Н.М. Новикова С.Л. Подвальный

ПРИКЛАДНАЯ МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
Часть 2

Учебное пособие

Воронеж 2013

УДК 681.3

Новикова Н.М. Прикладная математическая статистика учеб. пособие / Н.М. Новикова, С.Л. Подвальный. Воронеж: ФГБОУ ВПО «Воронежский государственный технический университет», 2013. Ч.2. 179 с.

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

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

Табл. 11. Ил. 30. Библиогр.: 14 назв.

Рецензенты кафедра цифровых технологий Воронежского государственного университета (зав. кафедрой д-р физ.-мат. наук, проф. С.Д. Кургалин);
д-р физ.-мат.наук, проф. В.В. Провоторов

ВВЕДЕНИЕ

Прикладная математическая статистика – это математическая дисциплина, основанная на теории вероятностей. Математическая статистика учит тому, как нужно обрабатывать наблюдения, чтобы получить из них наиболее полную информацию, и как оценить степень достоверности полученных выводов. Основу прикладной математической статистики составляют методы сбора и обработки статистических данных с целью использования полученных результатов для практических и научных выводов.

Прикладная математическая статистика решает следующие задачи:

- указать способы сбора и группировки статистических данных, полученных в результате специально поставленных экспериментов;
- разработать методы анализа статистических данных в зависимости от целей исследования. Сюда относятся: оценка неизвестной вероятности события, оценка неизвестной функции распределения, оценка параметров распределения, вид которого известен;
- оценить достоверность полученных результатов, используя проверку статистических гипотез о виде неизвестного распределения или о величине параметров распределения, вид которого известен.

Первая часть учебного пособия посвящена решению первых двух задач, а вторая часть – решению третьей задачи. Во второй части учебного пособия рассмотрены методы проверки статистических гипотез, элементы регрессионного, дисперсионного и кластерного анализа.

Правильность исходных предпосылок математической статистики, как и всякой другой прикладной теории, проверяется практикой. В настоящее время трудно найти

такую область знаний, где в той или иной мере не применялись бы методы математической статистики.

Сюда, наряду с естественными отраслями науки и техники, такими, как физика, химия, компьютерные науки, можно отнести и далекие от математики области: историю, психологию, генетику, социологию, лингвистику и другие. Поэтому необходимо овладение методами прикладной математической статистики, которым посвящено данное учебное пособие.

6. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

6.1.Статистические гипотезы

Статистической гипотезой (или просто **гипотезой**) называют любое утверждение о виде или свойствах распределения случайных величин, наблюдаемых в эксперименте.

Пусть эксперимент состоит в многократном измерении некоторой физической величины, точное значение *a* которой неизвестно и в процессе измерения не изменяется. На результаты измерений влияют многие случайные факторы (точность настройки измерительных приборов, погрешность округления при считывании данных и т.д.). Поэтому результат i-го измерения X_i можно записать в виде $X_i=a+\varepsilon_i$, где ε_i - случайная погрешность измерения. Считают, что общая ошибка ε_i складывается из большого числа ошибок, каждая из которых невелика. На основании центральной предельной теоремы (ЦПТ) предполагается, что случайные величины X_i имеют нормальное распределение. Такое предположение является статической гипотезой о виде распределения наблюдаемых случайных величин.

Если для исследуемого явления сформулирована гипотеза - обычно её называют **основной** или **нулевой** гипотезой и обозначают H_0 - то задача в том, чтобы по статистическим данным (или результатам соответствующих наблюдений) принять или отклонить эту гипотезу. Правило, по которому гипотеза H_0 принимается или отвергается, называется **статистическим критерием** (или просто **критерием**) проверки гипотезы H_0.

Если результат эксперимента описывается в терминах некоторой случайной величины (выборки) $\vec{X}$ и $F=\{F\}$ - семейство распределений рассматриваемой статистической модели (т.е. множества априори допустимых в данной

ситуации распределений выборки), то в общем случае гипотеза Но задаётся указанием соответствующего класса $F_0 \in F$, которому, по предположению, принадлежит истинное распределение выборки. Это записывается так: $H_0: F_x \in F_0$. Распределения из дополнительного класса $F_1 = F \setminus F_0$ называют **альтернативными распределениями** или **альтернативами.** Утверждение же вида H_1: $F_x \in F_1$ называют **альтернативной гипотезой.** В этих терминах в общем случае задача формулируется как задача проверки H_0 против альтернативы H_1. Гипотеза H_0 (или альтернатива H_1) называется **простой**, если соответствующий класс F_0 (F_1) содержит одно распределение. **Сложной** гипотезе соответствует более чем одно распределение. Гипотезы о параметрах распределения называются **параметрическими.**

6.2. Статистические критерии проверки гипотез

Критерий - это правило, которое для каждой реализации $\vec{x}$ выборки $\vec{X}$ должно приводить к одному из двух решений: решение γ_0 - принять гипотезу H_0 и решение γ_1 -отклонить H_0 (принять H_1). Каждому критерию соответствует разбиение выборочного пространства $\mathbb{X}$ на два взаимно дополнительных множества $\mathbb{X}_0$ и $\mathbb{X}_1$ ($\mathbb{X}_0 \cap \mathbb{X}_1 = \varnothing$, $\mathbb{X}_0 \cup \mathbb{X}_1 = \mathbb{X}$) , где $\mathbb{X}_0$ состоит из точек $\vec{x}$, для которых гипотеза H_0 принимается, а $\mathbb{X}_1$ из точек, для которых H_0 отвергается. Множество $\mathbb{X}_0$ называют **областью принятия гипотезы** H_0 (или **допустимой** областью), а $\mathbb{X}_1$ областью её отклонения или **критической областью**. Таким образом, выбор правила проверки гипотезы H_0 эквивалентен заданию критической области $\mathbb{X}_1$. Если критическая область $\mathbb{X}_1$ выбрана, то критерий формулируется так: пусть $\vec{x}$ - реализация выборки $\vec{X}$, тогда при $\vec{x} \in \mathbb{X}_1$ гипотезу H_0 отвергают [принимают H_1], если же $\vec{x} \in \mathbb{X}_0$, то гипотезу H_0 принимают.

В некоторых ситуациях рассматривают **рандомизированные** критерии, когда при наблюдении $\vec{x}$ гипотезу H_0 отвергают с некоторой вероятностью $\varphi(\vec{x})$ и принимают с дополнительной вероятностью $1-\varphi(\vec{x})$. Рандомизированный критерий характеризуется критической функцией $\varphi(\vec{x})$ $\{0\le\varphi(\vec{x})\le 1,\ \forall \vec{x}\in \mathbb{X}\}$. Если $\varphi(\vec{x})$ принимает только два значения: 0 и 1, то приходим к **нерандомизированному критерию** с критической областью $\mathbb{X}_1=\{\vec{x}:\ \varphi(\vec{x})=1\}$. Будем далее рассматривать нерандомизированные критерии, которые используют на практике.

6.3. Общий принцип выбора критической области критерия

В процессе проверки гипотезы H_0 можно прийти к правильному решению или совершить ошибку первого рода - отклонить H_0, и когда она верна, или ошибку второго рода - принять H_0, когда она ложна. Иными словами ошибка первого рода имеет место, если точка $\vec{x}$ попадает в критическую область $\mathbb{X}_1$ в то время, как верна нулевая гипотеза H_0, а ошибка второго рода, когда $\vec{x}\in \mathbb{X}_0$, но гипотеза H_0 не верна (верна альтернатива H_1).

Вероятности этих ошибок можно выразить через **функцию мощности** W(F) критерия $\mathbb{X}_1$: $W(F)=W(\mathbb{X}_1;F)=P(\vec{X}\in\mathbb{X}_1|F)$, $F\in F$, т.е. вероятность попадания в критическую область, когда F - истинное распределение выборки. Вероятности ошибок можно представить так:

1. Ошибка первого рода - принимается решение γ_1: гипотеза H_0 не справедлива, когда она на самом деле справедлива

$$P(\gamma_1|H_0)=P(\vec{X}\in\mathbb{X}_1|H_0)=W(F),\ F\in F_0\ . \qquad (6.1)$$

2. Ошибка второго рода - принятие решения γ_0, когда справедлива альтернатива H_1

$$P(\gamma_0|H_1)=P(\vec{X}\in \text{Ж}_0|H_1)=1-P(\vec{X}\in \text{Ж}_1|H_1)=1-W(F),\ F\in F_1. \quad (6.2)$$

Желательно провести проверку гипотезы так, чтобы свести к минимуму вероятности обоих типов ошибок. На практике в общем случае сделать это невозможно. Рациональный принцип выбора критической области можно сформулировать так: при заданном числе испытаний n устанавливается граница для вероятности ошибки первого рода и при этом выбирается критическая область Ж_1, для которой вероятность ошибки второго рода минимальна. Выбирается число α между 0 и 1 и налагается условие

$$W(F)\le\alpha,\ \forall F\in F_0\,. \quad (6.3)$$

При этом условии желательно сделать минимальной величину 1-W(F) для всех $F\in F_0$ (за счет выбора критической области Ж_1). Или, что то же самое, сделать максимальной мощность

$$W(F),\ \forall F\in F_0\,. \quad (6.4)$$

Величину α в формуле (6.3) называют **уровнем значимости**, критерий обозначают Ж_1.

