Министерство образования и науки Российской Федерации

Федеральное государственное автономное образовательное учреждение высшего профессионального образования «Национальный исследовательский технологический университет «МИСиС» Новотроицкий филиал

Д. Д. Изаак

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебно-методическое пособие

Новотроицк, 2015

УДК 519.2
ББК 22.17
И 32

Научный редактор

Бонди И. Л., *кандидат физико-математических наук*

Рецензенты:

Швалева А.В., *доцент, заведующая кафедрой математики и естествознания филиала ФГАОУ ВПО «Национальный исследовательский технологический университет «МИСиС»*

Изаак А.Д., *первый заместитель директора Федерального государственного бюджетного учреждения науки Математического института им. В.А. Стеклова Российской академии наук, кандидат физико-математических наук*

Изаак, Д.Д. Теория вероятностей и математическая статистика: Учебно-методическое пособие / Д. Д. Изаак. – Новотроицк: НФ НИТУ «МИСиС», 2015. – 152 с.

ISBN 978-5-9905230-2-9

Данное учебно-методическое пособие написано в соответствии с программой по теории вероятностей и математической статистике и будет полезно как студентам технических направлений 13.03.01, 13.03.02, 15.03.02, 18.03.01, 22.03.02, так и студентам направлений 38.03.01, 38.03.02 очной формы обучения. В пособии приведены теоретические сведения, решения типовых задач, задания к лабораторным работам, условия задач для самостоятельной работы с ответами, а также условия расчетно-графических работ.

Рекомендовано Методическим советом НФ НИТУ «МИСиС».

ISBN 978-5-9905230-2-9

Оглавление

Введение

Настоящее учебно-методическое пособие написано автором на основе пособий Л. З. Румшинского и В. А. Карасева «Организация эксперимента» с учетом многолетнего опыта работы со студентами технических направлений Новотроицкого филиала Национального исследовательского технологического университета «МИСиС».

В нем содержатся основные теоретические сведения, решения типовых примеров и задач, поясняющих теоретический материал и способствующих более глубокому его пониманию, задачи для самостоятельной работы с ответами и указаниями, условия расчетно-графических и лабораторных работ, таблицы.

В пособии приведено девять глав. В первых шести главах достаточно развернуто излагаются теоретические сведения: формулируются определения, демонстрируется вывод необходимых формул, предлагаются доказательства рассматриваемых теорем. Глава 7 содержит задачи, предлагаемые студентам для работы на практических занятиях (как с разбором у доски, так и для самостоятельного решения). Глава 8 содержит условия расчетно-графических работ, глава 9 – лабораторных.

Математическая статистика – это раздел математики, посвященный методам сбора, анализа и обработки статистических данных для научных и практических целей. Статистические данные представляют собой данные, полученные в результате обследования большого числа объектов и явлений. Обработка эмпирических данных, их систематизация, наглядное представление в форме графиков и таблиц, количественное описание посредством основных статистических показателей, формулировка выводов, имеющих прикладное значение для конкретной области человеческой деятельности, – все это относится к методам математической статистики.

Однако, обрабатывая экспериментальные данные, приходится проводить очень трудоемкие вычисления. С появлением компьютеров такие вычисления стало проводить намного проще. В связи с этим будущим инженерам необходимо уметь проводить статистические расчеты не только аналитически, но и с помощью различных компьютерных программ. Поэтому в пособии имеются лабораторные работы, посвященные изучению обработки экспериментальных данных с помощью программ MathCad, Excel, StatGraph и Stadia.

В результате освоения дисциплины обучающийся должен обладать такими компетенциями, как способность и готовность использовать основные законы математической статистики и обработки экспериментальных данных в профессиональной деятельности, способность к саморазвитию и многими другими. Все это было учтено при написании пособия.

Данное учебно-методическое пособие послужит хорошим помощником для студентов при изучении раздела математики «Теория вероятностей и математическая статистика» как под руководством преподавателя, так и самостоятельно.

Договоримся при нахождении квантилей пользоваться таблицами, приведенными в данном практикуме, для квантилей стандартного нормального распределения находить ближайшее значение, для остальных – проводить линейную интерполяцию. В случае точного совпадения экспериментальной точки с точкой дробления в задачах на критерий Пирсона– относить эту точку к левому от нее промежутку. При записи приближенного ответа, за редким исключением, указывать три знака после запятой.

ГЛАВА 1. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ, СЛУЧАЙНЫЕ СОБЫТИЯ, ИХ ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ

1.1 Элементы комбинаторики

Правило суммы. Если имеется m способов выбрать элемент a и независимо от них n способов выбрать элемент b, то тогда элемент или a или b можно выбрать $m+n$ способов.

Пример 1.1.1. В Крыму имеется m курортов, на Кавказском побережье – n. Сколько существует вариантов в выборе курорта, если вы хотите провести отпуск либо в Крыму, либо на Кавказе? Ответ: $m+n$.

Правило произведения. Если имеется m способов выбрать элемент a и независимо от них n способов выбрать элемент b, то пару (a,b) можно выбрать mn способами.

Пример 1.1.2. Сколько существует вариантов в выборе курорта в предыдущем примере, если первую половину отпуска вы хотите провести в Крыму, вторую – на Кавказе? Ответ: mn.

Пример 1.1.3. Сколько существует вариантов, если первую половину отпуска вы хотите провести на одном побережье, вторую – на другом? Ответ: $2mn$.

Сочетания без повторений. Из n элементов некоторого множества требуется выбрать k элементов, причем порядок выбора элементов не играет роли. Это можно сделать C_n^k способами (число сочетаний из n по k).

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

Пример 1.1.4. Первому игроку в «Очко» из колоды в 36 карт сдается две карты. Сколькими возможностями можно это сделать? Ответ: C_{36}^2.

Размещения без повторений.

Из n элементов некоторого множества требуется выбрать k элементов, причем порядок выбора играет роль. Это можно сделать A_n^k способами (число размещений из n по k).

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

Пример 1.1.5. В первый день конкурса исполнителей песни выступают k артистов, общее число конкурсантов n. Первые k артистов, а также порядок их выступления определяются жеребьевкой. Сколько для этого имеется возможностей? Ответ: A_n^k.

Перестановки без повторений. Упорядочить n элементов можно следующим числом способов: $P_n = A_n^n = n!$

Пример 1.1.6. Сколькими способами 20 партий можно расположить в избирательном бюллетене? Ответ: 20!

Размещения с повторениями. Из n различных типов элементов требуется выбрать k элементов, причем порядок играет роль. Это можно сделать n^k способами.

Пример 1.1.7. Сколько имеется вариантов шестизначного пароля из латинских букв? Ответ: 52^6.

Перестановки с повторениями. Из n типов элементов можно составить n^k слов длины k. Некоторые из них отличаются составом («кот», «ром»), а другие только порядком элементов («кот», «ток»). Соберем все слова, имеющие один и тот же состав. Пусть в каждое такое слово k_1 раз входит первый элемент, и так далее, k_n раз входит n-й элемент. Пусть $k = k_1 + ... + k_n$. Тогда число слов в этом подмножестве $P(k_1,...,k_n) = \frac{k!}{k_1! \cdot ... \cdot k_n!}$.

Пример 1.1.8. Сколько различных слов можно составить из букв «м», «т», «т», «о», «о», «о»? Ответ: $\frac{6!}{1! \cdot 2! \cdot 3!}$.

Рассмотрим разные примеры.

Пример 1.1.9. Сколько существует правильных шестисимвольных идентификаторов в Паскале?

Идентификатор в Паскале может состоять из 26-и латинских букв, 10-и цифр и символа подчеркивания, но начинаться должен с буквы. Ответ: $26 \cdot 37^5$.

Пример 1.1.10. В классической игре в «Lines» имеется поле из 9x9 клеток и шарики 8-и цветов. Сколько есть вариантов для появления на поле трех первых шариков?

1. Выберем 3 шарика в определенном порядке, это можно сделать 8^3 способами.
2. Размещаем отдельно выбранные шарики на поле. Это можно сделать C_{81}^3 способами.
3. По правилу произведения общее число вариантов равно $8^3 \cdot C_{81}^3$.

1.2 Эмпирическая основа теории вероятностей

При изучении различных явлений часто производят эксперименты, их результаты фиксируют в виде некоторых значений. Но при повторении испытаний результаты могут отличаться друг от друга, даже если эксперименты проводились в одинаковых условиях. В этом случае говорят, что результат измерения есть величина случайная. В результате обработки результатов возникает желание уменьшить влияние случайности на принятие решения.

Оказывается, что наблюдается важная закономерность: статистическая устойчивость средних арифметических. Если результаты некоторой серии испытаний обозначить через $x_1, \ldots, x_n$, то разброс средних значений

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

будет значительно меньше разброса результатов в отдельных сериях и будет уменьшатся с увеличением объемов серий. Это установлено эмпирически (опытным путем).

Пример 1.2.1. Пусть в результате некоторого физического опыта производилась запись некоторой величины, допустим, напряжения U. Опыты повторялись несколько раз и проводились в одинаковых условиях. Показания вольтметра заносились в таблицу:

№	1	2	3	4	5
U, V	2	3	2,2	2,7	2,1

Различия в показаниях вольтметра объясняются влиянием на результат испытания некоторых случайных помех. Таким образом, разброс отдельных значений U в пределах $[2{,}3]$.

$$\overline{x_1} = \frac{2+3+2{,}2+2{,}7+2{,}1}{5} = 2{,}4.$$

При повторении опыта еще 10 раз были получены следующие значения средних арифметических: $\overline{x_2} = 2{,}35$; $\overline{x_3} = 2{,}5$. Таким образом, разброс средних арифметических получился в пределах $[2{,}35;2{,}5]$. Причем, этот разброс был бы еще меньше, если бы увеличить количество испытаний в отдельных сериях (сделать их много больше пяти).

Статистическая устойчивость средних значений наблюдаемых величин служит эмпирической основой построения теории вероятностей – науки о математических моделях, включающих понятие случая.

1.3 Случайные величины, их математические ожидания

Рассмотрим множество Ω всех возможных взаимоисключающих друг друга исходов некоторого испытания. Это множество будем называть пространством элементарных исходов, а сами эти исходы будем толковать как точки $\omega \in \Omega$. Случайной величиной X называется наблюдаемая величина, которая принимает вполне определенные значения $X(\omega)$ при каждом исходе $\omega \in \Omega$. Таким образом, случайная величина X есть функция, определенная на пространстве элементарных исходов. Договоримся в дальнейшем обозначать случайные величины большими буквами, а отдельные значения случайных величин – маленькими.

Пример 1.3.1. Рассмотрим колоду в 36 карт. Определим достоинства карт согласно следующей таблице.

Карта	6	7	8	9	10	В	Д	К	Т
Достоинство	6	7	8	9	10	2	3	4	11

Под испытанием будем понимать вытаскивание наугад карты. Тогда пространство элементарных исходов Ω будет состоять из точек

$$\{\omega_1 = 6ч, \omega_2 = 6б, \ldots, \omega_{36} = Тп\}.$$

Достоинство вынутой наугад карты можно толковать как случайную величину:

$$X(\omega_1) = 6,\ X(\omega_2) = 6,\ \ldots,\ X(\omega_{36}) = 11.$$

Заметим, что на одном и том же пространстве элементарных исходов можно определить целую систему случайных величин. (В предыдущем примере по-разному задать достоинства карт.)

Будем теперь повторять испытания n раз и фиксировать наблюдаемые при этом значения случайной величины X. Ограничимся ситуацией, когда случайная величина может принимать лишь конечное число различных значений $x_1,\ \ldots,\ x_k$. (В примере $k = 9$.) Если при повторении n испытаний значение x_i встретилось n_i раз, то эмпирическое среднее значение величины X за n испытаний равно

$$\bar{x} = \frac{x_1 n_1 + \ldots + x_k n_k}{n} = \sum_{i=1}^{k} \frac{x_i n_i}{n}$$

или

$$\bar{x} = \frac{X(\omega_1) + \ldots + X(\omega_n)}{n} = \sum_{i=1}^{n} \frac{X(\omega_i)}{n}.$$

Если в примере с картами в результате трех испытаний были вытянуты 6 червей, 6 треф и туз треф, то

$$\bar{x} = \frac{6 \cdot 2 + 7 \cdot 0 + 8 \cdot 0 + 9 \cdot 0 + 10 \cdot 0 + 2 \cdot 0 + 3 \cdot 0 + 4 \cdot 0 + 11 \cdot 1}{3} = 7\frac{2}{3}$$

или

$$\bar{x} = \frac{6 + 6 + 11}{3} = 7\frac{2}{3}.$$

При достаточно большом числе испытаний n это среднее значение будет близко к некоторому теоретическому среднему – числу, зависящему только от случайной величины X и не зависящему от произведенной серии испытаний. Такое число назовем математическим ожиданием случайной величины X и обозначим через $M(X)$. Математическое ожидание $M(X)$ имеет ту же размерность, что и величина X. (В примере – это «количество очков».) Другими словами

$$M(X) = \lim_{n\to\infty} \overline{x}. \qquad (1.1)$$

или

$$M(X) = \lim_{n\to\infty} \sum_{i=1}^{n} \frac{X(\omega_i)}{n}. \qquad (1.1^{/})$$

Свойства математического ожидания

1. Если случайная величина X не принимает отрицательных значений, то и $M(X) \ge 0$.

   Нам дано, что для любого $\omega \in \Omega$ $X(\omega) \ge 0$, а значит, по формуле $(1.1^{/})$ и $M(X) \ge 0$.

2. Если на одном и том же пространстве Ω определена система случайных величин X_1, X_2, …, X_N, то при любых постоянных C_1, C_2, …, C_N математическое ожидание линейной комбинации случайных величин равно линейной комбинации их математических ожиданий:

$$M\left(\sum_{i=1}^{N} C_i X_i\right) = \sum_{i=1}^{N} C_i M(X_i).$$

   Это свойство называется свойством линейности математического ожидания. Докажем его:

$$M\left(\sum_{i=1}^{N} C_i X_i\right) = M(C_1 X_1 + \ldots + C_N X_N) = [(1.1^{/})] =$$

$$= \lim_{n\to\infty} \frac{C_1 X_1(\omega_1) + \ldots + C_N X_N(\omega_1) + \ldots + C_1 X_1(\omega_n) + \ldots + C_N X_N(\omega_n)}{n} =$$

$$= \lim_{n\to\infty}\left(\frac{C_1X_1(\omega_1)+\ldots+C_1X_1(\omega_n)}{n}+\ldots+\frac{C_NX_N(\omega_1)+\ldots+C_NX_N(\omega_n)}{n}\right)=$$

$$= C_1M(X_1)+\ldots+C_NM(X_N)=\sum_{i=1}^{N}C_iM(X_i)$$

3. Математическое ожидание постоянной величины равно той же постоянной:

$$M(C)=C.$$

Действительно, $M(C)=\dfrac{C+C+\ldots+C}{n}=\dfrac{Cn}{n}=C$.

1.4. Случайные события, их вероятности

Случайным событием называется любое подмножество A множества всевозможных исходов испытания. То есть, событие A есть подмножество множества Ω.

В примере, рассмотренном выше, вытаскивание карты достоинством 6 очков есть событие $A=\{\omega_1,\omega_2,\omega_3,\omega_4\}$, вытаскивание пиковой карты: $B=\{\omega_4,\omega_8,\ldots,\omega_{36}\}$.

Если при испытании осуществился исход ω, входящий в A, то говорят, что событие A осуществилось. В частности, если $A=\Omega$, событие A называют достоверным, если $A=\varnothing$, – невозможным.

Введем индикатор события A, равный 1 при появлении события A и 0 в противном случае.

$$I_A=I_A(\omega)=\begin{cases}1, & \omega\in A\\ 0, & \omega\notin A\end{cases}. \qquad (1.2)$$

Вероятностью события А назовем математическое ожидание его индикатора I_A:

$$p(A)=M(I_A). \qquad (1.3)$$

Укажем простейшие свойства вероятности, вытекающие непосредственно из ее определения:

1. $\forall A \ \ 0 \le p(A) \le 1$;

2. $p(\Omega) = 1$;

3. $p(\varnothing)=0$.

Пример 1.4.1. На заводе изготавливают детали. Исследования показали, что выбрав наугад 100 деталей, 5 из них оказываются бракованными. Найти вероятность того, что выбранная наугад деталь окажется бракованной.

$$I_Б = \begin{cases} 1, & \omega \in Б \\ 0, & \omega \notin Б \end{cases},$$

$$p(Б) = M(I_Б) \approx \frac{95 \cdot 0 + 5 \cdot 1}{100} = \frac{1}{20}.$$

Пусть множество исходов конечно $\Omega = \{\omega_1, ..., \omega_N\}$ и каждому исходу ω_k соответствует вероятность $p(\omega_k) = p_k$ $(k = \overline{1, N})$. Любую случайную величину $X = X(\omega)$ можно представить в виде линейной комбинации индикаторов этих исходов:

$$X(\omega) = \sum_{k=1}^{N} X(\omega_k) I_{\omega_k}(\omega). \qquad (1.4)$$

Действительно, при $\omega = \omega_i$ правая часть будет равна

$$X(\omega_1) \cdot 0 + X(\omega_2) \cdot 0 + ... + X(\omega_i) \cdot 1 + ... + X(\omega_N) \cdot 0.$$

Перейдем в формуле (1.4) к математическим ожиданиям:

$$M(X) = M\left(\sum_{k=1}^{N} X(\omega_k) I_{\omega_k}(\omega) \right) = \sum_{k=1}^{N} X(\omega_k) M(I_{\omega_k}(\omega)) = \sum_{k=1}^{N} X(\omega_k) p(\omega_k).$$

То есть

$$M(X) = \sum_{k=1}^{N} X(\omega_k) p_k. \qquad (1.5)$$

В частном случае, при $X = C$ формула (1.5) принимает вид:

$$M(C) = \sum_{k=1}^{N} C \cdot p_k,$$

$$C = C\sum_{k=1}^{N} p_k\ ,$$

$$\sum_{i=1}^{N} p_k = 1\,.$$

Также, в частном случае, когда $X = I_A$, формула (1.5) принимает вид:

$$p(A) = M(I_A) = \sum_{\omega_k \in A} p_k\ . \qquad (1.6)$$

Классическая модель

Если в рассматриваемой модели не только множество всех исходов конечно, но и всем исходам соответствуют одинаковые вероятности

$$\forall_{k=\overline{1,N}} \quad p_k = p\ ,$$

то такая модель называется классической. Заметим, что основы теории вероятностей зародились примерно в XVI веке, это было связано с популярностью азартных игр, в которых как раз имеет место классическая модель.

Рассмотрим, как будут выглядеть полученные ранее формулы для классической модели.

$$\sum_{k=1}^{N} p_k = 1 \Rightarrow p \cdot N = 1 \Rightarrow p = \frac{1}{N}.$$

А значит вероятность любого события A равна

$$p(A) = \sum_{\omega_k \in A} p = \frac{N_A}{N}\ , \qquad (1.7)$$

где N_A – число исходов, входящих в множество A, или, как обычно говорят, благоприятствующих событию A исходов. Формулу (1.7) иногда называют классическим определением вероятности.

Формула (1.5) для классической модели примет вид

$$M(X) = \sum_{k=1}^{N} \frac{X(\omega_k)}{N}\ . \qquad (1.8)$$

Равновозможность или равноправие исходов устанавливается из соображений симметрии событий. Так, например, при подбрасывании монеты, выпадение решки или орла можно принять за равновозможные исходы, так как предполагается, что монета сделана из однородного материала, имеет правильную геометрическую форму, а наличие пробоин не влияет на результат. Это одно из наиболее фундаментальных и в то же время «скользких» понятий, так как очень часто бывает сложно утверждать, являются ли исходы равновозможными. Например, учитывая то, что число мужчин и женщин на Земле примерно одинаковое, на первый взгляд можно сделать вывод, что первый прошедший мимо твоего дома человек окажется мужчиной или женщиной – события равновозможные. Однако, если рядом с вашим домом находится военное училище, это не так. Кроме того, исход события может зависеть от времени суток и еще от многих факторов, так что считать эти события равновозможными или нет – это на самом деле очень сложный вопрос.

Возвратимся к примеру с картами. Здесь имеет место классическая модель, а, значит, если случайная величина X – это достоинство вытащенной карты, то

$$M(X) = \frac{2+2+2+2+3+\ldots+11}{36} = 6\frac{2}{3}.$$

Пример 1.4.2. Найти вероятность того, что вытащенная наугад карта из колоды в 36 карт окажется красным тузом.

$$p(A) = \frac{2}{36} = \frac{1}{18}.$$

Пример 1.4.3. N гостей приглашены к круглому столу; их места указаны разложенными наугад визитными карточками. Какова вероятность того, что гости A и B окажутся рядом?

Заменим задачу равносильной. Представим, что гость A уже занял некоторое место, и будем искать вероятность того, что гость B окажется рядом: $p = \dfrac{2}{N-1}$.

Правило сложения вероятностей

Назовем события A и B несовместными, если они не могут произойти при одном и том же испытании, то есть если $A \bigcap B = \varnothing$. Суммой таких событий назовем событие заключающееся в том, что произойдет одно из них, сумме событий $A + B$ соответствует множество $A \bigcup B$.

Индикатор суммы несовместных событий A и B равен сумме их индикаторов:

$$I_{A+B} = I_A + I_B .$$

Это следует из следующей таблицы.

A	0	1	0
B	0	0	1
$A+B$	0	1	1

Переходя к математическим ожиданиям, получим:

$$p(A + B) = p(A) + p(B) . \qquad (1.9)$$

Это правило сложения вероятностей для двух несовместных событий A и B. Оно распространяется и на сумму любого числа попарно-несовместных событий:

$$p(A_1 + ... + A_N) = p(A_1) + ... + p(A_N) .$$

Говорят, что события A_1, …, A_N образуют полную группу, если они попарно-несовместны и одно из них обязательно должно произойти при любом испытании, то есть $A_1 + ... + A_N = \Omega$, и, следовательно, $p(A_1 + ... + A_N) = 1$.

Два события называются противоположенными, если они образуют полную группу. Событие, противоположенное событию A, обозначается $\overline{A}$.

$$p(A) + p(\overline{A}) = 1 .$$

1.5 Числовые характеристики случайных величин и их систем

Центр распределения, дисперсия, стандартное отклонение

Пусть $M(X)=\alpha$. Если все значения величины *X* ограничены числами *a* и *b*, то и $\alpha \in [a,b]$. Действительно, если $a \le X \le b$, то $M(a) \le M(X) \le M(b)$, то есть $a \le M(X) \le b$. Величину $\alpha = M(X)$ называют также центром распределения величины *X*.

Во многих задачах статистики представляет интерес рассеяние величины *X* относительно этого центра. В качестве числовой характеристики рассеяния применяют средний квадрат отклонения случайной величины от центра распределения, который называют дисперсией и обозначают $D(X)$. Таким образом, по определению, $D(X)=M[(X-\alpha)^2]$.

Заметим, что нельзя принять за характеристику рассеяния $M(X-\alpha)$ – среднее значение самого отклонения, так как $M(X-\alpha)=M(X)-\alpha=0$. Так как дисперсия имеет квадратичную размерность, то в окончательных выводах применяют стандартное (среднее квадратическое) отклонение $\sigma(X)=\sqrt{D(X)}$, которое имеет ту же размерность, что и величина *X*. Иногда вместо $D(X)$ пишут $\sigma^2(X)$.

Пример 1.5.1. Вновь вернемся к примеру с картами. Как было показано ранее, $M(X)=6\frac{2}{3}$, случайная величина *X* может принимать значения 2, 3, 4, 6, 7, 8, 9, 10, 11.

$$D(X)=\frac{\left(-4\frac{2}{3}\right)^2+\left(-3\frac{2}{3}\right)^2+\left(-2\frac{2}{3}\right)^2+\left(-\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2+\left(1\frac{1}{3}\right)^2+\left(2\frac{1}{3}\right)^2+\left(3\frac{1}{3}\right)^2+\left(4\frac{1}{3}\right)^2}{9}=8,(8)$$

$$\sigma(X)=\sqrt{8,(8)}\approx 3.$$

Смысл дисперсии лучше осознается в сравнении. Представим, что два стрелка ведут стрельбу по мишени. После проведения серии выстрелов для результатов каждого из них была подсчитана дисперсия. Тогда можно сделать вывод, что более метким является тот из них, у которого она меньше.

Получим еще одну формулу для вычисления дисперсии.

$$D(X) = M((X-\alpha)^2) = M(X^2 - 2X\alpha + \alpha^2) =$$

$$= M(X^2) - 2\alpha M(X) + \alpha^2 = M(X^2) - M^2(X) \qquad (1.10)$$

На практике для вычисления дисперсии в основном используется формула (1.10), а не определение. Рассмотрим дисперсию от индикатора.

$$D(I_A) = M(I_A^2) - p^2 = M(I_A) - p^2 = p - p^2 = p(1-p) = p(A)p(\overline{A})$$

Укажем основные свойства дисперсии.

1. $D(C) = 0$;
2. $D(X+C) = D(X)$;
3. $D(CX) = C^2 D(X)$.

Каждое из них легко доказывается с помощью формулы (1.10) и свойств математического ожидания. Докажем, например, второе свойство.

$$D(X+C) = M((X+C)^2) - M^2(X+C) =$$

$$= M(X^2 + 2CX + C^2) - (M(X) + M(C))^2 =$$

$$= M(X^2) + 2CM(X) + C^2 - M^2(X) - 2CM(X) - C^2 =$$

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

Учитывая свойства дисперсии, можно получить аналогичные свойства стандартного отклонения.

1. $\sigma(C) = \sqrt{D(C)} = 0$.
2. $\sigma(X+C) = \sqrt{D(X+C)} = \sqrt{D(X)} = \sigma(X)$.
3. $\sigma(CX) = \sqrt{D(CX)} = \sqrt{C^2 D(X)} = |C|\sqrt{D(X)} = |C|\sigma(X)$.

Роль дисперсии в качестве характеристики рассеяния ясно проявляется в неравенстве Чебышева:

$$p(|X-\alpha| > \varepsilon) \le \frac{D(X)}{\varepsilon^2}, \qquad (1.11)$$

где $\varepsilon > 0$. Докажем его. Пусть $A = \{\omega \mid |X(\omega) - \alpha| > \varepsilon\}$.

$$I_A(\omega)=\begin{cases}1, & |X(\omega)-\alpha|>\varepsilon\\ 0, & |X(\omega)-\alpha|\le\varepsilon\end{cases}$$

Докажем сначала, что

$$I_A(\omega)\le\frac{(X(\omega)-\alpha)^2}{\varepsilon^2}.$$

Для этого рассмотрим два случая.

I случай. $\omega\in A$.

Левая часть равна 1.

$|X(\omega)-\alpha|>\varepsilon$, $\frac{|X(\omega)-\alpha|}{\varepsilon}>1$, правая часть $\frac{(X(\omega)-\alpha)^2}{\varepsilon^2}>1$.

II случай. $\omega\notin A$.

Левая часть равна 0.

Правая часть $\frac{(X(\omega)-\alpha)^2}{\varepsilon^2}\ge 0$.

Перейдем к математическим ожиданиям.

$$M\left(I_A(\omega)\right)\le M\left(\frac{(X(\omega)-\alpha)^2}{\varepsilon^2}\right)$$

$$p(A)\le\frac{1}{\varepsilon^2}D(X)$$

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

Пример 1.5.2. И вновь вернемся к примеру с картами. Проверим, выполняется ли для него неравенство Чебышева при $\varepsilon=3{,}5$.

$$p(|X-\alpha|>\varepsilon)=p\left(\left|X-6\frac{2}{3}\right|>3{,}5\right)$$

Найдем вероятность этого события по классическому определению вероятности. Рассматриваемое событие произойдет, если будет вынут валет, дама или туз. То есть $p=\frac{12}{36}=\frac{1}{3}$. $\frac{D(X)}{\varepsilon^2}=\frac{8,(8)}{12{,}25}\approx 0{,}73$. Неравенство Чебышева выполняется: $\frac{1}{3}<0{,}73$.

Центр распределения и матрица ковариаций

Систему случайных величин X_1, …, X_m можно толковать как точку или вектор m-мерного пространства $X=(X_1,...,X_m)$. При этом вектор $M(X)=(\alpha_1,...,\alpha_m)$, где $\alpha_i = M(X_i)$ $(i=\overline{1,m})$, определяет центр распределения системы. Для характеристики рассеяния величин вводится матрица ковариаций

$$K=\begin{pmatrix} K_{11} & K_{12} & ... & K_{1m} \\ K_{21} & K_{22} & ... & K_{2m} \\ ... & ... & ... & ... \\ K_{m1} & K_{m2} & ... & K_{mm} \end{pmatrix},$$

где ковариация

$$K_{ij}=\operatorname{cov}(X_i,X_j)=M[(X_i-\alpha_i)(X_j-\alpha_j)]. \qquad (1.12)$$

Заметим, что ковариация

$$K_{ii}=D(X_i). \qquad (1.13)$$

Матрица ковариаций симметрична: $K_{ij}=K_{ji}$. Формулу (1.12) можно преобразовать подобно формуле (1.10):

$$M[(X_i-\alpha_i)(X_j-\alpha_j)]=M(X_iX_j)-\alpha_iM(X_j)-\alpha_jM(X_i)+\alpha_i\alpha_j=$$

$$=M(X_iX_j)-\alpha_i\alpha_j.$$

Пример 1.5.3. Подбрасываются две монеты. События A заключается в выпадении ровно одного герба, событие B – в выпадении хотя бы одного герба. Найти матрицу ковариаций для индикаторов событий A и B.

Решение.

$$D(I_A)=p(A)p(\overline{A})=\frac{1}{2}\cdot\frac{1}{2}=\frac{1}{4};$$

$$D(I_B)=p(B)p(\overline{B})=\frac{3}{4}\cdot\frac{1}{4}=\frac{3}{16};$$

$$K_{12}=M(I_AI_B)-M(I_A)M(I_B)$$

ω	I_A	I_B	$I_A I_B$
ГГ	0	1	0
ГР	1	1	1
РГ	1	1	1
РР	0	0	0

$$K_{12} = M(I_A) - M(I_A)M(I_B) = p(A)p(\overline{B}) = \frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}.$$

Итак

$$K = \begin{pmatrix} \frac{1}{4} & \frac{1}{8} \\ \frac{1}{8} & \frac{3}{16} \end{pmatrix}.$$

Если ковариация $K_{12} = 0$, то величины X_1 и X_2 называются некоррелированными. Если все величины X_1, …, X_m попарно некоррелированные, то

$$K = \begin{pmatrix} D(X_1) & 0 & ... & 0 \\ 0 & D(X_2) & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & D(X_m) \end{pmatrix}.$$

Дисперсия линейной комбинации случайных величин

Найдем дисперсию для линейной комбинации двух случайных величин.

$$D(C_1 X_1 + C_2 X_2) = M[(C_1 X_1 + C_2 X_2 - M(C_1 X_1 + C_2 X_2))^2] =$$

$$= M[(C_1 X_1 + C_2 X_2 - C_1 \alpha_1 - C_2 \alpha_2)^2] =$$

$$= M[(C_1(X_1 - \alpha_1) + C_2(X_2 - \alpha_2))^2] =$$

$$= C_1^2 D(X_1) + 2C_1 C_2 K_{12} + C_2^2 D(X_2)$$

Таким образом

$$D(C_1 X_1 + C_2 X_2) = C_1^2 K_{11} + 2C_1 C_2 K_{12} + C_2^2 K_{22}. \qquad (1.14)$$

Так как дисперсия не может быть отрицательна, то при любых C_1, C_2

$$C_1^2 K_{11} + 2C_1 C_2 K_{12} + C_2^2 K_{22} \geq 0.$$

Разделим обе части неравенства на $C_2^2 > 0$ и обозначим $\dfrac{C_1}{C_2} = C$:

$$K_{11}C^2 + 2K_{12}C + K_{22} \geq 0\,.$$

Квадратный трехчлен неотрицателен, если $K_{11} > 0$ и дискриминант $D \leq 0$. Следовательно, рассматриваемое условие равносильно

$$\begin{vmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{vmatrix} = K_{11}K_{22} - K_{12}^2 \geq 0\,. \qquad (1.15)$$

Если величины X_1 и X_2 некоррелированные, то есть $K_{12} = 0$, то формула (1.14) упрощается:

$$D(C_1X_1 + C_2X_2) = C_1^2 D(X_1) + C_2^2 D(X_2) \qquad (1.16)$$

В частности, если $C_1 = C_2 = 1$, получаем формулу сложения дисперсий:

$$D(X_1 + X_2) = D(X_1) + D(X_2)\,. \qquad (1.17)$$

В общем случае формула (1.14) выглядит так:

$$D\left(\sum_{i=1}^{m} C_i X_i\right) = \sum_{i=1}^{m}\sum_{j=1}^{m} C_i C_j K_{ij} \geq 0 \qquad (1.18)$$

Если все величины попарно некоррелированные, то формула (1.18) упрощается:

$$D\left(\sum_{i=1}^{m} C_i X_i\right) = \sum_{i=1}^{m} C_i^2 D(X_i)\,. \qquad (1.19)$$

При $C_1 = \ldots = C_m = 1$ получаем общую формулу сложения дисперсий:

$$D\left(\sum_{i=1}^{m} X_i\right) = \sum_{i=1}^{m} D(X_i)\,. \qquad (1.20)$$

Отметим еще один частный случай формулы (1.19). При $C_1 = \ldots = C_m = \dfrac{1}{m}$ линейная комбинация величин X_i равна их среднему арифметическому:

$$\overline{X} = \frac{X_1 + \ldots + X_m}{m}$$

и формула (1.19) дает дисперсию среднего арифметического попарно некоррелированных величин:

$$D(\overline{X})=\sum_{i=1}^{m}\frac{D(X_i)}{m^2}. \qquad (1.21)$$

В частности, если все случайные величины X_i имеют одинаковые дисперсии $D(X_i)=\sigma^2$, то

$$D(\overline{X})=\frac{\sigma^2}{m} \qquad (1.22)$$

и стандартное отклонение среднего арифметического в $\sqrt{m}$ раз меньше стандартного отклонения каждой величины:

$$\sigma(\overline{X})=\frac{\sigma}{\sqrt{m}}. \qquad (1.23)$$

Из последней формулы видно, что $\sigma(\overline{X})\to 0$ при $m\to\infty$, то есть рассеяние среднего арифметического затухает при $m\to\infty$. Это утверждение не противоречит эмпирически установленной устойчивости среднего арифметического.

