	0,074766	0,073407	0,072065	0,070740	0,069433	0,068144	0,066871
1,9	0,065616	0,064378	0,063157	0,061952	0,060765	0,059595	0,058441	0,057304	0,056183	0,055079
2,0	0,053991	0,052919	0,051864	0,050824	0,04980	0,048792	0,047800	0,046823	0,045861	0,044915
2,1	0,043984	0,043067	0,042166	0,041280	0,040408	0,039550	0,038707	0,037878	0,037063	0,036262
2,2	0,035475	0,034701	0,033941	0,033194	0,03246	0,031740	0,031032	0,030337	0,029655	0,028985
2,3	0,028327	0,027682	0,027048	0,026426	0,025817	0,025218	0,024631	0,024056	0,023491	0,022937
2,4	0,022395	0,021862	0,021341	0,020829	0,020328	0,019837	0,019356	0,018885	0,018423	0,017971
2,5	0,017528	0,017095	0,016670	0,016254	0,015848	0,015449	0,015060	0,014678	0,014305	0,01394
2,6	0,013583	0,013234	0,012892	0,012558	0,012232	0,011912	0,011600	0,011295	0,010997	0,010706
2,7	0,010421	0,010143	0,009871	0,009606	0,009347	0,009094	0,008846	0,008605	0,00837	0,00814
2,8	0,007915	0,007697	0,007483	0,007274	0,007071	0,006873	0,006679	0,006491	0,006307	0,006127
2,9	0,005953	0,005782	0,005616	0,005454	0,005296	0,005143	0,004993	0,004847	0,004705	0,004567
3,0	0,004432	0,004301	0,004173	0,004049	0,003928	0,003810	0,003695	0,003584	0,003475	0,00337
3,1	0,003267	0,003167	0,00307	0,002975	0,002884	0,002794	0,002707	0,002623	0,002541	0,002461
3,2	0,002384	0,002309	0,002236	0,002165	0,002096	0,002029	0,001964	0,001901	0,001840	0,001780

3,3	0,001723	0,001667	0,001612	0,001560	0,001508	0,001459	0,001411	0,001364	0,001319	0,001275
3,4	0,001232	0,001191	0,001151	0,001112	0,001075	0,001038	0,001003	0,000969	0,000936	0,000904
3,5	0,000873	0,000843	0,000814	0,000785	0,000758	0,000732	0,000706	0,000681	0,000657	0,000634
3,6	0,000612	0,00059	0,000569	0,000549	0,000529	0,000510	0,000492	0,000474	0,000457	0,000441
3,7	0,000425	0,000409	0,000394	0,000380	0,000366	0,000353	0,000340	0,000327	0,000315	0,000303
3,8	0,000292	0,000281	0,000271	0,000260	0,000251	0,000241	0,000232	0,000223	0,000215	0,000207
3,9	0,000199	0,000191	0,000184	0,000177	0,000170	0,000163	0,000157	0,000151	0,000145	0,000139
4,0	0,000134	0,000129	0,000124	0,000119	0,000114	0,000109	0,000105	0,000101	0,000097	0,000093

Приложение 2.