Обычно выбирают одно из следующих стандартных значений: α=0.05; 0.01; 0.1.

В терминах функции мощности W(F) можно сказать, что критерий тем лучше, чем больше его мощность при альтернативах. Действительно, если наблюдавшееся значение $\vec{x}$ выборки попадает в критическую область, то H_0 (нулевую гипотезу) отклоняют, и если истинной является альтернатива, то тем самым принимают правильное решение. Обычно критическая область задаётся с помощью некоторой статистики $T(\vec{X})$ и имеет следующий вид: $\text{Ж}_1=\{\vec{x}:\ T(\vec{x})\ge c\}$ или $\text{Ж}_1=\{\vec{x}: T(\vec{x})\le c\}$, или $\text{Ж}_1=\{\vec{x}: |\ T(\vec{x})\ge c\}$.

Функцию наблюдений $T(\vec{X})$ называют в этом случае **статистикой критерия**, а критическую область задают непосредственно в терминах её значений. Если $T=\{t:\ t=T(\vec{x}),\ \vec{x}\in \text{Ж}\}$ -множество всех возможных значений статистики T, то критическая область критерия есть некоторое подмножество $T_1\in T$, которое должно включать все маловероятные, при

гипотезе H_0, значения T. При заданном уровне значимости α для критической области используют обозначение $T_{1\alpha}$. Для функции мощности в этом случае имеем условие

$$W(F)=P(T(\vec{X})\in T_{1\alpha}|F)\le\alpha,\ \forall F\in F_0\ . \qquad (6.3')$$

Выбор статистики критерия произволен до некоторой степени, на практике для конкретных задач выбор статистики ясен. Главным для расчёта критерия, как следует из (6.3'), является отыскание распределения статистики $T(\vec{X})$ в случае справедливости гипотезы H_0. Чтобы полностью вычислить функцию мощности критерия, и тем самым исследовать и вероятность ошибки второго рода, требуется знать распределение статистики $T(\vec{X})$ и при альтернативах, что является весьма трудной задачей.

6.4. Понятие параметрической гипотезы

Важный класс статистических гипотез составляют гипотезы об истинном значении неизвестного параметра, определяющего заданное параметрическое семейство распределений. В этом случае класс $\mathbb{F}$ допустимых распределений наблюдаемой случайной величины ξ имеет вид $\mathbb{F}=\{F(x,\vec{\theta}),\ \ \vec{\theta}\in\Theta\}$. Функции этого класса находят в соответствии со значениями параметра $\vec{\theta}=(\theta_1,...,\theta_r)$ из некоторого параметрического множества Θ. Гипотезы поэтому по существу относятся к неизвестным параметрам распределения и называются **параметрическими**. Примерами параметрических гипотез являются следующие утверждения:

1) H_0: $\theta=\theta_0$, где $\theta_0\in\Theta$ - некоторое фиксированное значение параметра.

2) H_0: $\theta_1=\theta_2=...=\theta_r$.

3) H_0: $g(\vec{\theta})=g_0$, где $g(\vec{\theta})$ - некоторая (в общем случае векторная) функция θ, g_0 - фиксированное значение.

В общем случае параметрическая гипотеза задаётся указанием некоторого подмножества $\Theta_0\subset\Theta$, элементом

которого является, по предположению, неизвестная параметрическая точка θ.

Обозначение: H_0: $\theta \in \Theta_0$.

Альтернативная гипотеза имеет вид H_1: $\theta \in \Theta_1 = \Theta \setminus \Theta_0$; точки $\theta \in \Theta_1$ называют **альтернативными**.

Если множество Θ_0 (Θ_1) состоит из одной точки, то гипотезу H_0 (альтернативу H_1) называют **простой**, в противном случае гипотезу (или альтернативу) называют **сложной .** Например, гипотеза 1) - простая; 2) - сложная, а 3) - может быть как простой, так и сложной.

Поверка параметрической статистической гипотезы при помощи критерия значимости может быть разбита на следующие этапы:

1) сформировать проверяемую (H_0) и альтернативную (H_1) гипотезы ;

2) назначить уровень значимости α;

3) выбрать статистику $T(\vec{X})$ критерия для проверки гипотезы H_0;

4) определить выборочное распределение статистики T при условии, что верна гипотеза H_0;

5) в зависимости от формулировки альтернативной гипотезы определить критическую область $\bar{X}_1$ одним из неравенств: $T>t_{1-\alpha}$; $T<t_\alpha$ или совокупностью неравенств $T> t_{1-\alpha/2}$; $T< t_{\alpha/2}$

6) получить выборку наблюдений и вычислить выборочное значение статистики критерия T_b.

7) принять статистическое решение: если $T_b \in \bar{X}_0$, то принять H_0, т.е. считать, что гипотеза H_0 не противоречит результатам наблюдений; если $T_b \in \bar{X}_1$, то отклонить гипотезу H_0, как не согласующуюся с результатами наблюдений.

Замечание.

Обычно на этапах 4 - 7 используют статистику, квантили которой табулированы: статистику с нормальным распределением N(0,1); статистику Стьюдента, статистику χ^2

или статистику Фишера. Однако, в вычислении вероятности ошибок и интерпретацию решений удобно проводить для статистики, являющейся оценкой параметра θ, т.е. статистики $\vec{\theta}$.

Общие принципы построения критериев уже были рассмотрены, далее конкретизируем задачу. В случае параметрических гипотез функция мощности для произвольного критерия $\mathbb{X}_1$ обозначается:

$$W(\vec{\theta})=W(\mathbb{X}_1;\vec{\theta})=P_\theta(\vec{X}\in\mathbb{X}_1),\ \vec{\theta}\in\Theta.$$

В случае рандомизированного критерия, который задаётся критической функцией φ(x), имеем:

$$W(\vec{\theta})=W(\varphi;\theta)=M_\theta\varphi(\vec{X})$$

Условия (6.3) и (6.4) в новых обозначениях примут вид

$$W(\vec{\theta})\le\alpha,\ \forall\vec{\theta}\in\Theta_0, \qquad (6.5)$$

$$W(\vec{\theta})\to\max,\ \forall\vec{\theta}\in\Theta_1. \qquad (6.6)$$

6.5. Равномерно наиболее мощные критерии

Пусть $\mathbb{X}_{1\alpha}$ и $\mathbb{X}^*_{1\alpha}$ – два критерия одного и того же уровня значимости α для гипотезы H_0. Если

$$W(\mathbb{X}^*_{1\alpha};\theta)\le W(\mathbb{X}_{1\alpha};\theta),\ \theta\in\Theta_0, \text{ и} \qquad (6.7)$$

$$W(\mathbb{X}^*_{1\alpha};\theta)\ge W(\mathbb{X}_{1\alpha};\theta),\ \theta\in\Theta_1, \qquad (6.8)$$

причём строгое неравенство в (6.8) имеет место хотя бы при одном значении θ , то говорят, что критерий $\mathbb{X}^*_{1\alpha}$ **равномерно мощнее** критерия $\mathbb{X}_{1\alpha}$. В этом случае, очевидно следует отдать предпочтение критерию $\mathbb{X}^*_{1\alpha}$, так как он приводит к меньшим ошибкам. Если соотношения (6.7) и (6.8) выполняются для любого критерия $\mathbb{X}_{1\alpha}$, то $\mathbb{X}^*_{1\alpha}$ называют **равномерно наиболее мощным (р.н.м.) критерием** для проверки гипотезы H_0. В случае, если множество Θ состоит из одной точки (H_1 - простая гипотеза) вместо термина р.н.м.