1.6 Биномиальное распределение

Пусть производится *n* независимых испытаний. В результате каждого испытания может наступить только два исхода, которые будем называть «успех» и «неудача». Обозначим успех через *A*, неудачу – $\overline{A}$. Пусть вероятность события *A* во всех исходах одинаковая: $p(A)=p=const$, $p(\overline{A})=1-p=q=const$. Обозначим через *X* количество успехов в серии из *n* испытаний, $X=\overline{0,n}$. Заметим, что случайную величину *X* можно представить в следующем виде: $X=\sum_{i=1}^{n}I_{A_i}$, где A_i – успех в *i*-м испытании. Тогда

$$M(X)=M\left(\sum_{i=1}^{n}I_{A_i}\right)=\sum_{i=1}^{n}M(I_{A_i})=\sum_{i=1}^{n}p=np,$$

$$D(X)=D\left(\sum_{i=1}^{n}I_{A_i}\right)=\sum_{i=1}^{n}D(I_{A_i})=\sum_{i=1}^{n}p(A_i)\,p(\overline{A_i})=\sum_{i=1}^{n}pq=npq.$$

Итак,

$$M(X)=np,\ D(X)=npq,\ \sigma(X)=\sqrt{npq}.$$

Используя правила сложения и умножения вероятностей, можно доказать, что

$$p_n(X = m) = C_n^m p^m q^{n-m}.$$

Случайная величина *X*, имеющая такое распределение, называется распределенной по биномиальному закону или по закону Бернулли.

Если *n* велико, то пользоваться формулой Бернулли неудобно. Тогда применяются теоремы Муавра-Лапласа или формула Пуассона.

Локальная теорема Муавра-Лапласа. Если случайная величина *X* распределена по закону Бернулли, *n* – велико, тогда

$$p_n(X = m) \approx \frac{\varphi_0(X)}{\sqrt{npq}},$$

где $X = \dfrac{m - np}{\sqrt{npq}}$, $\varphi_0(X) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ – функция Гаусса (рис 1.1).

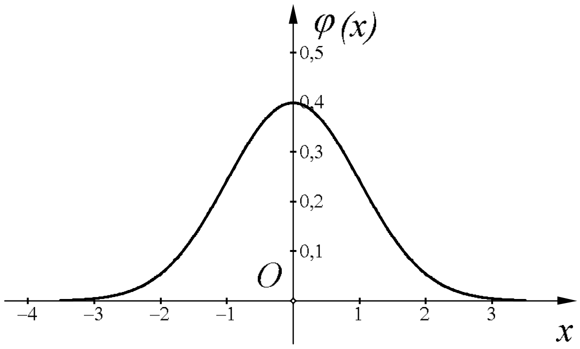


Рис. 1.1.

Интегральная теорема Муавра-Лапласа. Если случайная величина *X* распределена по закону Бернулли, *n* – велико, тогда

$$p_n(m_1 \le X \le m_2) \approx \Phi(x_2) - \Phi(x_1),$$

где $x_1 = \dfrac{m_1 - np}{\sqrt{npq}}$, $x_2 = \dfrac{m_2 - np}{\sqrt{npq}}$, $\Phi(x) = \int_0^x \varphi_0(t)dt = \dfrac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$ – функция Лапласа.

Следствие 1. Если границы m_1 и m_2 симметричны относительно np, то формула упрощается:

$$p_n(|x - np| \le r) \approx 2\Phi\left(\frac{r}{\sqrt{npq}}\right). \qquad (1.24)$$

Действительно,

$$p_n(|x-np|\le r)=p_n(np-r\le x\le np+r),$$

$$m_1=np-r,\ m_2=np+r,$$

$$x_1=-\frac{r}{\sqrt{npq}},\ x_2=\frac{r}{\sqrt{npq}},$$

$$p_n(m_1\le x\le m_2)=p_n(|x-np|\le r)\approx\Phi\left(\frac{r}{\sqrt{npq}}\right)+\Phi\left(\frac{r}{\sqrt{npq}}\right).$$

Следствие 2. $p_n\left(\left|\frac{x}{n}-p\right|\le\Delta\right)\approx 2\Phi\left(\Delta\sqrt{\frac{n}{pq}}\right)$.

Действительно, умножим обе части неравенства

$$\left|\frac{x}{n}-p\right|\le\Delta$$

на n:

$$|x-np|\le\Delta\cdot n.$$

Воспользуемся формулой (1.24) при $r=\Delta\cdot n$:

$$2\Phi\left(\frac{r}{\sqrt{npq}}\right)=2\Phi\left(\frac{\Delta\cdot n}{\sqrt{npq}}\right)=2\Phi\left(\Delta\sqrt{\frac{n}{pq}}\right).$$

Формула Пуассона. Найдем асимптотику распределения Бернулли при $n\to\infty$. Требуется выполнимость условий: n – велико, p – мало, $np=a=const$.

$$p_\infty(x=m)=\lim_{n\to\infty}p_n(x=m)=\lim_{n\to\infty}C_n^m p^m q^{n-m}=$$

$$=\lim_{n\to\infty}\frac{n!}{m!(n-m)!}\left(\frac{a}{n}\right)^m\left(1-\frac{a}{n}\right)^{n-m}=$$

$$=\lim_{n\to\infty}\frac{(n-m)!(n-(m-1))\cdot\ldots\cdot n}{(n-m)!\cdot n^m}\cdot\frac{a^m}{m!}\cdot\frac{\left(1-\frac{a}{n}\right)^n}{\left(1-\frac{a}{n}\right)^m}=$$

$$=\frac{a^m}{m!}\lim_{n\to\infty}\left(1-\frac{m-1}{n}\right)\cdot\ldots\cdot\left(1-\frac{m-m}{n}\right)\cdot\left(\left(1-\frac{a}{n}\right)^{-\frac{n}{a}}\right)^{-a}=\frac{a^m}{m!}e^{-a}.$$

В этом случае $M(X)=np=a$, $D(X)=npq=[q\approx1]\approx np=a$.

1.7 Условные математические ожидания

Пусть даны пространство элементарных исходов Ω и событие $A \subset \Omega$. Рассмотрим произведение $X(\omega) \cdot I_A(\omega)$. Оно будет равно X при $\omega \in A$ и нулю в противном случае. Отношение $\dfrac{M(X \cdot I_A)}{M(I_A)}$ называется условным математическим ожиданием при условии появления события A и обозначается:

$$M(X/A) = \frac{M(X \cdot I_A)}{M(I_A)} = \frac{M(X \cdot I_A)}{p(A)}. \qquad (1.25)$$

Здесь предполагается, что $p(A) > 0$.

Заметим, что любую случайную величину X можно представить в виде:

$$X = X \cdot I_A + X \cdot I_{\overline{A}}.$$

Перейдем к математическим ожиданиям:

$$M(X) = M(X \cdot I_A) + M(X \cdot I_{\overline{A}}) = \frac{M(X \cdot I_A)p(A)}{p(A)} + \frac{M(X \cdot I_{\overline{A}})p(\overline{A})}{p(\overline{A})} =$$

$$= M(X/A)p(A) + M(X/\overline{A})p(\overline{A}). \qquad (1.26)$$

Более общую формулу получим, если рассмотрим полную группу случайных событий A_1, …, A_N:

$$M(X) = \sum_{k=1}^{N} M(X/A_k)p(A_k). \qquad (1.27)$$

События A_1, …, A_N в формуле (1.27) называют гипотезами.

1.8 Условные вероятности. Правило умножения вероятностей

Совмещением или произведением событий A и B называется пересечение множеств A и B. Совмещение обозначается как AB. Перебрав всевозможные случаи, можно убедиться, что

$$I_{AB} = I_A I_B.$$

Перейдем к математическим ожиданиям:

$$p(AB)=M(I_{AB})=M(I_A I_B).$$

Применим формулу (1.25) к I_B:

$$M(I_B/A)=\frac{M(I_B I_A)}{p(A)}=\frac{p(AB)}{p(A)}.$$

Полученное отношение называют условной вероятностью события B при условии появления события A и обозначают:

$$p(B/A)=\frac{p(AB)}{p(A)}. \qquad (1.28)$$

Последнее равенство можно записать в виде:

$$p(AB)=p(A)p(B/A). \qquad (1.29)$$

Получили правило умножения вероятностей: вероятность произведения двух событий равна произведению вероятности одного из них на условную вероятность другого. Формула (1.29) допускает обобщение, например:

$$p(ABC)=p(A)p(B/A)p(C/AB).$$

Пример 1.8.1. Колода из 52-х карт перетасовывается и сдается 3 карты. Найти вероятность того, что все эти 3 карты – тузы.

Пусть событие A заключается в том, что первая сданная карта – туз, событие B – вторая карта туз, C – третья. Следует найти $p(ABC)$.

$$p(ABC)=p(A)p(B/A)p(C/AB)=\frac{4}{52}\cdot\frac{3}{51}\cdot\frac{2}{50}.$$

Заметим, что данную задачу можно было бы решить и по классическому определению вероятностей:

$$p(ABC)=\frac{A_4^3}{A_{52}^3}=\frac{4\cdot 3\cdot 2}{52\cdot 51\cdot 50}.$$

Из формулы (1.27) можно получить формулу полной вероятности:

$$p(B)=M(I_B)=\sum_{k=1}^{N}M(I_B/A_k)p(A_k)=\sum_{k=1}^{N}p(A_k)p(B/A_k). \quad (1.30)$$

В частности:

$$p(B) = p(A)p(B/A) + p(\overline{A})p(B/\overline{A}). \qquad (1.31)$$

Имеет место так называемая формула Байеса:

$$p(A_k/B) = \frac{p(A_kB)}{p(B)} = \frac{p(A_k)p(B/A_k)}{\sum_{k=1}^{N} p(A_k)p(B/A_k)}.$$

Пример 1.8.2. Из урны I, в которой находятся 10 белых и 15 черных шаров, в урну II, в которой находятся 10 белых и 8 черных шаров, перекладывают, не глядя, 2 шара. Затем, после перемешивания, вытаскивают наугад шар из второй урны. Найти вероятность того, что он белый.

Решение.

Здесь можно высказать три гипотезы:

A_1: переложили 2 белых шара;

A_2: переложили 1 белый и 1 черный шар;

A_3: переложили 2 черных шара.

Пусть событие B заключается в том, что из второй урны был вынут белый шар. Тогда:

$$p(A_1) = \frac{10}{25} \cdot \frac{9}{24} = \frac{3}{20};$$

$$p(B/A_1) = \frac{12}{20};$$

$$p(A_2) = \frac{10}{25} \cdot \frac{15}{24} + \frac{15}{25} \cdot \frac{10}{24} = \frac{10}{20};$$

$$p(B/A_2) = \frac{11}{20};$$

$$p(A_3) = \frac{15}{25} \cdot \frac{14}{24} = \frac{7}{20};$$

$$p(B/A_3) = \frac{10}{20}.$$

Тогда по формуле полной вероятности

$$p(B) = p(A_1)p(B/A_1) + p(A_2)p(B/A_2) + p(A_3)p(B/A_3) = 0{,}54.$$

Предположим, что в условиях данной задачи был вынут шар из второй урны, и он оказался белым. Найдем вероятность того, что были переложены 2 белых шара:

$$p(A_1 / B) = \frac{p(B / A_1)\, p(A_1)}{p(B)} = \frac{1}{6}.$$

1.9 Независимость случайных величин и случайных событий

Две случайные величины X и Y называются независимыми, если для любых двух функций $f(X)$ и $g(Y)$, для которых существуют математические ожидания, имеет место соотношение:

$$M[f(X)g(Y)] = M[f(X)]M[g(Y)]. \qquad (1.32)$$

Можно показать, что смысл вышеуказанного определения состоит в том, что если величины X и Y независимы, то никакая информация о значениях одной величины не влияет на математическое ожидание любой функции от другой величины.

Связь между понятиями независимости и некоррелированности

Теорема 1.9.1. Если величины X и Y независимы, то они некоррелированные.

Доказательство.

Нам дано, что $M[f(X)g(Y)] = M[f(X)]M[g(Y)]$ и следует доказать, что $K_{XY} = 0$.

$$K_{XY} = M[(X - M(X))(Y - M(Y))] = M[X - M(X)]M(Y - M(Y)) = 0 \cdot 0 = 0.$$

Замечание. Обратная теорема не верна. Для доказательства приведем контрпример. Рассмотрим две случайные величины, заданные таблицей.

Ω	p	X	$Y = X^2$
1	1/3	-1	1
2	1/3	0	0
3	1/3	1	1

$$M(X) = (-1) \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = 0;$$

$$M(Y) = 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \frac{2}{3};$$
$$M(XY) = M(X) = 0.$$

Эти величины являются некоррелированными:

$$K_{XY} = M[(X-0)(Y-2/3)] = M[XY - 2/3X] = M(XY) - 2/3M(X) = 0,$$

но не являются независимыми:

$$M(X^2Y) = M(Y) = 2/3 \neq M(X^2)M(Y) = 4/9.$$

Независимость нескольких случайных величин

Случайные величины X_1, …, X_m называются взаимно-независимыми или независимыми в совокупности, если для любых функций $f_i(X_i)$ $i = \overline{1,m}$, для которых существуют математические ожидания, имеет место соотношение:

$$M\left[\prod_{i=1}^{m} f_i(X_i)\right] = \prod_{i=1}^{m} M[f_i(X_i)].$$

Заметим, что из взаимной независимости случайных величин следует независимость любого их подмножества, в частности, попарная независимость. Но из попарной независимости не следует взаимная независимость. Отметим также, что если случайные величины попарно независимы, или тем более независимы в совокупности, то для них справедливы формулы (1.19), (1.20).

Независимость случайных событий

Случайные события A_1, …, A_m называются независимыми в совокупности, если взаимно независимы их индикаторы. Для таких событий имеют место формулы:

$$M(I_{A_1 \cdot \ldots \cdot A_m}) = M(I_{A_1} \cdot \ldots \cdot I_{A_m}) = \prod_{i=1}^{m} M(I_{A_i}),$$
$$p(A_1 \cdot \ldots \cdot A_m) = \prod_{i=1}^{m} p(A_i).$$

Для двух независимых событий A и B $p(AB) = p(A)p(B)$, следовательно их условные вероятности равны безусловным:

$$p(B/A) = \frac{p(AB)}{p(A)} = p(B),\ p(A/B) = \frac{p(AB)}{p(B)} = p(A).$$

ГЛАВА 2. НЕПРЕРЫВНОЕ РАСПРЕДЕЛЕНИЕ ВЕРОЯТНОСТЕЙ

2.1 Дискретное распределение вероятностей

Пусть в результате испытаний случайная величина X может принимать лишь конечное число значений $x_1, \ldots, x_N$. Напомним, что в этом случае

$$M(X) = \sum_{k=1}^{N} x_k p_k , \qquad (2.1)$$

где $p_k = p(A_k)$, $A_k = \{\omega \mid X(\omega) = x_k\}$, $\sum_{k=1}^{N} p_k = 1$. Таким образом, всю информацию о случайной величине X можно записать в виде таблицы распределения вероятностей:

X	x_1	$\ldots$	x_N
p	p_1	$\ldots$	p_N

При этом говорят, что случайная величина X имеет дискретное распределение вероятностей.

Формулу (2.1) можно обобщить:

$$M(g(X)) = \sum_{i=1}^{N} g(x_k) p_k . \qquad (2.2)$$

2.2 Непрерывное распределение вероятностей

Пусть область значений случайной величины X есть произвольное множество действительных чисел.

Говорят, что случайная величина X имеет непрерывное распределение, если найдется функция $\varphi(x)$ такая, что для любой функции $g(x)$

$$M(g(X)) = \int_{-\infty}^{+\infty} g(x)\varphi(x)dx . \qquad (2.3)$$

Рассматриваются такие функции $g(x)$, для которых этот интеграл абсолютно сходится.

В частности, пусть $g(x) = I_A(x)$, где

$$I_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}.$$

Тогда

$$M(I_A(X)) = p(X \in A) = \int_A \varphi(x)dx. \qquad (2.4)$$

Если же *A* есть интервал $(-\infty, x)$, то

$$p(X \in A) = p(X < x) = \int_{-\infty}^{x} \varphi(x)dx. \qquad (2.5)$$

Обозначим эту функцию через $F(x)$:

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

Функцию $F(x)$ называют функцией распределения случайной величины *X*. Тогда в точках непрерывности $\varphi(x)$

$$\varphi(x) = F^{/}(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{\int_{x}^{x+\Delta x} \varphi(x)dx}{\Delta x},$$

по формуле (2.4)

$$\varphi(x) = \lim \frac{p(X \in (x, x+\Delta x))}{\Delta x}.$$

Таким образом, $\varphi(x)$ – плотность вероятности. (В этих рассуждениях предполагалось, что $\Delta x > 0$.)

Укажем два основных свойства функции $\varphi(x)$:

1. $\varphi(x) \geq 0$;

2. $\int_{-\infty}^{+\infty} \varphi(x)dx = 1$ (условие нормировки).

График плотности вероятности $\varphi(x)$ называется кривой распределения. По кривой распределения можно вычислить вероятность события $X \in (x_1, x_2)$ как площадь криволинейной трапеции, расположенной над отрезком $[x_1, x_2]$.

Говорят, что случайная величина X распределена равномерно в интервале (a,b), если все ее возможные значения сосредоточены в этом интервале и плотность вероятности постоянна:

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

Пример 2.2.1. Случайная величина X имеет равномерное распределение в интервале (a,b). Найти $M(X)$, $D(X)$, $\sigma(X)$.

$$M(X) = \int_{-\infty}^{+\infty} x\varphi(x)dx = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{a+b}{2};$$

$$D(X) = M(X^2) - M^2(X) = \int_a^b x^2\varphi(x)dx - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12};$$

$$\sigma(X) = \frac{b-a}{2\sqrt{3}}.$$

Заметим, что при непрерывном распределении вероятностей вероятность попадания в любую отдельную точку равна нулю, поэтому вероятности здесь относятся лишь к интервалам.

Возникает вопрос: как на практике можно построить хотя бы приблизительно кривую распределения, по которой в свою очередь можно определить вероятность попадания случайной величины в интервал. Пусть все значения случайной величины X входят в отрезок $[a,b]$. Тогда этот отрезок разобьем на k участков и при проведении n испытаний (n достаточно велико, порядка сотен, лучше тысяч) подсчитывают частоты попадания ее значений на каждый участок: $\frac{n_i}{n}$ (n_i – число попаданий в i-ый участок).

$$\frac{n_i}{n} \approx p(X \in (x_i, x_{i+1}))$$

Если эти частоты разделить на длины соответствующих участков h_i, то графическое изображение полученных данных даст гистограмму. В гистограмме площадь каждого столбика равна $\frac{n_i}{n}$, а сумма всех площадей равна 1. При достаточно больших числах n и k гистограмма будет близка к кривой распределения.

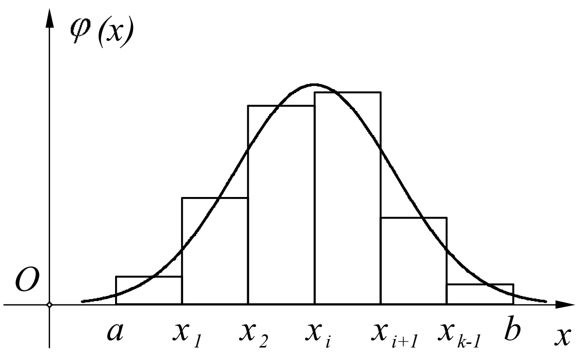


Рис. 2.1.

Используя свойства функции $\varphi(x)$, можно установить свойства функции $F(x)$.

1. Функция $F(x)$ является непрерывной;
2. Функция $F(x)$ неубывает от $F(-\infty)=0$ до $F(+\infty)=1$.
3. В точках непрерывности плотности $\varphi(x)$ $F^{/}(x)=\varphi(x)$.

На практике функция $F(x)$ часто задается таблично. По этой таблице можно найти вероятность попадания случайной величины X в интервал (x_1, x_2) без интегрирования:

$$p(X \in (x_1, x_2)) = p(X < x_2) - p(X < x_1) = F(x_2) - F(x_1).$$

Квантилью распределения случайной величины X называется функция x_p, которая является обратной к функции распределения: $p = F(x)$. В частном случае при $p = \dfrac{1}{2}$ квантиль называют медианой: $x_{\frac{1}{2}}$.

2.3 Непрерывное распределение систем случайных величин. Условие их независимости.

Говорят, что система случайных величин, или случайный вектор $X = (X_1, X_2, ..., X_m)$ имеет непрерывное распределение с совместной плотностью вероятности $\varphi(x_1, x_2, ..., x_m)$, если для любой функции $g(x_1, x_2, ..., x_m)$ математическое ожидание можно записать в виде

$$M(g(X_1, ..., X_m)) = \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} g(x_1, ,. x_m)\varphi(x_1, ..., x_m)dx_1...dx_m \,. \qquad (2.6)$$

При этом для любой области A вероятность события $x \in A$ есть интеграл по этой области от плотности вероятности:

$$p(x \in A) = p((x_1, ..., x_m) \in A) = M(I_A) = \int ... \int_A \varphi(x_1, ..., x_m)dx_1...dx_m \,. \qquad (2.7)$$

Заметим, что $\varphi(x_1, ..., x_m) \geq 0$ и

$$\int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \varphi(x_1, ..., x_m)dx_1...dx_m = 1.$$

Если случайный вектор $X = (X_1, ..., X_m)$ имеет непрерывное распределение с плотностью $\varphi(x_1, ..., x_m)$, то любая его координата, например X_1, имеет непрерывное распределение с плотностью

$$\varphi_1(x_1) = \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \varphi(x_1, ..., x_m)dx_2...dx_m \,.$$

Действительно, если подставить в формулу (2.6) $g = g(x_1)$, то

$$M(g(X_1)) = \int_{-\infty}^{+\infty} g(x_1)dx_1 \int_{-\infty}^{+\infty} ... \int_{-\infty}^{+\infty} \varphi(x_1, ..., x_m)dx_2...dx_m \,.$$

Теорема 2.3.1. Для взаимной независимости случайных величин X_1, …, X_m с плотностью вероятности $\varphi(x_1, ..., x_m)$ необходимо и достаточно, чтобы она разлагалась в произведение плотностей отдельных величин:

$$\varphi(x_1, ..., x_m) = \varphi_1(x_1) \cdot ... \cdot \varphi_m(x_m) \,.$$

2.4 Стандартное нормальное распределение

Случайная величина U имеет стандартное нормальное распределение, если ее плотность задается формулой

$$\varphi_0(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}, \; u \in R \,. \qquad (2.8)$$

Кривая стандартного нормального распределения имеет вид

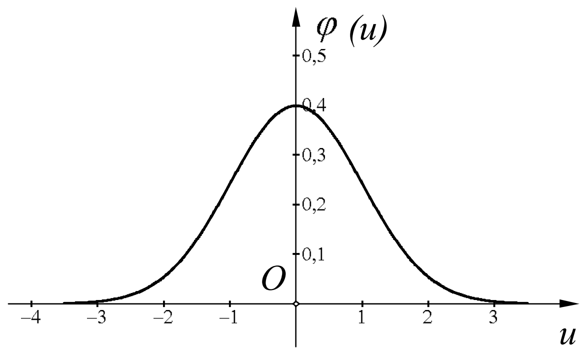


Рис. 2.2.

Найдем основные характеристики этого распределения. Распределение симметрично, так как функция $\varphi_0(u)$ – четная. Это значит, что $M(U)=0$. Впрочем, это можно проверить и интегрированием.

$$M(U)=\int\limits_{-\infty}^{+\infty} u\cdot\varphi_0(u)du=0.$$

Последний интеграл равен нулю как интеграл от нечетной функции по симметричному промежутку.

Вычислим дисперсию случайной величины U.

$$D(U)=M(U^2)-M^2(U)=M(U^2)=\int\limits_{-\infty}^{+\infty} u^2\cdot\varphi_0(u)du=$$

$$=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} u^2 e^{-\frac{u^2}{2}}du=\begin{bmatrix} w=u & dw=du \\ dv=ue^{-\frac{u^2}{2}}du & v=-e^{-\frac{u^2}{2}} \end{bmatrix}=$$

$$=\frac{1}{\sqrt{2\pi}}\left[-ue^{-\frac{u^2}{2}}\Bigg|_{-\infty}^{+\infty}+\int\limits_{-\infty}^{+\infty} e^{-\frac{u^2}{2}}du\right]=\frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} e^{-\frac{u^2}{2}}du=1,$$

$$\left(\lim_{u\to+\infty} u\cdot e^{-\frac{u^2}{2}}=\lim_{u\to+\infty}\frac{u}{e^{\frac{u^2}{2}}}=\lim_{u\to+\infty}\frac{1}{u\cdot e^{\frac{u^2}{2}}}=0.\right)$$

Итак, $D(U)=1$, а, значит и $\sigma(U)=1$.

Расчет вероятностей в стандартном нормальном распределении выполняется с помощью таблиц интеграла вероятностей:

$$\Phi(t)=\int_0^t \varphi_0(u)du\,,\ t>0\,. \qquad (2.9)$$

$$\Phi(-t)=-\Phi(t)\,.$$

Интеграл вероятностей $\Phi(t)$ связан с функцией распределения формулой

$$F(t)=\frac{1}{2}+\Phi(t)\,.$$

Поэтому вероятности рассчитываются так:

$$p(t_1<U<t_2)=F(t_2)-F(t_1)=\Phi(t_2)-\Phi(t_1)\,. \qquad (2.10)$$

Для симметричного интервала:

$$p(|U|<t)=2p(0<U<t)=2\Phi(t)\,. \qquad (2.11)$$

2.5 Нормальное распределение с параметрами (α,σ)

Говорят, что случайная величина *X* имеет нормальное распределение с параметрами (α,σ) (или нормальна (α,σ)), если она может быть представлена в виде $X=\alpha+\sigma U$, где U – имеет стандартное нормальное распределение, $\sigma>0$. Отсюда следует, что $M(X)=\alpha$, $D(X)=\sigma^2$, $\sigma(X)=\sigma$. Так что параметры α и σ представляют собой математическое ожидание и стандартное отклонение величины *X*. Преобразование $\dfrac{X-\alpha}{\sigma}=U$ называется нормировкой величины *X*, поэтому говорят, что случайная величина *X* имеет нормальное распределение, если ее нормировка приводит к величине со стандартным нормальным распределением.

Расчет вероятностей для любой нормально распределенной величины проводится с помощью интеграла (2.9):

$$p(x_1<X<x_2)=p\left(\frac{x_1-\alpha}{\sigma}<\frac{X-\alpha}{\sigma}<\frac{x_2-\alpha}{\sigma}\right)=$$

$$=p\left(\frac{x_1-\alpha}{\sigma}<U<\frac{x_2-\alpha}{\sigma}\right)=\Phi\left(\frac{x_2-\alpha}{\sigma}\right)-\Phi\left(\frac{x_1-\alpha}{\sigma}\right).$$

Нормальное распределение с параметрами (0;1) является стандартным.

2.6 Стандартно нормальный вектор

Говорят, что вектор $U=(U_1,...,U_m)$ имеет стандартное нормальное распределение (стандартно нормален), если все его координаты взаимно независимы и имеют стандартное нормальное распределение (2.8). Плотность вероятности вектора U есть

$$\varphi(u_1,...,u_m)=\varphi_0(u_1)\cdot\ldots\cdot\varphi_0(u_m),$$

его центр распределения совпадает с началом координат, а его матрица ковариаций является единичной матрицей:

$$K_{ij}=M[(U_i-\alpha_i)(U_j-\alpha_j)]=M(U_iU_j)=\begin{cases} M(U_i^2)=1, & i=j \\ M(U_i)M(U_j), & i\neq j \end{cases}.$$

Теорема 2.6.1. Приведенное выше определение стандартно-нормального вектора не зависит от выбора системы координат. То есть если в некоторой декартовой системе координат координаты вектора U взаимно независимы и имеют стандартное нормальное распределение, то это же будет верно при любом повороте системы координат.

2.7 Свойства нормально распределенных величин

Теорема 2.7.1. Если случайные величины X_i взаимно независимы и распределены нормально (α_i,σ_i) $(i=\overline{1,m})$, то любая их линейная комбинация

$$X=\sum_{i=1}^{m}\lambda_i X_i$$

имеет нормальное распределение с параметрами (α,σ), где

$$\alpha=\sum_{i=1}^{m}\lambda_i\alpha_i\ ;\ \sigma=\sqrt{\sum_{i=1}^{m}\lambda_i^2\sigma_i^2}\ .$$

Теорема 2.7.2. Для того, чтобы две нормально распределенных случайных величины были некоррелированными, необходимо и достаточно, чтобы они были независимыми.

Заметим, что из независимости двух случайных величин всегда следует их некоррелированность, так что, по сути, теорема 2.7.2 утверждает, что для двух нормально распределенных величин из их некоррелированности следует их независимость.

2.8 Распределения Пирсона, Стьюдента и Фишера

Особую роль нормальному распределению отводят по той причине, что оно часто встречается при решении многих статистических задач. Кроме того, широкое применение нормального распределения связано с центральной предельной теоремой Ляпунова.

Теорема 2.8.1. Если имеется последовательность взаимно независимых случайных величин X_1, …, X_n, … с любыми распределениями, то за редким исключением распределение нормированной суммы первых n из этих величин

$$\tau_n = \frac{(X_1 + ... + X_n) - M(X_1 + ... + X_n)}{\sigma(X_1 + ... + X_n)}$$

стремится к стандартному нормальному распределению при $n \to \infty$.

Большую роль также играют еще три распределения, тесно связанные с нормально распределенными величинами.

Распределение Пирсона

Пусть даны k взаимно независимых случайных величин U_1, …, U_k, каждая из которых имеет стандартное нормальное распределение. Распределением Пирсона с k степенями свободы называется распределение суммы квадратов этих величин:

$$\chi_k^2 = V = U_1^2 + ... + U_k^2.$$

Кривые распределения Пирсона для $k = 15$ и $k = 20$ показаны на рисунке 2.3. Аналогичную форму имеет любая кривая распределения Пирсона для $k > 2$.

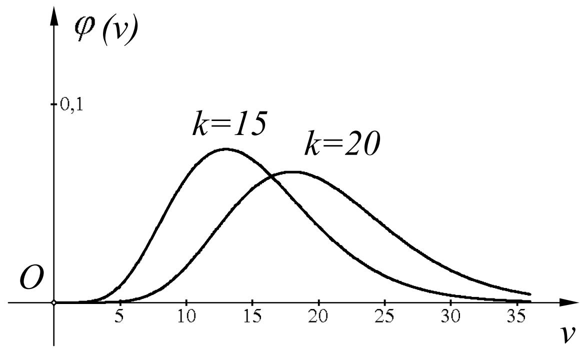


Рис. 2.3.

Найдем основные характеристики величины V.

$$M(V) = M(U_1^2 + ... + U_k^2) = M(U_1^2) + ... + M(U_k^2) = 1 + ... + 1 = k .$$

$$M(U^4) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} u^4 e^{-\frac{u^2}{2}} du = \begin{bmatrix} w = u^3 & dw = 3u^2 du \\ dv = ue^{-\frac{u^2}{2}} du & v = -e^{-\frac{u^2}{2}} \end{bmatrix} =$$

$$= \frac{1}{\sqrt{2\pi}} \left(-u^3 e^{-\frac{u^2}{2}} \Big|_{-\infty}^{+\infty} + 3 \int_{-\infty}^{+\infty} u^2 e^{-\frac{u^2}{2}} du \right) = 0 + 3 \cdot 1 = 3;$$

$$D(U^2) = M(U^4) - M^2(U^2) = M(U^4) - 1 = 3 - 1 = 2;$$

$$D(V) = D(U_1^2 + ... + U_k^2) = [U_i - \text{взаимно независимы}] =$$

$$= D(U_1^2) + ... + D(U_k^2) = 2 + ... + 2 = 2k .$$

Итак, $M(V) = k$, $D(V) = 2k$, $\sigma(V) = \sqrt{2k}$.

Расчет вероятностей в случае распределения Пирсона проводится так же, как и в случае нормального распределения, с помощью таблиц. Причем таблицы часто приводятся лишь для $k \le 60$. При больших k сумма $U_1^2 + ... + U_k^2$, как следует из теоремы Ляпунова, имеет распределение близкое к нормальному с параметрами $(k, \sqrt{2k})$.

Распределение Пирсона обладает следующим свойством: если m взаимно независимых случайных величин V_1, …, V_m имеют распределение Пирсона с k_1, …, k_m степенями свободы соответственно, то их сумма

$V_1 + \ldots + V_m$ также имеет распределение Пирсона с числом степеней свободы $k = k_1 + \ldots + k_m$:

$$V_1 + \ldots + V_m = U_{11}^2 + \ldots + U_{1k_1}^2 + \ldots + U_{m1}^2 + \ldots + U_{mk_m}^2 = \chi_k^2 .$$

Распределение Стьюдента

Пусть даны $k+1$ взаимно независимых величин U, U_1, …, U_k, каждая из которых имеет стандартное нормальное распределение. Распределением Стьюдента с k степенями свободы называется распределение случайной величины T:

$$T = \frac{U}{\sqrt{\dfrac{U_1^2 + \ldots + U_k^2}{k}}} = \frac{U}{\sqrt{\dfrac{\chi_k^2}{k}}} .$$

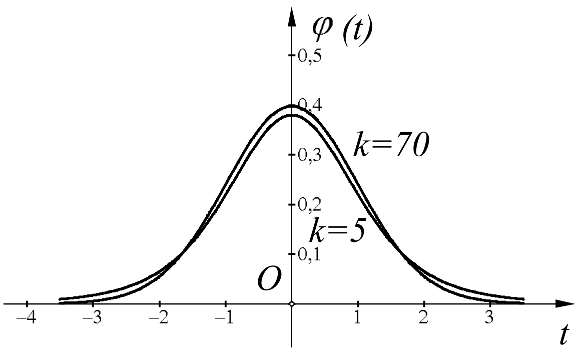


Рис. 2.4.