Таблица значений функции Лапласа

$$\Phi(x)=\int_0^x \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy$$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,00000	0,50	0,19146	1,00	0,34134	1,50	0,43319	2,00	0,47725	3,00	0,49865
0,01	0,00399	0,51	0,19497	1,01	0,34375	1,51	0,43448	2,02	0,47831	3,05	0,49886
0,02	0,00798	0,52	0,19847	1,02	0,34614	1,52	0,43574	2,04	0,47932	3,10	0,49903
0,03	0,01197	0,53	0,20194	1,03	0,34849	1,53	0,43699	2,06	0,48030	3,15	0,49918
0,04	0,01595	0,54	0,20540	1,04	0,35083	1,54	0,43822	2,08	0,48124	3,20	0,49931
0,05	0,01994	0,55	0,20884	1,05	0,35314	1,55	0,43943	2,10	0,48214	3,25	0,49942
0,06	0,02392	0,56	0,21226	1,06	0,35543	1,56	0,44062	2,12	0,48300	3,30	0,49952
0,07	0,02790	0,57	0,21566	1,07	0,35769	1,57	0,44179	2,14	0,48382	3,35	0,49960
0,08	0,03188	0,58	0,21904	1,08	0,35993	1,58	0,44295	2,16	0,48461	3,40	0,49966
0,09	0,03586	0,59	0,22240	1,09	0,36214	1,59	0,44408	2,18	0,48537	3,45	0,49972
0,10	0,03983	0,60	0,22575	1,10	0,36433	1,60	0,44520	2,20	0,48610	3,50	0,49977
0,11	0,04380	0,61	0,22907	1,11	0,36650	1,61	0,44630	2,22	0,48679	3,55	0,49981
0,12	0,04776	0,62	0,23237	1,12	0,36864	1,62	0,44738	2,24	0,48745	3,60	0,49984
0,13	0,05172	0,63	0,23565	1,13	0,37076	1,63	0,44845	2,26	0,48809	3,65	0,49987
0,14	0,05567	0,64	0,23891	1,14	0,37286	1,64	0,44950	2,28	0,48870	3,70	0,49989
0,15	0,05962	0,65	0,24215	1,15	0,37493	1,65	0,45053	2,30	0,48928	3,75	0,49991
0,16	0,06356	0,66	0,24537	1,16	0,37698	1,66	0,45154	2,32	0,48983	3,80	0,49993
0,17	0,06749	0,67	0,24857	1,17	0,37900	1,67	0,45254	2,34	0,49036	3,85	0,49994
0,18	0,07142	0,68	0,25175	1,18	0,38100	1,68	0,45352	2,36	0,49086	3,90	0,49995
0,19	0,07535	0,69	0,25490	1,19	0,38298	1,69	0,45449	2,38	0,49134	3,95	0,49996
0,20	0,07926	0,70	0,25804	1,20	0,38493	1,70	0,45543	2,40	0,49180	4,00	0,49997
0,21	0,08317	0,71	0,26115	1,21	0,38686	1,71	0,45637	2,42	0,49224	4,05	0,49997
0,22	0,08706	0,72	0,26424	1,22	0,38877	1,72	0,45728	2,44	0,49266	4,10	0,49998
0,23	0,09095	0,73	0,26730	1,23	0,39065	1,73	0,45818	2,46	0,49305	4,15	0,49998
0,24	0,09483	0,74	0,27035	1,24	0,39251	1,74	0,45907	2,48	0,49343	4,20	0,49999
0,25	0,09871	0,75	0,27337	1,25	0,39435	1,75	0,45994	2,50	0,49379	4,25	0,49999
0,26	0,10257	0,76	0,27637	1,26	0,39617	1,76	0,46080	2,52	0,49413	4,30	0,49999
0,27	0,10642	0,77	0,27935	1,27	0,39796	1,77	0,46164	2,54	0,49446	4,35	0,49999