критерий используют термин **наиболее мощный критерий.** Равномерно наиболее мощный критерий не всегда существует, так как экстремальная задача (6.6) при ограничениях (6.5) имеет решения только в некоторых специальных случаях.

Часто ограничиваются подклассом несмещённых критериев, для которых одновременно с (6.5) выполняется следующее условие

$$W(\theta)\geq\alpha\ ,\ \forall\theta\in\Theta_1.$$

В ряде задач для которых р.н.м. критерии не существуют, могут иметь место р.н.м. несмещённые критерии.

6.6. Выбор из двух простых гипотез. Критерий Неймана-Пирсона

Проверяется простая параметрическая гипотеза против простой альтернативы. Параметрическое множество состоит из двух точек $\Theta=\{\theta_0;\theta_1\}$. Основная (проверяемая) гипотеза утверждает - H_0: $q=q_0$, а альтернатива H_1: $q=q_1$. Необходимо построить правило, позволяющее на основе значений выборки принять или отвергнуть H_0.

Решение задачи. Запишем вероятности ошибок

$$P(\gamma_1|H_0)=P(\gamma_1|\theta_0)=\int\limits_{\bar{x}_1} L(\vec{x};\theta_0)d\vec{x}$$

$L(x;\theta)$ - функция правдоподобия.

$$P(\gamma_0|H_1)=P(\gamma_0|\theta_1)=\int\limits_{\bar{x}_0} L(\vec{x};\theta_1)d\vec{x}=1-\int\limits_{\bar{x}_1} L(\vec{x};\theta_0)d\vec{x}$$

Зафиксируем значение вероятности ошибки первого рода $P(\gamma_1|H_0)=\alpha$. Будем искать критерий $\bar{X}_{1\alpha}$ обеспечивающий (min) минимум вероятности ошибки 2-го рода. Он будет при условии

$$\max_{\bar{x}_{1\alpha}}\int\limits_{\bar{x}_1} L(\vec{x};\theta_1)d\vec{x}=\max_{\bar{x}_{1\alpha}}\int\limits_{\bar{x}_1}\frac{L(\vec{x};\theta_1)}{L(\vec{x};\theta_0)}L(\vec{x};\theta_0)d\vec{x},$$

$\frac{L(\vec{x};\theta_1)}{L(\vec{x};\theta_0)}=l(\vec{X})$ – отношение правдоподобия.

Учитывая, что $l(\vec{X})$ и $L(\vec{x};\theta)$ – положительные, то максимум интеграла будет достигаться, если

$$\mathbb{X}_{1\alpha}=\{\vec{x}: l(\vec{x})\geq c\}.$$

Значение с выбирается из равенства:

$$\int\limits_{\mathbb{X}=\{\vec{x}:l(\vec{x})\geq c\}} L(\vec{x};\theta_0)d\vec{x}=\alpha$$

Покажем, что такое разбиение приводит к наиболее мощному критерию.

Теорема Неймана - Пирсона. Пусть функции $F_0(x)=F(x;\theta_0)$ и $F_1(x)=F(x;\theta_1)$ – возможные распределения случайной величины ξ. Пусть они непрерывны по x. Отношение правдоподобия $l(\vec{X})=\dfrac{L(\vec{x};\theta_1)}{L(\vec{x};\theta_0)}$ задаётся таким образом. Тогда при заданной вероятности ошибки 1-го рода существует наиболее мощные критерий $\mathbb{X}^*_{1\alpha}$, определяющий критическую область следующим образом

$$\mathbb{X}^*_{1\alpha}=\{\vec{x}: l(\vec{x})\geq c\}.$$

Доказательство: Рассмотрим любой другой критерий $\mathbb{X}_{1\alpha}$ уровня значимости α. Тогда

$$W(\mathbb{X}_{1\alpha};\theta_1)=\int\limits_{\mathbb{X}_{1\alpha}=\mathbb{X}_{1\alpha}\mathbb{X}^*_{1\alpha}+\mathbb{X}_{1\alpha}\overline{\mathbb{X}}^*_{1\alpha}} L(\vec{x};\theta_1)d\vec{x}=\int\limits_{\mathbb{X}_{1\alpha}\mathbb{X}^*_{1\alpha}} L(\vec{x};\theta_1)d\vec{x}+\int\limits_{\mathbb{X}_{1\alpha}\overline{\mathbb{X}}^*_{1\alpha}} L(\vec{x};\theta_1)d\vec{x}$$

Функция мощности для критерия $\mathbb{X}^*_{1\alpha}$ выражается аналогично

$$W(\mathbb{X}^*_{1\alpha};\theta_1)=\int\limits_{\mathbb{X}_{1\alpha}\mathbb{X}^*_{1\alpha}} L(\vec{x};\theta_1)d\vec{x}+\int\limits_{\overline{\mathbb{X}}_{1\alpha}\mathbb{X}^*_{1\alpha}} L(\vec{x};\theta_1)d\vec{x}$$

Из второго равенства находим:

$$\int\limits_{\mathbb{X}_{1\alpha}\mathbb{X}^*_{1\alpha}} L(\vec{x};\theta_1)d\vec{x}=W(\mathbb{X}^*_{1\alpha};\theta_1)-\int\limits_{\overline{\mathbb{X}}_{1\alpha}\mathbb{X}^*_{1\alpha}} L(\vec{x};\theta_1)d\vec{x},$$

подставляем в первое равенство и получаем

$$W(\mathbb{X}_{1\alpha};\theta_1)=W(\mathbb{X}^*_{1\alpha};\theta_1)+\int\limits_{\mathbb{X}_{1\alpha}\overline{\mathbb{X}}^*_{1\alpha}} L(\vec{x};\theta_1)d\vec{x}-\int\limits_{\overline{\mathbb{X}}_{1\alpha}\mathbb{X}^*_{1\alpha}} L(\vec{x};\theta_1)d\vec{x}=$$

умножим и разделим на $L(x;\theta_0)$ и получим

$$= W(\overline{X}^*_{1\alpha};\theta_1) + \int\limits_{X_{1\alpha}\overline{X}^*_{1\alpha}} l(\vec{x})L(\vec{x};\theta_0)d\vec{x} - \int\limits_{\overline{X}_{1\alpha}X^*_{1\alpha}} l(\vec{x})L(\vec{x};\theta_0)d\vec{x}$$

В соответствии с условиями теоремы:

$$\overline{X}^*_{1\alpha} = \{\vec{x} : l(\vec{x}) \geq c\}; \quad \overline{\overline{X}}^*_{1\alpha} = \{\vec{x} : l(\vec{x}) < c\}$$

получаем

$$W(\overline{X}_{1\alpha};\theta_1) < W(\overline{X}^*_{1\alpha};\theta_1) + c\left(\int\limits_{X_{1\alpha}\overline{X}^*_{1\alpha}} L(\vec{x};\theta_0)d\vec{x} - \int\limits_{\overline{X}_{1\alpha}X^*_{1\alpha}} L(\vec{x};\theta_0)d\vec{x}\right). \quad (*)$$

Рассмотрим интегралы в скобках. Первый интеграл, как и второй, можно представить в виде:

$$\int\limits_{X_{1\alpha}\overline{X}^*_{1\alpha}} L(\vec{x};\theta_0)d\vec{x} = \int\limits_{X_{1\alpha}} L(\vec{x};\theta_0)d\vec{x} - \int\limits_{X_{1\alpha}X^*_{1\alpha}} L(\vec{x};\theta_0)d\vec{x}$$

$$\int\limits_{\overline{X}_{1\alpha}X^*_{1\alpha}} L(\vec{x};\theta_0)d\vec{x} = \int\limits_{X^*_{1\alpha}} L(\vec{x};\theta_0)d\vec{x} - \int\limits_{X_{1\alpha}X^*_{1\alpha}} L(\vec{x};\theta_0)d\vec{x}$$

По условиям теоремы уровень значимости равен α. Интегралы в выражении (*) в правой части совпадают и скобки равны 0, отсюда получаем $W(\overline{X}_{1\alpha};\theta_1) < W(\overline{X}^*_{1\alpha};\theta_1)$, т.е. $\overline{X}^*_{1\alpha}$ более мощный критерий по сравнению с $\overline{X}_{1\alpha}$. В силу произвольности $\overline{X}_{1\alpha}$ соотношение выполняется для всех критериев с уровнем значимости α, т.е. $\overline{X}^*_{1\alpha}$ - наиболее мощный критерий. ▩

Замечание.