Распределение Стьюдента симметрично, $M(T) = 0$, его кривая распределения внешне похожа на кривую стандартного распределения. Можно показать, что при $k \to \infty$ распределение Стьюдента стремится к стандартному нормальному распределению.

Часто приводятся квантили распределения абсолютной величины T. Эти квантили обозначают $|t|_p(k)$. В силу симметрии распределения Стьюдента для $t > 0$:

$$p(|T| < t) = 2p(0 < T < t) = 2\left[p(T < t) - \frac{1}{2}\right] = 2p(T < t) - 1 .$$

Поэтому квантили t_p и $|t|_p$ при одном и том же k связаны так:

$$t_p = |t|_{2p-1} .$$

Распределение Фишера

Пусть даны две независимых случайных величины V_1 и V_2, имеющих распределение Пирсона с k_1 и k_2 степенями свободы соответственно. Распределением Фишера с k_1 и k_2 степенями свободы называется распределение величины F:

$$F = \frac{V_1 / k_1}{V_2 / k_2}.$$

Кривая распределения Фишера похожа на кривую распределения Пирсона.

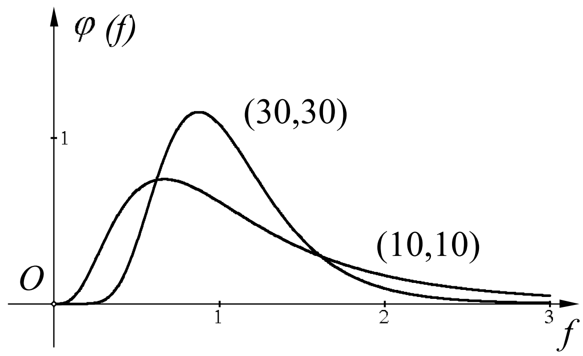


Рис. 2.5.

ГЛАВА 3. ТОЧЕЧНЫЕ ОЦЕНКИ ПАРАМЕТРОВ

3.1 Оценки истинного значения измеряемой величины

Пусть производятся измерения некоторой величины β. Результаты измерений отличаются друг от друга непредсказуемым образом, поэтому результат измерения Y есть случайная величина. Каждое значение y_i этой величины, полученное в результате измерения, дает приближенное значение величины β. Чтобы увеличить точность наших суждений об истинном значении величины β, измерения повторяются n раз и получают ряд результатов: y_1, …, y_n – приближенных значений величины β. Поставим задачу построения по этим результатам приближенного значения величины β с возможно большей точностью.

Обычно рекомендуют брать среднее арифметическое

$$\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}. \qquad (3.1)$$

Это объясняется справедливостью следующей теоремы.

Теорема 3.1.1. Сумма квадратов отклонений значений y_i от $\overline{y}$ меньше, чем сумма квадратов их отклонений от любого другого числа:

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 < \sum_{i=1}^{n} (y_i - c)^2 \text{ , } c \neq \overline{y}.$$

Доказательство.

$$A = \sum_{i=1}^{n} (y_i - c)^2 = \sum_{i=1}^{n} (y_i^2 - 2cy_i + c^2) = \left(\sum_{i=1}^{n} y_i^2 \right) - 2c\overline{y}n + c^2 n;$$

$$B = \sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (y_i^2 - 2\overline{y}y_i + \overline{y}^2) = \left(\sum_{i=1}^{n} y_i^2 \right) - 2\overline{y}^2 n + \overline{y}^2 n;$$

$$A - B = n(c^2 - 2c\overline{y} + \overline{y}^2) = n(c - \overline{y})^2 > 0, \text{ при } c \neq \overline{y}.$$

Пусть дана система случайных величин Y_1, …, Y_n. Будем считать, что:

1. все эти величины взаимно независимы;
2. имеют одинаковые математические ожидания $M(Y_i)=\beta$;
3. имеют одинаковые дисперсии $D(Y_i)=\sigma^2$.

В этом случае систему случайных величин называют **случайной выборкой**. На практике эти три условия означают:

1. измерения проводятся независимо друг от друга, то есть результаты одних измерений не влияют на результаты других;
2. результаты измерений свободны от систематических ошибок;
3. измерения равноточны, то есть производятся на одном и том же приборе при одинаковых условиях.

В математической статистике любую функцию $f(Y_1,...,Y_n)$ от случайной выборки Y_1, …, Y_n, значения которой применяются в качестве приближенного значения параметра β, называют **точечной оценкой** или просто **оценкой** этого параметра и обозначают через $\tilde{\beta}$. Если математическое ожидание оценки $\tilde{\beta}$ совпадает с истинным значением параметра β, то оценку $\tilde{\beta}$ называют **несмещенной**: $M(\tilde{\beta})=\beta$. Если $\tilde{\beta}=\sum_{i=1}^{n}\lambda_i Y_i$, то есть $\tilde{\beta}$ есть линейная комбинация случайных величин Y_i, то оценка $\tilde{\beta}$ называется **линейной**.

3.2 Наилучшая оценка истинного значения при равноточных измерениях

Теорема 3.2.1. Среднее арифметическое случайной выборки есть наилучшая оценка среди всех линейных несмещенных оценок.

Доказательство.

В теореме, по сути, утверждается три факта:

1. среднее арифметическое есть линейная оценка;

2. среднее арифметическое есть несмещенная оценка;

3. среднее арифметическое есть наилучшая оценка.

Докажем последовательно все три утверждения.

1. Среднее арифметическое $\overline{Y} = \sum_{i=1}^{n} \frac{Y_i}{n}$ действительно есть линейная оценка, в качестве коэффициентов выступают числа $\lambda_i = \frac{1}{n}$.

2. Докажем, что среднее арифметическое есть несмещенная оценка:

$$M(\overline{Y}) = M\left(\sum_{i=1}^{n} \frac{Y_i}{n}\right) = \sum_{i=1}^{n} \frac{M(Y_i)}{n} = \beta .$$

3. Докажем, что среднее арифметическое есть наилучшая оценка, то есть имеет наименьшую дисперсию. Другими словами, докажем, что $D(\tilde{\beta}) \geq D(\overline{Y})$, где $\tilde{\beta}$ – произвольная линейная несмещенная оценка.

$$\tilde{\beta} = \sum_{i=1}^{n} \lambda_i Y_i \; ; \; M(\tilde{\beta}) = \beta .$$

$$M(\tilde{\beta}) = \sum_{i=1}^{n} \lambda_i M(Y_i) = \beta \sum_{i=1}^{n} \lambda_i = \beta ;$$

$$\sum_{i=1}^{n} \lambda_i = 1. \qquad (3.2)$$

Так как Y_1, …, Y_n – взаимно независимы, то

$$D(\tilde{\beta}) = \sum_{i=1}^{n} \lambda_i^2 D(Y_i) = \sigma^2 \sum_{i=1}^{n} \lambda_i^2 .$$

Как было показано в главе 1 (см. формулу (1.22))

$$D(\overline{Y}) = \frac{\sigma^2}{n}.$$

Таким образом, остается доказать, что

$$\sum_{i=1}^{n} \lambda_i^2 \geq \frac{1}{n}.$$

Действительно,

$$\sum_{i=1}^{n}\left(\lambda_i-\frac{1}{n}\right)^2=\sum_{i=1}^{n}\lambda_i^2-2\cdot\frac{1}{n}\cdot 1+\frac{1}{n}=\sum_{i=1}^{n}\lambda_i^2-\frac{1}{n}\geq 0\,.$$

Причем равенство здесь достигается только при всех $\lambda_i=\frac{1}{n}$, то есть для среднего арифметического.

3.3 Наилучшая оценка истинного значения при неравноточных измерениях. Введение весов измерений.

Пусть вновь дана система случайных величин $Y_1,\ \ldots,\ Y_n$. По-прежнему предположим, что эти величины взаимно независимы и имеют одинаковые ожидания $M(Y_i)=\beta$. (Выполняются условия 1,2.) Но пусть теперь они имеют различные дисперсии $D(Y_i)=\frac{\sigma^2}{w_i}$. (Пункт 3 не выполняется, измерения не равноточны.) Величина w_i называется весом случайной величины Y_i. Отметим, что веса случайных величин обратно пропорциональны их дисперсиям. (Чем меньше дисперсия, тем точнее результат измерения и поэтому ему придается больший вес.) Теперь мы уже не можем утверждать, что среднее арифметическое случайных величин будет наилучшей среди всех линейных несмещенных оценок.

Теорема 3.3.1. Наилучшей линейной несмещенной оценкой для данной системы случайных величин будет так называемая средневзвешенная оценка:

$$\overline{Y}_{\text{взв}}=\frac{\sum_{i=1}^{n}Y_i w_i}{\sum_{i=1}^{n}w_i}\,.$$

Доказательство.

Так же как и для теоремы 3.2 доказательство будет состоять из трех частей.

1. Средневзвешенная оценка является линейной, в качестве коэффициентов выступают числа $\lambda_i = \dfrac{w_i}{\sum\limits_{i=1}^{n} w_i}$.

2. Средневзвешенная оценка является несмещенной:

$$M(\overline{Y}_{взв}) = \frac{\sum\limits_{i=1}^{n} w_i M(Y_i)}{\sum\limits_{i=1}^{n} w_i} = \frac{\sum\limits_{i=1}^{n} w_i \beta}{\sum\limits_{i=1}^{n} w_i} = \beta .$$

3. Покажем, что средневзвешенная оценка является наилучшей, то есть имеет наименьшую дисперсию. Другими словами, следует показать, что $D(\tilde{\beta}) \geq D(\overline{Y}_{взв})$, где $\tilde{\beta}$ – произвольная линейная несмещенная оценка, то есть

$$\tilde{\beta} = \sum_{i=1}^{n} \lambda_i Y_i \text{ ; } M(\tilde{\beta}) = \beta .$$

В доказательстве теоремы 3.2 было показано, что в этом случае

$$\sum_{i=1}^{n} \lambda_i = 1 .$$

Так как Y_1, …, Y_n – взаимно независимы, то

$$D(\tilde{\beta}) = \sum_{i=1}^{n} \lambda_i^2 D(Y_i) = \sum_{i=1}^{n} \lambda_i^2 \frac{\sigma^2}{w_i} = \sigma^2 \sum_{i=1}^{n} \frac{\lambda_i^2}{w_i} .$$

$$D(\overline{Y}_{взв}) = \sigma^2 \sum_{i=1}^{n} \frac{\left(\dfrac{w_i}{\sum\limits_{i=1}^{n} w_i} \right)^2}{w_i} = \sigma^2 \sum_{i=1}^{n} \frac{w_i}{\left(\sum\limits_{i=1}^{n} w_i \right)^2} = \sigma^2 \left(\sum_{i=1}^{n} w_i \right) \left(\frac{1}{\left(\sum\limits_{i=1}^{n} w_i \right)^2} \right) =$$

$$= \frac{\sigma^2}{\sum\limits_{i=1}^{n} w_i}$$

Остается показать, что

$$\sum_{i=1}^{n}\frac{\lambda_i^2}{w_i} \geq \frac{1}{\sum\limits_{i=1}^{n} w_i}.$$

Действительно,

$$\sum_{i=1}^{n}\left(\frac{\lambda_i}{\sqrt{w_i}} - \frac{\sqrt{w_i}}{\sum\limits_{i=1}^{n} w_i}\right)^2 = \sum_{i=1}^{n}\left(\frac{\lambda_i^2}{w_i} - \frac{2\lambda_i}{\sum\limits_{i=1}^{n} w_i} + \frac{w_i}{\left(\sum\limits_{i=1}^{n} w_i\right)^2}\right) =$$

$$= \sum_{i=1}^{n}\frac{\lambda_i^2}{w_i} - \frac{2}{\sum\limits_{i=1}^{n} w_i} + \frac{1}{\sum\limits_{i=1}^{n} w_i} = \sum_{i=1}^{n}\frac{\lambda_i^2}{w_i} - \frac{1}{\sum\limits_{i=1}^{n} w_i} \geq 0.$$

Теорема доказана.

Величина σ^2 называется дисперсией случайной величины с единичным весом. Напомним, что $D(Y_i) = \dfrac{\sigma^2}{w_i}$, а значит, $D(Y_i) = \sigma^2$ при $w_i = 1$. Вес средневзвешенного равен сумме всех весов w_i, так как $D(\overline{Y}_{взв}) = \dfrac{\sigma^2}{\sum\limits_{i=1}^{n} w_i}$. Коэффициент σ^2 в формуле $D(Y_i) = \dfrac{\sigma^2}{w_i}$ обычно бывает неизвестен, а известны только отношения

$$D(Y_1) : D(Y_2) : \ldots : D(Y_n) = \frac{1}{w_1} : \frac{1}{w_2} : \ldots : \frac{1}{w_n}.$$

ГЛАВА 4. ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

4.1 Постановка задачи

Будем искать такой интервал $(\tilde{\beta}-\varepsilon_1,\tilde{\beta}+\varepsilon_2)$, который накроет истинное значение параметра β с заранее заданной вероятностью р, значение p близко к 1. Эту вероятность называют доверительной вероятностью или надежностью, а сам интервал – доверительным интервалом для параметра β. При этом обычно требуют, чтобы вероятности выхода за границы доверительного интервала в обе стороны были равны:

$$p(\beta<\tilde{\beta}-\varepsilon_1)=p(\beta>\tilde{\beta}+\varepsilon_2)=\frac{1-p}{2}.$$

Это дополнительное требование обеспечивает единственность решения задачи.

4.2 Построение доверительного интервала для стандартного отклонения

Пусть Y_1, …, Y_n – случайные величины, причем:

1. они взаимно независимы;
2. $M(Y_i)=\beta$;
3. $D(Y_i)=\sigma^2$;
4. каждая из этих величин распределена нормально с параметрами (β,σ), то есть $Y_i=\beta+\sigma U_i$, где U_i имеет стандартное нормальное распределение.

Предположим сначала, что истинное значение измеряемой величины β известно. За оценку дисперсии σ^2 примем $\tilde{\sigma}^2=\sum\limits_{i=1}^{n}\frac{(Y_i-\beta)^2}{n}$. Докажем, что эта оценка является несмещенной, то есть $M(\tilde{\sigma}^2)=\sigma^2$. Действительно

$$M(\tilde{\sigma}^2) = \sum_{i=1}^{n} \frac{M[(Y_i - \beta)^2]}{n} = \sum_{i=1}^{n} \frac{D(Y_i)}{n} = \sigma^2.$$

Кроме того, эта оценка представима в виде

$$\tilde{\sigma}^2 = \sum_{i=1}^{n} \frac{(Y_i - \beta)^2}{n} = \sum_{i=1}^{n} \frac{(\beta + \sigma U_i - \beta)^2}{n} = \frac{\sigma^2}{n} \sum_{i=1}^{n} U_i^2 = \frac{\sigma^2 V}{n},$$

где V имеет распределение Пирсона с n степенями свободы.

Зададим доверительную вероятность p и по таблице распределения Пирсона найдем квантили a_1 и a_2 для вероятностей $\frac{1-p}{2}$ и $\frac{1+p}{2}$:

$$p(V < a_1) = \frac{1-p}{2};$$

$$p(V > a_2) = 1 - \frac{1+p}{2} = \frac{1-p}{2}.$$

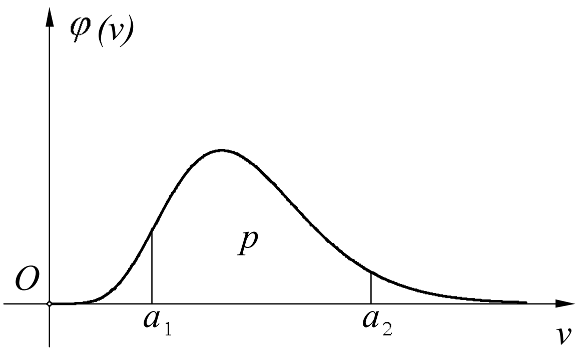


Рис. 4.1.

Рассмотрим цепочку равносильных неравенств:

$$a_1 < V < a_2;$$

$$\frac{a_1 \sigma^2}{n} < \frac{V \sigma^2}{n} < \frac{a_2 \sigma^2}{n};$$

$$\frac{a_1 \sigma^2}{n} < \tilde{\sigma}^2 < \frac{a_2 \sigma^2}{n}.$$

Все эти неравенства выполняются с вероятностью p.

Из последнего неравенства находим доверительный интервал для σ с надежностью p.

$$\tilde{\sigma}\sqrt{\frac{n}{a_2}} < \sigma < \tilde{\sigma}\sqrt{\frac{n}{a_1}}.$$

Пусть теперь истинное значение измеряемой величины β неизвестно. В этом случае нельзя пользоваться оценкой дисперсии

$$\tilde{\sigma}^2 = \sum_{i=1}^{n}\frac{(Y_i-\beta)^2}{n},$$

так как она содержит β. Рассмотрим оценку

$$s_*^2 = \sum_{i=1}^{n}\frac{(Y_i-\overline{Y})^2}{n},$$

которая получается из предыдущей заменой истинного значения β на среднее арифметическое результатов измерения. Однако эта оценка оказывается смещенной. Докажем это. Для этого рассмотрим следующую разность:

$$s_*^2 - \tilde{\sigma}^2 = \sum_{i=1}^{n}\frac{(Y_i-\overline{Y})^2}{n} - \sum_{i=1}^{n}\frac{(Y_i-\beta)^2}{n} =$$

$$= \sum_{i=1}^{n}\left(\frac{Y_i^2}{n} - \frac{2Y_i\overline{Y}}{n} + \frac{\overline{Y}^2}{n} - \frac{Y_i^2}{n} + \frac{2Y_i\beta}{n} - \frac{\beta^2}{n}\right) =$$

$$= -2\overline{Y}^2 + \overline{Y}^2 + 2\beta\overline{Y} - \beta^2 = -(\overline{Y}^2 - 2\beta\overline{Y} + \beta^2) = -(\overline{Y}-\beta)^2.$$

$$M(s_*^2) = M(\tilde{\sigma}^2 - (\overline{Y}-\beta)^2) = M(\tilde{\sigma}^2) - M[(\overline{Y}-\beta)^2]$$

Так как оценка $\tilde{\sigma}^2$ является несмещенной, то есть $M(\tilde{\sigma}^2) = \sigma^2$, то

$$M(s_*^2) = \sigma^2 - D(\overline{Y}).$$

Ранее было показано, что $D(\overline{Y}) = \dfrac{\sigma^2}{n}$.

$$M(s_*^2) = \sigma^2 - \frac{\sigma^2}{n} = \sigma^2\frac{n-1}{n}$$

Из последней формулы ясно как построить несмещенную оценку дисперсии. Достаточно положить

$$s^2 = \frac{n}{n-1}s_*^2 = \sum_{i=1}^{n}\frac{(Y_i-\overline{Y})^2}{n-1}. \qquad (4.1)$$

Эта оценка является несмещенной: $M(s^2)=\sigma^2$. Она называется эмпирической дисперсией. Число $k=n-1$ называется числом степеней свободы этой эмпирической дисперсии. Заметим, что эмпирическую дисперсию можно вычислить по следующей формуле:

$$s^2=\sum_{i=1}^{n}\frac{(Y_i-\overline{Y})^2}{n-1}=\sum_{i=1}^{n}\frac{Y_i^2-2\overline{Y}Y_i+\overline{Y}^2}{n-1}=\frac{\sum_{i=1}^{n}Y_i^2-n\overline{Y}^2}{n-1},$$

то есть

$$s^2=\frac{\sum_{i=1}^{n}Y_i^2-n\overline{Y}^2}{n-1}. \qquad (4.2)$$

Величина $s=\sqrt{s^2}$ называется эмпирическим стандартом.

Рассмотрим формулу (4.1) и выразим эмпирическую дисперсию через величины U_i.

$$s^2=\sum_{i=1}^{n}\frac{(Y_i-\overline{Y})^2}{k}=\frac{1}{k}\sum_{i=1}^{n}[(Y_i-\beta)-(\overline{Y}-\beta)]^2$$

Напомним, что $Y_i=\beta+\sigma U_i$, а $\overline{Y}=\beta+\sigma\overline{U}$, так как

$$\overline{U}=\sum_{i=1}^{n}\frac{U_i}{n}=\sum_{i=1}^{n}\frac{Y_i-\beta}{\sigma n}=\frac{1}{\sigma}(\overline{Y}-\beta).$$

Тогда

$$s^2=\frac{1}{k}\sum_{i=1}^{n}(\sigma U_i-\sigma\overline{U})^2=\frac{\sigma^2}{k}\sum_{i=1}^{n}(U_i^2-2U_i\overline{U}+\overline{U}^2)=$$
$$=\frac{\sigma^2}{k}\left(\sum_{i=1}^{n}U_i^2-2\overline{U}^2n+n\overline{U}^2\right)=\frac{\sigma^2}{k}\left(\sum_{i=1}^{n}U_i^2-n\overline{U}^2\right).$$

Покажем, что вышеуказанная скобка имеет распределение Пирсона. Повернем оси координат в пространстве так, чтобы прежний единичный вектор $\vec{e}_1$ перешел в единичный вектор

$$\vec{e}_1^{\,/}=\frac{\vec{e}_1}{\sqrt{n}}+\frac{\vec{e}_2}{\sqrt{n}}+\ldots+\frac{\vec{e}_n}{\sqrt{n}},$$

$$\left(|\overline{e}_1^{\,/}|=\sum_{i=1}^{n}\frac{1}{(\sqrt{n})^2}=1\right).$$

Тогда стандартно нормальный вектор $U(U_1, U_2, ..., U_n)$ перейдет в стандартно нормальный вектор $U^{/}(U_1^{/}, U_2^{/}, ..., U_n^{/})$ (см. теорему 2.6.1), где

$$U^{/} = U_1^{/}\vec{e}_1^{\,/} + U_2^{/}\vec{e}_2^{\,/} + ... + U_n^{/}\vec{e}_n^{\,/}.$$

Заметим, что

$$\sum_{i=1}^{n} U_i^2 = \sum_{i=1}^{n} \left(U_i^{/}\right)^2$$

как квадрат расстояния от одной и той же точки до начала координат.

Известно, что при повороте системы координат если например первый базисный вектор определен равенством

$$\vec{e}_1^{\,/} = \frac{\vec{e}_1}{\sqrt{n}} + \frac{\vec{e}_2}{\sqrt{n}} + ... + \frac{\vec{e}_n}{\sqrt{n}},$$

то и

$$U_1^{/} = \frac{U_1}{\sqrt{n}} + \frac{U_2}{\sqrt{n}} + ... + \frac{U_n}{\sqrt{n}}.$$

Последнее равенство можно переписать так:

$$U_1^{/} = \sqrt{n} \cdot \overline{U}.$$

Тогда

$$s^2 = \frac{\sigma^2}{k}\left(\sum_{i=1}^{n} U_i^2 - n\overline{U}^2\right) = \frac{\sigma^2}{k}\left(\sum_{i=1}^{n}\left(U_i^{/}\right)^2 - \left(U_1^{/}\right)^2\right) = \frac{\sigma^2}{n-1}\sum_{i=2}^{n}\left(U_i^{/}\right)^2 = \frac{\sigma^2}{n-1}V,$$

где V имеет распределение Пирсона с $k = n-1$ степенями свободы, так как величины $U_i^{/}$ $(i = \overline{2,n})$ взаимно независимы и имеют стандартное нормальное распределение.

Далее рассуждаем так же как и прежде. Зададим доверительную вероятность p и по таблице распределения Пирсона найдем квантили a_1 и a_2 для вероятностей $\frac{1-p}{2}$ и $\frac{1+p}{2}$. Тогда следующие неравенства будут выполняться с вероятностью p:

$$a_1 < V < a_2;$$

$$\frac{\sigma^2}{k}a_1 < \frac{\sigma^2}{k}V < \frac{\sigma^2}{k}a_2;$$

$$\frac{\sigma^2}{k}a_1 < s^2 < \frac{\sigma^2}{k}a_2;$$

$$s\sqrt{\frac{k}{a_2}} < \sigma < s\sqrt{\frac{k}{a_1}}.$$

Это и есть доверительный интервал для стандартного отклонения в предположении, что истинное значение измеряемой величины неизвестно. Иногда принимают следующие обозначения:

$$z_н = \sqrt{\frac{k}{a_2}};\ z_в = \sqrt{\frac{k}{a_1}}.$$

Тогда доверительный интервал запишется в виде:

$$s \cdot z_н < \sigma < s \cdot z_в.$$

В некоторых таблицах приводятся значения $z_н$ и $z_в$ для данных *p* и *k*.

4.3 Построение доверительного интервала для истинного значения измеряемой величины (для математического ожидания).

Пусть вновь дана система случайных величин Y_1, …, Y_n, удовлетворяющая прежним четырем условиям. Напомним, что наилучшей линейной несмещенной оценкой истинного значения измеряемой величины β служит среднее арифметическое, которое также распределено нормально с параметрами $\left(\beta, \frac{\sigma}{\sqrt{n}}\right)$. Это следует из теоремы 2.7.1.

Как было показано ранее

$$\overline{Y} - \beta = \sigma\overline{U} = \frac{\sigma U_1^{/}}{\sqrt{n}},$$

где $U_1^{/}$ имеет стандартное нормальное распределение,

$$s = \sigma\sqrt{\frac{V}{k}},$$

где $V = \left(U_2^/\right)^2 + \ldots + \left(U_n^/\right)^2$ имеет распределение Пирсона.

Рассмотрим отношение

$$T = \frac{\overline{Y} - \beta}{s/\sqrt{n}} = \frac{U_1^/ \sigma\sqrt{n}}{\sqrt{n}\sigma\sqrt{\frac{V}{k}}} = \frac{U_1^/}{\sqrt{\frac{V}{k}}}.$$

Оно имеет распределение Стьюдента с числом степеней свободы $k = n - 1$.

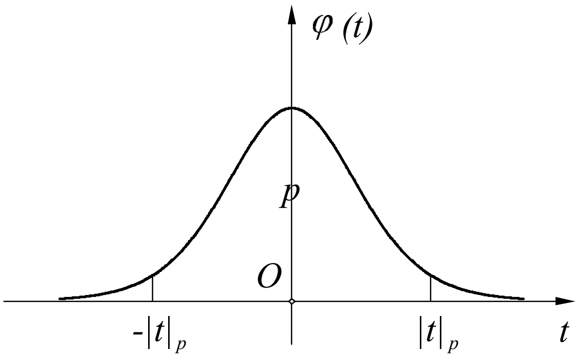


Рис. 4.2.

Если задана доверительная вероятность *p*, то по таблице распределения модуля отношения Стьюдента можно найти квантиль $|t|_p$. Тогда с вероятностью *p* будут выполняться следующие неравенства:

$$|T| < |t|_p;$$

$$\left|\frac{\overline{Y} - \beta}{s/\sqrt{n}}\right| < |t|_p;$$

$$|\overline{Y} - \beta| < |t|_p \frac{s}{\sqrt{n}};$$

$$\overline{Y} - |t|_p \frac{s}{\sqrt{n}} < \beta < \overline{Y} + |t|_p \frac{s}{\sqrt{n}}.$$

Заметим, что в силу симметрии распределения Стьюдента вероятность выхода за границы полученного интервала в обе стороны будут равны $\frac{1-p}{2}$.

ГЛАВА 5. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

5.1 Проверка гипотезы о равенстве дисперсий. Критерий Фишера

Пусть даны две случайные выборки $Y_1, \ldots, Y_{n_1}$ и $Y_1^/, \ldots, Y_{n_2}^/$, каждая из которых удовлетворяет вышеуказанным четырем условиям, причем каждая из величин первой выборки распределена нормально с параметрами (β_1, σ_1), второй – (β_2, σ_2). Требуется проверить гипотезу о равенстве дисперсий: $\sigma_1^2 = \sigma_2^2$. Для проверки гипотезы сравним эмпирические дисперсии

$$s_1^2 = \sum_{i=1}^{n_1} \frac{(Y_i - \overline{Y})^2}{k_1} = \frac{V_1 \sigma_1^2}{k_1},$$

$$s_2^2 = \sum_{i=1}^{n_2} \frac{(Y_i^/ - \overline{Y}^/)^2}{k_2} = \frac{V_2 \sigma_2^2}{k_2},$$

где $k_1 = n_1 - 1$, $k_2 = n_2 - 1$. Рассмотрим отношение большей эмпирической дисперсии к меньшей:

$$F = \frac{s_1^2}{s_2^2} = \frac{V_1 \sigma_1^2 k_2}{k_1 V_2 \sigma_2^2}.$$

Если гипотеза верна, то

$$F = \frac{V_1 / k_1}{V_2 / k_2},$$

то есть F имеет распределение Фишера, причем $F > 1$.

Пусть задано число α ($0 < \alpha < 1$), которое назовем уровнем значимости. Обозначим через $p = 1 - \dfrac{\alpha}{2}$. Тогда по заданным p, k_1, k_2 можно по таблицам найти квантиль распределения Фишера F_p.

Критерий Фишера требует отвергнуть гипотезу о равенстве дисперсий с уровнем значимости α, если вычисленное по экспериментальным данным значение F будет больше квантили: $F > F_p$, и не отвергать в противном случае (рис 5.1).

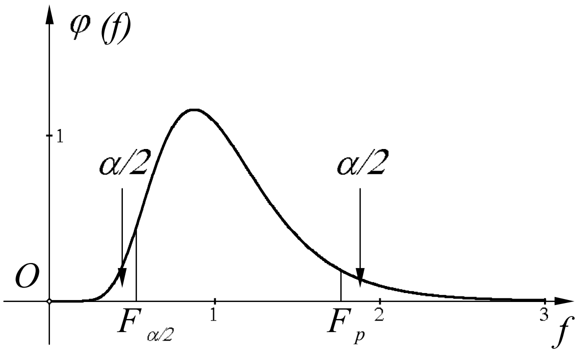


Рис. 5.1.

В качестве уровня значимости часто берут $\alpha = 0{,}05$, и таблицы квантилей распределения Фишера часто приводятся для вероятности $p = 0{,}975$ ($0{,}975 = 1 - 0{,}05 : 2$). Поэтому мы делили большую дисперсию на меньшую. Если у нас есть таблицы для вероятности $p = \frac{\alpha}{2} = 0{,}025$, то можно делить меньшую дисперсию на большую и сравнивать экспериментальное значения F с квантилью $F_{\frac{\alpha}{2}}$.

Если гипотеза о равенстве дисперсий не противоречит экспериментальным данным, то можно найти сводную оценку дисперсии по двум сериям измерений. Если для первой серии подсчитана эмпирическая дисперсия s_1^2 с числом степеней свободы k_1, а для второй серии – s_2^2 с числом степеней свободы k_2, то в качестве сводной дисперсии принимают

$$s_{св}^2 = \frac{s_1^2 k_1 + s_2^2 k_2}{k_1 + k_2}.$$

Из этой формулы видно, что сводная оценка дисперсии есть средневзвешенное ранее найденных эмпирических дисперсий, причем в качестве весов выступают величины k_1 и k_2.

Покажем, что сводная дисперсия является несмещенной оценкой.

$$M(s_{св}^2) = M\left(\frac{s_1^2 k_1 + s_2^2 k_2}{k_1 + k_2}\right) = \frac{k_1 M(s_1^2) + k_2 M(s_2^2)}{k_1 + k_2}$$

Так как s_1^2 и s_2^2 – несмещенные оценки, то

$$M(s_{св}^2) = \frac{k_1 \sigma_1^2 + k_2 \sigma_2^2}{k_1 + k_2}.$$

Так как мы предположили, что $\sigma_1^2 = \sigma_2^2 = \sigma$, то

$$M(s_{св}^2) = \sigma^2.$$

Покажем еще одно сходство между эмпирической и сводной дисперсиями. Напомним, что

$$s_1^2 = \frac{V_1\sigma_1^2}{k_1},\ s_2^2 = \frac{V_2\sigma_2^2}{k_2}.$$

Тогда

$$s_{св}^2 = \frac{s_1^2 k_1 + s_2^2 k_2}{k_1 + k_2} = \frac{\sigma^2}{k}(V_1 + V_2) = \frac{\sigma^2}{k}V,$$

где V имеет распределение Пирсона с числом степеней свободы $k = k_1 + k_2$.

5.2 Проверка гипотезы о равенстве математических ожиданий. Критерий Стьюдента.

Пусть вновь даны две случайные выборки такие же, как в предыдущем параграфе. Но требуется проверить гипотезу о равенстве математических ожиданий: $\beta_1 = \beta_2$.

Предположим сначала, что стандартные отклонения σ_1 и σ_2 известны. Напомним, что $\overline{Y}$ имеет нормальное распределение с параметрами $\left(\beta_1, \frac{\sigma_1}{\sqrt{n_1}}\right)$, $\overline{Y}'$ имеет нормальное распределение с параметрами $\left(\beta_2, \frac{\sigma_2}{\sqrt{n_2}}\right)$. При условии $\beta_1 = \beta_2$ разность $\overline{Y} - \overline{Y}'$ также распределена нормально с параметрами $\left(0, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$. Напомним также, что если величина X имеет нормальное распределение с параметрами (β, σ), то $\frac{X - \beta}{\sigma} = U$ имеет стандартное нормальное распределение. Таким образом

$$\frac{\overline{Y}-\overline{Y}'}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}=U \qquad (5.1)$$

имеет стандартное распределение. По заданному уровню значимости α находим квантиль $|u|_p$ для вероятности $p=1-\alpha$. (Квантиль модуля величины со стандартным нормальным распределением.) Сравним с ней $|U|$. Если окажется, что $|U|<|u|_p$, то гипотеза принимается, как не противоречащая результатам опыта.

На практике же как правило точные значения σ_1 и σ_2 не известны. Значительно чаще стандартные отклонения неизвестны, но можно считать, что они равны между собой, то есть обе серии измерений равноточны ($\sigma_1=\sigma_2=\sigma$). В этом случае формула (5.1) перепишется в виде:

$$\frac{\overline{Y}-\overline{Y}'}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}=U.$$

Умножим обе части равенства на $\frac{\sigma}{s_{св}}$, где $s_{св}=\sigma\sqrt{\frac{V}{k}}$ – эмпирический стандарт:

$$\frac{\overline{Y}-\overline{Y}'}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\cdot\frac{\sigma}{s_{св}}=U\cdot\frac{\sigma}{s_{св}}=\frac{U}{\sqrt{\frac{V}{k}}}=T.$$

Здесь T имеет распределение Стьюдента с числом степеней свободы $k=k_1+k_2$.