0,28	0,11026	0,78	0,28230	1,28	0,39973	1,78	0,46246	2,56	0,49477	4,40	0,49999
0,29	0,11409	0,79	0,28524	1,29	0,40147	1,79	0,46327	2,58	0,49506	4,45	0,50000
0,30	0,11791	0,80	0,28814	1,30	0,40320	1,80	0,46407	2,60	0,49534	4,50	0,50000
0,31	0,12172	0,81	0,29103	1,31	0,40490	1,81	0,46485	2,62	0,49560	4,55	0,50000
0,32	0,12552	0,82	0,29389	1,32	0,40658	1,82	0,46562	2,64	0,49585	4,60	0,50000
0,33	0,12930	0,83	0,29673	1,33	0,40824	1,83	0,46638	2,66	0,49609	4,65	0,50000
0,34	0,13307	0,84	0,29955	1,34	0,40988	1,84	0,46712	2,68	0,49632	4,70	0,50000
0,35	0,13683	0,85	0,30234	1,35	0,41149	1,85	0,46784	2,70	0,49653	4,75	0,50000
0,36	0,14058	0,86	0,30511	1,36	0,41309	1,86	0,46856	2,72	0,49674	4,80	0,50000
0,37	0,14431	0,87	0,30785	1,37	0,41466	1,87	0,46926	2,74	0,49693	4,85	0,50000
0,38	0,14803	0,88	0,31057	1,38	0,41621	1,88	0,46995	2,76	0,49711	4,90	0,50000
0,39	0,15173	0,89	0,31327	1,39	0,41774	1,89	0,47062	2,78	0,49728	4,95	0,50000
0,40	0,15542	0,90	0,31594	1,40	0,41924	1,90	0,47128	2,80	0,49744	5,00	0,50000
0,41	0,15910	0,91	0,31859	1,41	0,42073	1,91	0,47193	2,82	0,49760		
0,42	0,16276	0,92	0,32121	1,42	0,42220	1,92	0,47257	2,84	0,49774		
0,43	0,16640	0,93	0,32381	1,43	0,42364	1,93	0,47320	2,86	0,49788		
0,44	0,17003	0,94	0,32639	1,44	0,42507	1,94	0,47381	2,88	0,49801		
0,45	0,17364	0,95	0,32894	1,45	0,42647	1,95	0,47441	2,90	0,49813		
0,46	0,17724	0,96	0,33147	1,46	0,42785	1,96	0,47500	2,92	0,49825		
0,47	0,18082	0,97	0,33398	1,47	0,42922	1,97	0,47558	2,94	0,49836		
0,48	0,18439	0,98	0,33646	1,48	0,43056	1,98	0,47615	2,96	0,49846		
0,49	0,18793	0,99	0,33891	1,49	0,43189	1,99	0,47670	2,98	0,49856		

Приложение 3

Критические точки распределения Стьюдента

k \ α	0,1	0,05	0,02	0,01	0,001
1	6,3138	12,7062	31,8205	63,6567	636,6192
2	2,9200	4,3027	6,9646	9,9248	31,5991
3	2,3534	3,1824	4,5407	5,8409	12,924
4	2,1318	2,7764	3,7469	4,6041	8,6103
5	2,0150	2,5706	3,3649	4,0321	6,8688
6	1,9432	2,4469	3,1427	3,7074	5,9588
7	1,8946	2,3646	2,9980	3,4995	5,4079
8	1,8595	2,3060	2,8965	3,3554	5,0413
9	1,8331	2,2622	2,8214	3,2498	4,7809
10	1,8125	2,2281	2,7638	3,1693	4,5869
11	1,7959	2,2010	2,7181	3,1058	4,4370
12	1,7823	2,1788	2,6810	3,0545	4,3178
13	1,7709	2,1604	2,6503	3,0123	4,2208
14	1,7613	2,1448	2,6245	2,9768	4,1405
15	1,7531	2,1314	2,6025	2,9467	4,0728
16	1,7459	2,1199	2,5835	2,9208	4,0150
17	1,7396	2,1098	2,5669	2,8982	3,9651
18	1,7341	2,1009	2,5524	2,8784	3,9216
19	1,7291	2,0930	2,5395	2,8609	3,8834
20	1,7247	2,0860	2,5280	2,8453	3,8495
21	1,7207	2,0796	2,5176	2,8314	3,8193
22	1,7171	2,0739	2,5083	2,8188	3,7921
23	1,7139	2,0687	2,4999	2,8073	3,7676
24	1,7109	2,0639	2,4922	2,7969	3,7454
25	1,7081	2,0595	2,4851	2,7874	3,7251
26	1,7056	2,0555	2,4786	2,7787	3,7066
27	1,7033	2,0518	2,4727	2,7707	3,6896
28	1,7011	2,0484	2,4671	2,7633	3,6739
29	1,6991	2,0452	2,4620	2,7564	3,6594
30	1,6973	2,0423	2,4573	2,7500	3,6460
35	1,6896	2,0301	2,4377	2,7238	3,5911

40	1,6839	2,0211	2,4233	2,7045	3,5510
45	1,6794	2,0141	2,4121	2,6896	3,5203
50	1,6759	2,0086	2,4033	2,6778	3,4960
55	1,6730	2,004	2,3961	2,6682	3,4764
60	1,6706	2,0003	2,3901	2,6603	3,4602
70	1,6669	1,9944	2,3808	2,6479	3,4350
80	1,6641	1,9901	2,3739	2,6387	3,4163
90	1,6620	1,9867	2,3685	2,6316	3,4019
100	1,6602	1,9840	2,3642	2,6259	3,3905
110	1,6588	1,9818	2,3607	2,6213	3,3812
120	1,6577	1,9799	2,3578	2,6174	3,3735
∞	1,6448	1,9600	2,3263	2,5758	3,2905