Критерий $\overline{X}^*_{1\alpha}$, построенный в соответствии с указанными условиями, называется критерием Неймана-Пирсона. Фиксируется вероятность ошибки 1-го рода и минимизируется вероятность ошибки 2-го рода.

7. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ О ПАРАМЕТРАХ НОРМАЛЬНОГО РАСПРЕДЕЛЕНИЯ

Когда основная гипотеза и альтернатива проостые, можно определить закон распределения выборки или какой-либо функции выборки и построить нерандомизированное правило принятия решений. Это правило обеспечит заданную вероятность ошибки первого рода (вероятность отвергнуть основную гипотезу, если она справедлива). Можно рассчитать и вероятность ошибки второго рода. При проверке сложных гипотез стремятся выбрать такую статистику, чтобы можно было бы свести хотя бы одну из гипотез к простой. Эту гипотезу и берут в качестве основной. Критерий также строится на основе заданной вероятности ошибки 1-го рода. На примерах выборок из нормального распределения рассмотрим те задачи, которые возникают при проверке простых и сложных гипотез. Эти задачи могут быть решены, применяя критерий Неймана - Пирсона или критерий правдоподобия.

7.1. Проверка гипотезы о математическом ожидании нормального распределения

Пусть случайная величина ξ имеет нормальное распределение с известной дисперсией σ^2 и неизвестным средним θ $\sim N(\theta,\sigma^2)$, $\theta\in\{\theta_0,\theta_1\}$, $\theta_0<\theta_1$. Необходимо построить критерий, позволяющий на основе значений выборки решить, какое значение имеет параметр θ.

$$H_0: \theta=\theta_0;\ H_1: \theta=\theta_1.$$

Будем использовать критерий Неймана-Пирсона. Необходимо построить отношение $l(\vec{x}) = \dfrac{L(\vec{x};\theta_1)}{L(\vec{x};\theta_0)}$ и сравнить с некоторым порогом $c\sim const$.

$l(\vec{x})\geq c$ - принимается решение $\gamma_1\sim H_1:\theta=\theta_1$,

$l(\vec{x})<c$ - принимается решение $\gamma_0\sim H_0:\theta=\theta_0$.

Значение c находится из условия: $P(\gamma_1|H_0)=\alpha$

$$l(\vec{x}) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(x_i-\theta_1)^2-(x_i-\theta_0)^2\right]\right\} = \exp\left\{\frac{n}{\sigma^2}(\theta_1-\theta_0)\bar{x}-\frac{n}{2\sigma^2}(\theta_1^2-\theta_0^2)\right\}$$

В виду монотонности экспоненты можно перейти к следующему неравенству

$$\ln l(x) \geq \ln c,$$

т.е.

$$[\frac{n}{\sigma^2}(\theta_1-\theta_0)\bar{x}-\frac{n}{2\sigma^2}(\theta_1^2-\theta_0^2)\ ] \geq \ln c.$$

В качестве статистики критерия при проверке простой параметрической гипотезы выбирают ту же статистику, что и для оценки параметра θ, т.е. выборочное среднее. Поэтому из этого неравенства определим выборочное среднее $\bar{x}$, после преобразований получим:

$$\bar{x} \geq \frac{\sigma^2 \ln c}{n(\theta_1-\theta_0)} + \frac{\theta_1+\theta_0}{2} \qquad (*)$$

Обозначим через h правую часть равенства (*), и получаем следующий алгоритм

$$\begin{aligned} &\bar{x} \geq h \rightarrow \gamma_1 \\ &\bar{x} < h \rightarrow \gamma_0 \end{aligned} . \qquad (7.1)$$

Необходимо найти h из условия $P(\gamma_1|H_0)=\alpha$. Выборочное среднее имеет нормальный закон распределения с параметрами $N(\theta, \sigma/\sqrt{n})$.

Определим ошибки первого и второго рода

$$\alpha=P(\gamma_1|H_0)=P(\bar{x}>h|\theta_0)= \frac{1}{\sqrt{2\pi}} \int_{\frac{(h-\theta_0)\sqrt{n}}{\sigma}}^{\infty} e^{-\frac{u^2}{2}} du = 1-\Phi\left(\frac{h-\theta_0}{\sigma}\sqrt{n}\right) \quad (7.2)$$

$$\beta= P(\gamma_0|H_1)=P(\bar{x}\leq h|\theta_1)= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{(h-\theta_1)\sqrt{n}}{\sigma}} e^{-\frac{u^2}{2}} du = \Phi\left(\frac{h-\theta_1}{\sigma}\sqrt{n}\right). \qquad (7.3)$$

Обозначим u_γ то значение, для которого

$$1-\Phi(u_\gamma)=\gamma\ ,$$

u_γ носит название **квантиль нормального распределения**. Тогда из (7.2), (7.3) и из того, что $u_\gamma=-u_{1-\gamma}$ вытекает

$$\frac{h-\theta_0}{\sigma}\sqrt{n}=u_\alpha$$

$$\frac{h-\theta_1}{\sigma}\sqrt{n}=-u_\beta, \tag{7.4}$$

отсюда определим h, т.е.

$$h=\theta_0+u_\alpha\frac{\sigma}{\sqrt{n}}=\theta_1-u_\beta\frac{\sigma}{\sqrt{n}}$$

Из этого выражения найдём n

$$n=\frac{\sigma^2(u_\alpha+u_\beta)^2}{(\theta_1-\theta_0)^2} \tag{7.5}$$

Равенство (7.5) даёт тот объём выборки, который при оптимальном критерии обеспечивает ошибки 1-го и 2-го рода (α и β). Если правая часть (7.5) - не целая, то за n надо брать ближайшее большее целое число

Проиллюстрируем полученные результаты

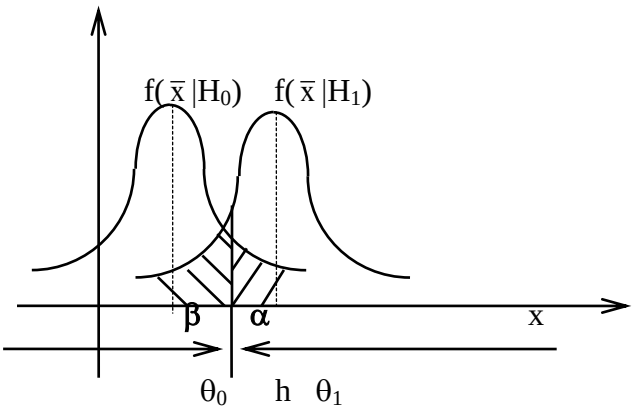


Рис.7.1

В соответствии с выражением (7.4) пороговое значение находится правее θ_0. Справа от h находится критическая область, слева – допустимая.

На рисунке приведено графическое представление плотности вероятности выборочного среднего при основной гипотезе и альтернативе. Вероятность ошибки 1-го рода представлена заштрихованной областью. Функция мощности (это вероятность попадания выборки в критическую область) выражается через ошибки 1-го и 2-го рода следующим образом:

$$W(X_1,\theta_1)=1-P(\gamma_0/H_1)=1-\beta,$$
$$W(X_1,\theta_0)=P(\gamma_1/H_0)=\alpha.$$

Значение h называется критическим значением критерия. Слева от него находится допустимая область, справа – критическая область. На практике обычно считают известным уровень значимости α и объем выборки n, а h - критическое значение определяют из таблиц или с помощью пакетов MATHCAD и STATISTICA.

Уровень значимости α связан с критическим значением h приближенной формулой

$$\alpha \approx 1 - A(h),$$

где A(h) – функция распределения той статистики, которая используется при проверке гипотез. При большом объёме выборки критическое значение h совпадает с (1 – α) квантилью соответствующего распределения.

Самостоятельно:

Построить алгоритм проверки гипотезы о математическом ожидании нормальной генеральной совокупности при условии . $\theta_0>\theta_1$, $\theta_0 \neq \theta_1$.

7.2. Проверка гипотезы о дисперсии нормального распределения

Пусть в модели $\xi \sim N(a,\sigma^2)$ следует проверить простую гипотезу о неизвестной дисперсии, т.е. рассматриваются две

гипотезы H_0: $\theta=\theta_0$; H_1: $\theta=\theta_1>\theta_0$. Необходимо построить критерии Неймана - Пирсона для принятия решения.