По заданному уровню значимости α находим квантиль распределения модуля отношения Стьюдента $|t|_p$, где $p=1-\alpha$.

<u>Критерий Стьюдента</u> требует отвергнуть гипотезу о равенстве математических ожиданий, если при заданном уровне значимости α вычисленное значение отношения Стьюдента окажется по модулю больше квантили: $|T|>|t|_p$, в противном случае гипотеза принимается (рис 5.2).

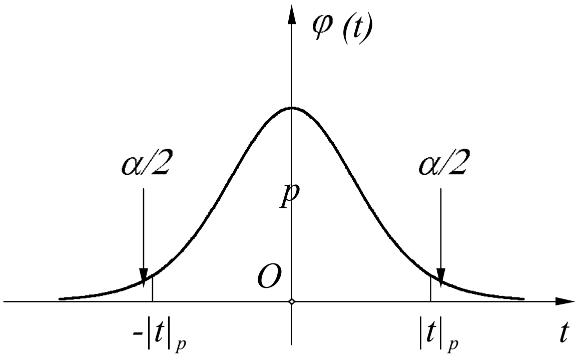


Рис. 5.2.

Подчеркнем, что применение критерия Стьюдента для проверки гипотезы о равенстве математических ожиданий требует предварительной проверки гипотезы о равенстве дисперсий, например, по критерию Фишера.

Если гипотеза о равенстве математических ожиданий не противоречит экспериментальным данным, то можно найти сводную оценку математического ожидания. Если для обеих серий найдены средние арифметические $\overline{Y}$ и $\overline{Y}'$, то в качестве сводной оценки математического ожидания принимают

$$\overline{Y}_{св} = \frac{n_1\overline{Y} + n_2\overline{Y}'}{n_1 + n_2}.$$

Из этой формулы видно, что сводная оценка математического ожидания есть средневзвешенное ранее найденных средних арифметических с весами n_1 и n_2. Заметим, что сводная оценка $\overline{Y}_{св}$ является несмещенной.

Если приняты гипотезы о равенствах $\sigma_1^2 = \sigma_2^2$ и $\beta_1 = \beta_2$, то можно обе выборки объединить в одну и найти еще одну оценку дисперсии $s_{об}^2$, так называемую объединенную оценку.

$$s_{об}^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n_1} Y_i^2 + \sum_{i=1}^{n_2} Y_i'^2 - n\left(\overline{Y}_{св}\right)^2\right)$$

Отметим, что последняя формула аналогична формуле (4.2) для эмпирической дисперсии.

5.3 Критерий согласия Пирсона

Критерий согласия Пирсона служит для проверки гипотезы о согласованности результатов эксперимента с данным распределением непрерывной случайной величины. Если плотность данного распределения обозначить через $\varphi(x)$, то вероятность p_i попадания величины X в интервал (x_i, x_{i+1}) будет равна:

$$p_i = p(x_i < X < x_{i+1}) = \int_{x_i}^{x_{i+1}} \varphi(x)dx. \qquad (5.2)$$

Эксперимент определяет частоту $\frac{n_i}{n}$ попадания в тот же интервал. При условии истинности нашей гипотезы и при достаточно большом числе n экспериментов каждая частота $\frac{n_i}{n}$ должна быть близка к вероятности p_i.

Для применения критерия Пирсона весь диапазон возможных значений случайной величины X заранее, до эксперимента, разбивают на L частей, вычисляют по формуле (5.2) вероятности попадания в каждый интервал p_i, и затем по результатам эксперимента подсчитывают частоты попадания в те же интервалы. Взвешенную сумму квадратов отклонений частот $\frac{n_i}{n}$ от вероятностей p_i обозначают:

$$\chi^2 = n\sum_{i=1}^{L} \frac{\left(\frac{n_i}{n} - p_i\right)^2}{p_i} = \sum_{i=1}^{L} \frac{(n_i - np_i)^2}{np_i}. \qquad (5.3)$$

Если наша гипотеза верна, то при достаточно общих условиях распределение величины χ^2, определяемой формулой (5.3), стремится к распределению Пирсона с $k = L - 1$ числом степеней свободы при $n \to \infty$.

<u>Критерий Пирсона</u> требует отвергнуть нашу гипотезу, если вычисленное по формуле (5.3) значение χ^2 окажется больше квантили $\chi_p^2(k)$ при заданном уровне значимости $\alpha = 1 - p$, в противном случае гипотеза принимается (рис 5.3).

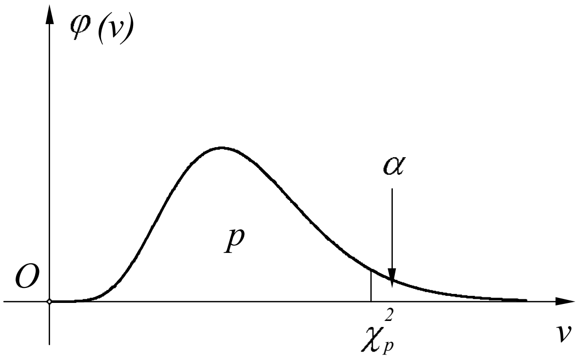


Рис. 5.3.

5.4 Критерий согласия Пирсона при оцениваемых параметрах распределения.

На практике при формулировке гипотезы о законе распределения непрерывной величины X указывают обычно не точное распределение, а лишь его тип, например «X имеет нормальное распределение». Параметры этого распределения оценивают при этом по той же серии экспериментов, по которой строят критерий согласия. В этой ситуации весьма трудно проверить условия применимости критерия Пирсона. Принято считать, что критерием Пирсона все же можно пользоваться, уменьшая число степеней свободы распределения Пирсона на число оцениваемых параметров. Например, если проверяют гипотезу о нормальном распределении, причем оба параметра α и σ оценивают по результатам эксперимента, то число степеней свободы надо брать равным $k = (L-1) - 2 = L - 3$.

На практике точки дробления x_i выбирают таким образом, чтобы либо все промежутки (x_i, x_{i+1}) имели одинаковую длину, либо все вероятности p_i равнялись бы между собой, то есть $p_i = \dfrac{1}{L}$.

ГЛАВА 6. РЕГРЕССИОННЫЙ АНАЛИЗ

6.1 Простая линейная регрессия. Постановка задачи

На практике часто возникает такая ситуация, когда есть предположение о функциональной зависимости величины y от величины x:

$$y = f(x, \alpha_1, ..., \alpha_k)$$

причем вид функции известен, а параметры α_1, …, α_k неизвестны. В этом случае вышеуказанное уравнение называют уравнением регрессии.

Пусть уравнение регрессии имеет вид $y = a + bx$, где параметры a и b неизвестны. Независимое переменное x называют фактором, а зависимое переменное y – откликом. Функции $\varphi_1(x) = 1$ и $\varphi_2(x) = x$ называют базисными. Предполагают, что y – случайная величина, а x – неслучайная, то есть отдельные значения x можно получать точно (или с очень большой степенью точности, при которой погрешностью можно пренебречь). Задача заключается в том, чтобы приближенно найти значения a и b при условии, что нам известны результаты n экспериментов, в результате которых найдены числа y_1, …, y_i, …, y_n, которые являются приближенными значениями отклика при n различных значениях фактора x_1, …, x_i, …, x_n, которые нам известны точно.

Сделаем предположение, что все эксперименты производятся независимо друг от друга, равноточны (то есть производятся с помощью одних и тех же приборов и при одинаковых условиях) и свободны от систематических ошибок. Поэтому будем считать, что все случайные величины y_1, …, y_i, …, y_n взаимно независимы, распределены нормально, причем все они имеют одинаковую, но неизвестную нам дисперсию $D(y_i) = \sigma^2$ и имеют математические ожидания $M(y_i)$ равные истинным значениям отклика: $M(y_i) = a + bx_i = \beta_i$. Отметим, что векторы

$$\varphi_1 = (\varphi_1(x_1), ..., \varphi_1(x_i), ..., \varphi_1(x_n)) = (1, ..., 1, ..., 1)$$

$$\varphi_2 = (\varphi_2(x_1),...,\varphi_2(x_i),...,\varphi_2(x_n)) = (x_1,...,x_i,...,x_n)$$

линейно независимы, так как точки x_1, …, x_i, …, x_n различны.

6.2 Построение регрессионной модели по методу наименьших квадратов

При методе наименьших квадратов значения коэффициентов находятся из требования, чтобы сумма квадратов отклонений эмпирических значений отклика от найденных из уравнения регрессии была бы наименьшей, то есть требуется найти $\min Z$, где

$$Z = \sum_{i=1}^{n}(y_i - (a + bx_i))^2 .$$

Составим систему уравнений, находя частные производные функции Z по каждому из параметров и приравнивая их к нулю.

$$\frac{\partial Z}{\partial a} = -2\sum_{i=1}^{n}(y_i - (a + bx_i))$$

$$\frac{\partial Z}{\partial b} = -2\sum_{i=1}^{n}(y_i - (a + bx_i))x_i$$

$$\begin{cases} \sum_{i=1}^{n}(y_i - (a + bx_i)) = 0 \\ \sum_{i=1}^{n}(y_i - (a + bx_i))x_i = 0 \end{cases} \qquad (6.1)$$

$$\begin{cases} na + \left(\sum_{i=1}^{n} x_i\right)b = \sum_{i=1}^{n} y_i \\ \left(\sum_{i=1}^{n} x_i\right)a + \left(\sum_{i=1}^{n} x_i^2\right)b = \sum_{i=1}^{n} x_i y_i \end{cases} \qquad (6.2)$$

Это система линейных уравнений. Можно показать, что ее определитель отличен от нуля, а, значит, она имеет единственное решение $a = \hat{a}$, $b = \hat{b}$. Пусть $\hat{y} = \hat{a} + \hat{b}x$, а $\hat{y}_i = \hat{a} + \hat{b}x_i$. Тогда

$$\min Z = \sum_{i=1}^{n} (y_i - (\hat{a} + \hat{b} x_i))^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 .$$

Отметим, что каждый из найденных коэффициентов $\hat{a}$ и $\hat{b}$ представляет собой линейную комбинацию случайных величин y_i $(i = \overline{1,n})$ и поэтому имеет нормальное распределение. Можно доказать, что $M(\hat{a}) = a$ и $M(\hat{b}) = b$, где a и b – истинные значения коэффициентов, а дисперсии пропорциональны неизвестной дисперсии σ^2: $D(\hat{a}) = k_a \sigma^2$ и $D(\hat{b}) = k_b \sigma^2$. Для любого $i = \overline{1,n}$ $\hat{y}_i$ также имеет нормальное распределение и $M(\hat{y}_i) = M(\hat{a} + \hat{b} x_i) = a + bx_i = \beta_i$ есть истинное значение отклика в точке x_i.

6.3 Различные вариации отклика

При неизвестной ошибке эксперимента для проверки адекватности модели понадобятся некоторые новые понятия. Рассмотрим 3 вариации отклика.

Общая вариация отклика:

$$ss_{общ} = \sum_{i=1}^{n} (y_i - \overline{y})^2 .$$

Это сумма квадратов отклонений эмпирических значений от их среднего арифметического.

Вариация, относящаяся к модели, или объяснимая вариация:

$$ss_{мод} = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 .$$

Это сумма квадратов отклонений значений отклика, найденных из уравнения регрессии, от среднего арифметического эмпирических значений. Эта вариация объясняется изменением значений фактора при различных опытах.

Остаточная вариация отклика:

$$ss_{ост} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 .$$

Это сумма квадратов отклонений эмпирических значений от найденных из уравнения регрессии. Остаточная вариация может быть вызвана влиянием факторов, неучтенных при проведении опытов, неточностью самих опытов, неудачным выбором типа модели, и т.д. и свидетельствует о неточности найденной модели.

Если модель построена по методу наименьших квадратов, то

$$ss_{общ} = ss_{мод} + ss_{ост}. \tag{6.3}$$

Докажем это.

$$ss_{общ} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 =$$

$$= \sum_{i=1}^{n}((y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})) =$$

$$= ss_{ост} + ss_{мод} + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)\hat{y}_i - 2\sum_{i=1}^{n}(y_i - \hat{y}_i)\bar{y}$$

Остается доказать, что обе последние суммы равны нулю. Так как $\hat{a}$ и $\hat{b}$ удовлетворяют системе (6.1), то

$$\begin{cases} \sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i)) = 0 \\ \sum_{i=1}^{n}(y_i - (\hat{a} + \hat{b}x_i))x_i = 0 \end{cases},$$

$$\begin{cases} \sum_{i=1}^{n}(y_i - \hat{y}_i) = 0 \\ \sum_{i=1}^{n}(y_i - \hat{y}_i)x_i = 0 \end{cases}. \tag{6.4}$$

Если обе части первого равенства умножить на $\hat{a}$, а второго равенства на $\hat{b}$ и затем сложить их, то получим

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{a}+\hat{b}x_i)=0$$

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)=0$$

Если же обе части первого равенства умножить на $\overline{y}$, то получим

$$\sum_{i=1}^{n}(y_i-\hat{y}_i)\overline{y}=0$$

Итак, равенство (6.3) доказано.

Число степеней свободы сумм рассмотренного вида определяется как разность между числом экспериментов и максимальным числом независимых линейных соотношений, связывающих соответствующие отклонения. Под линейным соотношением понимается равенство вида

$$\sum_{i=1}^{n}c_i\Delta y_i=0, \tag{6.5}$$

где $(c_1,...,c_i,...,c_n)$ некоторый набор коэффициентов, а Δy_i $(i=\overline{1,n})$ – отклонения отклика. Если это соотношение выполняется для m различных наборов коэффициентов, которые образуют линейно независимую систему векторов, но любые $m+1$ наборов коэффициентов, удовлетворяющих равенству (6.5), оказываются линейно зависимыми векторами, то число степеней свободы суммы $\sum_{i=1}^{n}(\Delta y_i)^2$ равно $n-m$.

Для суммы $ss_{общ}$ существует единственное соотношение вида (6.5). Пусть $c_i=1$ $(i=\overline{1,n})$. Тогда $\sum_{i=1}^{n}(y_i-\overline{y})=0$. Действительно

$$\sum_{i=1}^{n}(y_i-\overline{y})=\sum_{i=1}^{n}y_i-n\overline{y}=\sum_{i=1}^{n}y_i-n\cdot\frac{\sum_{i=1}^{n}y_i}{n}=0.$$

Если же допустить, что при некоторых c_i $(i=\overline{1,n})$ $\sum_{i=1}^{n}c_i(y_i-\overline{y})=0$, то нетрудно доказать (опираясь на взаимную независимость y_i $(i=\overline{1,n})$), что все c_i будут равны между собой, а, значит, векторы $(c_1,...,c_i,...,c_n)$ и $(1,...,1,...,1)$

будут линейно зависимыми. Следовательно, число степеней свободы суммы $ss_{общ}$ равно $n-1$.

Чтобы найти число степеней свободы для $ss_{ост}$, надо обратиться к равенствам (6.4). Мы видим, что для $\Delta y_i = y_i - \hat{y}_i$ соотношения вида (6.5) выполняются, как для набора коэффициентов $c_i = 1 = \varphi_1(x_i)$ ($i = \overline{1,n}$), так и для набора $c_i = x_i = \varphi_\alpha(x_i)(i = \overline{1,n})$, причем векторы φ_1 и φ_2, как отмечалось выше, линейно независимы. Можно доказать, что всякий вектор $(c_1,...,c_i,...,c_n)$, также удовлетворяющий (6.5), связан с векторами φ_1 и φ_2 линейной зависимостью. Таким образом, число степеней свободы суммы $ss_{ост}$ равно $n-2$. Однако, этот результат связан с тем, что в нашем частном случае число оцениваемых параметров равнялось двум. В более общем случае, который мы рассмотрим ниже, когда число параметров равно *k*, число степеней свободы $ss_{ост}$ равняется $n-k$.

И, наконец, число степеней свободы суммы $ss_{мод}$ равно $(n-1)-(n-k) = k-1$. (В нашем частном случае $k-1=1$.)

Разделив каждую из рассмотренных вариаций на соответствующее число степеней свободы, получим случайные величины

$$s^2_{общ} = \frac{ss_{общ}}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$$
$$s^2_{мод} = \frac{s_{мод}}{k-1} = \frac{1}{k-1}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$
$$s^2_{ост} = \frac{ss_{ост}}{n-k} = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

В частности, $s^2_{ост}$ является несмещенной оценкой неизвестной дисперсии σ^2 и ее называют эмпирической дисперсией регрессии, а

$$s_{ост} = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

называется средней квадратичной ошибкой регрессии.

6.4 Проверка адекватности модели

Эмпирическим коэффициентом корреляции называется

$$R=\sqrt{\frac{ss_{мод}}{ss_{общ}}}=\sqrt{1-\frac{ss_{ост}}{ss_{общ}}}=\sqrt{1-\frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{y})^2}{\sum_{i=1}^{n}(y_i-\overline{y})^2}}.$$

Эмпирический коэффициент корреляции R изменяется от 0 до 1. Чем ближе R к 1,тем больше доля объяснимой вариации в общей вариации, а, следовательно, модель выбрана удачней.

Квадрат эмпирического коэффициента корреляции

$$R^2=\frac{ss_{мод}}{ss_{общ}}=1-\frac{ss_{ост}}{ss_{общ}}$$

называют эмпирическим коэффициентом детерминации.

Если в последней формуле $ss_{ост}$ заменить на $s^2_{ост}=\frac{1}{n-k}ss_{ост}$ и $ss_{общ}$ заменить на $s^2_{общ}=\frac{1}{n-1}ss_{общ}$, то получится приведенное значение коэффициента детерминации

$$R^2_{прив}=1-\frac{s^2_{ост}}{s^2_{общ}}=1-\frac{(n-1)ss_{ост}}{(n-k)ss_{общ}},$$

которое является несмещенной оценкой коэффициента детерминации. При удачном построении R, R^2 и $R^2_{прив}$ близки к единице.

Для проверки адекватности модели используют критерий Фишера. Можно доказать, что

$$ss_{общ}=\sigma^2\sum_{i=1}^{n-1}V_i^2+z,$$

где случайные величины V_1, …, V_n взаимно независимы и распределены стандартно нормально, а математическое ожидание случайной величины z:

$$M(z)=B=\sum_{i=1}^{n}(\beta_i-\overline{\beta})^2,$$

где β_i неизвестные нам истинные значения отклика в точках x_i, а $\overline{\beta}$ – их среднее арифметическое. При этом

$$ss_{мод} = \sigma^2 \sum_{i=1}^{k-1} V_i^2 + z \text{, а } ss_{ост} = \sigma^2 \sum_{i=k}^{n-1} V_i^2 .$$

Таким образом

$$ss_{ост} = \sigma^2 \chi^2 (n-k),$$

где $\chi^2(n-k)$ имеет распределение Пирсона с $n-k$ степенями свободы, а

$$s_{ост}^2 = \sigma^2 \frac{1}{n-k} \chi^2 (n-k).$$

Отсюда, в частности следует, что

$$M(s_{ост}^2) = \sigma^2 \frac{1}{n-k} M(\chi^2(n-k)) = \sigma^2 .$$

Таким образом, доказано отмеченное ранее утверждение, что $s_{ост}^2$ есть несмещенная оценка σ^2.

Благодаря наличию второго слагаемого $ss_{мод}$ имеет распределение $\sigma^2\chi^2(k-1)$ лишь в частном случае, когда $B = \sum_{i=1}^{n} (\beta_i - \overline{\beta})^2 = 0$, то есть когда все значения β_i $(i = \overline{1,n})$ равны между собой. Рассмотрим отношение

$$F = \frac{s_{мод}^2}{s_{ост}^2} \text{, где}$$

$$s_{мод}^2 = \frac{1}{k-1} ss_{мод} = \sigma^2 \frac{1}{k-1} \sum_{i=1}^{k-1} V_i^2 + \frac{1}{k-1} \cdot z .$$

В частном случае, когда $B = 0$, z также рано 0, и это отношение

$$F_0 = \frac{s_{мод}^2}{s_{ост}^2} = \frac{\frac{1}{k-1} \sum_{i=1}^{k-1} V_i^2}{\frac{1}{n-k} \sum_{i=k}^{n-1} V_i^2}$$

имеет распределение Фишера с $k-1$ и $n-k$ степенями свободы. Если модель адекватна, то найденное из эксперимента значение F^* не меньше соответствующего F_0^*: $F^* \ge F_0^*$. К сожалению, значение F_0^* нам неизвестно, а известно лишь распределение случайной величины F_0. Поэтому рассуждают следующим образом. Задают уровень значимости α (например, $\alpha = 0{,}05$), находят по таблицам $F(\alpha, k-1, n-k)$. При этом вероятность, что $F_0 \in (0, F(\alpha, k-1, n-k)) = 1-\alpha$, а вероятность, что $F_0 \in (F(\alpha, k-1, n-k), \infty)$ равна α (рис 6.1).

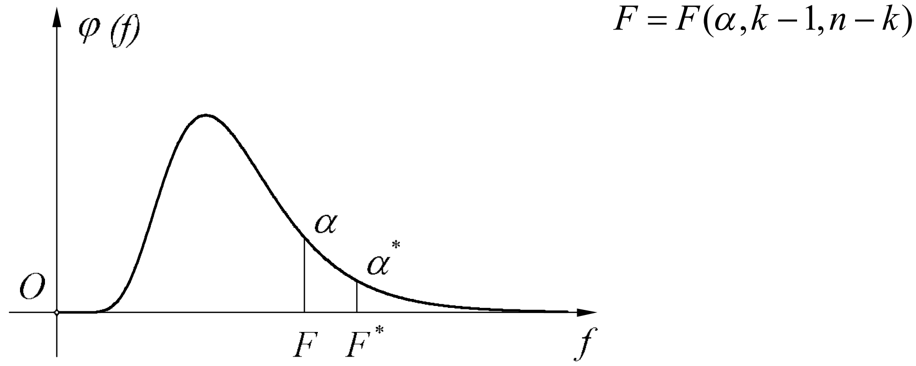


Рис. 6.1.

Если окажется, что $F^* \le F(\alpha, k-1, n-k)$, то гипотеза об адекватности модели экспериментальным данным отвергается (на уровне значимости α). Если же $F^* > F(\alpha, k-1, n-k)$, то с вероятностью ошибиться, меньшей чем α, можно утверждать, что $F^* \ge F_0^*$. В этом случае гипотезу об адекватности модели мы не отвергаем, как непротиворечащую экспериментальным данным. Однако исследование модели необходимо продолжить.

Замечание. Компьютерные программы наряду со значением F^* выдают и его уровень значимости, то есть $\alpha^* = p(F_0 \ge F^*)$. Это позволяет делать вывод об адекватности модели, не прибегая к таблицам. Если окажется, что $\alpha^* \ge \alpha$, то гипотеза об адекватности отвергается, а если $\alpha^* < \alpha$, то нет.

6.5 Цензурирование и прогнозирование

Сравним найденные из уравнения регрессии значения отклика $\hat{y}_i$ ($i=\overline{1,n}$) с соответствующими эмпирическими значениями y_i ($i=\overline{1,n}$). Если эти значения резко отличаются друг от друга, то модель неприемлема. Если для всех $i=\overline{1,n}$ эти обозначения близки друг к другу, то модель принимается. Отметим, что в том случае, когда мы отвергли построенную простую линейную модель, полезно изобразить набор точек (x_i, y_i) ($i=\overline{1,n}$) на графике. Иногда бывает возможно подметить некоторую зависимость между ними, отличную от линейной. Это указывает на необходимость перехода к некоторой нелинейной модели (например, квадратичной). Возможен случай, когда лишь отдельные значения y_i значительно отклоняются от соответствующих значений $\hat{y}_i$. Таким значения y_i называются выбросами. Возникает подозрение, что при проведении *i*-го опыта исследователем допущена ошибка или на получение неправильного результата повлияли случайные внешние факторы. Для устранения эффекта выброса рекомендуется провести корректировку построенной модели. Отдельное неправдоподобное значение y_i следует удалить из набора эмпирических значений отклика и в дальнейшем использовать результаты лишь $n-1$ эксперимента. Затем надо заново найти коэффициенты модели и после этого новую модель проверить на адекватность описанным выше способом. Эта процедура называется цензурированием. Она может значительно улучшить показатели адекватности.

В случае большого числа экспериментов можно удалить из набора данных более одной точки.

Удачно подобранная модель позволяет прогнозировать значения отклика при значениях фактора, отличного от x_1, …, x_n. За приближенное значение отклика в точке *x* принимается $\hat{y}(x)=\hat{a}+\hat{b}x$.

6.6 Регрессии, линейные относительно параметров. Постановка задачи.

Пусть дано уравнение регрессии

$$y = \alpha_1 \varphi_1(x) + \ldots + \alpha_k \varphi_k(x) = \sum_{m=1}^{k} \alpha_m \varphi_m(x),$$

где $\varphi_m(x)$ ($m = \overline{1,k}$) – базисные функции, а коэффициенты (параметры) α_m ($m = \overline{1,k}$) неизвестны. Нам также даны результаты n экспериментов, то есть набор значений отклика y_1, …, y_n, соответствующий набору различных значений фактора x_1, …, x_n. Кроме ранее сделанных допущений, предположим также, что векторы

$$\varphi_1 = (\varphi_1(x_1), \ldots, \varphi_1(x_n)),\ \ldots,\ \varphi_m = (\varphi_m(x_1), \ldots, \varphi_m(x_n))$$

линейно независимы (в противном случае можно было бы уменьшить число базисных функций) и что вектор $\varphi_1 = (1, \ldots, 1)$ (что справедливо для полиномиальных моделей). Оба эти утверждения в случае линейной регрессии справедливы.

Найдем коэффициенты $\hat{\alpha}_l$ ($l = \overline{1,k}$) методом наименьших квадратов, решив на этот раз систему из k линейных уравнений с k неизвестными и с определителем системы, отличным от нуля. По-прежнему можно доказать, что все коэффициенты $\hat{\alpha}_l$ нормально распределены и для $l = \overline{1,k}$ $M(\hat{\alpha}_l) = \alpha_l$ и $D(\hat{\alpha}_l) = k_l \sigma^2$, где σ^2 неизвестная нам дисперсия, а коэффициенты k_l ($l = \overline{1,k}$) – неслучайные величины. Другие результаты, изложенные нами для простой линейной регрессии, остаются также справедливыми. Как отмечалось выше, $s_{ост}^2$ является несмещенной оценкой σ^2. Следовательно, эмпирическая дисперсия и средняя квадратичная ошибка коэффициента $\hat{\alpha}_l$ равны соответственно

$$s_l^2(\hat{\alpha}_l) = k_l s_{ост}^2;\ s_l(\hat{\alpha}_l) = \sqrt{k_l}\, s_{ост}.$$

Итак, эмпирические дисперсии всех коэффициентов с точностью до неслучайного множителя распределены по Пирсону с одним и тем же числом степеней свободы $n-k$.

В уравнении регрессии под буквой x мы до сих пор понимали действительное число, то есть точку числовой оси. Однако, под x можно понимать и точку плоскости $x=(x_1,x_2)$ и точку трехмерного пространства $x=(x_1,x_2,x_3)$ и даже точку m-мерного Евклидова пространства $x=(x_1,x_2,...,x_m)$. Другими словами, базисные функции могут быть и функциями одного аргумента и функциями многих аргументов. В первом случае регрессии называются простыми или однофакторными, а во втором – многофакторными. Изложенные выше результаты справедливы и для многофакторных регрессий.

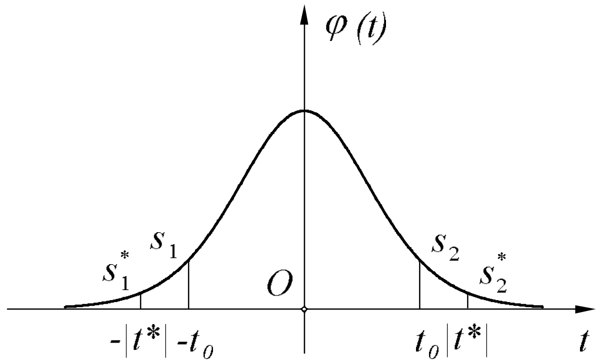
6.7 Проверка значимости коэффициентов регрессии

Коэффициент уравнения регрессии называется незначимым, если его математическое ожидание (истинное значение) равно нулю. Если предположить, что $M(\hat{\alpha}_l)=0$, то тогда случайная величина $t_l=\dfrac{\hat{\alpha}_l}{s(\hat{\alpha}_l)}$, равная отношению этого коэффициента к его средней квадратичной ошибке, имеет распределение Стьюдента с $n-k$ степенями свободы. Действительно, в силу гипотезы

$$t_l=\frac{\hat{\alpha}_l}{s(\hat{\alpha}_l)}=\frac{\hat{\alpha}-\alpha_l}{\sqrt{k_l}\cdot s_{ocm}}=\frac{\hat{\alpha}-\alpha_l}{\sqrt{k_l}\cdot\sigma\cdot\sqrt{\dfrac{1}{n-k}\cdot\chi^2(n-k)}}=$$

$$=\frac{\dfrac{\hat{\alpha}_l-\alpha_l}{\sqrt{k_l}\cdot\sigma}}{\sqrt{\dfrac{1}{n-k}\chi^2(n-k)}}=\frac{U_l}{\sqrt{\dfrac{1}{n-k}\chi^2(n-k)}},$$

где U_l имеет стандартное нормальное распределение.



$$s_1 + s_2 = \alpha$$
$$s_1^* + s_2^* = \alpha^*$$
$$t_0 = t(\alpha, n-k)$$

Рис. 6.2.

Вероятность, что $t_l \notin (-t(\alpha,n-k);t(\alpha,n-k))$ равна α, а вероятность, что $t_l \in (-t(\alpha,n-k),t(\alpha,n-k))$ равна $1-\alpha$ (рис 6.2). Если при заданном уровне значимости α модуль найденного значения $|t_l^*| \geq t(\alpha,n-k)$, то гипотеза отвергается. При этом вероятность отвергнуть правильную гипотезу равна α. В этом случае коэффициент считается значимым (для принятого уровня значимости).

Многие статистические программы наряду со значениями коэффициентов $\hat{\alpha}_l$, их средних квадратичных ошибок $s(\hat{\alpha}_l)$, значениями статистик t_l^*, выдают также и уровни значимости этих статистик, то есть $\alpha_l^* = p(|t| \geq |t_l^*|)$. Это позволяет решать вопрос о значимости коэффициента, не пользуясь таблицами. Если окажется, что $\alpha_l^* \leq \alpha$, то коэффициент надо считать значимым, а если мы выберем $\alpha_l^* > \alpha$, то коэффициент признается незначимым. Если для некоторых коэффициентов $\hat{\alpha}_l$ уровень значимости α_l^* окажется значительным, то с очень большой вероятностью можно считать их истинные значения равными нулю и отбросить в уравнении регрессии соответствующие слагаемые. Затем значения остальных коэффициентов следует пересчитать заново, учитывая, что число базисных функций k уменьшается на число удаленных слагаемых. Новая модель окажется проще прежней, однако ее заново придется проверить на адекватность.

Пусть, например, уравнение регрессии имеет вид

$$y = a_0 + a_{11}x_1 + a_{12}x_1^2 + a_{21}x_2 + a_{22}x_2^2 .$$

Если для фактора x_j оба коэффициента a_{j1} и a_{j2} окажутся незначимыми, то это означает, что зависимость отклика от j-го фактора ничтожна и ею можно пренебречь. Если же незначимым окажется только коэффициент a_{j2}, стоящий перед x_j^2, то это означает, что зависимость отклика от j-го фактора близка к линейной.

6.8 Понятие об активном эксперименте. Полный факторный план

До сих пор при построении регрессионных моделей мы предполагали, что нам даны два набора: набор значений фактора x_1, …, x_n и соответствующий набор значений отклика y_1, …, y_n, полученный экспериментально. Однако часто (особенно в случае многофакторных регрессий) бывает желательно самим заранее планировать эксперимент, то есть заранее специальным образом выбирать точки, в которых он ставится. В этом случае эксперимент называется активным, а планом эксперимента называется множество точек x_1, …, x_n, в которых производится эксперимент. В случае активного эксперимента возможно подбирать область изменения факторов, ставить повторные эксперименты при одном и том же значении фактора и использовать полный факторный план.

Полным факторным планом называется план, в котором предусмотрены все комбинации значений факторов, то есть план, точками которого служат все узлы некоторой декартовой сетки координат. Например, при $m = 2$ (рис. 6.3):

$$x_1 = (a_1, b_1),\ x_2 = (a_2, b_1),\ x_3 = (a_3, b_1),\ x_4 = (a_4, b_1),$$

$$x_5 = (a_1, b_2),\ x_6 = (a_2, b_2),\ x_7 = (a_3, b_2),\ x_8 = (a_4, b_2).$$

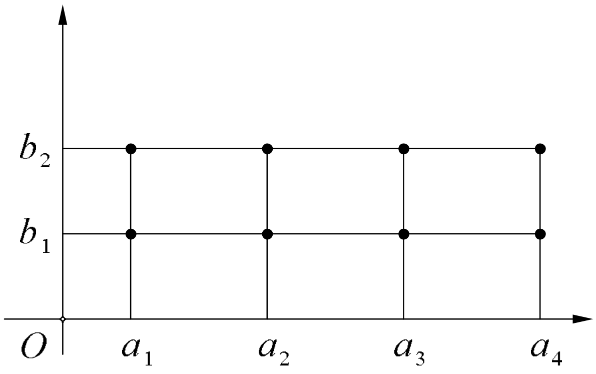


Рис. 6.3.