Приложение 4
Критические точки для χ_n^2 распределения

$n \backslash \alpha$	**0.995**	**0.990**	**0.975**	**0.950**	**0.900**	**0.750**	**0.500**	**0.250**	**0.100**	**0.050**	**0.025**	**0.010**
1	0.00004	0.00016	0.00098	0.00393	0.01579	0.10153	0.45494	1.32330	2.70554	3.84146	5.02389	6.63490
2	0.01003	0.02010	0.05064	0.10259	0.21072	0.57536	1.38629	2.77259	4.60517	5.99146	7.37776	9.21034
3	0.07172	0.11483	0.21580	0.35185	0.58437	1.21253	2.36597	4.10834	6.25139	7.81473	9.34840	11.34487
4	0.20699	0.29711	0.48442	0.71072	1.06362	1.92256	3.35669	5.38527	7.77944	9.48773	11.14329	13.27670
5	0.41174	0.55430	0.83121	1.14548	1.61031	2.67460	4.35146	6.62568	9.23636	11.07050	12.83250	15.08627
6	0.67573	0.87209	1.23734	1.63538	2.20413	3.45460	5.34812	7.84080	10.64464	12.59159	14.44938	16.81189
7	0.98926	1.23904	1.68987	2.16735	2.83311	4.25485	6.34581	9.03715	12.01704	14.06714	16.01276	18.47531
8	1.34441	1.64650	2.17973	2.73264	3.48954	5.07064	7.34412	10.21885	13.36157	15.50731	17.53455	20.09024
9	1.73493	2.08790	2.70039	3.32511	4.16816	5.89883	8.34283	11.38875	14.68366	16.91898	19.02277	21.66599
10	2.15586	2.55821	3.24697	3.94030	4.86518	6.73720	9.34182	12.54886	15.98718	18.30704	20.48318	23.20925
11	2.60322	3.05348	3.81575	4.57481	5.57778	7.58414	10.34100	13.70069	17.27501	19.67514	21.92005	24.72497
12	3.07382	3.57057	4.40379	5.22603	6.30380	8.43842	11.34032	14.84540	18.54935	21.02607	23.33666	26.21697
13	3.56503	4.10692	5.00875	5.89186	7.04150	9.29907	12.33976	15.98391	19.81193	22.36203	24.73560	27.68825
14	4.07467	4.66043	5.62873	6.57063	7.78953	10.16531	13.33927	17.11693	21.06414	23.68479	26.11895	29.14124
15	4.60092	5.22935	6.26214	7.26094	8.54676	11.03654	14.33886	18.24509	22.30713	24.99579	27.48839	30.57791
16	5.14221	5.81221	6.90766	7.96165	9.31224	11.91222	15.33850	19.36886	23.54183	26.29623	28.84535	31.99993
17	5.69722	6.40776	7.56419	8.67176	10.08519	12.79193	16.33818	20.48868	24.76904	27.58711	30.19101	33.40866
18	6.26480	7.01491	8.23075	9.39046	10.86494	13.67529	17.33790	21.60489	25.98942	28.86930	31.52638	34.80531
19	6.84397	7.63273	8.90652	10.11701	11.65091	14.56200	18.33765	22.71781	27.20357	30.14353	32.85233	36.19087
20	7.43384	8.26040	9.59078	10.85081	12.44261	15.45177	19.33743	23.82769	28.41198	31.41043	34.16961	37.56623
21	8.03365	8.89720	10.28290	11.59131	13.23960	16.34438	20.33723	24.93478	29.61509	32.67057	35.47888	38.93217
22	8.64272	9.54249	10.98232	12.33801	14.04149	17.23962	21.33704	26.03927	30.81328	33.92444	36.78071	40.28936
23	9.26042	10.19572	11.68855	13.09051	14.84796	18.13730	22.33688	27.14134	32.00690	35.17246	38.07563	41.63840
24	9.88623	10.85636	12.40115	13.84843	15.65868	19.03725	23.33673	28.24115	33.19624	36.41503	39.36408	42.97982
25	10.51965	11.52398	13.11972	14.61141	16.47341	19.93934	24.33659	29.33885	34.38159	37.65248	40.64647	44.31410
26	11.16024	12.19815	13.84390	15.37916	17.29188	20.84343	25.33646	30.43457	35.56317	38.88514	41.92317	45.64168
27	11.80759	12.87850	14.57338	16.15140	18.11390	21.74940	26.33634	31.52841	36.74122	40.11327	43.19451	46.96294
28	12.46134	13.56471	15.30786	16.92788	18.93924	22.65716	27.33623	32.62049	37.91592	41.33714	44.46079	48.27824
29	13.12115	14.25645	16.04707	17.70837	19.76774	23.56659	28.33613	33.71091	39.08747	42.55697	45.72229	49.58788
30	13.78672	14.95346	16.79077	18.49266	20.59923	24.47761	29.33603	34.79974	40.25602	43.77297	46.97924	50.89218