В этом случае отношение правдоподобия

$$l(x)=\frac{\sigma_0^n}{\sigma_1^n}\exp\left\{\frac{1}{2}\left(\frac{1}{\theta_0^2}-\frac{1}{\theta_1^2}\right)\sum_{i=1}^{n}x_i^2\right\}$$

при условии, что a=0 приводит к статистике

$$\sum_{i=1}^{n}x_i^2 \geq c$$

Известно, что сумма квадратов случайных величин $\eta_j=\frac{X_j-a}{\sigma}$ будет иметь χ^2-распределение $\left[\sum_{j=1}^{n}\eta_j^2 \sim \chi_n^2\right]$ с n степенями свободы, поэтому для решения задачи будем испытывать статистику

$$T=\sum_{i=1}^{n}(x_i-a)^2=nS^2, \qquad T\in(0,\infty)$$

$$\frac{nS^2}{\theta}=\chi_n^2 \quad \text{и} \quad T=\theta\chi_n^2.$$

Алгоритм проверки гипотезы

$$T\geq c\rightarrow\gamma_1 \qquad T<c\rightarrow\gamma_0$$

Найдём значение с из условия.

Ошибка первого рода:

$$P(\gamma_1|H_0)=P(T\geq c|\theta_0)=P(\theta_0\chi_n^2\geq c)=P(\chi_n^2\geq\frac{c}{\theta_0})=1-F_{\chi_n^2}\left(\frac{c}{\theta_0}\right)=\alpha,$$

отсюда:

$$c=\chi_{n,1-\alpha}^2\cdot\theta_0$$

Здесь $F_{\chi_n^2}$ – функция распределения χ^2 с n степенями свободы; $\chi_{n,1-\alpha}^2$ – квантиль χ^2-распределения порядка $(1-\alpha)$.

Ошибка второго рода:

$$P(\gamma_0|H_1)=P(T<c|\theta_1)=P(\theta_1\chi_n^2<c)=P(\chi_n^2<\frac{c}{\theta_1})=\beta,$$

$$\beta = F_{\chi^2_n}\left(\frac{\chi^2_{n,1-\alpha}\cdot\theta_0}{\theta_1}\right).$$

Замечание.

Статистика $T=\sum_{i=1}^{n}(x_i-a)^2$ предполагает, что математическое ожидание известно, поэтому $T=\theta\chi^2_n$. Если математическое ожидание неизвестно, то можно использовать статистику $T=\sum_{i=1}^{n}(x_i-\overline{X})^2$. Для неё справедливо представление $T=\theta\chi^2_{n-1}$.

Самостоятельно:

Построить алгоритм проверки гипотезы о дисперсии нормального распределения при условии $\theta_0<\theta_1$.

7.3. Проверка сложных статистических гипотез. Гипотеза о равенстве математических ожиданий нормальных распределений

Пусть имеются две выборки из нормальных распределений

$$\vec{X}=(X_1,...,X_n),\quad \xi \sim N(\theta_1,\sigma_1)$$
$$\vec{Y}=(Y_1,...,Y_n),\quad \eta \sim N(\theta_2,\sigma_2)$$

H_0: $\theta_1=\theta_2$ H_1: $\theta_1\neq\theta_2$ - сложные гипотезы.

Необходимо построить правила, позволяющие на основе значений выборок $\vec{X}$ и $\vec{Y}$, принять или отвергнуть основную гипотезу.

Будем пользоваться статистикой

$$T=\frac{\vec{X}-\vec{Y}}{\sigma};\qquad \sigma=\sqrt{\frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{n}},$$

$\vec{X}\sim N\left(\theta_1;\frac{\sigma_1}{\sqrt{n}}\right)$ – выборочное среднее имеет нормальное распределение;

$\vec{Y} \sim N\left(\theta_2; \frac{\sigma_2}{\sqrt{n}}\right)$ - аналогично.

$$\vec{X} - \vec{Y} \sim N\left(\theta_1 - \theta_2; \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}}\right);$$

$$T = \frac{\vec{X} - \vec{Y}}{\sigma} \sim N\left(\frac{\theta_1 - \theta_2}{\sigma}; 1\right); \qquad \frac{\theta_1 - \theta_2}{\sigma} \equiv \theta_0.$$

При этом основная гипотеза и альтернатива могут быть сформулированы следующим образом: H_0: $\theta_0=0$ - простая гипотеза; H_1: $\theta_0\neq 0$ - сложная гипотеза.

Решение этой задачи иллюстрируется рисунками, приведенными ниже.

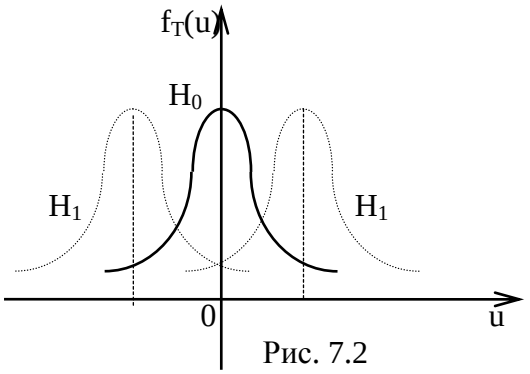


Рис. 7.2

В связи с этим мы должны рассмотреть три случая

a) $\theta_1>\theta_2$, $\theta_0>0$

b) $\theta_1<\theta_2$, $\theta_0<0$

c) $\theta_1\neq\theta_2$, $\theta_0>0$, $\theta_0<0$ ($\theta_0<>0$)

В каждом случае критическая область выбирается по-своему.

a) $\theta_0>0$ ($\theta_1>\theta_2$) критическая область правосторонняя. Алгоритм принятия решения в этом случае, как и в задаче проверки гипотезы о математическом ожидании нормального распределения, имеет вид

$t \geq h \rightarrow \gamma_0$ $\qquad$ $t < h \rightarrow \gamma_1$

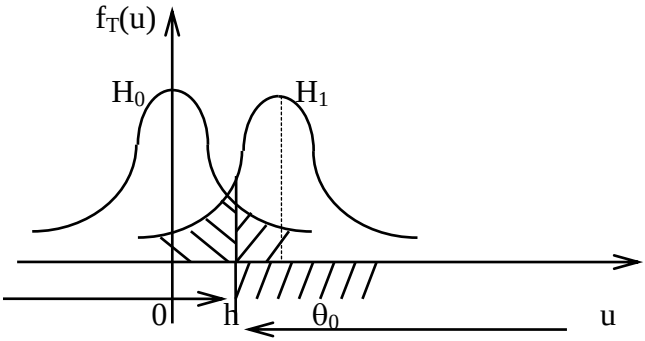
h находят из условия

$$P(\gamma_1|H_0)=\alpha \qquad (*)$$

при H_0, $T\sim N(0,1)$, поэтому

$$P\left(T \geq h \middle| H_0\right) = \frac{1}{\sqrt{2\pi}} \int_h^{\infty} e^{-t^2/2}\, dt = 1 - \Phi(h) = \alpha ,$$

отсюда $h=u_{1-\alpha}$.



допустимая область критическая область правосторонняя $\overline{X}_1^+$

Рис. 7.3

Найдём вероятность ошибки 2-го рода

$P(\gamma_0|H_1)=P(T<h|H_1)=\Phi(h-\theta_0)=\beta$, при H_1 $T\sim N(\theta_0;1)$.

Вероятность ошибки зависит от разности параметров.

Если $\frac{\theta_1-\theta_2}{\sigma}=\theta_0\to 0$, то $P(\gamma_0|H_1)=1-P(\gamma_1|H_0)\sim\beta=1-\alpha$.

Если параметры расходятся, т.е. $\theta_0\to\infty$, то $P(\gamma_0|H_1)\to 0$ ($\beta\to 0$ - ошибка второго рода). Функция мощности при альтернативе будет иметь вид

$$W(\theta_0)=1-P(\gamma_0|H_1)=1-\Phi(h-\theta_0).$$

Исследуем поведение функции мощности при альтернативе для различных значений θ_0.