Или при $m = 3$ (рис. 6.4.):

$$x_1 = (a_1, b_1, c_1),\ x_2 = (a_1, b_1, c_2),\ x_3 = (a_1, b_2, c_1),\ x_4 = (a_1, b_2, c_2),$$

$$x_5 = (a_1, b_3, c_1),\ x_6 = (a_1, b_3, c_2),\ x_7 = (a_2, b_1, c_1),\ x_8 = (a_2, b_1, c_2),$$

$$x_9 = (a_2, b_2, c_1),\ x_{10} = (a_2, b_2, c_2),\ x_{11} = (a_2, b_3, c_1),\ x_{12} = (a_2, b_3, c_2).$$

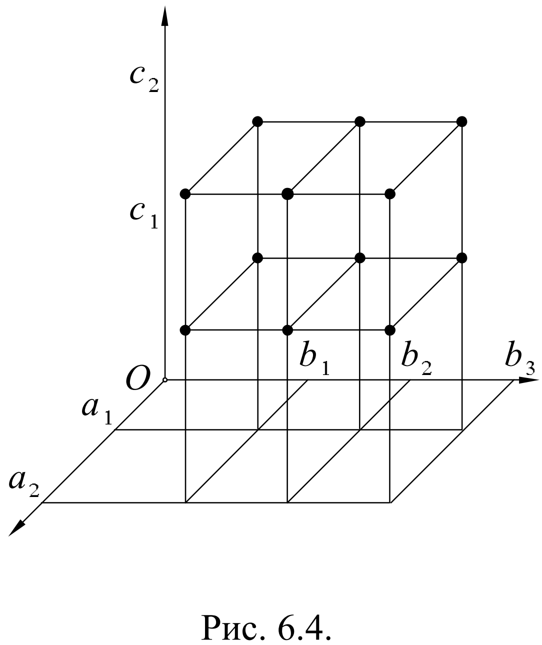


Рис. 6.4.

6.9 Проверка адекватности модели при повторных экспериментах

При возможности планировать эксперимент желательно в каждой точке x_i $(i = \overline{1, n})$ проводить по несколько измерений. Проще всего в каждой точке повторять эксперимент одно и тоже число раз, например, m. Допустим, что в результате таких mn независимых экспериментов получились значения отклика y_{ij}, где $i=\overline{1,n}$, $j=\overline{1,m}$, причем $M(y_{ij})=\beta_i$, где β_i - истинное значение отклика в точке x_i и $D(y_{ij})=\sigma^2$. Для каждого $i=\overline{1,n}$ найдем $\overline{y_i}=\frac{1}{m}\sum_{j=1}^{m}y_{ij}$, причем $M(\overline{y_i})=\beta_i$ и $D(\overline{y_i})=\frac{1}{m^2}\sum_{j=1}^{m}D(y_{ij})=\frac{m\sigma^2}{m^2}=\frac{\sigma^2}{m}=\sigma_1^2$. По набору $\overline{y_i}$ $i=\overline{1,n}$ построим модель методом наименьших квадратов. Для проверки ее адекватности сравним две оценки неизвестной дисперсии σ^2. Первая из них связана с построенной моделью, а вторая нет.

1. Известно, что $s_{ост}^2=\frac{1}{n-k}ss_{ост}=\frac{1}{n-k}\sum_{i=1}^{n}(\overline{y_i}-\hat{y}_i)^2$ есть несмещенная оценка σ_1^2, а, следовательно, $m\cdot s_{ост}^2$ есть несмещенная оценка $\sigma^2=m\sigma_1^2$. Кроме того, известно, что $\frac{1}{\sigma_1^2}ss_{ост}$ имеет распределение Пирсона с $n-k$ степенями свободы.

2. Известно также, что для каждого $i=\overline{1,n}$ $s_i^2=\frac{1}{m-1}\sum_{j=1}^{m}(y_{ij}-\overline{y_i})^2$ есть несмещенная оценка дисперсии σ^2, причем $\frac{1}{\sigma^2}\sum_{j=1}^{m}(y_{ij}-\overline{y_i})^2$ распределена по Пирсону с $m-1$ степенью свободы. Обозначим через $s_{сред}^2$ среднее арифметическое величин s_i^2: $s_{сред}^2=\frac{1}{n}\sum_{i=1}^{n}s_i^2=\frac{1}{n(m-1)}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\overline{y_i})^2$. Тогда $s_{сред}^2$ также есть несмещенная оценка σ^2, причем $\frac{1}{\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij}-\overline{y_i})^2$ распределена по Пирсону с $n(m-1)$ степенями свободы (так как все y_{ij} независимы друг от друга).

Если $ms_{ост}^2\le s_{сред}^2$, то гипотеза об адекватности не отвергается. Если же $ms_{ост}^2>s_{сред}^2$, то находим отношение

$$F=\frac{ms_{ост}^2}{s_{сред}^2}=\frac{m\dfrac{1}{n-k}\sigma_1^2\chi^2(n-k)}{\dfrac{1}{n(m-1)}\sigma^2\chi^2(n(m-1))}=\frac{\dfrac{1}{n-k}\chi^2(n-k)}{\dfrac{1}{n(m-1)}\chi^2(n(m-1))}$$

В нашем случае $F>1$. Однако, если построенная модель адекватна, то F не может быть слишком большим, так как в числителе и знаменателе стоят оценки одной и той же дисперсии σ^2. Случайная величина F имеет распределение Фишера с $n-k$ и $n(m-1)$ степенями свободы.

Зададим уровень значимости α. Если найденное из экспериментальных данных значение F^* будет больше найденного по таблицам $F(\alpha,n-k,n(m-1))$, то гипотеза об адекватности отвергается на уровне значимости α. Если же $F^*\le F(\alpha,n-k,n(m-1))$, то считают, что гипотеза об адекватности модели не противоречит результатам эксперимента.

ГЛАВА 7. ЗАДАЧИ

7.1 Непосредственное вычисление математических ожиданий. Классическое определение вероятности.

Задачи для разбора у доски

1.1. Выборочное пространство состоит из пяти исходов, их вероятности даны во втором столбце таблицы 7.1; в третьем и четвертом столбцах указаны значения величин X и Y. Вычислите математические ожидания $M(X)$, $M(Y)$, $M(X+Y)$, $M(XY)$ и вероятности событий $A=(X>2)$, $B=(|Y-5|<3)$.

Таблица 7.1 – условие к задаче 1.1.

ω	p	X	Y
ω_1	0,1	3	1
ω_2	0,2	-1	3
ω_3	0,4	2	5
ω_4	0,1	5	7
ω_5	0,2	0	9
Σ	1	–	–

1.2. На десяти карточках записаны цифры 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Карточки перемешивают и одну за другой выбирают две карточки, выкладывая двузначное число от 01 до 98. Найдите вероятность того, что это двузначное число делится на 18.

1.3. В некоторой партии из 100 изделий имеется 2 дефектных. Случайным образом отбирается 20 изделий. Какова вероятность того, что оба дефектных изделия попадут в выборку?

Задачи для самостоятельного решения

1.4. В урне a белых и b черных шаров. Из урны вынимаются два шара. Найдите вероятность того, что оба шара будут белыми.

1.5. Абонент забыл две последние цифры телефонного номера, но, помня, что они различны, набрал их наугад. Какова вероятность того, что он набрал нужный номер?

1.6. Имеется колода в 36 карт. Наугад берется 5 карт. Какова вероятность того, что среди них будет а) ровно два туза; б) хотя бы один туз?

1.7. Бросается 23 монеты. Определить вероятность выпадения орла на каждой из монет.

7.2 Правила сложения и умножения вероятностей. Условные вероятности.

Задачи для разбора у доски

2.1. В урне 10 шаров, из них 3 белых. Вынимают 3 шара. Найдите таблицу распределения числа X вынутых белых шаров при а) повторной выборке; б) бесповторной выборке.

2.2. Вероятность попадания в цель из первого орудия равна 0,9; из второго 0,8; из третьего 0,7. Найдите вероятность того, что при одновременном залпе из трех орудий будет иметь место а) ровно одно поражение цели; б) хотя бы одно поражение цели.

2.3. На рисунке 7.1 указан порядок соединения элементов системы и для каждого из них дана надежность (вероятность безотказной работы за определенный промежуток времени). Рассчитайте надежность системы при условии независимости выхода элементов из строя.

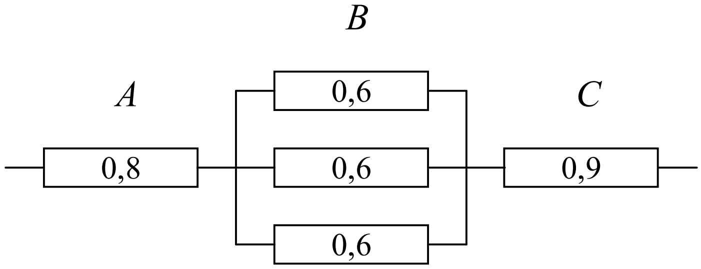


Рис. 7.1.

Задачи для самостоятельного решения

2.4. Имеется колода в 36 карт. Достоинства карт определяются таблицей 7.2. Игроку в «Двадцать одно» сдается две карты. Какова вероятность того, что он сразу получит: а) 21 очко («очко»); б) 22 очка («Москву»); в) либо «очко», либо «Москву»?

Таблица 7.2 – достоинства карт к задаче 2.4.

Карта	6	7	8	9	10	В	Д	К	Т
Дост-во	6	7	8	9	10	2	3	4	11

2.5. Вероятность попадания стрелка в цель равна 0,4. Стрелок делает 4 выстрела. Какова вероятность того, что будет иметь место ровно одно поражение цели?

2.6. Система содержит три последовательно соединенных элемента, причем надежность каждого равна *p*. Имеется еще три резервных элемента той же надежности. Из предложенных на рисунке 7.2 схем выберете схему дублирования с большей надежностью.

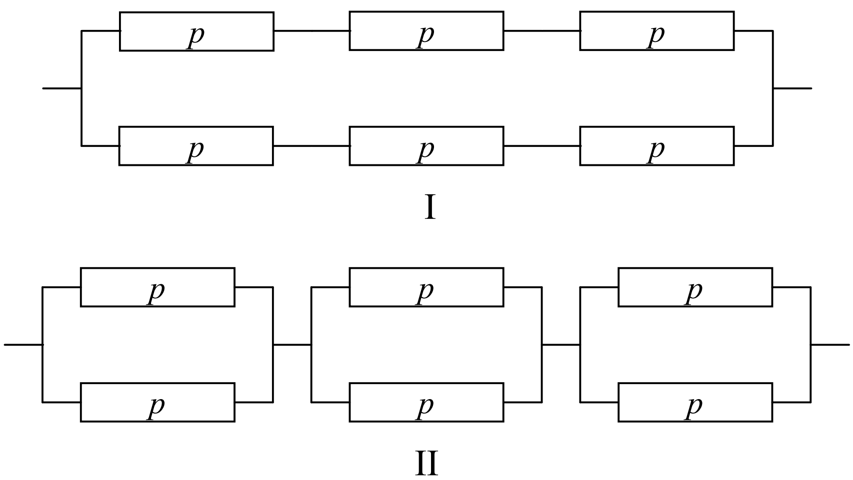


Рис. 7.2.

7.3 Формула полной вероятности и формула Байеса

Задачи для разбора у доски

3.1. Случайно оказались смешанными две партии изделий, причем известно, что число изделий в первой партии втрое больше, чем во второй, а дефектные изделия составляют 3% в первой партии и 2% –

во второй. Какова вероятность того, что взятое наугад из смеси изделие будет дефектным?

3.2. В условиях предыдущей задачи взятое изделие подвергли проверке и оно оказалось дефектным. Какова вероятность того, что изделие было взято из первой партии?

3.3. На письменном вступительном экзамене по математике шесть абитуриентов усаживают за стол в случайном порядке. Среди этих абитуриентов двое окончили одну и ту же школу. Какова вероятность того, что они будут сидеть рядом?

3.4. В первой урне лежит 10 белых шаров и 6 черных, во второй 5 белых и 2 черных. Из первой урны во вторую перекладывают 2 шара; затем, после перемешивания, из второй урны вынимают 2 шара. Найдите распределение числа X вынутых белых шаров и математическое ожидание величины X.

3.5. В группе 25 студентов, из них 5 отличников (отвечают на 5), 10 хорошистов (отвечают на 4 или на 5 с равной вероятностью) и 10 посредственных (отвечают на 3, 4 или 2 с равной вероятностью). Найдите вероятность того, что опрошенные два студента получат в сумме точно 8 баллов.

Задачи для самостоятельного решения

3.6. В магазин поступают 20% телевизоров с завода A, 30% с завода B и 50% с завода C. Вероятности того, что телевизоры, изготовленные на заводах A, B и C, сломаются во время гарантийного срока, равны соответственно 0,01, 0,08 и 0,1. Найдите вероятность того, что купленный телевизор сломается в период гарантии.

3.7. В первой урне лежат 20 белых и 5 черных шаров, во второй 9 белых и 5 черных. Из первой урны во вторую переложили, не глядя 1 шар; затем, после перемешивания, из второй урны вынули 1 шар, который оказался белым. Найдите вероятность того, что из первой урны во вторую был переложен черный шар.

7.4 Числовые характеристики случайных величин и их систем

Задачи для разбора у доски

4.1. В урне 10 шаров, из них 3 белых. Вынимают 3 шара а) по схеме повторной выборки; б) по схеме бесповторной выборки. Найдите математическое ожидание, дисперсию и стандартное отклонение числа X вынутых белых шаров.

4.2. Два стрелка независимо друг от друга делают по одному выстрелу. Вероятности их попадания в цель равны соответственно 0,8 и 0,6. X – число попаданий в цель. Вычислите $M(X)$, $D(X)$ и $\sigma(X)$.

4.3. Эксперимент имеет 5 исходов, вероятности которых указаны в таблице 7.3. Там же приведены значения случайных величин X, Y, Z. Рассчитайте для них матрицу ковариаций.

Таблица 7.3 – условие к задача 4.3.

ω	p	X	Y	Z
ω_1	0,1	3	1	5
ω_2	0,2	-1	3	4
ω_3	0,4	2	5	3
ω_4	0,1	5	7	2
ω_5	0,2	0	9	1
Σ	1	–	–	–

4.4. Для трех величин X, Y, Z известна матрица ковариаций K. Вычислите $D(X+2Y+3Z)$, если $K=\begin{pmatrix} 1 & 2 & -3 \\ 2 & 5 & -4 \\ -3 & -4 & 15 \end{pmatrix}$.

Задачи для самостоятельного решения

4.5. Стрелок делает 3 независимых выстрела по цели, вероятность его попадания при каждом выстреле равна 0,6. X – число попаданий.

Найдите таблицу распределения величины X и вычислите $M(X)$, $D(X)$ и $\sigma(X)$.

4.6. Эксперимент имеет 6 исходов, вероятности которых указаны в таблице 7.4. В этой же таблице приведены значения случайных величин X, Y, Z. Найдите значение λ, центр распределения величин, их матрицу ковариаций, вычислите $D(X+2Y-3Z-4)$.

Таблица 7.4 – условие к задаче 4.6.

ω	p	X	Y	Z
ω_1	0,1	1	3	5
ω_2	0,15	0	1	8
ω_3	0,3	– 2	0	7
ω_4	0,1	1	4	0
ω_5	0,1	2	3	4
ω_6	λ	1	2	5

7.5 Непрерывные случайные величины

Задачи для разбора у доски

5.1. Плотность распределения величины X задана на рисунке 7.3. Вычислите $\sigma(X)$ и $p(|X|>\sigma(X))$.

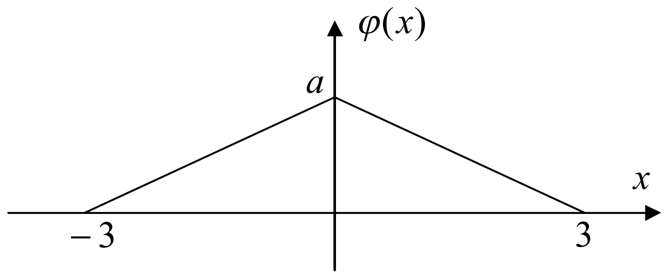


Рис. 7.3.

5.2. Задана плотность распределения величины X; а) постройте график плотности вероятности; б) проверьте, выполняется ли условие нормировки; в) найдите функцию распределения и постройте ее график; г) найдите $M(X)$, $D(X)$ и $\sigma(X)$; д) найдите вероятность попадания случайной величины в интервал $(-1, \ln 2)$ с помощью плотности вероятности и с помощью функции распределения; е) найдите медиану и квантиль для вероятности $1-e^{-1}$.

$$\varphi(x)=\begin{cases} 0, & x<0 \\ 3e^{-3x}, & x\ge 0 \end{cases}.$$

5.3. Система случайных величин *X*, *Y* распределена равномерно в квадрате $0 \le X \le 1$, $0 \le Y \le 1$, величина *T* есть модуль их разности $T = |X-Y|$. Вычислите $M(T)$ и вероятность $p(T < M(T))$.

Задачи для самостоятельного решения

5.4. Плотность распределения величины *X* задана на рисунке 7.4. Вычислите $M(X)$ и $p(X > M(X))$.

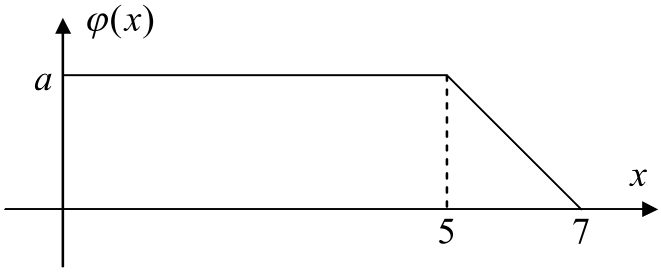


Рис. 7.4.

5.5. Задана плотность распределения величины *X*; а) постройте график плотности вероятности; б) проверьте, выполняется ли условие нормировки; в) найдите функцию распределения и постройте ее график; г) найдите $M(X)$, $D(X)$ и $\sigma(X)$; д) найдите вероятность попадания случайной величины в интервал (1;2) с помощью плотности вероятности и с помощью функции распределения; е) найдите медиану и квантиль для вероятности 0,2.

$$\varphi(x)=\begin{cases} 0, & x<1 \\ \dfrac{4}{x^5}, & x\ge 1 \end{cases}.$$

7.6 Непрерывные случайные величины. Разбор нулевого варианта РГР №1

Случайная величина X в промежутке $(q_1;q_2)$ распределена с постоянной плотностью C; с вероятностью R попадает в промежуток $(z_1;z_2)$, и имеет там плотность распределения $\varphi(x)$ вида $\varphi(x) = A\,|\,x - z_3\,|$; значения некоторых параметров указаны в условии. Требуется:

6.1. найти недостающие значения параметров;

6.2. найти плотность распределения и функцию распределения случайной величины X и построить их графики;

6.3. вычислить $M(X)$, $D(X)$, $\sigma(X)$;

6.4. найти $p(|\,X - M(X)\,| < \sigma(X))$;

6.5. найти медиану случайной величины X.

Результаты занести в таблицу 7.5.

Таблица 7.5 – шаблон для ответов к РГР №1.

Вар-т	q_1	q_2	z_1	z_2	z_3	R	C	A	$M(X)$	$D(X)$	$\sigma(X)$	p	Мед.
0	–3		–1	0	1	0,4	0,6						

7.7 Сравнение двух выборок. Разбор нулевого варианта РГР №2

Даны две серии измерений некоторой случайной величины.

7.1. Найдите средние арифметические $\overline{y_1}$, $\overline{y_2}$ и эмпирические стандарты s_1, s_2 для каждой из выборок.

7.2. Постройте доверительные интервалы для β_1, β_2, σ_1, σ_2. (Доверительная вероятность 0,95).

7.3. Проверьте гипотезу о равенстве дисперсий или стандартных отклонений по критерию Фишера. (Уровень значимости $\alpha = 0{,}05$).

7.4. Если $\sigma_1 = \sigma_2$, то найдите $s_{св}$, постройте доверительный интервал для σ. (Доверительная вероятность 0,95).

7.5. Проверьте гипотезу о равенстве математических ожиданий по критерию Стьюдента. (Уровень значимости $\alpha = 0{,}05$).

7.6. Если $\beta_1 = \beta_2$, но найдите $\overline{y_{св}}$, постройте доверительный интервал для β, найдите $s_{об}$.

7.7. Объедините две выборки в одну и проверьте гипотезу о том, что экспериментальные данные имеют нормальный закон распределения. Рассматривать интервалы равной вероятности. Число интервалов равно L.

7.8. Постройте гистограмму.

1 серия измерений. $n_1 = 26$.

0,44 1,29 1,25 1,06 1,24 0,87 0,96 1,25 0,88
1,09 0,90 0,75 1,09 0,73 1,04 1,07 1,20 1,15
0,94 0,83 1,19 0,88 0,77 0,84 0,79 0,80

2 серия измерений. $n_2 = 32$.

0,89 0,97 0,33 0,93 0,84 1,43 1,34 0,88 0,75
0,96 1,09 0,83 0,95 0,33 0,56 1,20 1,12 0,92
0,73 1,30 0,70 1,27 0,82 0,86 1,30 1,00 1,12
0,69 1,03 0,58 1,26 1,16

L=8.

7.8 Построение регрессионных моделей

Задачи для разбора у доски

8.1. Постройте линейную модель регрессии и изобразите на одном графике экспериментальные данные и найденную модель.

X	Y	W
2	1	3
3	2	1
6	3	1

8.2. Постройте квадратичную модель регрессии. Непосредственным подсчетом сумм квадратов отклонений убедитесь, что найденная модель лучше модели $y = x^2$.

X	Y	W
0	0	1
1	1/2	1
2	9/2	1
3	9	1

Задачи для самостоятельного решения

8.3. Постройте линейные модели регрессии.

а)

X	Y	W
1	1	2
2	2	2
5	3	1

б)

X	Y	W
0	0	4
4	2	1
8	6	2

в)

X	Y	W
1	10	4
5	5	2
9	0	2

г)

X	Y	W
0	5	1
2	2	2
4	1	5

7.9 Построение регрессионных моделей и проверка их адекватности. Разбор нулевого варианта РГР №3

9.1. Для каждого значения фактора X было проведено несколько измерений величины Y (таблица 7.6).

а) Постройте линейную модель регрессии.

б) Проверьте ее адекватность.

в) Постройте квадратичную модель регрессии.

г) Проверьте ее адекватность.

Таблица 7.6 – условие к задаче 9.1.

X	*Y*				
10	50,7	50,6	50,6		
20	49,9	50,0	50,1	50,2	
40	48,3	48,4	48,4		
60	47,4	47,4	46,8	47,4	
80	46,0	45,9	46,2	46,1	45,5
120	43,2	43,3	43,1	43,5	43,0
140	41,9	42,2	41,9	42,2	
150	41,7	41,7	42,0		
180	40,2	40,4	40,5	40,1	
200	39,5	39,9	39,7		

7.10 Разные задачи

Теория вероятностей

10.1. Производится бомбометание по трем складам боеприпасов, причем сбрасывается одна бомба. Вероятность попадания в первый склад 0,01; во второй 0,008; в третий 0,025. При попадании в один из складов взрываются все три. Найти вероятность того, что склады будут взорваны.

10.2. Детали поступают в ОТК с двух конвейеров, производительность второго конвейера в два раза больше производительности первого. В каждой сотне деталей с первого конвейера в среднем две бракованные, а в каждой сотне со второго конвейера 3 бракованные. Найти вероятность того, что наудачу взятая деталь в ОТК будет бракованной.

10.3. По самолету производится три одиночных выстрела. Вероятность попадания при первом выстреле равна 0,4; при втором 0,5; при третьем 0,7. Для вывода самолета из строя заведомо достаточно трех попаданий. При одном попадании самолет выходит из строя с вероятностью 0,2; при двух попаданиях с вероятностью 0,6. Найти вероятность того, что в результате трех выстрелов самолет будет выведен из строя.

10.4. В условиях предыдущей задачи оказалось, что самолет был выведен из строя. Найдите вероятность того, что имело место одно попадание.

10.5. Эксперимент имеет 6 исходов ω_i вероятности которых указаны в таблице 7.7; в этой же таблице приведены значения случайной величины X. Найдите λ, $M(X)$, $D(X)$, $\sigma(X)$, $p(A)$, где $A = (|X - M(X)| < \sigma(X))$.

Таблица 7.7 – условие к задаче 10.5.

ω	p	X
1	0,3	–2
2	0,1	–1
3	0,2	0
4	0,1	2
5	0,15	4
6	λ	6

10.6. На десяти карточках написаны цифры 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. Три карточки выбирают случайным образом и раскладывают в порядке выбора. Найдите вероятность того, что полученное при этом число (от 012 до 987) делится на 36.

10.7. Подбрасывают три монеты, X – число выпадающих гербов. Найти таблицу распределения величины X и вычислить $p(X \geq 1)$.

10.8. В урне лежит 5 шаров, из них 3 белых. Из урны вынимают 3 шара без возвращения, X – число вынутых белых шаров. Найдите таблицу распределения вероятностей величины X и вычислите $M(X)$, $D(X)$, $\sigma(X)$.

10.9. В урне лежит 5 шаров, из них два белых. Из урны вынимают 3 шара по схеме возвращенного шара, X – число вынутых белых шаров. Найдите таблицу распределения вероятностей величины X и вычислите $M(X)$, $D(X)$, $\sigma(X)$.

10.10. В системе имеется элемент с надежностью $p = 0{,}5$. Имеются запасные элементы той же надежности. Сколько раз надо дублировать указанный элемент (параллельно), чтобы повысить надежность построенного блока до 0,9?

10.11. В первой урне 20 шаров, из них 5 белых, во второй – 18 шаров, из них 8 белых. Из первой урны во вторую перекладывают два шара и, после перемешивания, вынимают из второй урны 1 шар. Найти вероятность того, что он – белый.

10.12. В предыдущей задаче оказалось, что вынутый шар из второй урны оказался белым. Найдите вероятность того, что во вторую урну были переложены шары разного цвета.

Непрерывное распределение

10.13. Плотность распределения задается формулой $\varphi(x) = \dfrac{1}{\pi(1+x^2)}$. Постройте график плотности, найдите вероятность того, что величина X попадет на участок $(-1;1)$.

10.14. Задана плотность распределения величины X; а) постройте график плотности вероятности; б) проверьте, выполняется ли условие нормировки; в) найдите функцию распределения и постройте ее график; г) найдите $M(X)$, $D(X)$ и $\sigma(X)$; д) найдите вероятность попадания случайной величины в интервал (1;2) с помощью плотности вероятности и с помощью функции распределения; е) найдите медиану и квантиль для вероятности 0,2.

$$\varphi(x) = \begin{cases} 0, & x < -1 \\ \dfrac{2}{7}(2-|x|), & -1 \le x \le 2 \\ 0, & x > 2 \end{cases}.$$

10.15. Плотность распределения величины *X* задана на рисунке *6*. Найдите *a* и вычислите $\sigma(X)$ и $p(X > \sigma(X))$.

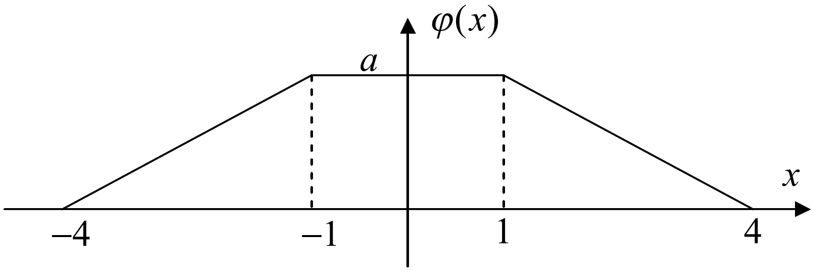


Рис. 7.5.

10.16. Задана плотность распределения величины *X*. Найдите *C*, $M(X)$, $\sigma(X)$, $p(X \geq 2\sigma(X))$.

$$\varphi(x) = \begin{cases} 0, & x < 0 \\ Cxe^{-5x}, & x \geq 0 \end{cases}.$$

Сравнение двух выборок и критерий согласия Пирсона

10.17. Выполните задания 1-7 из параграфа 7.7 для следующих выборок:

а) 1 серия: 1; 2; 3; 2; 2. 2 серия: 1; 3; 2; 4.

Число интервалов равной вероятности $L = 5$.

б) 1 серия: 4; 6; 8; 4. 2 серия: 3; 5; 7.

Число интервалов равной вероятности $L = 7$.

10.18. Выполните задания 7-8 из параграфа 7.7, но взяв интервалы равной длины.

Регрессионный анализ

10.19. Постройте линейную модель регрессии.

X	*Y*	*W*
0	0	1
1	2	2
4	3	2
5	6	1

10.19. Проверьте значимость коэффициентов линейной и квадратичной регрессионных моделей в задаче 9.1.

7.11. ПРИЛОЖЕНИЯ

1. Квантили распределения абсолютной величины отношения Стьюдента $|t|_{0,95}(k)$, коэффициенты $z_н$ и $z_в$, определяющие нижнюю и верхнюю границы доверительного интервала для среднего квадратического отклонения ($p = 0,95$), квантили $\chi^2_{0,95}(k)$ распределения Пирсона.

k	$\|t\|_{0,95}(k)$	$z_н$	$z_в$	$\chi^2_{0,95}(k)$
2	4,303	0,521	6,28	5,99
3	3,182	0,566	3,73	7,81
4	2,776	0,599	2,87	9,49
5	2,571	0,624	2,45	11,1
6	2,447	0,644	2,20	12,6
7	2,365	0,661	2,04	14,1
8	2,306	0,675	1,92	15,5
9	2,262	0,688	1,83	16,9
10	2,228	0,699	1,75	18,3
12	2,179	0,717	1,65	21,0
14	2,145	0,732	1,58	23,7
16	2,120	0,745	1,52	26,3
18	2,101	0,756	1,48	28,9
20	2,086	0,765	1,44	31,4
22	2,074	0,773	1,41	33,9
24	2,064	0,780	1,39	36,4
26	2,056	0,788	1,37	38,9
28	2,048	0,793	1,35	41,4
30	2,042	0,799	1,34	43,8
35	2,030	0,811	1,30	49,8
40	2,021	0,821	1,28	55,8
60	2,000	0,846	1,22	79,1

2. Квантили F-распределения Фишера, $p = 0{,}975$.

	3	4	5	6	8	10	12	15	20	24	30	40
3	15,44	15,10	14,88	14,73	14,54	14,42	14,34	14,22	14,17	14,12	14,08	14,04
4	9,98	9,60	9,36	9,20	8,98	8,84	8,75	8,66	8,56	8,51	8,46	8,41
5	7,76	7,39	7,15	6,98	6,76	6,62	6,5?	6,43	6,33	6,28	6,23	6,18
6	6,60	6,23	5,99	5,82	5,60	5,46	5,37	5,27	5,17	5,12	5,07	5,01
7	5,89	5,52	5,29	5,12	4,90	4,76	4,67	4,57	4,47	4,42	4,36	4,31
8	5,42	5,05	4,82	4,65	4,43	4,30	4,20	4,10	4,00	3,95	3,89	3,84
9	5,08	4,72	4,48	4,32	4,10	3,96	3,87	3,77	3,67	3,61	3,56	3,51
10	4,83	4,47	4,24	4,07	3,85	3,72	3,62	3,52	3,42	3,37	3,31	3,26
11	4,63	4,28	4,04	3,88	3,66	3,53	3,43	3,33	3,23	3,17	3,12	3,06
12	4,47	4,12	3,89	3,73	3,51	3,37	3,28	3,18	3,07	3,02	2,96	2,91
13	4,35	4,00	3,77	3,60	3,39	3,25	3,15	3,05	2,95	2,89	2,84	2,78
14	4,24	3,89	3,66	3,50	3,29	3,15	3,05	2,95	2,84	2,79	2,73	2,67
15	4,15	3,80	3,58	3,41	3,20	3,06	2,96	2,86	2,76	2,70	2,64	2,59
16	4,08	3,73	3,50	3,34	3,12	2,99	2,89	2,79	2,68	2,63	2,57	2,51
18	3,95	3,61	3,38	3,22	3,01	2,87	2,77	2,67	2,56	2,50	2,44	2,38
20	3,86	3,51	3,29	3,13	2,91	2,77	2,68	2,57	2,46	2,41	2,35	2,29
22	3,78	3,44	3,22	3,05	2,84	2,70	2,60	2,50	2,39	2,33	2,27	2,21
24	3,72	3,38	3,15	2,99	2,78	2,64	2,54	2,44	2,33	2,27	2,21	2,15
26	3,67	3,33	3,10	2,94	2,73	2,59	2,49	2,39	2,28	2,22	2,16	2,09
28	3,63	3,29	3,06	2,90	2,69	2,55	2,45	2,34	2,23	2,17	2,11	2,05
30	3,59	3,25	3,03	2,87	2,65	2,51	2,41	2,31	2,20	2,14	2,07	2,01
40	3,46	3,13	2,90	2,74	2,53	2,39	2,29	2,18	2,07	2,01	1,94	1,88
60	3,34	3,01	2,79	2,63	2,41	2,27	2,17	2,06	1,94	1,88	1,82	1,74

Первая строка – k_1, первый столбец – k_2.

3. Квантили F-распределения Фишера, $p = 0{,}95$.

	3	4	5	6	8	10	12	15	20	24	30	40
3	9,28	9,12	9,01	8,94	8,85	8,79	8,74	8,70	8,66	8,64	8,62	8,59
4	6,59	6,39	6,26	6,16	6,04	5,96	5,91	5,86	5,80	5,77	5,75	5,72
5	5,41	5,19	5,05	4,95	4,82	4,74	4,68	4,62	4,56	4,53	4,50	4,46
6	4,76	4,53	4,39	4,28	4,15	4,06	4,00	3,94	3,87	3,84	3,81	3,77
7	4,35	4,12	3,97	3,87	3,73	3,64	3,57	3,51	3,44	3,41	3,38	3,34
8	4,07	3,84	3,69	3,58	3,44	3,35	3,28	3,22	3,15	3,12	3,08	3,04
9	3,86	3,63	3,48	3,37	3,23	3,14	3,07	3,01	2,94	2,90	2,86	2,83
10	3,71	3,48	3,33	3,22	3,07	2,98	2,91	2,85	2,77	2,74	2,70	2,66
11	3,59	3,36	3,20	3,09	2,95	2,85	2,79	2,72	2,65	2,61	2,57	2,53
12	3,49	3,26	3,11	3,00	2,85	2,75	2,69	2,62	2,54	2,51	2,47	2,43
13	3,41	3,18	3,03	2,92	2,77	2,67	2,60	2,53	2,46	2,42	2,38	2,34
14	3,34	3,11	2,96	2,85	2,70	2,60	2,53	2,46	2,39	2,35	2,31	2,27
15	3,29	3,06	2,90	2,79	2,64	2,54	2,48	2,40	2,33	2,29	2,25	2,20
16	3,24	3,01	2,85	2,74	2,59	2,49	2,42	2,35	2,28	2,24	2,19	2,15
18	3,16	2,93	2,77	2,66	2,51	2,41	2,34	2,27	2,19	2,15	2,11	2,06
20	3,10	2,87	2,71	2,60	2,45	2,35	2,28	2,20	2,12	2,08	2,04	1,99
22	3,05	2,82	2,66	2,55	2,40	2,30	2,23	2,15	2,07	2,03	1,98	1,94
24	3,01	2,78	2,62	2,51	2,36	2,25	2,18	2,11	2,03	1,98	1,94	1,89
26	2,98	2,74	2,59	2,47	2,32	2,22	2,15	2,07	1,99	1,95	1,90	1,85
28	2,95	2,71	2,56	2,45	2,29	2,19	2,12	2,04	1,96	1,91	1,87	1,82
30	2,92	2,69	2,53	2,42	2,27	2,16	2,09	2,01	1,93	1,89	1,84	1,79
40	2,84	2,61	2,45	2,34	2,18	2,08	2,00	1,92	1,84	1,79	1,74	1,69
60	2,76	2,53	2,37	2,25	2,10	1,99	1,92	1,84	1,75	1,70	1,65	1,59

Первая строка – k_1, первый столбец – k_2.