Приложение 5

Критические точки для распределения Фишера (F-распределение)

Уровень значимости α = 0,10

k_2	k_1										
	2	3	4	5	6	7	8	9	10	11	12
2	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,40	9,41
3	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,22
4	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,91	3,90
5	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,28	3,27
6	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94	2,92	2,90
7	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,68	2,67
8	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,52	2,50
9	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,40	2,38
10	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,30	2,28
11	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,23	2,21
12	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,17	2,15
13	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14	2,12	2,10
14	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10	2,07	2,05

Уровень значимости $\alpha = 0{,}05$

k_2	k_1										
	2	3	4	5	6	7	8	9	10	11	12
2	19,0	19,2	19,3	19,3	19,3	19,3	19,4	19,4	19,4	19,4	19,4
3	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74
4	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91
5	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68
6	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57
8	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28
9	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07
10	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91
11	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79
12	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69
13	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60
14	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53

Уровень значимости α = 0,01

k_2	k_1										
	2	3	4	5	6	7	8	9	10	11	12
2	99,0	99,2	99,3	99,3	99,3	99,4	99,4	99,4	99,4	99,4	99,4
3	30,8	29,5	28,7	28,2	27,9	27,7	27,5	27,3	27,2	27,1	27,1
4	18,0	16,7	16,0	15,5	15,2	15,0	14,8	14,7	14,6	14,5	14,4
5	13,3	12,1	11,4	11,0	10,7	10,5	10,3	10,2	10,1	10,0	9,9
6	10,9	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72
7	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,54	6,47
8	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,73	5,67
9	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,18	5,11
10	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,77	4,71
11	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,46	4,40
12	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,22	4,16
13	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96
14	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80

Рекомендуемая литература

1. Ширяев А.Н. Вероятность. - М.: Наука, 1980. - 576 с.

2. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 2005. – 479 с. : ил.

3. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. – М.: Высшая школа, 2005. – 404 с. : ил.

4. Вентцель Е.С., Овчаров Л.А. Теория вероятностей и ее инженерные приложения. – М.: Высшая школа, 2000. – 480 с.

5. Вентцель Е.С., Овчаров Л.А. Теория случайных процессов и ее инженерные приложения. – М.: Высшая школа, 2000. – 383 с.

6. О.М.Полещук Основы теории вероятностей и математической статистики: учеб. пособие.-М: ГОУ ВПО МГУЛ, 2007.-140 с. : ил.

7. О.М.Полещук Элементы теории вероятностей и случайных процессов: учеб. пособие.-М:ГОУ ВПО МГУЛ, 2006.-68 с. : ил.

8. О.М.Полещук Основы теории вероятностей, математической статистики и случайных процессов: учеб. пособие.- М:ГОУ ВПО МГУЛ, 2012.-256 с. : ил.

9. О.М.Полещук, Е.Г.Комаров Математическая статистика: практикум-М:ГОУ ВПО МГУЛ, 2013.-90 с. : ил.