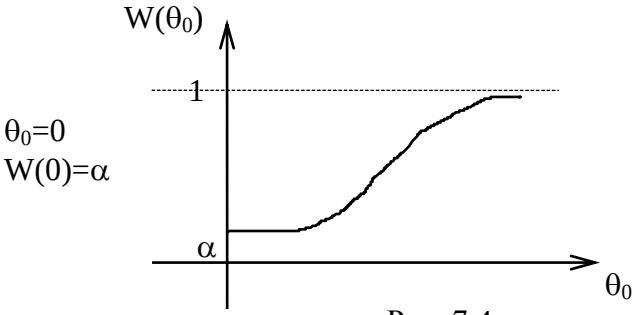


Рис. 7.4

При $\theta_0 \to \infty$ $W(0) \to 1$

b) $\theta_0 < 0$ ($\theta_1 < \theta_2$). Алгоритм принятия решения запишется в виде

$$<h \to \gamma_1 \; t \geq h \to \gamma_0.$$

Найдём h из следующего выражения

$$P(\gamma_1|H_0)=P(T<h|H_0)=\Phi(h)=\alpha$$

$$h=u_\alpha.$$

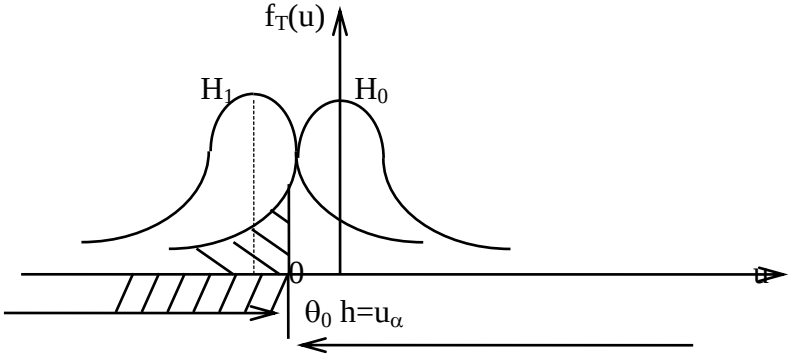


Рис. 7.5

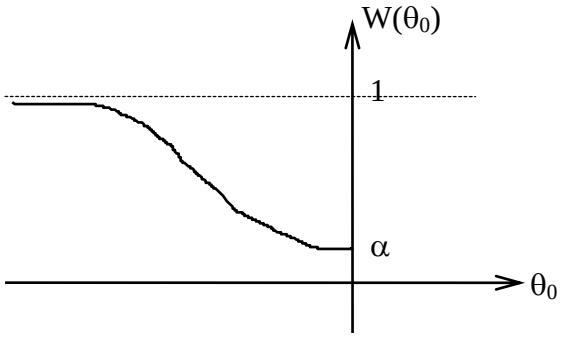
Найдём вероятность ошибки 2-го рода

$P(\gamma_0|H_1)=P(T \geq h|H_1)=1-P(T<h|H_1)=1-\Phi(h-\theta_0).$

Функция мощности имеет вид

$W(\theta_0)=1- P(\gamma_0|H_1)=\Phi(h-\theta_0).$

Рассмотрим поведение функции мощности



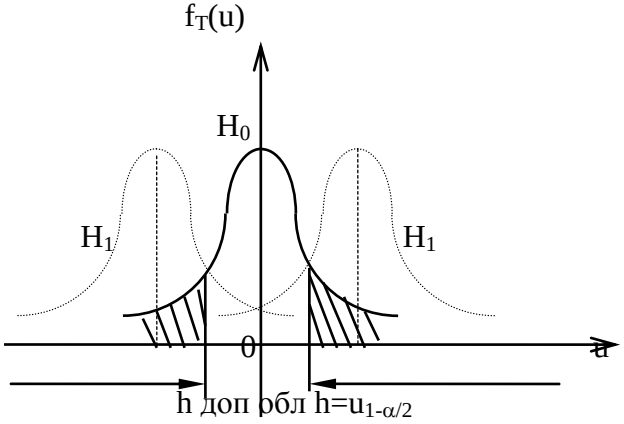
$\theta_0=0$
$W(0)=\alpha$

При $\theta_0 \to -\infty$ $W(\theta_0) \to 1$.

Рис. 7.6

с) $\theta_0 = 0$. Алгоритм принятия решения запишется в виде

$|t| \geq h \to \gamma_1$ $|t| < h \to \gamma_0$



критическая область двухсторонняя

Рис. 7.7

$P(\gamma_1|H_0)=P(|T| \geq h|H_0)=1-P(|T|<h|H_0)=1-\Phi(h)+\Phi(-h)=2-2\Phi(h)=\alpha.$

Используя свойство $\Phi(-h)=1-\Phi(h)$.

$$h= u_{1-\alpha/2}.$$

Вероятность ошибки 2-го рода определяется следующим образом:

$P(\gamma_0|H_1)=P(|T|<h|H_1)=\Phi(h-\theta_0)+\Phi(-h-\theta_0)=\Phi(h-\theta_0)+\Phi(h+\theta_0)-1.$

$W(\theta_0)=1-P(\gamma_0|H_1)=2-\Phi(h-\theta_0)-\Phi(h+\theta_0)$.

График функции мощности представлен на рисунке, как и ранее $W(0)=P(\gamma_1|H_0)=\alpha$, при $\theta_0\to\pm\infty$ функция мощности стремится к 1 ($W(\theta_0)\to 1$).

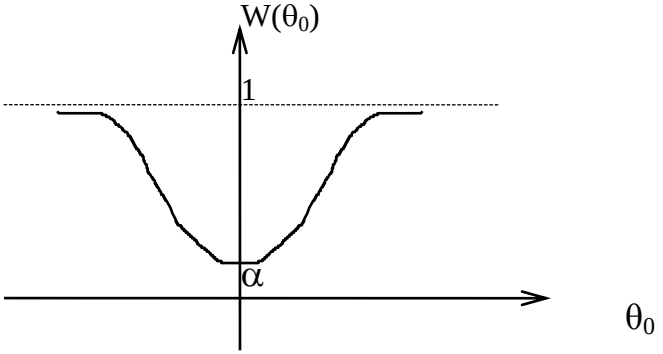


Рис 7.8

Из рассмотрения функций мощности для односторонних и двустороннего критерия можно сделать вывод, что двусторонний критерий $\overline{X}_{1\alpha}$ всегда менее мощный, чем один из односторонних критериев $\overline{X}^{+}_{1\alpha}$ или $\overline{X}^{-}_{1\alpha}$.

Пример. С помощью одного и того же прибора, среднеквадратическая ошибка измерений которого $\sigma_0=1$, получена по 5 измерений для двух величин.

Для первой величины X_i= 4; 5; 6; 7; 8.

Для второй величины Y_i= 5; 5; 5; 4; 6.

Проверить гипотезу о равенстве измеренных величин при уровне значимости $\alpha=0.05$. Ошибки измерения считаются нормальными.

Решение. Воспользуемся результатами решения задачи проверки гипотез о равенстве математических ожиданий нормальной генеральной совокупности.

1. Воспользуемся статистикой $T=\dfrac{\vec{X}-\vec{Y}}{\sigma}$; $\sigma=\sqrt{\dfrac{\sigma_1^2}{n_1}+\dfrac{\sigma_2^2}{n_2}}$

$\sigma_1=\sigma_2=\sigma_0=1$ $n_1=n_2=5$ $\sigma=\sqrt{\frac{1}{5}+\frac{1}{5}}=\sqrt{\frac{2}{5}}$

$\bar{x}=6;\quad \bar{y}=5.$

2. Используем случай (c)

$$h=u_{1-\alpha/2}=u_{0.975}=1.96$$

$$|t_{экс}|<>h$$

$$t_{экс}=\frac{6-5}{\sqrt{2/5}}=\frac{1}{\sqrt{2/5}}\approx 1.58.$$

$1.58<1.96$, т.е. верна гипотеза H_0, принимается решение γ_0 - средние значения совпадают, следовательно, измерялась одна и та же величина.

7.4. Проверка гипотезы о равенстве дисперсий нормальных распределений

Пусть имеются две выборки из нормальных распределений:

$$\vec{X}=(X_1,...,X_n)\sim\xi\sim N(a_1,\theta_1^{1/2})\text{ и}$$

$$\vec{Y}=(Y_1,...,Y_n)\sim\eta\sim N(a_2,\theta_2^{1/2}).$$

Выдвигаются гипотезы: H_0: $\theta_1=\theta_2$, H_1: $\theta_1\neq\theta_2$ (сложные гипотезы). Нужно построить правило, позволяющее на основе значений выборок принять или отвергнуть гипотезу H_0. Воспользуемся статистикой, которая при гипотезе H_0 имеет известный закон распределения, а именно статистикой:

$$T=\frac{S_x^2 n(m-1)}{S_y^2 m(n-1)}$$

или $T\cdot\frac{\theta_2}{\theta_1}=F$ - распределение Фишера, при $\theta_1=\theta_2$, $T=F$.(S_x^2, S_y^2 - соответствующие выборочные дисперсии).