4. Нормированная функция Лапласа $\Phi_0(z) = \frac{1}{\sqrt{2\pi}} \int_0^z e^{-\frac{x^2}{2}} dx$.

z	Сотые доли z									
	0	1	2	3	4	5	6	7	8	9
0,0	0,0 000	040	080	120	160	199	239	279	319	359
0,1	398	438	478	517	557	596	636	675	714	753
0,2	793	832	871	910	948	987	026	064	103	141
0,3	0,1 179	217	255	293	331	368	406	443	480	517
0,4	554	591	628	664	700	736	772	808	844	879
0,5	915	950	985	019	054	088	123	157	190	224
0,6	0,2 257	291	324	357	389	422	454	486	517	549
0,7	580	611	642	673	703	734	764	794	823	852
0,8	881	910	939	967	995	023	051	078	106	133
0,9	0,3 159	186	212	238	261	289	315	340	365	389
1,0	413	437	461	485	508	533	554	577	599	621
1,1	643	665	686	708	729	749	770	790	810	830
1,2	849	869	888	907	925	944	962	980	997	015
1,3	0,4 032	049	066	082	099	115	131	147	162	177
1,4	192	207	222	236	251	265	279	292	306	319
1,5	332	345	357	370	382	394	406	418	429	441
1,6	452	463	474	484	495	505	515	525	535	545
1,7	554	564	573	582	591	599	608	616	625	633
1,8	641	649	656	664	671	678	686	693	699	706
1,9	713	719	726	732	738	744	750	756	761	767
2,0	772	778	783	788	793	798	803	808	812	817
2,1	821	826	830	834	838	842	846	850	854	857
2,2	860	864	867	871	874	877	880	883	886	889
2,3	892	895	898	900	903	906	908	911	913	915
2,4	918	920	922	924	926	928	930	932	934	936
2,5	937	939	941	942	944	946	947	949	950	952
2,6	953	954	956	957	958	959	960	962	963	964
2,7	965	966	967	968	969	970	971	971	972	973
2,8	974	975	975	976	977	978	978	979	980	980
2,9	981	981	982	983	983	984	984	985	985	986
3,0	986	986	987	987	988	988	988	989	989	989

5. Плотности вероятностей «важных» распределений.

$\Gamma(p)=\int\limits_0^{\infty} e^{-z} z^{p-1} dz$ $p \neq 0,-1,-2,\ldots$	Гамма-функция
$\beta(a,b)=\int\limits_0^{1} x^{a-1}(1-x)^{b-1} dx$ $a,b>0$	Бета-функция
$\varphi_Y(x)=\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\alpha)^2}{2\sigma^2}}$ $x \in R$	Гауссово или нормальное
$\varphi_V(x)=\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$ $k=1,2,\ldots;\ x \geq 0$	Пирсона
$\varphi_T(x)=\frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi n}\cdot\Gamma\left(\frac{k}{2}\right)\left(1+\frac{x^2}{k}\right)^{\frac{k+1}{2}}}$ $k=1,2,\ldots;\ x \in R$	Стьюдента
$\varphi_F(x)=\frac{\left(\frac{k_1}{k_2}\right)^{\frac{k_1}{2}}}{\beta\left(\frac{k_1}{2},\frac{k_2}{2}\right)}\cdot\frac{x^{\frac{k_1}{2}-1}}{\left(1+\frac{k_1 x}{k_2}\right)^{\frac{k_1+k_2}{2}}}$ $k_1,k_2=1,2,\ldots;\ x \geq 0$	Фишера

7.12. ОТВЕТЫ

Занятие 7.1

1) $M(X)=1{,}4$; $M(Y)=5{,}2$; $M(X+Y)=6{,}6$; $M(XY)=7{,}2$; $p(A)=0{,}2$; $p(B)=0{,}7$. 2) $\frac{1}{18}$. 3) $\frac{19}{495}$. 4) $\frac{a(a-1)}{(a+b)(a+b-1)}$. 5) $\frac{1}{90}$. 6) а) $p_1=\frac{310}{3927}\approx 0{,}079$; б) $p_2=\frac{784}{1683}\approx 0{,}466$. 7) $\frac{1}{2^{23}}$.

Занятие 7.2

1) а)

X	p
0	0,343
1	0,441
2	0,189
3	0,027

б)

X	p
0	≈0,292
1	0,525
2	0,175
3	≈0,008

2) а) $p_1=0{,}092$; б) $p_2=0{,}994$. 3) $P=0{,}67392$. 4) а) $p_1=\frac{8}{315}\approx 0{,}025$; б) $p_2=\frac{1}{105}\approx 0{,}010$; в) $p_3\approx 0{,}035$. 5) $p=0{,}3456$. 6) $P_1=p^3(2-p^3)$; $P_2=p^3(2-p)^3$; $P_2\ge P_1$.

Занятие 7.3

1) $\frac{11}{400}$. 2) $\frac{9}{11}$. 3) $\frac{1}{3}$. 4) $M(X)=1{,}3889$.

X	p
0	0,0729
1	0,4653
2	0,4618

5) $p\approx 0{,}221$. 6) $p=0{,}076$. 7) $\frac{9}{49}$.

Занятие 7.4

1) а) $M_1(X)=0{,}9$; $D_1(X)=0{,}63$; $\sigma_1(X)\approx 0{,}794$; б) $M_2(X)\approx 0{,}900$; $D_2(X)\approx 0{,}490$; $\sigma_2(X)\approx 0{,}700$. 2) $M(X)=1{,}4$; $D(X)=0{,}4$; $\sigma(X)\approx 0{,}632$. 3) $K=\begin{pmatrix} 3{,}24 & -0{,}08 & 0{,}04 \\ -0{,}08 & 5{,}96 & -2{,}98 \\ 0{,}04 & -2{,}98 & 1{,}49 \end{pmatrix}$. 4)98. 5) $M(X)=1{,}8$; $D(X)=0{,}72$; $\sigma(X)\approx 0{,}849$.

6) $\lambda=0{,}25$; $D=103{,}4$; $M\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}=\begin{pmatrix} 0{,}05 \\ 1{,}65 \\ 5{,}45 \end{pmatrix}$; $K=\begin{pmatrix} 2{,}05 & 1{,}72 & -1{,}92 \\ 1{,}72 & 1{,}83 & -2{,}59 \\ -1{,}92 & -2{,}59 & 4{,}95 \end{pmatrix}$.

Занятие 7.5

1) $\sigma(X)=\sqrt{1{,}5}$; $p\approx 0{,}350$. 2) г) $M(X)=\frac{1}{3}$; $D(X)=\frac{1}{9}$; $\sigma(X)=\frac{1}{3}$; д) $p=\frac{7}{8}$; е) $x_{1/2}=\frac{1}{3}\ln 2$; $x_p=\frac{1}{3}$. 3) $M(T)=\frac{1}{3}$; $p=\frac{5}{9}$. 4) $M(X)=\frac{109}{36}$; $p=\frac{107}{216}$. 5) г) $M(X)=\frac{4}{3}$; $D(X)=\frac{2}{9}$; $\sigma(X)=\frac{\sqrt{2}}{3}$; д) $p=\frac{15}{16}$; е) $x_{1/2}=\sqrt[4]{2}$; $x_p=\sqrt[4]{1{,}25}$.

Занятие 7.6

Вар-т	q_1	q_2	z_1	z_2	z_3	R	C	A	$M(X)$	$D(X)$	$\sigma(X)$	p	Мед.
0	–3	**–2**	–1	0	1	0,4	0,6	**4/15**	**–1,722**	**0,990**	**0,995**	**0,566**	**–13/6**

Занятие 7.7

1) $\overline{y_1}\approx 0{,}973$; $\overline{y_2}\approx 0{,}942$; $s_1\approx 0{,}207$; $s_2\approx 0{,}276$. 2) $0{,}163<\sigma_1<0{,}286$; $0{,}221<\sigma_2<0{,}367$; $0{,}889<\beta_1<1{,}057$; $0{,}843<\beta_2<1{,}041$. 3) $F_Э\approx 1{,}768$; $F_m=2{,}1785$. 4) $s_{св}\approx 0{,}247$; $0{,}208<\sigma<0{,}305$. 5) $T_Э\approx 0{,}478$; $|t|_p=2{,}0042$. 6) $\overline{y_{св}}\approx 0{,}956$; $0{,}891<\beta<1{,}021$; $s_{об}\approx 0{,}246$. 7) y_i: 0,67322; 0,79120; 0,87722; 0,95586; 1,03451; 1,12053; 1,23850; n_i: 5; 8; 8; 9; 5; 8; 5; 10; $\chi^2\approx 3{,}793$; $\chi_p^2=11{,}1$.

Занятие 7.8

1) $y=\frac{x}{2}+\frac{1}{10}$. 2) $y=x^2+0{,}1x-0{,}15$; $Q_1=0{,}45$; $Q_2=0{,}5$. 3) а) $y=\frac{13}{27}x+\frac{20}{27}$; б) $y=-\frac{2}{19}x+\frac{14}{19}$; в) $y=-\frac{5}{4}x+\frac{45}{4}$; г) $y=-\frac{7}{8}x+\frac{35}{8}$.

Занятие 7.9

1) а) $y=-0{,}060x+50{,}881$; б) $s_{ад}^2\approx 0{,}635$; $s_{св}^2\approx 0{,}039$; $F_Э\approx 16{,}123$; $F_т=2{,}29$; в) $y=0{,}000106x^2-0{,}082x+51{,}591$; г) $s_{ад}^2\approx 0{,}111$; $s_{св}^2\approx 0{,}039$; $F_Э\approx 2{,}809$; $F_т=2{,}37$.

Разные задачи

Теория вероятностей

1) 0,043. 2) $\frac{2}{75}$. 3) 0,458. 4) $\frac{9}{57}$. 5) $\lambda=0{,}15$; $M(X)=1$; $D(X)=8{,}5$; $\sigma(X)=2{,}8$; $p(A)=0{,}4$. 6) $\frac{11}{360}$. 7) $p=\frac{7}{8}$. 8) $M(X)=1{,}8$; $D(X)=0{,}36$; $\sigma(X)=0{,}6$. 9) $M(X)=1{,}2$; $D(X)=0{,}72$; $\sigma(X)\approx 0{,}849$. 10) 4. 11) $\frac{323}{760}=0{,}425$. 12) $\frac{135}{323}\approx 0{,}418$.

Непрерывное распределение

13) $p=0{,}5$. 14) г) $M(X)=\frac{4}{21}$; $D(X)=\frac{409}{882}\approx 0{,}464$; $\sigma(X)\approx 0{,}681$; д) $p=\frac{1}{7}$; е) $x_{1/2}=2-\frac{\sqrt{14}}{2}\approx 0{,}129$; $x_p=-2+\frac{2}{5}\sqrt{15}\approx -0{,}451$. 15) $a=\frac{1}{5}$; $\sigma(X)=\sqrt{\frac{17}{6}}\approx 1{,}683$; $p\approx 0{,}179$. 16) $C=25$; $M(X)=\frac{2}{5}$; $\sigma(X)=\frac{\sqrt{2}}{5}\approx 0{,}283$; $p=\mathrm{e}^{-2\sqrt{2}}(2\sqrt{2}+1)\approx 0{,}226$.

Сравнение двух выборок и критерий согласия Пирсона

17) а) $\overline{y_1}=2$; $\overline{y_2}=2{,}5$; $s_1=\frac{1}{\sqrt{2}}\approx 0{,}707$; $s_2=\sqrt{\frac{5}{3}}\approx 1{,}291$; $0{,}424<\sigma_1<2{,}029$; $0{,}731<\sigma_2<4{,}815$; $1{,}122<\beta_1<2{,}878$; $0{,}446<\beta_2<4{,}554$; $F_Э=\frac{10}{3}\approx 3{,}333$; $F_т=9{,}98$; $s_{св}=1$; $0{,}661<\sigma<2{,}040$; $T_Э=\frac{\sqrt{5}}{3}\approx 0{,}745$; $|t|_p=2{,}365$; $\overline{y_{св}}=\frac{20}{9}\approx 2{,}222$; $1{,}434<\beta<3{,}011$; $s_{об}=0{,}972$; y_i: 1,406; 1,979; 2,465; 3,039;

n_i: 2; 0; 4; 2; 1; $\chi^2 = \dfrac{44}{9} \approx 4{,}889$; $\chi_p^2 = 5{,}99$. б) $\overline{y_1} = 5{,}5$; $\overline{y_2} = 5$; $s_1 = \sqrt{\dfrac{11}{3}} \approx 1{,}915$; $s_2 = 2$; $1{,}084 < \sigma_1 < 7{,}142$; $1{,}042 < \sigma_2 < 12{,}560$; $2{,}453 < \beta_1 < 8{,}547$; $0{,}031 < \beta_2 < 9{,}969$; $F_Э = \dfrac{12}{11} \approx 1{,}091$; $s_{св} = \sqrt{\dfrac{19}{5}} \approx 1{,}949$; $1{,}216 < \sigma < 4{,}776$; $T_Э \approx 0{,}336$; $|t|_p = 2{,}571$; $\overline{y_{св}} = \dfrac{37}{7} \approx 5{,}286$; $3{,}391 < \beta < 7{,}180$; $s_{об} = 1{,}799$; y_i: 3,360; 4,260; 4,962; 5,610; 6,311; 7,211; n_i: 1; 2; 0; 1; 1; 1; 1; $\chi^2 = 2$; $\chi_p^2 = 9{,}49$. 18) n_i: 3; 2; 4; 15; 10; 10; 10; 4; p_i: 0,023; 0,053; 0,116; 0,186; 0,220; 0,192; 0,124; 0,085; $\chi^2 = 8{,}719$; $\chi_p^2 = 11{,}1$.

Регрессионный анализ

19) $y = 0{,}837x + 0{,}574$. 20) $s_{11} \approx 2{,}126 \cdot 10^{-3}$; $s_{12} \approx 0{,}249$; $s_{21} \approx 1{,}702 \cdot 10^{-3}$; $s_{22} \approx 3{,}582 \cdot 10^{-3}$; $s_{23} \approx 0{,}154$; $t_1 = 2{,}306$; $t_2 = 2{,}365$; все коэффициенты значимы.

ГЛАВА 8. РАСЧЕТНО-ГРАФИЧЕСКИЕ РАБОТЫ

РГР №1. Непрерывные случайные величины

Случайная величина X в промежутке $(q_1;q_2)$ распределена с постоянной плотностью C; с вероятностью R попадает в промежуток $(z_1;z_2)$, и имеет там плотность распределения $\varphi(x)$ вида $\varphi(x)=A\,|\,x-z_3\,|$; значения некоторых параметров указаны в условии. Требуется:

1. найти недостающие значения параметров;
2. найти плотность распределения и функцию распределения случайной величины X и построить их графики;
3. вычислить $M(X)$, $D(X)$, $\sigma(X)$;
4. найти $p(|\,X-M(X)\,|<\sigma(X))$;
5. найти медиану случайной величины X.
6. Результаты занести в таблицу:

Таблица 8.1 – Шаблон для ответов РГР №1

Вар-т	q_1	q_2	z_1	z_2	z_3	R	C	A	$M(X)$	$D(X)$	$\sigma(X)$	p	Мед.

Таблица 8.2 – Содержание вариантов к РГР №1

Вар-т	q_1	q_2	z_1	z_2	z_3	R	C	A
1	–1	0	–3	–2	–4		0,6	
2	–1	0	–3	–2	–3		0,7	
3	–3	–1	0	1	2			0,2
4	–2		–3	–2	–4	0,3	0,7	
5	1	3	–2	0	–3			0,075
6	–2		–3	–2	–4	0,3	0,7	
7	–1	1	–3	–1	–1	0,3		
8	–2	–1	–1	1	–2	0,3		
9	–2	–1	0	2	–1			0,1
10	0		–3	–1	–2	0,4	0,3	
11	–2	0	0	2	1			0,3

РГР №2. Сравнение двух выборок

Даны две серии измерений некоторой случайной величины.

1. Найдите средние арифметические $\overline{y_1}$, $\overline{y_2}$ и эмпирические стандарты s_1, s_2 для каждой из выборок.
2. Постройте доверительные интервалы для β_1, β_2, σ_1, σ_2. (Доверительная вероятность 0,95).
3. Проверьте гипотезу о равенстве дисперсий или стандартных отклонений по критерию Фишера. (Уровень значимости $\alpha = 0{,}05$, доверительная вероятность 0,975).
4. Если $\sigma_1 = \sigma_2$, то найдите $s_{ñâ}$, постройте доверительный интервал для σ. (Доверительная вероятность 0,95).
5. Проверьте гипотезу о равенстве математических ожиданий по критерию Стьюдента. (Уровень значимости $\alpha = 0{,}05$, доверительная вероятность 0,95).
6. Если $\beta_1 = \beta_2$, но найдите $\overline{y_{ñâ}}$, постройте доверительный интервал для β, найдите $s_{îá}$.
7. Объедините две выборки в одну и проверьте гипотезу о том, что экспериментальные данные имеют нормальный закон распределения. Рассматривать интервалы равной вероятности. Число интервалов равно L.
8. Постройте гистограмму.

Вариант 1. L=7.

1 серия измерений. $n_1 = 28$.

7,2	7,1	3,7	5,3	6,4	5,2	9,7	8,8	6,3	5,9
6,9	4,5	9,0	5,5	6,1	7,5	8,9	3,7	6,3	6,1
6,3	7,2	6,2	3,5	9,0	6,4	7,5	9,8		

2 серия измерений. $n_2 = 31$.

6,7	5,1	7,4	5,9	9,8	6,6	8,8	9,3	7,9	5,6	
7,2	6,2	6,8	5,4	6,8	8,2	9,3	8,0	6,0	6,0	
7,6	7,5	8,9	4,9	5,8	8,5	8,9	8,7	6,4	6,6	5,7

Вариант 2. L=7.

1 серия измерений. $n_1 = 38$.

23,2	20,1	18,8	24,1	21,6	22,8	22,1	25,2	24,8	20,6
25,9	27,0	21,9	23,5	21,7	21,1	21,3	18,3	21,0	23,8
17,4	17,3	17,9	20,6	18,4	24,2	20,7	22,0	18,3	22,6
20,2	21,5	16,5	21,3	21,5	17,9	26,2	29,1		

2 серия измерений. $n_2 = 21$.

24,2	30,4	21,7	21,2	20,7	25,4	17,8	19,0	21,5	18,7
22,1	25,6	19,6	19,9	21,9	26,4	21,9	24,1	21,1	23,5
23,9									

Вариант 3. L=8.

1 серия измерений. $n_1 = 36$.

4,4	4,3	2,9	3,8	5,1	5,6	3,7	4,7	5,6	4,4
4,6	4,0	4,7	5,0	5,4	4,3	6,2	6,4	4,5	5,1
4,4	5,4	5,8	4,8	6,2	5,0	5,6	4,8	4,7	4,0
5,3	5,4	2,5	5,4	5,4	6,9				

2 серия измерений. $n_2 = 22$.

5,0	4,3	3,9	6,0	3,4	4,1	4,7	3,4	4,2	5,1
3,5	3,1	4,3	3,7	3,7	6,2	4,8	3,5	4,3	6,2
2,7	7,1								

Вариант 4. L=7.

1 серия измерений. $n_1 = 34$.

0,37	0,67	0,64	0,94	0,82	0,60	0,70	0,84	0,81	0,67
0,77	0,52	0,70	0,54	0,76	0,47	0,86	0,62	0,54	0,88
0,85	0,84	0,62	0,65	0,70	0,97	0,55	0,72	0,74	0,93
0,65	0,61	0,66	0,78						

2 серия измерений. $n_2 = 23$.

0,84	0,85	0,84	0,64	0,97	0,78	0,81	0,73	0,93	0,60
0,79	0,69	0,85	0,73	0,72	0,84	0,54	0,95	0,42	0,90
0,90	0,87	0,78							

Вариант 5. L=8.

1 серия измерений. $n_1 = 32$.

23,5	26,5	22,8	26,0	23,7	25,8	22,7	18,9	24,6	26,1
26,2	20,9	23,0	31,9	23,2	21,3	21,8	24,1	26,7	19,3
27,1	26,6	29,5	20,5	21,8	26,8	23,9	20,3	21,5	24,8
25,8	20,5								

2 серия измерений. $n_2 = 24$.

24,6	20,4	20,1	25,7	24,0	24,2	19,5	18,3	21,2	18,8
15,4	22,1	24,2	20,4	19,2	15,2	18,0	21,5	22,2	21,4
17,0	17,8	25,4	21,5						

Вариант 6. L=6.

1 серия измерений. $n_1 = 30$.

4,9	4,8	3,6	4,8	6,5	8,6	5,5	7,7	6,1	7,0
7,5	7,9	5,7	6,2	6,6	5,0	7,2	5,9	7,2	4,0
6,3	6,1	5,0	6,7	3,6	6,4	3,9	3,5	5,4	6,5

2 серия измерений. $n_2 = 25$.

4,5	5,8	5,0	7,1	5,7	6,2	6,2	6,6	5,6	4,9
5,5	5,7	5,3	6,7	5,5	7,1	6,2	3,0	6,3	7,9
5,9	9,5	6,7	7,3	5,5					

Вариант 7. L=7.

1 серия измерений. $n_1 = 28$.

0,22	0,75	0,77	0,59	0,85	0,99	0,79	1,10	1,04	0,51
0,83	0,72	0,65	1,03	0,92	0,53	0,63	0,85	0,91	0,72
0,62	0,58	0,81	0,91	0,81	1,02	1,15			
0,35									

2 серия измерений. $n_2 = 26$.

0,65	0,96	0,85	0,98	0,69	1,01	0,79	0,62	0,71	0,84
1,13	0,81	1,02	0,65	0,54	0,78	0,69	0,65	0,61	0,61
0,72	1,01	0,54	0,58	0,70	0,82				

Вариант 8. L=6.

1 серия измерений. $n_1 = 26$.

29,8	29,5	30,4	30,4	28,5	35,6	29,3	28,0	26,4	24,2
32,3	26,2	22,9	25,9	27,5	20,2	28,4	22,7	21,3	23,3
23,2	29,7	24,0	26,5	28,5	24,7				

2 серия измерений. $n_2 = 27$.

27,9	32,9	29,1	31,1	28,9	34,1	35,7	23,7	33,9	25,2
25,3	29,3	29,7	29,9	36,7	24,7	32,5	30,4	26,4	31,1
28,8	34,7	32,9	36,1	30,5	39,7	30,8			

Вариант 9. L=7.

1 серия измерений. $n_1 = 24$.

6,7	6,6	5,7	5,8	6,5	6,5	7,5	5,6	6,3	5,9
3,9	5,8	8,0	7,3	7,9	5,5	9,4	6,3	6,2	7,8
7,1	9,4	7,6	7,3						

2 серия измерений. $n_2 = 28$.

4,9	6,0	3,7	6,6	5,5	4,7	8,0	6,7	4,7	4,4
5,3	5,7	5,1	5,7	6,5	6,3	6,9	5,1	6,2	4,2
2,0	5,7	5,6	7,3	5,1	6,6	6,1	5,6		

Вариант 10. L=6.

1 серия измерений. $n_1 = 22$.

0,82	0,61	1,10	0,51	0,44	0,65	0,88	0,58	0,68	0,89
1,07	1,15	0,96	1,01	0,49	0,99	0,90	1,10	0,74	0,88
1,09	1,22								

2 серия измерений. $n_2 = 29$.

1,12	0,90	1,03	1,35	1,38	0,77	1,05	0,90	0,60	0,74
1,04	1,00	1,32	0,52	1,13	0,68	0,90	1,04	0,66	0,95
0,66	0,99	0,95	1,19	0,90	1,26	1,12	0,99	1,14	

Вариант 11. L=8.

1 серия измерений. $n_1 = 30$.

28,2	32,0	30,3	34,8	25,2	34,1	32,3	38,3	31,7	29,1
26,3	31,9	34,2	34,0	23,4	34,5	31,4	28,3	30,2	27,4
33,9	32,3	28,7	28,9	35,7	29,3	24,9	28,1	30,7	22,6

2 серия измерений. $n_2 = 30$.

29,0	26,3	24,5	27,9	38,4	28,7	31,0	31,5	29,7	36,0
35,0	36,1	24,7	36,3	33,2	31,8	24,0	38,1	34,6	33,2
28,3	33,0	31,3	25,6	33,8	33,3	31,8	35,3	28,0	30,0

РГР №3. Построение регрессионных моделей и проверка их адекватности

Для каждого значения фактора *X* было проведено несколько измерений величины *Y*.

а) Постройте линейную модель регрессии;

б) проверьте ее адекватность;

в) постройте квадратичную модель регрессии;

г) проверьте ее адекватность.

Таблица 8.3 – Содержание вариантов к РГР №3

Вариант 1					
X	*Y*				
0,00	4,96	4,94	4,96	4,94	
0,10	5,42	5,46	5,42	5,39	
0,15	5,63	5,64	5,63	5,64	
0,25	6,00	6,03	6,02		
0,30	6,19	6,25	6,20		
0,55	7,04	7,04	7,04		
0,70	7,46	7,41	7,43		
0,75	7,54	7,54	7,53	7,60	
0,80	7,66	7,65	7,64	7,64	
0,90	7,87	7,79	7,81	7,86	

Продолжение таблицы 8.3

Вариант 2					
X	*Y*				
5	53	52	53		
10	63	63	62	64	
25	91	91	91	90	
30	96	94	97	97	
40	108	108	107		
55	117	118	116		
70	117	116	117	117	
75	114	114	114	115	
90	103	102	101	106	
100	89	89	91		
Вариант 3					
X	*Y*				
50	48,8	48,4	48,6	48,7	48,2
80	46,3	46,3	46,2	46,2	
90	45,6	45,3	45,8		
110	44,9	44,8	44,9	44,7	
130	43,3	43,4	43,4		
160	41,8	41,5	42,0		
170	41,5	41,5	41,7	41,4	
180	40,5	40,6	40,6		
200	39,9	39,8	40,2	40,4	
230	38,9	38,8	38,7	39,0	38,6
Вариант 4					
X	*Y*				
0,10	5,33	5,32	5,31		
0,20	5,70	5,69	5,73	5,73	5,71
0,25	5,93	5,91	5,87	5,94	
0,30	6,07	6,08	6,12	6,09	6,09
0,40	6,38	6,38	6,41		
0,65	7,13	7,15	7,09		
0,80	7,43	7,44	7,44	7,45	7,44
0,85	7,56	7,52	7,53	7,58	
0,90	7,63	7,66	7,62	7,59	7,61
1,05	7,82	7,86	7,86		

Продолжение таблицы 8.3

Вариант 5					
X	*Y*				
0	37	37	37	39	37
10	60	63	63		
20	81	83	82		
30	94	97	97	97	96
50	115	115	112		
60	121	118	118		
70	117	118	118	121	118
75	115	117	114		
85	110	111	111		
100	90	90	89	93	88
Вариант 6					
X	*Y*				
10	50,6	50,7	50,7		
30	49,2	48,8	49,4	49,1	
50	47,2	47,8	47,9	48,1	47,7
70	46,8	46,2	46,4	46,1	46,6
100	44,4	44,5	44,8		
110	43,9	43,9	43,9		
150	42,1	41,8	41,8	41,7	41,6
170	40,7	41,0	40,7	41,0	40,5
190	40,3	39,7	39,8	39,9	
220	38,4	38,7	38,8		
Вариант 7					
X	*Y*				
0,00	4,96	4,95	5,00	4,94	
0,05	5,21	5,23	5,23		
0,20	5,83	5,81	5,82	5,88	5,86
0,30	6,28	6,20	6,22		
0,40	6,59	6,60	6,55	6,54	6,54
0,50	6,86	6,89	6,90	6,88	6,87
0,65	7,32	7,34	7,26		
0,70	7,42	7,39	7,36	7,35	7,44
0,80	7,65	7,64	7,63		
0,90	7,81	7,77	7,85	7,90	

Продолжение таблицы 8.3					
Вариант 8					
X	*Y*				
5	52	51	52	52	52
25	90	90	88		
35	102	104	102	105	103
40	110	109	106	106	
45	111	113	112		
50	114	115	112		
60	119	119	117	117	
65	117	118	118	120	116
80	112	115	110		
95	98	97	97	100	97
Вариант 9					
X	*Y*				
10	51,4	51,4	50,9		
40	49,7	49,5	49,2	49,1	49,0
50	48,2	48,7	48,7		
60	47,6	47,8	47,7		
100	45,0	45,3	45,3	45,1	44,9
120	43,6	43,7	43,8	43,9	44,1
150	42,2	42,1	42,3		
180	40,9	40,9	41,0	40,6	41,1
190	40,8	40,6	40,7		
200	40,1	40,1	40,2		
Вариант 10					
X	*Y*				
0,00	4,89	4,96	4,88	4,89	4,89
0,05	5,13	5,15	5,13	5,15	
0,20	5,70	5,76	5,71		
0,30	6,06	6,06	6,05	6,07	
0,40	6,40	6,40	6,38		
0,55	6,87	6,86	6,86		
0,60	7,00	7,01	6,96	7,02	
0,65	7,10	7,10	7,11		
0,85	7,55	7,49	7,56	7,54	
0,90	7,68	7,64	7,62	7,63	7,63

Окончание таблицы 8.3					
Вариант 11					
X	*Y*				
5	49	50	51	50	
20	81	82	80		
35	103	103	105		
40	108	108	108	107	
45	111	115	112	110	112
55	119	118	118	118	119
60	120	120	117	118	
70	116	118	118		
75	118	114	118		
95	98	97	99	100	

ГЛАВА 9. ЛАБОРАТОРНЫЕ РАБОТЫ

Лабораторная работа 1. Непрерывные распределения

***Используемое ПО*:** MathCad.

***Цель работы*:** научиться с помощью программы MathCad находить основные характеристики непрерывного распределения.

Задания

1. Построить график плотности вероятности.
2. Проверить, выполняется ли условие нормировки.
3. Найти функцию распределения и построить ее график.
4. Найти математическое ожидание, дисперсию и среднее квадратичное отклонение случайной величины.
5. Найти вероятность попадания случайной величины в указанный интервал с помощью плотности вероятности и с помощью функции распределения.
6. Найти медиану и квантиль, соответствующую указанной вероятности.

Таблица 9.1 – Содержание вариантов к лабораторной работе 1

№	Содержание варианта	№	Содержание варианта
1	$\varphi(x)=\begin{cases}0, & x<0\\ \frac{1}{9}x^2, & 0\le x\le 3\\ 0, & x>3\end{cases}$ (1,2) p=0,1	2	$\varphi(x)=\begin{cases}0, & x<1\\ \frac{4}{x^5}, & x\ge 1\end{cases}$ (2,3) p=0,2
3	$\varphi(x)=\begin{cases}0, & x<1\\ \frac{9}{x^{10}}, & x\ge 1\end{cases}$ (3,4) p=0,3	4	$\varphi(x)=\begin{cases}0, & x<0\\ e^{-x}, & x\ge 0\end{cases}$ (1,2) p=0,4

Окончание таблицы 9.1			
5	$\varphi(x)=\begin{cases}0, & x<0\\ 2\cos 2x, & 0\le x\le\frac{\pi}{4}\\ 0, & x>\frac{\pi}{4}\end{cases}$ (4,5) p=0,6	6	$\varphi(x)=\begin{cases}0, & x<0\\ \frac{2}{9}(3x-x^2), & 0\le x\le 3\\ 0, & x>3\end{cases}$ (1,2) p=0,7
7	$\varphi(x)=0{,}5e^{-\lvert x\rvert}$ (1,5) p=0,8	8	$\varphi(x)=\begin{cases}0, & x<0\\ 0{,}5\sin x, & 0\le x\le\pi\\ 0, & x>\pi\end{cases}$ (1,2) p=0,1
9	$\varphi(x)=\begin{cases}0, & x<-1\\ \frac{2}{7}(2-\lvert x\rvert), & -1\le x\le 2\\ 0, & x>2\end{cases}$ (1,2) p=0,2	10	$\varphi(x)=\begin{cases}0, & x<0\\ 5e^{-5x}, & x\ge 0\end{cases}$ (3,6) p=0,3

Некоторые теоретические сведения

1-2. Функция плотности вероятности непрерывной случайной величины считается заданной корректно, если выполняются два условия:

а) $\varphi(x)\ge 0$;

б) $\int\limits_{-\infty}^{\infty}\varphi(x)dx=1$ (условие нормировки).

3. Функция распределения $F(x)$ случайной величины X определяется как

$$F(x)=\int\limits_{-\infty}^{x}\varphi(t)dt\,.$$

4. Если случайная величина X имеет непрерывное распределение, то для любой функции $g(x)$

$$M(g(X)) = \int_{-\infty}^{\infty} g(x)\varphi(x)dx$$

(при условии, что интеграл сходится абсолютно).