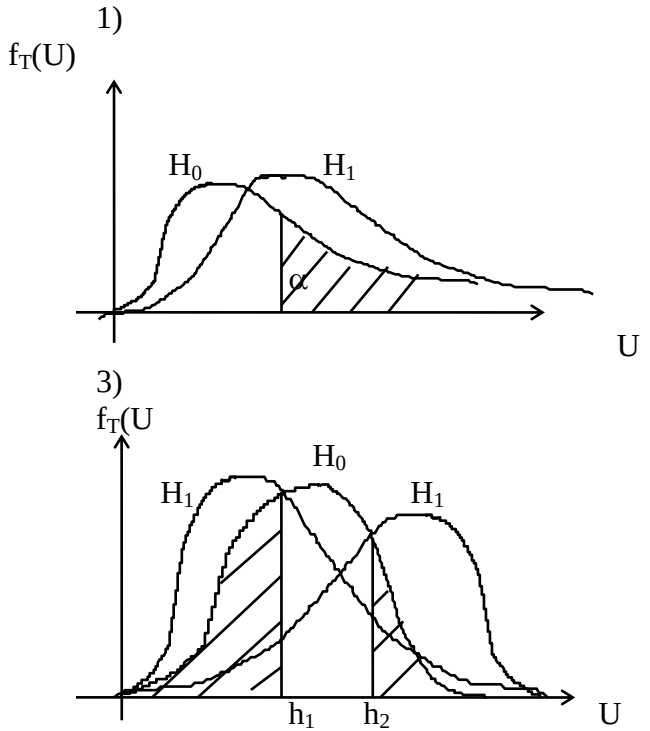
При альтернативе возможны следующие варианты.

1) $\theta_1>\theta_2$, $\frac{\theta_2}{\theta_1}<1$;

2) $\theta_1<\theta_2$, $\frac{\theta_2}{\theta_1}>1$;

3) $\theta_1<>\theta_2$, $\frac{\theta_2}{\theta_1}<>1$.

Для каждого случая критическая область выбирается по-разному.



$h_1=U_{\alpha/2,(n-1)(m-1)}$; $h_2=U_{1-\alpha/2,(n-1)(m-1)}$;

Рис. 7.9

2) Этот случай можно свести к 1), если выбрать статистику $T=\frac{S_y^2 m(n-1)}{S_x^2 n(m-1)}$.

Выпишем критерии, вероятности ошибок и функции мощности для указанных случаев.

1) $t \geq h \rightarrow \gamma_1 \quad t < h \rightarrow \gamma_0$

h найдем из условия: ошибка 1-го рода равна заданному значению α.

$$P(\gamma_1|H_0)=P(T\geq h|H_0)=1-P(T<h|H_0)=1-F_F(h)=\alpha\ .$$

(F_F - распределение Фишера с (n-1), (m-1) степенями свободы)

Отсюда $h=U_{1-\alpha,(n-1)(m-1)}$ – квантиль распределения Фишера порядка $(1-\alpha)$ с указанными степенями свободы.

Вычислим вероятность ошибки 2-го рода:

$$P(\gamma_0|H_1) = P(T<h|H_1) = P(T\cdot\frac{\theta_2}{\theta_1}<h\frac{\theta_2}{\theta_1}|H_1) = F_F(\frac{\theta_2}{\theta_1}h)$$

и функцию мощности:

$$W\left(\frac{\theta_2}{\theta_1}\right) = 1 - P(\gamma_0|H_1) = [1 - F_F(\frac{\theta_2}{\theta_1}h)]\ .$$

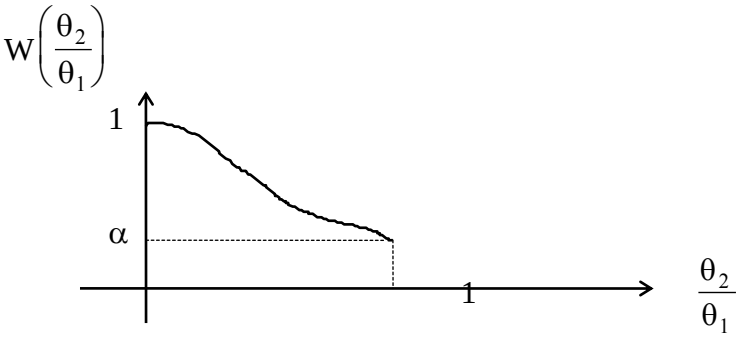


Рис. 7.10

Поведение функции мощности

Если $\theta_2/\theta_1 \rightarrow \infty$, $W\left(\frac{\theta_2}{\theta_1}\right) \rightarrow 1$

Если $\theta_2/\theta_1 \rightarrow 1$, $W\left(\frac{\theta_2}{\theta_1}\right) \rightarrow \alpha$

3) $t \geq h_2$ и $t < h_1 \rightarrow \gamma_1$

$t \leq h_2$ и $t > h_1 \rightarrow \gamma_0$

Найдем h_1 и h_2 из условия:

$$P(\gamma_1|H_0)=\alpha$$

$$P(\gamma_1|H_0)=P(T\geq h_2|H_0)+P(T<h_1|H_0)=\alpha/2+\alpha/2,$$

отсюда $h_1=U_{\alpha/2,(n-1)(m-1)}$, $h_2=U_{1-\alpha/2,(n-1)(m-1)}$ – квантили распределения Фишера.

Вычислим вероятность ошибки 2-го рода:

$$P(\gamma_0|H_1)=P(h_1<T\leq h_2|H_1)=P(\frac{\theta_2}{\theta_1}h_1<F\leq\frac{\theta_2}{\theta_1}h_2|H_1)=F_F(\frac{\theta_2}{\theta_1}h_2)-$$

$$-F_F(\frac{\theta_2}{\theta_1}h_1)$$

и функцию мощности:

$$W\left(\frac{\theta_2}{\theta_1}\right)=1-F_F(\frac{\theta_2}{\theta_1}h_2)+F_F(\frac{\theta_2}{\theta_1}h_1).$$

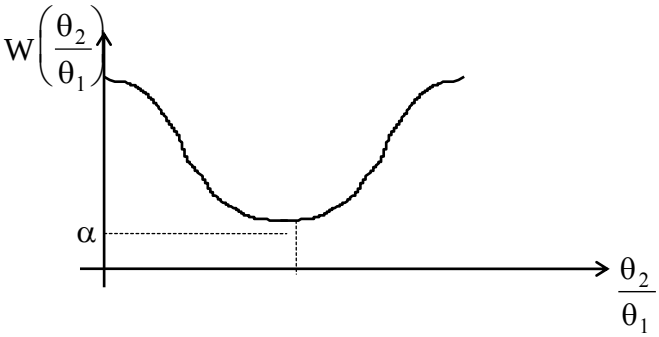


Рис. 7.11

В данном случае критическая область двусторонняя. Поведение функции мощности можно представить так.

Если $\theta_2/\theta_1 \to 0$, $W\left(\frac{\theta_2}{\theta_1}\right)\to 1$

Если $\theta_2/\theta_1 \to \infty$, $W\left(\frac{\theta_2}{\theta_1}\right)\to 1$

Если $\theta_2/\theta_1 \to 1$, $W\left(\frac{\theta_2}{\theta_1}\right)\to \alpha$

Самостоятельно:

1. Решить задачу проверки статистических гипотез о математическом ожидании нормальной генеральной совокупности H_0: $\theta=\theta_0$, H_1: $\theta=\theta_1$ (случай $\theta_1<\theta_0$,).
2. Систематизировать задачи о проверке статистических гипотез
   a) критерий (критическая область)
   b) значение α.
   c) вероятность ошибки 2-го рода и функция мощности.

Задачи и решения

При проверке сложных параметрических гипотез следует использовать следующие правила для облегчения решения задач

Сравнение двух дисперсий нормальных генеральных совокупностей

По независимым выборкам, объемом n_1 и n_2 , извлеченным из нормальных генеральных совокупностей, найдены несмещенные выборочные дисперсии S_X^2 и S_Y^2. Требуется сравнить эти дисперсии.