Таким образом,

$$M(X) = \int_{-\infty}^{\infty} x\varphi(x)dx,$$

$$D(X) = M(X^2) - M^2(X) = \int_{-\infty}^{\infty} x^2\varphi(x)dx - M^2(X),$$

$$\sigma(X) = \sqrt{D(X)}.$$

5. Вероятность попадания случайной величины в интервал есть интеграл от плотности вероятности по этому промежутку:

$$p(a < X < b) = \int_{a}^{b} \varphi(x)dx,$$

или приращение функции распределения на этом промежутке:

$$p(a < X < b) = F(b) - F(a).$$

6. Квантиль – это функция, обратная функции распределения. То есть, если $p = F(x)$, то квантиль $x_p = F^{-1}(p)$. В частном случае, когда $p = 0{,}5$, квантиль называют медианой. Таким образом, для нахождения квантили, соответствующей вероятности *p*, следует решить уравнение

$$F(x) - p = 0.$$

Пояснения к работе с программой

1. Для выполнения работы понадобятся панели инструментов «*Графики*», «*Матанализ*», «*Арифметика*», «*Греческий алфавит*», «*Программирование*», «*Булево*». Вывести их на экран можно через пункт меню «*Вид / Панель инструментов*».

2. Программу желательно составлять так, чтобы она обладала универсальностью. В том случае, если придется изменить данные: функцию плотности, интервал или вероятность, то пусть в программе это придется

сделать всего один раз – в самом ее начале.

3. MathCad различает большие и маленькие буквы.

4. Задать функцию кусочно можно с помощью команд панели программирования «*Add Line*» и «*If*».

5. Построить декартов график можно с помощью команды панели графиков «*Декартов график*». При этом следует указать в нижнем поле ввода имя независимой переменной и в левом – функцию.

6. Если известно, что корень уравнения $f(x)=0$ находится на отрезке $[a,b]$, то его можно найти командой **root**. Первый ее параметр – функция $f(x)$, второй – имя независимой переменной, в нашем случае *x*, третий и четвертый параметры – границы отрезка *a* и *b*. Так, например, один из корней уравнения $x^2=3$ можно найти так: **root(*x*^{2}–3,*x*,1,2)**.

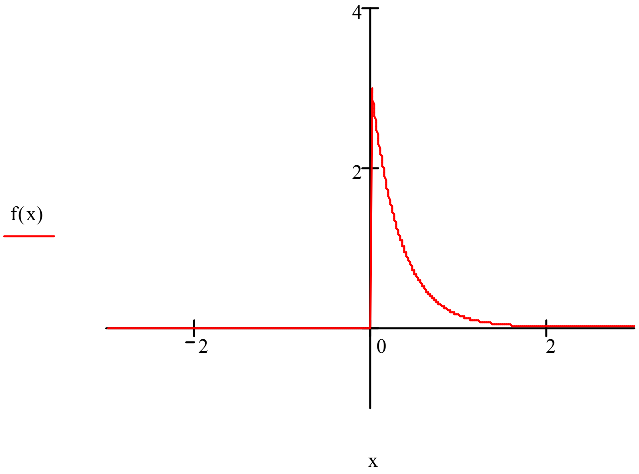
Программа MathCad.
Разбор нулевого варианта

Пусть $\varphi(x)=\begin{cases}3e^{-3x}, x\geq 0\\ 0, \quad x<0\end{cases}$, интервал (1,2), вероятность *p*=0,1.

0. Условие задачи

$$f(x) := \begin{vmatrix} 0 & \text{if } x < 0 \\ 3 \cdot e^{-3 \cdot x} & \text{if } x \geq 0 \end{vmatrix} \qquad a := 1 \qquad b := 2 \qquad p := 0.1$$

1. График функции плотности

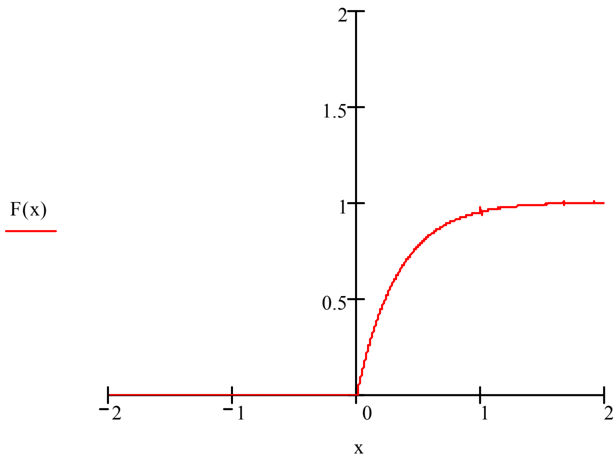


2. Условие нормировки

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

3. Функции распределения и ее график

$$F(x) := \int_{-\infty}^{x} f(t)\,dt$$



4. Математическое ожидание, дисперсия и стандартное отклонение

$$M := \int_{-\infty}^{\infty} x \cdot f(x)\,dx \qquad M = 0.333$$

$$D := \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx - M^2 \qquad D = 0.111$$

$$\sigma := \sqrt{D} \qquad \sigma = 0.333$$

5. Вероятность попадания в интервал

$$p1 := \int_{a}^{b} f(x)\,dx \qquad p1 = 0.047$$

$$p2 := F(b) - F(a) \qquad p2 = 0.047$$

6. Медиана и квантиль

$$\text{root}(F(x) - 0.5, x, 0, 2) = 0.231 \qquad \text{root}(F(x) - p, x, 0, 2) = 0.035$$

Лабораторная работа 2. Сравнение двух выборок

Используемое ПО: MathCad, Excel, StatGraph, Stadia.

Цель работы: научиться с помощью вышеуказанных программ находить основные характеристики случайных выборок и сравнивать их.

Задания

1. Найти средние арифметические и эмпирические стандарты для каждой из выборок.
2. Построить доверительные интервалы для математических ожиданий и стандартных отклонений.
3. Проверить гипотезу о равенстве дисперсий.
4. Если гипотеза о равенстве дисперсий принята, найти сводную оценку стандартного отклонения.
5. Проверить гипотезу о равенстве математических ожиданий.
6. Если гипотеза о равенстве математических ожиданий принята, найти сводную оценку математического ожидания и объединенную оценку стандартного отклонения.
7. Объединить две выборки в одну и проверить гипотезу о том, что экспериментальные данные имеют нормальный закон распределения. Рассматривать интервалы равной длины. Число интервалов равно L.
8. Построить гистограмму.

Указания:

1) Во всех пунктах брать уровень значимости $\alpha = 0{,}05$.
2) Выполнить пункты 1-7 в программе MathCad. В пункте 7 искать только теоретическую квантиль.
3) Выполнить пункты 1, 2, 3, 5, 7 в программе Excel. Доверительные интервалы для математических ожиданий искать с помощью встроенной команды и без нее. В пункте 7 искать только теоретическую квантиль.
4) Выполнить пункты 1, 2, 3, 5, 7, 8 в программе StatGraph.
5) Выполнить пункты 1, 2, 3, 5, 7, 8 в программе Stadia.

Экспериментальные данные следует взять из условий РГР №2.

Некоторые теоретические сведения

1. Несмещенными оценками математического ожидания и стандартного отклонения являются среднее арифметическое и эмпирический стандарт:

$$\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}, \; s = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \overline{y})^2}{n-1}}.$$

2. Доверительные интервалы для математического ожидания и стандартного отклонения соответственно имеют вид:

$$\overline{y} - |t|_p \frac{s}{\sqrt{n}} < \beta < \overline{y} + |t|_p \frac{s}{\sqrt{n}}, \tag{2.1}$$

$$s \cdot \sqrt{\frac{k}{a_2}} < \sigma < s \cdot \sqrt{\frac{k}{a_1}}.$$

Здесь $p = 1 - \alpha$, $k = n - 1$, a_1 и a_2 – квантили распределения Пирсона для вероятностей $\frac{1-p}{2}$ и $\frac{1+p}{2}$ соответственно. Видно, что доверительный интервал для математического ожидания симметричен относительно среднего арифметического, поэтому иногда считают, что интервал найден, если указаны среднее арифметическое и радиус интервала $R = |t|_p \frac{s}{\sqrt{n}}$.

3. Пусть $F_{Э1}$ – отношение большей эмпирической дисперсии к меньшей, $F_{Э2}$ – меньшей к большей. Пусть F_{m1} и F_{m2} – квантили распределения Фишера $F(p_1, k_1, k_2)$ и $F(p_2, k_2, k_1)$. Здесь $p_1 = 1 - \frac{\alpha}{2}$, $p_2 = \frac{\alpha}{2}$, k_1 – число степеней свободы большей эмпирической дисперсии, k_2 – меньшей. Критерий Фишера требует отвергнуть гипотезу о равенстве дисперсий, если $F_{Э1} > F_{m1}$ ($F_{Э2} < F_{m2}$). В противном случае гипотеза принимается.

4. Если гипотеза о равенстве дисперсий не противоречит экспериментальным данным, можно найти сводную оценку дисперсии

$$s_{св}^2 = \frac{s_1^2 k_1 + s_2^2 k_2}{k_1 + k_2},$$

где s_1^2 и s_2^2 – эмпирические дисперсии с числом степеней свободы k_1 и k_2 соответственно.

5. Пусть

$$T_Э = \frac{|\overline{y_1} - \overline{y_2}|}{s_{св}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

а $|t|_p$ – квантиль распределения модуля отношения Стьюдента для вероятности $p = 1 - \alpha$ и числа степеней свободы $k = k_1 + k_2$. Критерий Стьюдента требует отвергнуть гипотезу о равенстве математических ожиданий, если $T_Э > |t|_p$. В противном случае гипотеза принимается.

6. Если гипотеза о равенстве математических ожиданий не противоречит экспериментальным данным, можно найти сводную оценку математического ожидания и объединенную оценку дисперсии:

$$\overline{y_{св}} = \frac{\overline{y_1}n_1 + \overline{y_2}n_2}{n_1 + n_2},\ s_{об}^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n_1} y_i^2 + \sum_{i=1}^{n_2} y_i^{/2} - n \cdot \overline{y_{св}}^2\right).$$

Здесь $n = n_1 + n_2$, $\{y_i\}$ и $\{y_i^/\}$ – данные выборки. Заметим, что $\overline{y_{св}}$ и $s_{об}$ есть обычное среднее арифметическое и обычный эмпирический стандарт для объединенной выборки.

7. Для проверки гипотезы о том, что экспериментальные данные распределены по нормальному закону, можно применить критерий согласия Пирсона. Для этого весь диапазон изменения случайной величины разбивают на L участков точками $x_1 = -\infty$, x_2, …, $x_{L+1} = +\infty$, подсчитывают частоты попадания в эти участки $\frac{N_1}{n}$, $\frac{N_2}{n}$, …, $\frac{N_L}{n}$, а также теоретические вероятности попадания в эти же участки:

$$p_i = \frac{1}{\sqrt{2\pi}} \int_{x_i}^{x_{i+1}} e^{-\frac{U^2}{2}} dU.$$

Затем вычисляют значение величины

$$\chi^2 = \sum_{i=1}^{L} \frac{(N_i - np_i)^2}{np_i}$$

и сравнивают его с квантилью распределения Пирсона $\chi_p^2(k)$ для вероятности $p = 1 - \alpha$ и числа степеней свободы $k = L - 3$. Критерий Пирсона требует отвергнуть гипотезу о том, что экспериментальные данные распределены по нормальному закону, если $\chi^2 > \chi_p^2(k)$, в противном случае гипотеза принимается. На практике точки дробления x_i выбирают так, чтобы либо все интервалы имели одинаковую длину, либо все вероятности p_i равнялись бы между собой.

8. Если по оси абсцисс отметить точки x_i из пункта 7, только вместо x_1 и x_{L+1} взять наименьшее и наибольшее числа в выборке, и на каждом отрезке построить прямоугольник высотой $\frac{N_i}{n \cdot h_i}$, где h_i – длина *i*-го участка, то графическое изображение полученных данных даст гистограмму.

Пояснения к работе с программой MathCad

1. Для выполнения работы понадобятся панели инструментов «*Арифметика*», «*Греческий алфавит*», «*Программирование*», «*Матрицы*». Вывести их на экран можно через пункт меню «*Вид / Панель инструментов*».

2. Программу желательно составлять так, чтобы она обладала универсальностью. В том случае, если придется изменить выборки, то пусть в программе при этом ничего не придется менять.

3. Перед тем, как приступить к работе с MathCad'ом, следует подготовить два текстовых файла с выборками, например, с помощью программы Блокнот. Набирать данные следует либо в одну строку, разделяя их пробелами, либо в один столбец. В качестве десятичного разделителя следует использовать точку. Сохраните эти текстовые файлы в той же папке, где будет храниться файл, созданный в MathCad'е.

4. Укажем некоторые команды, которые нам понадобятся для выполнения работы.

READPRN("FileName.txt")

Считать данные из текстового файла.

length(x)

Длина массива x.

mean(x)

Среднее арифметическое массива x.

stdev(x)

Смещенная оценка стандартного отклонения $\tilde{s} = \sqrt{\sum_{i=1}^{n} \frac{(y_i - \overline{y})^2}{n}}$.

qchisq(p,k)

Квантиль распределения Пирсона для вероятности p и числа степеней свободы k.

qt(p,k)

Обычная квантиль (не модуля) отношения Стьюдента для вероятности p и числа степеней свободы k. Заметим, что из обычной квантили qt(p,k) можно получить нужную нам квантиль модуля так: $-qt\left(\frac{1-p}{2}, k\right)$.

qF(p,k_1,k_2)

Квантиль распределения Фишера для вероятности p и числа степеней свободы k_1 и k_2.

stack(x,y)

Объединение двух массивов-столбцов в один.

augment(x,y)

Объединение двух массивов-строк в один.

x:=x^{T}

Транспонирование массива x. Некоторые команды программы MathCad применимы к массивам-строкам, некоторые к массивам-столбцам. С помощью команды транспонирования (панель «*Матрицы*») можно преобразовывать рассматриваемый массив.

Пояснения к работе с программой Excel

1. Список всех доступных команд можно получить с помощью кнопки быстрого доступа «*Вставка функции*» или через меню «*Вставка / Функция…*».

2. Укажем некоторые команды, которые нам понадобятся для выполнения работы.

СРЗНАЧ(A_i:B_j)

Среднее арифметическое диапазона ячеек A_i:B_j.

СТАНДОТКЛОН(A_i:B_j)

Эмпирический стандарт диапазона ячеек A_i:B_j.

КОРЕНЬ(A_i)

Квадратный корень.

ХИ2ОБР(α,k)

Квантиль распределения Пирсона для уровня значимости α и числа степеней свободы k.

СТЬЮДРАСПОБР(α,k)

Квантиль распределения модуля отношения Стьюдента для уровня значимости α и числа степеней свободы k.

FРАСПОБР(α,k_1,k_2)

Квантиль распределения Фишера для уровня значимости α и числа степеней свободы k_1 и k_2.

ДОВЕРИТ(α,s,n)

Радиус доверительного интервала для математического ожидания. Здесь α – уровень значимости, s – эмпирический стандарт, n – число элементов в выборке. Заметим, что эта команда находит радиус, заменяя в формуле (2.1) квантиль модуля отношения Стьюдента квантилью модуля стандартного нормального распределения. Учитывая то, что при $n \to \infty$ распределение Стьюдента стремится к стандартному нормальному, при достаточно больших n эти квантили практически совпадают.

Пояснения к работе с программой StatGraph

1. Следует подготовить три столбца с экспериментальными данными: два с соответствующими выборками и третий – с объединенной выборкой.

2. В качестве десятичного разделителя следует использовать запятую.

3. Для выполнения пунктов 1, 2, 3, 5 используйте пункт меню «*Compare / Two Samples / Two-sample Comparison…*». В окошке «*Two-sample Comparison*» используйте кнопку быстрого доступа «*Tabular Options*» для

получения доступа к разделам «*Comparison Of Means*» и «*Comparison Of Standard Deviations*».

4. Для выполнения пунктов 7, 8 используйте пункт меню «*Describe / Distribution Fitting / Uncensored Data…*». В окошке «*Uncensored Data*» используйте кнопки быстрого доступа «*Tabular Options*» и «*Graphical Options*» для получения доступа к разделам «*Test For Normality*» и «*Frequency Histogram*» соответственно.

Пояснения к работе с программой Stadia

1. Следует подготовить три столбца с экспериментальными данными: два с соответствующими выборками и третий – с объединенной выборкой.

2. В качестве десятичного разделителя следует использовать точку.

3. Для выполнения всех пунктов воспользуемся пунктом меню «*Статист-F9*».

4. Для выполнения пунктов 1, 2 воспользуемся кнопкой «*Описательная статистика*».

5. Для выполнения пунктов 3, 5 воспользуемся кнопкой «*Стьюдента и Фишера*».

6. Для выполнения пунктов 7, 8 воспользуемся кнопкой «*Гистограмма/Нормальность*».

Разбор нулевого варианта

Вариант 0. L=8.

1 серия измерений. $n_1 = 26$.

0,44	1,29	1,25	1,06	1,24	0,87	0,96	1,25	0,88	1,09
0,90	0,75	1,09	0,73	1,04	1,07	1,20	1,15	0,94	0,83
1,19	0,88	0,77	0,84	0,79	0,80				

2 серия измерений. $n_2 = 32$.

0,89	0,97	0,33	0,93	0,84	1,43	1,34	0,88	0,75	0,96
1,09	0,83	0,95	0,33	0,56	1,20	1,12	0,92	0,73	1,30
0,70	1,27	0,82	0,86	1,30	1,00	1,12	0,69	1,03	0,58
1,26	1,16								

Программа MathCad

Разбор нулевого варианта

0. Подготовка

$X := READPRN("00Vib1.txt")$ $Y := READPRN("00Vib2.txt")$

$n1 := length(X)$ $n2 := length(Y)$

$k1 := n1 - 1$ $k2 := n2 - 1$

$a := 0.05$ $p := 1 - a$

X =

	0
0	0.44
1	1.29
2	1.25
3	1.06
4	1.24
5	0.87
6	0.96
7	1.25
8	0.88
9	1.09
10	0.9
11	0.75
12	1.09
13	0.73
14	1.04
15	1.07

Y =

	0
0	0.89
1	0.97
2	0.33
3	0.93
4	0.84
5	1.43
6	1.34
7	0.88
8	0.75
9	0.96
10	1.09
11	0.83
12	0.95
13	0.33
14	0.56
15	1.2

1. Средние и стандарты

$x1 := mean(X)$ $x1 = 0.973$

$y1 := mean(Y)$ $y1 = 0.942$

$s1 := stdev(X) \cdot \sqrt{\frac{n1}{k1}}$ $s1 = 0.207$

$s2 := stdev(Y) \cdot \sqrt{\frac{n2}{k2}}$ $s2 = 0.276$

2. Доверительные интервалы

Для сигма 1

$p1 := \frac{1-p}{2}$ $p2 := \frac{1+p}{2}$ $a1 := qchisq(p1, k1)$ $a2 := qchisq(p2, k1)$

$sig11 := s1 \cdot \sqrt{\frac{k1}{a2}}$ $sig12 := s1 \cdot \sqrt{\frac{k1}{a1}}$ $sig11 = 0.163$ $sig12 = 0.286$

Для сигма 2

$a1 := qchisq(p1, k2)$ $a2 := qchisq(p2, k2)$

$sig21 := s2 \cdot \sqrt{\frac{k2}{a2}}$ $sig22 := s2 \cdot \sqrt{\frac{k2}{a1}}$ $sig21 = 0.221$ $sig22 = 0.366$

Для бета 1

$tp := -qt\left(\frac{1-p}{2}, k1\right)$ $r1 := tp \cdot \frac{s1}{\sqrt{n1}}$ $beta11 := x1 - r1$ $beta11 = 0.889$

$beta12 := x1 + r1$ $beta12 = 1.057$

Для бета 2

$tp := -qt\left(\frac{1-p}{2}, k2\right)$ $r2 := tp \cdot \frac{s2}{\sqrt{n2}}$ $beta21 := y1 - r2$ $beta21 = 0.843$

$beta22 := y1 + r2$ $beta22 = 1.041$

3. Гипотеза о равенстве дисперсий

$$F1 := \begin{cases} \frac{s1^2}{s2^2} & \text{if } (s1 > s2) \\ \frac{s2^2}{s1^2} & \text{if } (s2 \ge s1) \end{cases} \qquad K1 := \begin{cases} k1 & \text{if } (s1 > s2) \\ k2 & \text{if } (s2 \ge s1) \end{cases} \qquad K2 := \begin{cases} k1 & \text{if } (s2 > s1) \\ k2 & \text{if } (s1 \ge s2) \end{cases}$$

$$F1 = 1.768 \qquad F2 := qF\left(1 - \frac{a}{2}, K1, K2\right) \qquad F2 = 2.174$$

$$Flag := \begin{cases} \text{"No"} & \text{if } (F1 > F2) \\ \text{"Yes"} & \text{if } (F2 \ge F1) \end{cases} \qquad Flag = \text{"Yes"}$$

4. Сводная оценка стандартного отклонения

$$Ssv := \frac{s1^2 \cdot k1 + s2^2 \cdot k2}{k1 + k2} \qquad Ssv := \sqrt{Ssv} \qquad Ssv = 0.247$$

5. Гипотеза о равенстве математических ожиданий

$$T1 := \frac{|x1 - y1|}{Ssv \cdot \sqrt{\frac{1}{n1} + \frac{1}{n2}}} \qquad T1 = 0.478 \qquad T2 := -qt\left(\frac{1-p}{2}, k1 + k2\right) \qquad T2 = 2.003$$

$$Flag := \begin{cases} \text{"No"} & \text{if } (T1 > T2) \\ \text{"Yes"} & \text{if } (T2 \ge T1) \end{cases} \qquad Flag = \text{"Yes"}$$

6. Сводная оценка математического ожидания, объединенный стандарт

$$Ysv := \frac{x1 \cdot n1 + y1 \cdot n2}{n1 + n2} \qquad Ysv = 0.956 \qquad z := stack(X, Y)$$

$$Sob := stdev(z) \cdot \sqrt{\frac{length(z)}{length(z) - 1}} \qquad Sob = 0.246$$

z =

	0
0	0.44
1	1.29
2	1.25
3	1.06
4	1.24
5	0.87
6	0.96
7	1.25
8	0.88
9	1.09
10	0.9
11	0.75
12	1.09
13	0.73
14	1.04
15	1.07

7. Квантиль распределения Пирсона для критерия согласия

$$\chi := qchisq(p, 5) \qquad \chi = 11.07$$

Программа Excel

Разбор нулевого варианта

0,44	0,89	26	**x1-сред.**	**y1-сред.**	**S1**	**S2**
1,29	0,97	32	0,973077	0,941875	0,207283	0,275616
1,25	0,33	8	**Сигма1-1**	**Сигма1-2**	**Сигма2-1**	**Сигма2-2**
1,06	0,93		0,162563	0,286135	0,220962	0,366426
1,24	0,84		**Квантили ст. норм. распр.**			
0,87	1,43		**Бета1-1**	**Бета1-2**	**Бета2-1**	**Бета2-2**
0,96	1,34		0,893402	1,052752	0,846381	1,037369
1,25	0,88		**Квантили распр. Стьюд.**			
0,88	0,75		**Бета1-1**	**Бета1-2**	**Бета2-1**	**Бета2-2**
1,09	0,96		0,889354	1,0568	0,842505	1,041245
0,9	1,09		**F-эксп.**	**F-квант.**	**F-эксп.**	**F-квант.**
0,75	0,83		0,565611	0,460064	1,767999	2,173607
1,09	0,95		**T-эксп.**	**T-квант.**		
0,73	0,33		0,47757	2,003239		
1,04	0,56		**X2-квант.**			
1,07	1,2		11,07048			
1,2	1,12					
1,15	0,92					
0,94	0,73					
0,83	1,3					
1,19	0,7					
0,88	1,27					
0,77	0,82					
0,84	0,86					
0,79	1,3					
0,8	1					
	1,12					
	0,69					
	1,03					
	0,58					
	1,26					
	1,16					

Программа StatGraph

Разбор нулевого варианта

```
Comparison of Means
-------------------
95,0% confidence interval for mean of Col_1: 0,973077 +/- 0,0837235
95,0% confidence interval for mean of Col_2: 0,941875 +/- 0,0993703
95,0% confidence intervals for the difference between the means:
   assuming equal variances: 0,0312019 +/- 0,130882
   not assuming equal variances: 0,0312019 +/- 0,127129

t tests to compare means

   Null hypothesis: mean1 = mean2

   (1) Alt. hypothesis: mean1 NE mean2
          assuming equal variances: t = 0,47757   P-value = 0,634815
          not assuming equal variances: t = 0,491724   P-value = 0,624846
   (2) Alt. hypothesis: mean1 > mean2
          assuming equal variances: t = 0,47757   P-value = 0,317408
          not assuming equal variances: t = 0,491724   P-value = 0,312423
   (3) Alt. hypothesis: mean1 < mean2
          assuming equal variances: t = 0,47757   P-value = 0,682592
          not assuming equal variances: t = 0,491724   P-value = 0,687577

The StatAdvisor
---------------
   This option runs a t-test to compare the means of the two samples.
It also constructs confidence intervals for each mean and for the
difference between the means.  Of particular interest is the
confidence interval for the difference between the means, which
extends from -0,0996797 to 0,162084.  Since the interval contains the
value 0.0, there is not a statistically significant difference between
the means of the two samples at the 95,0% confidence level.  The
t-tests can also be used to arrive at the same conclusion.  P-values
below 0,05 indicate significant differences between the two means.
NOTE: the interval used above assumes that the variances of the two
samples are equal.  This was determined by running an F-test to
compare the standard deviations of the two samples.  You can see the
results of that test by selecting Comparison of Standard Deviations
from the Tabular Options menu.
```

```
Comparison of Standard Deviations
---------------------------------
                    Col_1               Col_2
------------------------------------------------------------
Standard deviation  0,207283            0,275616
Variance            0,0429662           0,0759641
Df                  25                  31
     Ratio of Variances = 0,565611

95,0% Confidence Intervals
     Standard deviation of Col_1: [0,162563;0,286135]
     Standard deviation of Col_2: [0,220962;0,366426]
     Ratio of Variances: [0,26847;1,22942]

F-tests to Compare Standard Deviations

   Null hypothesis: sigma1 = sigma2

   (1) Alt. hypothesis: sigma1 NE sigma2
          F = 0,565611   P-value = 0,147687
   (2) Alt. hypothesis: sigma1 > sigma2
          F = 0,565611   P-value = 0,926156
   (3) Alt. hypothesis: sigma1 < sigma2
          F = 0,565611   P-value = 0,0738437

The StatAdvisor
---------------
   This option runs an F-test to compare the variances of the two
samples.  It also constructs confidence intervals for each standard
deviation and for the ratio of the variances.  Of particular interest
is the confidence interval for the ratio of the variances, which
extends from 0,26847 to 1,22942.  Since the interval contains the
value 1.0, there is not a statistically significant difference between
the standard deviations of the two samples at the 95,0% confidence
level.  The F-tests can also be used to arrive at the same conclusion.
P-values below 0,05 indicate significant differences between the two
standard deviations.  IMPORTANT NOTE:  the F-tests and confidence
intervals shown here depend on the samples having come from normal
distributions.  To test this assumption, select Summary Statistics
from the list of Tabular Options and check the standardized skewness
and standardized kurtosis values.
```

```
Tests for Normality for Col_4

Computed Chi-Square goodness-of-fit statistic = 10,1379
P-Value = 0,859327

Shapiro-Wilks W statistic = 0,964943
P-Value = 0,193093

Z score for skewness = 0,992315
P-Value = 0,321042

Z score for kurtosis = 0,389011
P-Value = 0,697264



The StatAdvisor
---------------
   This pane shows the results of several tests run to determine
whether Col_4 can be adequately modeled by a normal distribution.  The
chi-square test divides the range of Col_4 into 19 equally probable
classes and compares the number of observations in each class to the
number expected.  The Shapiro-Wilks test is based upon comparing the
quantiles of the fitted normal distribution to the quantiles of the
data.  The standardized skewness test looks for lack of symmetry in
the data.  The standardized kurtosis test looks for distributional
shape which is either flatter or more peaked than the normal
distribution.
   The lowest P-value amongst the tests performed equals 0,193093.
Because the P-value for this test is greater than or equal to 0.10, we
can not reject the idea that Col_4 comes from a normal distribution
with 90% or higher confidence.
```

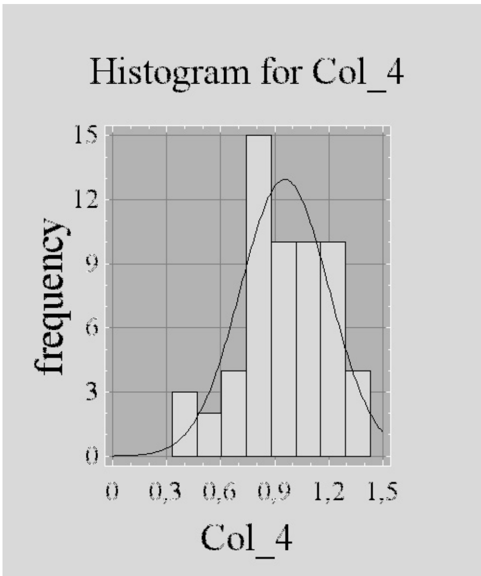


Рисунок 9.1

Программа Stadia

Разбор нулевого варианта

ОПИСАТЕЛЬНАЯ СТАТИСТИКА. Файл: var0.std

Переменная	Размер	<---Диапазон--->		Среднее---Ошибка		Дисперс	Ст.откл	Сумма
vib1	26	0,44	1,29	0,9731	0,04065	0,04297	0,2073	25,3
vib2	32	0,33	1,43	0,9419	0,04872	0,07596	0,2756	30,14

Переменная	Медиана	<--Квартили-->		ДовИнтСр.	<-ДовИнтДисп->	Ош.СтОткл
vib1	0,95	0,8225	1,16	0,08268	0,02643 0,0819	0,07776
vib2	0,94	0,7675	1,15	0,09813	0,04883 0,1343	0,09802

Переменная	Асимметр.	Значим	Эксцесс	Значим
vib1	-0,363	0,1986	2,812	0,4812
vib2	-0,3621	0,1795	2,747	0,4587

КРИТЕРИЙ ФИШЕРА И СТЬЮДЕНТА. Файл: var0.std

Переменные: vib1, vib2

Статистика Фишера=0,5656, Значимость=0,0737, степ.своб = 31,25

Гипотеза 0: <Нет различий между выборочными дисперсиями>

Статистика Стьюдента=0,492, Значимость=0,6302, степ.своб = 56

Гипотеза 0: <Нет различий между выборочными средними>

ГИСТОГРАММА И ТЕСТ НОРМАЛЬНОСТИ. Файл: var0.std

Х-лев.	Х-станд	Частота	%	Накопл.	%
0,33	-2,547	3	5,172	3	5,172
0,4675	-1,987	2	3,448	5	8,621
0,605	-1,428	4	6,897	9	15,52
0,7425	-0,8681	15	25,86	24	41,38
0,88	-0,3087	10	17,24	34	58,62
1,017	0,2508	10	17,24	44	75,86
1,155	0,8103	10	17,24	54	93,1
1,293	1,37	4	6,897	58	100
1,43	1,929				

Колмогоров=0,05836, Значимость=1,582, степ.своб = 58

Гипотеза 0: <Распределение не отличается от нормального>

Омега-квадрат=0,04288, Значимость=0,6628, степ.своб = 58

Гипотеза 0: < Распределение не отличается от нормального>

Хи-квадрат=8,909, Значимость=0,1127, степ.своб = 5

Гипотеза 0: < Распределение не отличается от нормального>

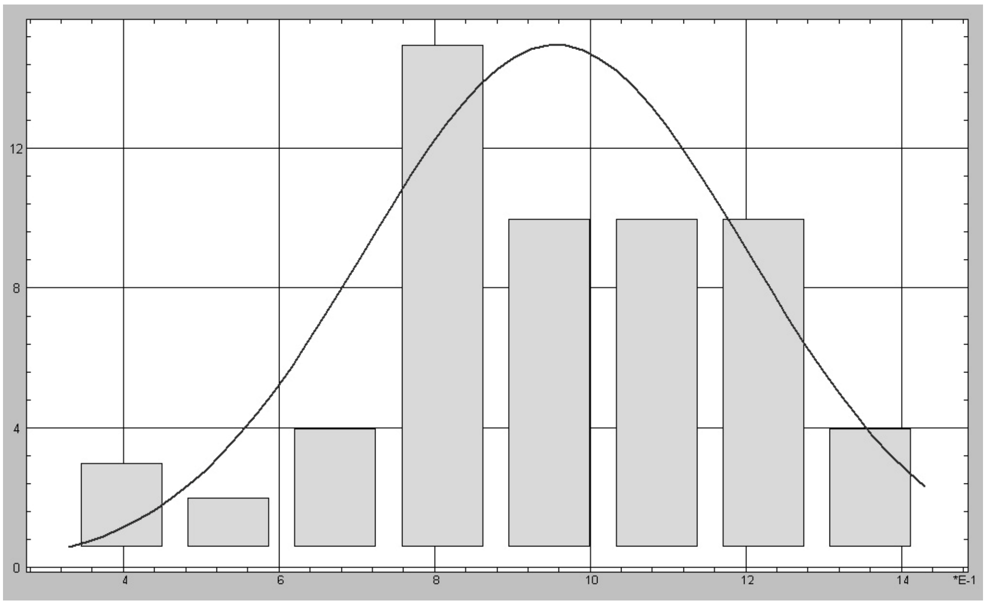


Рисунок 9.2

Лабораторная работа 3. Регрессионный анализ

Используемое ПО: MathCad, Excel, StatGraph, Stadia.

Цель работы: научиться с помощью вышеуказанных программ строить регрессионные модели, проверять их адекватность и значимость коэффициентов.

Задания

1. Найти линейное, квадратичное и кубическое регрессионные уравнения, уравнение вида $y = e^{ax+b}$. Изобразить на одном графике найденную экспоненциальную функцию и результаты эксперимента. Построить график остатков для экспоненциальной модели. Данные считать из файла. В качестве значений y брать средние арифметические. Все веса считать равными 1 (Программа MathCad).
2. Найти линейное уравнение регрессии. В качестве значений y брать средние арифметические. Все веса считать равными 1. Проверить адекватность построенной модели и значимость коэффициентов (Программа Excel).
3. Найти линейное и квадратичное уравнения регрессии с учетом весов. Для каждой регрессионной модели построить график найденной функции и результатов эксперимента. Проверить адекватность каждой построенной модели и значимость всех коэффициентов (Программа StatGraph).
4. Найти линейное, квадратичное и кубическое регрессионные уравнения, уравнение вида $y = e^{ax+b}$. Изобразить на одном графике найденную квадратичную функцию и результаты эксперимента. В качестве значений y брать средние арифметические. Все веса считать равными 1. Проверить адекватность каждой построенной модели и значимость всех коэффициентов (Программа Stadia).
5. Напишите программу в MathCad для построения полноценных линейной и квадратичной регрессионных моделей с учетом весов, проверки адекватности построенных моделей и значимости всех коэффициентов.

Экспериментальные данные следует взять из условия РГР №3.

Некоторые теоретические сведения

1. Для построения линейной регрессионной модели $y = ax + b$ по методу наименьших квадратов коэффициенты *a* и *b* ищут исходя из условия минимизации взвешенной суммы квадратов отклонений результатов эксперимента от точек на прямой: $MinZ = \sum_{i=1}^{n}(ax_i + b - y_i)^2 w_i$. Нахождение минимума этой функции после нахождения частных производных по параметрам *a* и *b* сведется к решению системы линейных уравнений:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 w_i + b\sum_{i=1}^{n} x_i w_i = \sum_{i=1}^{n} x_i y_i w_i \\ a\sum_{i=1}^{n} x_i w_i + b\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} y_i w_i \end{cases}.$$

2. Для построения квадратичной регрессионной модели $y = ax^2 + bx + c$ по методу наименьших квадратов коэффициенты *a*, *b* и *c* ищут исходя из условия минимизации взвешенной суммы квадратов отклонений результатов эксперимента от точек на параболе: $MinZ = \sum_{i=1}^{n}(ax_i^2 + bx_i + c - y_i)^2 w_i$. Нахождение минимума этой функции после нахождения частных производных по параметрам *a*, *b* и *c* сведется к решению системы линейных уравнений:

$$\begin{cases} a\sum_{i=1}^{n} x_i^4 w_i + b\sum_{i=1}^{n} x_i^3 w_i + c\sum_{i=1}^{n} x_i^2 w_i = \sum_{i=1}^{n} x_i^2 y_i w_i \\ a\sum_{i=1}^{n} x_i^3 w_i + b\sum_{i=1}^{n} x_i^2 w_i + c\sum_{i=1}^{n} x_i w_i = \sum_{i=1}^{n} x_i y_i w_i \\ a\sum_{i=1}^{n} x_i^2 w_i + b\sum_{i=1}^{n} x_i w_i + c\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} y_i w_i \end{cases}.$$

3. Аналогично строятся и другие регрессионные модели.

4. Заметим, что из рассматриваемых программ только StatGraph позволяет сразу искать регрессионные модели с учетом весов.

5. Для проверки адекватности построенных моделей в том случае, если

имел место повторный эксперимент, используют следующую схему. Вычисляют $s_{ад}^2 = \sum_{i=1}^{n} \frac{(\Delta Y_i)^2 w_i}{k_{ад}}$, где $\Delta Y_i = (ax_i + b - y_i)$ для линейной модели (a и b – найденные коэффициенты регрессионной модели), $\Delta Y_i = ax_i^2 + bx_i + c - y_i$ для квадратичной модели (a, b и c – найденные коэффициенты регрессионной модели), $k_{ад} = n - l$, где l – число оцениваемых параметров, то есть $l = 2$ для линейной модели, $l = 3$ – для квадратичной. Величина $s_{ад}^2$ (дисперсия адекватности) иногда обозначается $s_{ост}^2$ (остаточная дисперсия). Затем вычисляют $s_{св}^2 = \frac{\sum_{i=1}^{n} s_i^2 k_i}{\sum_{i=1}^{n} k_i}$, где $s_i^2 = \sum_{j=1}^{w_i} \frac{(y_{ij} - \overline{y_i})^2}{k_i}$, $k_i = w_i - 1$.

Если $s_{ад}^2 < s_{св}^2$, то модель является адекватной. Если нет, то находят $F_э = \frac{s_{ад}^2}{s_{св}^2}$ и сравнивают с квантилью распределения Фишера $F_т = F(p, k_{ад}, k_{св})$, где $k_{св} = \sum_{i=1}^{n} k_i$, $p = 1 - \alpha$, а α – уровень значимости. Если $F_э < F_т$, то модель является адекватной на уровне значимости α, в противном случае – нет.

6. Для проверки адекватности построенных моделей в том случае, если не имел место повторный эксперимент, используют следующую схему. Сначала вычисляют $s_{ад}^2$ (см. пункт 5), все веса принимаются равными 1. Затем вычисляют $s_{мод}^2 = \sum_{i=1}^{n} \frac{(\Delta \overline{Y}_i)^2}{l-1}$, где $\Delta \overline{Y}_i = ax_i + b - \overline{y}$ для линейной модели, $\Delta \overline{Y}_i = ax_i^2 + bx_i + c - \overline{y}$ – для квадратичной, а $\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}$. Величина $s_{мод}^2$ (дисперсия модели) иногда обозначается $s_{рег}^2$ (регрессионная дисперсия). После этого сравнивают величину $F_э = \frac{s_{мод}^2}{s_{ад}^2}$ с квантилью распределения Фишера $F_т = F(p, l-1, n-l)$. Если окажется, что $F_э < F_т$, то

гипотеза об адекватности модели экспериментальным данным отвергается (на уровне значимости α), в противном случае гипотеза принимается.

7. Коэффициент уравнения регрессии называется незначимым, если его математическое ожидание (истинное значение) равно нулю. Для проверки значимости найденных коэффициентов регрессионной модели находят отношения этих коэффициентов к их стандартным ошибкам. Если через A обозначить матрицу системы для нахождения коэффициентов регрессионной модели, то стандартную ошибку i-го коэффициента d_i можно найти так:

$$s(d_i) = \sqrt{s_{ад}^2 \cdot b_{ii}},$$

где b_{ii} – соответствующий элемент матрицы $B = A^{-1}$.

Если модуль этого отношения больше квантили модуля отношения Стьюдента для числа степеней свободы $k_{ад}$, то есть $\left|\frac{d_i}{s(d_i)}\right| > |t|_p \ (k = k_{ад})$, то коэффициент считается значимым, в противном случае – нет.

Пояснения к работе с программой MathCad (задание 1)

1. Следует подготовить два текстовых файла, например, в «Блокноте» со значениями *x* и средних арифметических *y*.

2. Укажем некоторые команды, которые нам понадобятся для выполнения работы.

slope(*x*,*y*)

Возвращает коэффициент при *x* в линейной регрессионной модели.

intercept(*x*,*y*)

Возвращает свободный член в линейной регрессионной модели.

regress(*x*,*y*,*l*–1)

Возвращает вектор, последние *l* координат которого – коэффициенты степенной регрессионной модели степени *l*–1.

3. MathCad позволяет искать практически любые регрессионные модели, даже не линейные относительно параметров. Для этого необходимо задать вектор, координатами которого являются приближающая функция и ее частные производные по параметрам (в примере – это вектор *F*). Затем следует задать начальное приближение для параметров (в примере – это вектор *vb*). Начальное приближение можно брать наугад, но затем всегда следует проверять правильность построенной модели графически, так как в случае неудачно подобранного начального приближения программа может выдать неправильный результат. Далее надо воспользоваться командой **genfit(*x*,*y*,*vb*,*F*)**, которая возвратит коэффициенты регрессионной модели.

4. Индексация в массивах в MathCad'е начинается с нуля.

Пояснения к работе с программой Excel

1. Для построения линейной регрессионной модели в программе Excel следует воспользоваться командой **ЛИНЕЙН(*Y*,*X*)**. И так как результатом этой команды является массив (из двух чисел), то, указав диапазон ячеек для *x* и для *y*, следует нажать комбинацию клавиш Ctrl+Shift+Enter.

2. Если же указать диапазон ячеек из двух столбцов и пяти строк, то программа выдаст дополнительную статистику, а именно:

Коэффициент *a* линейной модели	Коэффициент *b* линейной модели
Стандартная ошибка коэффициента *a*	Стандартная ошибка коэффициента *b*
Коэффициент детерминации	Стандартная ошибка для *y* ($s_{ад}$)
F-наблюдаемое ($F_э$)	Степени свободы ($k_{ад} = n - l$)
Регрессионная сумма квадратов ($(\Delta\overline{Y}_i)^2$)	Остаточная сумма квадратов ($(\Delta Y_i)^2$)

Пояснения к работе с программой StatGraph

1. Для построения линейной регрессионной модели воспользуйтесь пунктом меню «*Relate / Multiple Regression…*», для построения квадратичной модели – «*Relate / Polynomial Regression…*».

2. Столбец «p-Value» из первой таблицы показывает уровни значимости коэффициентов регрессионной модели. Если уровень значимости некоторого коэффициента не превосходит заданного уровня значимости α, то этот коэффициент считается значимым, в противном случае – незначимым. Столбец «p-Value» из второй таблицы показывает уровень значимости $F_Э$. Если уровень значимости $F_Э$ не превосходит заданного уровня значимости α, то модель считается адекватной, в противном случае – неадекватной.

Пояснения к работе с программой Stadia

1. Следует подготовить два столбца с экспериментальными данными, один – со значениями *x*, второй – со средними арифметическими для *y*.

2. В качестве десятичного разделителя следует использовать точку.

3. Для выполнения задания воспользуемся пунктом меню «*Статист-F9*». В появившемся окне следует нажать на кнопку «*Простая регрессия/ тренд*», затем выбрать вид регрессионной модели.

4. Строка «Значим.» из первой таблицы показывает уровни значимости коэффициентов регрессионной модели. Вывод относительно адекватности построенной модели выдается текстовым сообщением.

Пояснения к работе с программой MathCad (задание 5)

1. Для выполнения работы понадобятся панели инструментов «*Арифметика*», «*Матанализ*», «*Программирование*», «*Матрицы*», «*Булево*». Вывести их на экран можно через пункт меню «*Вид / Панель инструментов*».

2. Вновь будем составлять по возможности универсальную программу для своего задания.

3. Перед тем, как приступить к работе с MathCad’ом, следует подготовить два текстовых файла, например, с помощью программы

Блокнот. Один – со значениями *x* (в нашем примере эти значения записаны в столбец), второй – со значениями *y* (в нашем примере для фиксированного *x* соответствующие значения *y* записаны в строки). В качестве десятичного разделителя следует использовать точку. Сохраните эти текстовые файлы в той же папке, где будет хранится файл, созданный в MathCad'е.

4. Для выполнения работы нам понадобятся кнопки «*Add Line*», «*Локальное присвоение*», «*Цикл For*», «*Оператор If*» (панель «*Программирование*»); «*Суммирование*» (панель «*Матанализ*»); «*Булево равенство*» (панель «*Булево*»); «*Создать матрицу или вектор*», «*Нижний индекс*» (панель «*Матрицы*»).

5. Команда **lsolve(*a*,*b*)** возвращает вектор, являющийся решением системы линейных уравнений с матрицей системы *a* и вектором свободных членов *b*.

6. Напомним, что индексация в массивах в MathCad'е начинается с нуля.

Разбор нулевого варианта

Таблица 9.1 – Содержание нулевого варианта к лабораторной работе 3

Нулевой вариант					
X	*Y*				
10	50,7	50,6	50,6		
20	49,9	50,0	50,1	50,2	
40	48,3	48,4	48,4		
60	47,4	47,4	46,8	47,4	
80	46,0	45,9	46,2	46,1	45,5
120	43,2	43,3	43,1	43,5	43,0
140	41,9	42,2	41,9	42,2	
150	41,7	41,7	42,0		
180	40,2	40,4	40,5	40,1	
200	39,5	39,9	39,7		

Программа MathCad (задание 1)

Разбор нулевого варианта

Подготовка

X := READPRN("Var0-x.txt")   Y := READPRN("Var0-ys.txt")

	X = 0
0	10
1	20
2	40
3	60
4	80
5	120
6	140
7	150
8	180
9	200

	Y = 0
0	50.633
1	50.05
2	48.367
3	47.25
4	45.94
5	43.22
6	42.05
7	41.8
8	40.3
9	39.7

Линейное уравнение регрессии

slope(X, Y) = –0.060   intercept(X, Y) = 50.886

Ответ y=-0.060x+50.886

Квадратичное и кубическое уравнения

Z1 := regress(X, Y, 2)   Z2 := regress(X, Y, 3)

$$Z1 = \begin{pmatrix} 3 \\ 3 \\ 2 \\ 51.566 \\ -0.081 \\ 1.073\times 10^{-4} \end{pmatrix} \qquad Z2 = \begin{pmatrix} 3 \\ 3 \\ 3 \\ 51.316 \\ -0.067 \\ -6.525\times 10^{-5} \\ 5.453\times 10^{-7} \end{pmatrix}$$

Ответ y=0.0001073x^2-0.081x+51.566

Ответ y=0.0000005453x^3-0.00006525x^2-0.067x+51.316

Экспоненциальное уравнение регрессии, график

$$F(w,u) := \begin{pmatrix} e^{u_0\cdot w+u_1} \\ w\cdot e^{u_0\cdot w+u_1} \\ e^{u_0\cdot w+u_1} \end{pmatrix} \qquad vb := \begin{pmatrix} 0.1 \\ 1 \end{pmatrix} \qquad p := \mathrm{genfit}(X, Y, vb, F) \qquad p = \begin{pmatrix} -1.336\times 10^{-3} \\ 3.935 \end{pmatrix}$$

$g(r) := F(r,p)_0$ $i := 0..9$

Ответ y=e^(-0.001336x+3.935)

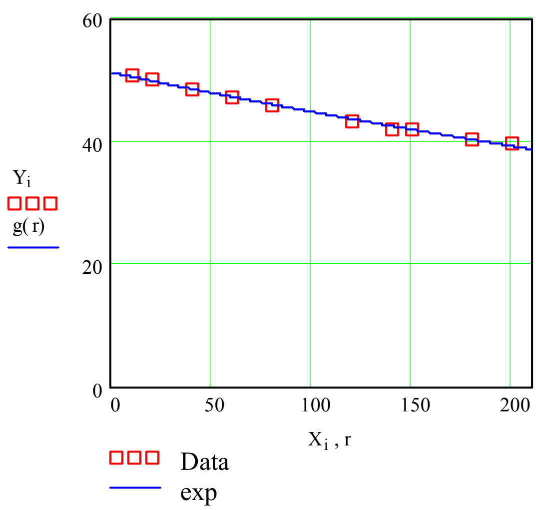
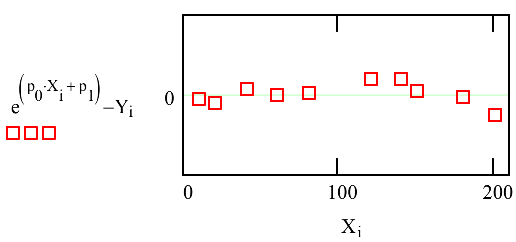


График остатков



Программа Excel

Разбор нулевого варианта

X	Y				
10	50,633333				
20	50,05				
40	48,366667				
60	47,25				
80	45,94				
120	43,22				
140	42,05				
150	41,8				
180	40,3				
200	39,7				
Линейная модель:					
-0,0595463	50,8856341	**Коэффициенты**	-28,984815	208,593654	**Отношения коэффициентов к их ошибкам**
0,0020544	0,24394622	**Их стандартные ошибки**	2,30600563	**|t|(k=8)**	**Оба коэф. знач., т.к. 29,0>2,3 и 208,6>2,3**
0,99056737	0,41598385	**Коэф. детерм. / Станд. ош. для y**			
840,119521	8	**F-наблюдаемое / Степени свободы**	5,31764499	**F(1;8)**	**Модель адекватна, т.к. 5,3<840,1**
145,376437	1,38434052	**Регрес. / Остаточ. суммы квадратов**			

Программа StatGraph

Разбор нулевого варианта

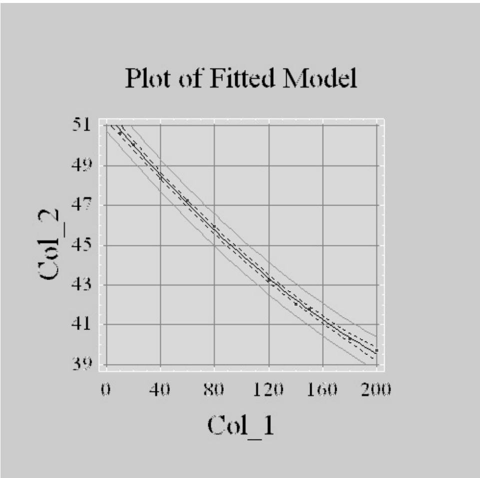


Рисунок 9.3

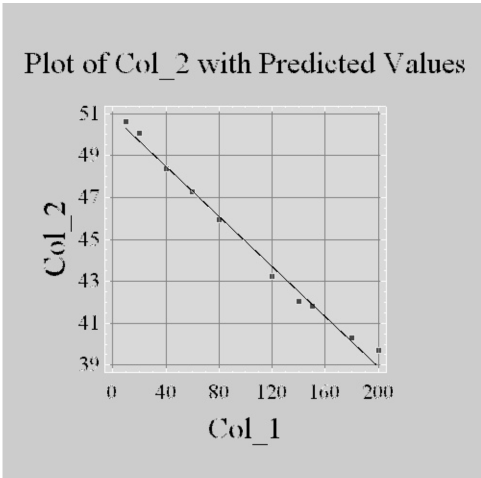


Рисунок 9.4

Multiple Regression Analysis

Dependent variable: Col_2

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	50,8814	0,248811	204,498	0,0000
Col_1	-0,0598933	0,00212574	-28,1753	0,0000

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	504,362	1	504,362	793,85	0,0000
Residual	5,0827	8	0,635337		
Total (Corr.)	509,444	9			

R-squared = 99,0023 percent
R-squared (adjusted for d.f.) = 98,8776 percent
Standard Error of Est. = 0,79708
Mean absolute error = 0,300386
Durbin-Watson statistic = 0,658416

The StatAdvisor

The output shows the results of fitting a multiple linear regression model to describe the relationship between Col_2 and 1 independent variables. The equation of the fitted model is

Col_2 = 50,8814 - 0,0598933*Col_1

Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant relationship between the variables at the 99% confidence level.

The R-Squared statistic indicates that the model as fitted explains 99,0023% of the variability in Col_2. The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 98,8776%. The standard error of the estimate shows the standard deviation of the residuals to be 0,79708. This value can be used to construct prediction limits for new observations by selecting the Reports option from the text menu. The mean absolute error (MAE) of 0,300386 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the DW value is less than 1.4, there may be some indication of serial correlation. Plot the residuals versus row order to see if there is any pattern which can be seen.

In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0,0000, belonging to Col_1. Since the P-value is less than 0.01, the highest order term is statistically significant at the 99% confidence level. Consequently, you probably don't want to remove any variables from the model.

Polynomial Regression Analysis

Dependent variable: Col_2

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	51,5913	0,154053	334,894	0,0000
Col_1	-0,0815422	0,00358172	-22,7662	0,0000
Col_1^2	0,000106205	0,0000170235	6,23871	0,0004

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	508,67	2	254,335	2297,88	0,0000
Residual	0,774776	7	0,110682		
Total (Corr.)	509,444	9			

R-squared = 99,8479 percent
R-squared (adjusted for d.f.) = 99,8045 percent
Standard Error of Est. = 0,33269
Mean absolute error = 0,25194
Durbin-Watson statistic = 1,9561

The StatAdvisor

The output shows the results of fitting a second order polynomial model to describe the relationship between Col_2 and Col_1. The equation of the fitted model is

Col_2 = 51,5913-0,0815422*Col_1 + 0,000106205*Col_1^2

Since the P-value in the ANOVA table is less than 0.01, there is a statistically significant relationship between Col_2 and Col_1 at the 99% confidence level.

The R-Squared statistic indicates that the model as fitted explains 99,8479% of the variability in Col_2. The adjusted R-squared statistic, which is more suitable for comparing models with different numbers of independent variables, is 99,8045%. The standard error of the estimate shows the standard deviation of the residuals to be 0,33269. This value can be used to construct prediction limits for new observations by selecting the Forecasts option from the text menu. The mean absolute error (MAE) of 0,25194 is the average value of the residuals. The Durbin-Watson (DW) statistic tests the residuals to determine if there is any significant correlation based on the order in which they occur in your data file. Since the DW value is greater than 1.4, there is probably not any serious autocorrelation in the residuals.

In determining whether the order of the polynomial is appropriate, note first that the P-value on the highest order term of the polynomial equals 0,000428817. Since the P-value is less than 0.01, the highest order term is statistically significant at the 99% confidence level. Consequently, you probably don't want to consider any model of lower order.

Программа Stadia

Разбор нулевого варианта

ПРОСТАЯ РЕГРЕССИЯ. Файл var0.std

Переменные: x, y

Модель: линейная Y = a0+a1*x

Коэфф.	a0	a1
Значение	50,89	-0,05955
Ст. ошиб.	0,2439	0,002054
Значим.	0	0

Источник	Сум. квадр.	Степ. св	Средн. квадр.
Регресс.	145,4	1	145,4
Остаточн.	1,384	8	0,173
Вся	146,8	9	

Множеств R	R^2	R^2прив	Ст. ошиб.	F	Значим.
0,99527	0,99057	0,98939	0,41598	840,1	0

Гипотеза 1: <Регрессионная модель адекватна экспериментальным данным>

Модель: парабола Y = a0+a1*x+a2*x^2

Коэфф.	a0	a1	a2
Значение	51,57	-0,08147	0,0001073
Ст. ошиб.	0,1431	0,003461	1,644E-5
Значим.	0	0	0,0006

Источник	Сум. квадр.	Степ. св	Средн. квадр.
Регресс.	146,6	2	73,28
Остаточн.	0,1955	7	0,02793
Вся	146,8	9	

Множеств R	R^2	R^2прив	Ст. ошиб.	F	Значим.
0,99933	0,99867	0,99829	0,16711	2624	0

Гипотеза 1: <Регрессионная модель адекватна экспериментальным данным>

Модель: полином Y = сумма{ai*x^i}

Коэфф.	a0	a1	a2	a3
Значение	51,32	-0,0669	-6,525E-5	5,453E-7
Ст. ошиб.	0,1635	0,00729	8,081E-5	2,519E-7
Значим.	0	0,0003	0,5453	0,072

Источник	Сум. квадр.	Степ. св	Средн. квадр.
Регресс.	146,7	3	48,88
Остаточн.	0,1098	6	0,01829
Вся	146,8	9	

Множеств R	R^2	R^2прив	Ст. ошиб.	F	Значим.
0,99963	0,99925	0,99888	0,13526	2672	0

Гипотеза 1: <Регрессионная модель адекватна экспериментальным данным>

Модель: экспонента Y = EXP(a0+a1*x)

Коэфф.	a0	a1
Значение	3,934	-0,001326
Ст. ошиб.	0,004036	3,399E-5
Значим.	0	0

Источник	Сум. квадр.	Степ. св.	Средн. квадр.
Регресс.	0,07213	1	0,07213
Остаточн.	0,0003788	8	4,736E-5
Вся	0,07251	9	

Множеств R	R^2	R^2прив	Ст. ошиб.	F	Значим.
0,99738	0,99478	0,99412	0,0068816	1523	0

Гипотеза 1: <Регрессионная модель адекватна экспериментальным данным>

Программа Mathcad (задание 5)

Разбор нулевого варианта

Подготовка

x := READPRN("Var0-x.txt")

y := READPRN("Var0-y.txt")

n := 10   p := 0.95

$$ysr := \left|\begin{array}{l} \text{for } i \in 0..n-1 \\ \quad ysr_i \leftarrow \sum_{j=0}^{w_i-1} \frac{y_{i,j}}{w_i} \\ ysr \end{array}\right.$$

$$w := \begin{pmatrix} 3 \\ 4 \\ 3 \\ 4 \\ 5 \\ 5 \\ 4 \\ 3 \\ 4 \\ 3 \end{pmatrix}$$

x =

	0
0	10
1	20
2	40
3	60
4	80
5	120
6	140
7	150
8	180
9	200

y =

	0	1	2	3	4
0	50.7	50.6	50.6	0	0
1	49.9	50	50.1	50.2	0
2	48.3	48.4	48.4	0	0
3	47.4	47.4	46.8	47.4	0
4	46	45.9	46.2	46.1	45.5
5	43.2	43.3	43.1	43.5	43
6	41.9	42.2	41.9	42.2	0
7	41.7	41.7	42	0	0
8	40.2	40.4	40.5	40.1	0
9	39.5	39.9	39.7	0	0

Построение линейной модели

$$a1 := \begin{bmatrix} \sum_{i=0}^{n-1} (x_i)^2 \cdot w_i & \sum_{i=0}^{n-1} x_i \cdot w_i \\ \sum_{i=0}^{n-1} x_i \cdot w_i & \sum_{i=0}^{n-1} w_i \end{bmatrix} \qquad a2 := \begin{pmatrix} \sum_{i=0}^{n-1} x_i \cdot w_i \cdot ysr_i \\ \sum_{i=0}^{n-1} ysr_i \cdot w_i \end{pmatrix}$$

a := lsolve(a1, a2)

$$a = \begin{pmatrix} -0.060 \\ 50.881 \end{pmatrix}$$

$$b := a1^{-1}$$

Ответ: y=-0.060x+50.881

Построение квадратичной модели

$$A1 := \begin{bmatrix} \sum_{i=0}^{n-1} (x_i)^4 \cdot w_i & \sum_{i=0}^{n-1} (x_i)^3 \cdot w_i & \sum_{i=0}^{n-1} (x_i)^2 \cdot w_i \\ \sum_{i=0}^{n-1} (x_i)^3 \cdot w_i & \sum_{i=0}^{n-1} (x_i)^2 \cdot w_i & \sum_{i=0}^{n-1} (x_i) \cdot w_i \\ \sum_{i=0}^{n-1} (x_i)^2 \cdot w_i & \sum_{i=0}^{n-1} (x_i) \cdot w_i & \sum_{i=0}^{n-1} w_i \end{bmatrix} \qquad A2 := \begin{bmatrix} \sum_{i=0}^{n-1} (x_i)^2 \cdot w_i \cdot ysr_i \\ \sum_{i=0}^{n-1} x_i \cdot w_i \cdot ysr_i \\ \sum_{i=0}^{n-1} w_i \cdot ysr_i \end{bmatrix}$$

A := lsolve(A1, A2)

$$A = \begin{pmatrix} 1.062 \times 10^{-4} \\ -0.082 \\ 51.591 \end{pmatrix}$$

$$B := A1^{-1}$$

Ответ: y=0.0001062x^2-0.082x+51.591

Проверка адекватности обеих моделей

$$Sad1 := \sum_{i=0}^{n-1} \left(a_0 \cdot x_i + a_1 - ysr_i\right)^2 \cdot w_i \qquad Sad1 := \frac{1}{n-2} \cdot Sad1 \qquad Sad1 = 0.6353$$

$$Sad2 := \sum_{i=0}^{n-1} \left[A_0 \cdot (x_i)^2 + A_1 \cdot x_i + A_2 - ysr_i\right]^2 \cdot w_i \qquad Sad2 := \frac{1}{n-3} \cdot Sad2 \qquad Sad2 = 0.1107$$

$$Ssv := \frac{\displaystyle\sum_{i=0}^{n-1} \left[\sum_{j=0}^{w_i-1} \frac{\left(y_{i,j} - ysr_i\right)^2}{w_i - 1}\right] \cdot (w_i - 1)}{\displaystyle\sum_{i=0}^{n-1} (w_i - 1)} \qquad Ssv = 0.0394$$

$$F1 := \frac{Sad1}{Ssv} \qquad F1 = 16.1234 \qquad F2 := \frac{Sad2}{Ssv} \qquad F2 = 2.8089$$

$$Fkv1 := qF\left[p, n-2, \sum_{i=0}^{n-1} (w_i - 1)\right] \qquad Fkv2 := qF\left[p, n-3, \sum_{i=0}^{n-1} (w_i - 1)\right]$$

$Fkv1 = 2.291$ $\qquad$ $Fkv2 = 2.359$

$$FlagLi1 := \left| \begin{array}{l} \text{"Yes" if } (Sad1 < Ssv) \\ \text{"No" if } (Sad1 > Ssv) \end{array}\right. \qquad FlagSq1 := \left| \begin{array}{l} \text{"Yes" if } (Sad2 < Ssv) \\ \text{"No" if } (Sad2 > Ssv) \end{array}\right.$$

$$FlagLi2 := \left| \begin{array}{l} \text{"Yes" if } F1 < Fkv1 \\ \text{"No" if } F1 > Fkv1 \end{array}\right. \qquad FlagSq2 := \left| \begin{array}{l} \text{"Yes" if } F2 < Fkv2 \\ \text{"No" if } F2 > Fkv2 \end{array}\right.$$

$$FlagLi := \left| \begin{array}{l} \text{"Yes" if } (FlagLi1 = \text{"Yes"}) \\ FlagLi2 \text{ if } FlagLi1 \neq \text{"Yes"} \end{array}\right. \qquad FlagSq := \left| \begin{array}{l} \text{"Yes" if } (FlagSq1 = \text{"Yes"}) \\ FlagSq2 \text{ if } FlagSq1 \neq \text{"Yes"} \end{array}\right.$$

FlagLi = "No" $\qquad$ FlagSq = "No"

Ответ: обе модели не адекватны

Проверка значимости коэффициентов

Стандартные ошибки коэффициентов

$s11 := \sqrt{Sad1} \cdot \sqrt{b_{0,0}}$ $\quad s11 = 2.126 \times 10^{-3}$ $\quad s12 := \sqrt{Sad1} \cdot \sqrt{b_{1,1}}$ $\quad s12 = 0.249$

$s21 := \sqrt{Sad2} \cdot \sqrt{B_{0,0}}$ $\quad s21 = 1.702 \times 10^{-5}$ $\quad s22 := \sqrt{Sad2} \cdot \sqrt{B_{1,1}}$ $\quad s22 = 3.582 \times 10^{-3}$

$s23 := \sqrt{Sad2} \cdot \sqrt{B_{2,2}}$ $\quad s23 = 0.154$

Отношения коэффициентов к их ошибкам

$t11 := \frac{a_0}{s11} \qquad t12 := \frac{a_1}{s12} \qquad t21 := \frac{A_0}{s21} \qquad t22 := \frac{A_1}{s22} \qquad t23 := \frac{A_2}{s23}$

Проверка значимости

$Tkv1 := -qt\left(\frac{1-p}{2}, n-2\right)$ $\quad Tkv1 = 2.306$ $\qquad Tkv2 := -qt\left(\frac{1-p}{2}, n-3\right)$ $\quad Tkv2 = 2.365$

$$Flag11 := \left| \begin{array}{l} \text{"a1\_Znachim"} \;\; \text{if } (|t11| > Tkv1) \\ \text{"a1\_Ne\_Znachim"} \;\; \text{if } (|t11| < Tkv1) \end{array} \right. \qquad Flag12 := \left| \begin{array}{l} \text{"a2\_Znachim"} \;\; \text{if } (|t12| > Tkv1) \\ \text{"a2\_Ne\_Znachim"} \;\; \text{if } (|t12| < Tkv1) \end{array} \right.$$

$$Flag21 := \left| \begin{array}{l} \text{"A1\_Znachim"} \;\; \text{if } (|t21| > Tkv2) \\ \text{"A1\_Ne\_Znachim"} \;\; \text{if } (|t21| < Tkv2) \end{array} \right. \qquad Flag22 := \left| \begin{array}{l} \text{"A2\_Znachim"} \;\; \text{if } (|t22| > Tkv2) \\ \text{"A2\_Ne\_Znachim"} \;\; \text{if } (|t22| < Tkv2) \end{array} \right.$$

$$Flag23 := \left| \begin{array}{l} \text{"A3\_Znachim"} \;\; \text{if } (|t23| > Tkv2) \\ \text{"A3\_Ne\_Znachim"} \;\; \text{if } (|t23| < Tkv2) \end{array} \right.$$

Flag11 = "a1_Znachim"

Flag12 = "a2_Znachim"

Flag21 = "A1_Znachim"

Flag22 = "A2_Znachim"

Flag23 = "A3_Znachim"

Ответ: все коэффициенты значимы

Библиографический список

1. Румшинский Л. З. Организация эксперимента [Текст] : пособие / Л. З. Румшинский. – М. : ротапринт МИСиС, 1984. – 140 с.
2. Вентцель Е. С. Теория вероятностей [Текст] : учебник / Е. С. Вентцель. – М. : Наука, 1969. – 576 с.
3. Карасев В. А. Организация эксперимента [Текст] : пособие / В. А. Карасев, Л. З. Румшинский. – М. : ротапринт МИСиС, 1986. – 86 с.
4. Тюрин Ю. Н. Анализ данных на компьютере [Текст] : учебник / Ю. Н. Тюрин, А. А. Макаров. – М. : Инфра-М, 2003. – 544 с.
5. Изаак Д. Д. Математическая статистика [Текст] : лабораторный практикум / Д. Д. Изаак, А. В. Швалева. – Орск : изд-во ОГТИ, 2013. – 51 с.
6. Большев Л. Н. Таблицы математической статистики [Текст] : изд-е 3-е / Л. Н. Большев, Смирнов Н.В.. – М. : Наука, 1983 – 416 с.
7. Чертежи сделаны с помощью авторской программы Изаака Д.Д. «Януш» версии 5.2.

ИЗААК ДМИТРИЙ ДАВИДОВИЧ

ISBN 978-5-9905230-2-9

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Учебно-методическое пособие
для студентов обучающихся по направлениям
13.03.01 «Теплоэнергетика и теплотехника»;
13.03.02 «Электроэнергетика и электротехника»;
22.03.02 «Металлургия»;
15.03.02 «Технологические машины и оборудование»;
18.03.01 «Химическая технология»

Подписано в печать 16.12.2015		
Формат 60x90 $^1/_{16}$	Печать офсетная	Уч.-изд.л. 9,43
Рег.№ 76	Тираж 100 экз.	

Национальный исследовательский технологический университет «МИСиС»
Новотроицкий филиал
462359, Оренбургская обл., г. Новотроицк, ул. Фрунзе, 8.
E-mail: nfmisis@yandex.ru