Правило1: При уровне значимости α для проверки нулевой гипотезы H_0: $D(X)=D(Y)$ при альтернативе H_1: $D(X)>D(Y)$, надо вычислить наблюдаемое значение критерия (отношение большей выборочной дисперсии к меньшей)

$$F_{набл} = S_Б^2 / S_М^2$$

и по табл. П.3.5 критических точек распределения Фишера-Снедекора по заданному α и числам степеней свободы $k_1=n_1-1$, $k_2=n_2-1$ найти критическую точку

$$F_{кр}(\alpha,k_1,k_2)$$

(k_1 – число степеней свободы большей выборочной дисперсии).

Если $F_{набл}<F_{кр}$ – нет оснований отвергнуть H_0
Если $F_{набл}>F_{кр}$ – H_0 отвергают

Правило 2: При альтернативе H_1: $D(X)\neq D(Y)$ критическую точку

$$F_{кр}(\alpha/2,k_1,k_2)$$

ищут по уровню значимости $\alpha/2$ и числам степеней свободы k_1 и k_2 (k_1 – число степеней свободы большей дисперсии)
Если $F_{набл}<F_{кр}$ – H_0 принимают
Если $F_{набл}>F_{кр}$ – H_0 отвергают

Сравнение двух средних совокупностей, дисперсии которых известны (большие независимые выборки)

Обозначим через n и m объемы больших независимых выборок (n>30, m>30). По ним найдены выборочные средние $\overline{x}$ и $\overline{y}$. Генеральные дисперсии D(X) и D(Y) известны.

Правило 1: Для проверки при заданном α нулевой гипотезы H_0: $M(X)=M(X)$ при альтернативе H_1: $M(X)\neq M(X)$ надо вычислить наблюдаемое значение критерия

$$z_{набл} = \frac{\overline{x} - \overline{y}}{\sqrt{D(X)/n + D(Y)/m}}$$

и по табл. П.3.3 функций Лапласа найти критическую точку $z_{кр}$ из равенства $\Phi(z_{кр})=(1-\alpha)/2$.
Если $|z_{набл}| < z_{кр}$ - H_0 принимают.
Если $|z_{набл}| > z_{кр}$ - H_0 отвергают.

Правило 2: При альтернативе H_1: $M(X)>M(X)$ находят критическую точку $z_{кр}$ по табл. П.3.3 функций Лапласа из равенства $\Phi(z_{кр})=(1-2\alpha)/2$.
Если $z_{набл} < z_{кр}$ - H_0 принимают.
Если $z_{набл} > z_{кр}$ - H_0 отвергают.

Правило 3. При альтернативе H_1: $M(X)<M(X)$ находят вспомогательную точку $z_{кр}$ по правилу 2.
Если $z_{набл} > -z_{кр}$ - H_0 принимают.
Если $z_{набл} < -z_{кр}$ - H_0 отвергают.

Сравнение двух средних нормальных генеральных совокупностей, дисперсии которых неизвестны и одинаковы (малые независимые выборки)

Пусть n<30 и m<30 – объемы независимых выборок, по которым найдены выборочные средние $\overline{x}$, $\overline{y}$ и несмещенные выборочные дисперсии S_X^2 и S_Y^2. Генеральные дисперсии неизвестны, но предполагаются одинаковыми.

Правило 1. Для того, чтобы при заданном α проверить гипотезы H_0: M(X)=M(X) при альтернативе H_1: M(X)≠M(X), надо вычислить наблюдаемое значение критерия

$$T_{набл} = \frac{\overline{x} - \overline{y}}{\sqrt{(n-1)S_X^2 + (m-1)S_Y^2}} \sqrt{\frac{n\,m\,(n+m-2)}{n+m}}$$

и по табл. П.3.4 критических точек распределения Стьюдента по заданному α и числу степеней свободы k=n + m – 2 найти критическую точку $t_{двуст.\ крит}(\alpha,k)$. Если $|T_{набл}| < t_{двуст.\ крит}(\alpha,k)$ – H_0 принимают.

Если). Если $|T_{набл}| < t_{двуст.\ крит}(\alpha,k)$ – H_0 отвергают.

Правило 2. При альтернативе H_1: M(X)>M(X) находят критическую точку $t_{прав.\ кр}(\alpha,k)$ k=n + m – 2.).

Если $T_{набл} < t_{прав.\ кр}(\alpha,k)$ – H_0 принимают.

Если $T_{набл} > t_{прав.\ кр}(\alpha,k)$ – H_0 отвергают.

Правило 3. При альтернативе H_1: M(X)<M(X) находят критическую точку $t_{прав.\ кр}(\alpha,k)$ по правилу 2 и полагают $t_{лев.\ кр} = -t_{лев.\ кр}$.

Если $T_{набл} > -t_{левв.\ кр}$ – H_0 принимают.

Если $T_{набл} < -t_{лев.\ кр}$ – H_0 отвергают.

Задача 47

Считается, что новое антикоррозийное покрытие имеет эффективность 99%, если среди 20 испытанных образцов нет ни одного с признаками коррозии; в противном случае эффективность покрытия принимается равной 90%. Пусть p –

вероятность появления признаков коррозии у одного образца. Предположим, что образцы обрабатываются и испытываются независимо друг от друга. Рассмотрим нулевую гипотезу H_0: p=0,10 и альтернативную гипотезу H_1: p=0,01. Ответить на следующие вопросы:

а) Какая статистика критерия используется в данной задаче, каковы её распределение и область изменения?

б) Какова критическая область критерия?

в) В чем состоят ошибки первого и второго рода и чему равны их вероятности?

Решение:

99%- эффективность покрытия, 20 испытаний (нет коррозии)

90%- эффективность покрытия

p- вероятность появления коррозии

H_0: p=0,10

H_1: p=0,01

а) число образов с признаками коррозии имеет биноминальное распределение ~ Bi(20,p). V={0,1,2…20}- область изменения

б) альтернативная гипотеза H_1: p=0,01. Предполагается уменьшение вероятности появления коррозии $V_k = \{0\}$

в) ошибка I рода: принимается решение, что антикоррозийное покрытие имеет эффективность 99% в то время как его эффективность 90%

ошибка II рода: принимается решение, что антикоррозийное покрытие имеет эффективность 90%, в то время как его эффективность составляет 99%

$$\alpha = p\left(x \in V_k / H_0\right) = p\left(K = 0 / p = 0{,}1\right) = C_{20}^0 \left(0{,}1\right)^0 \left(0{,}9\right)^{20} \approx 0{,}112$$

$$\beta = p\left(x \notin V_k / H_1\right) = 1 - p\left(x \in V_k / H_1\right) = 1 - C_{20}^0 \left(0{,}01\right)^0 \left(0{,}99\right)^{20} \approx 0{,}182$$

Ответ: (а), (б), (в).

$$\alpha \approx 0{,}112$$

$$\beta \approx 0{,}182$$

Задача 48

Из продукции автомата, обрабатывающего болты с номинальным значением контролируемого размера $m_0 = 40мм$, была взята выборка болтов объема n=36. Выборочное среднее контролируемого размера $\bar{x} = 40,2мм$. Результаты предыдущих измерений дают основание предполагать, что действительные размеры болтов образуют нормально распределенную совокупность с дисперсией $\sigma^2 = 1мм^2$. Можно ли по результатам проведенного выборочного обследования утверждать, что контролируемый размер в продукции автомата не имеет положительного смещения по отношению к номинальному размеру? Принять α=0,01. Какова критическая область в этом случае?

Решение:

$n = 36; \quad m_0 = 40; \quad \overline{x} = 40,2; \quad \sigma^2 = 1$

$H_0 : m = 40$

$H_1 : m > 40$

$\alpha = 0,01$

$$u = \frac{\overline{x} - m}{\frac{\sigma}{\sqrt{n}}} = \frac{\overline{x} - 40}{\frac{1}{6}} = 6\left(\overline{x} - 40\right)$$

Критическая область $V_k = \{u > u_{1-\alpha}\}$

$u_{1-\alpha} = u_{0,99} = 2,326 \Rightarrow V_k \{u > 2,326\}$

$$u_{выб} = \frac{40,2 - 40}{\frac{1}{6}} = 1,2$$

$u_{выб} \notin V_k \Rightarrow$ *принимаем* H_0

$6\left(\overline{x} - 40\right) > 2,326$

$\overline{x} > 40,387$

Ответ: H_0 *принимаем*: $V_k = \{\overline{x} > 40,387\}$.

Задача 49

В соответствии с техническими условиями среднее время безотказной работы для приборов из большой партии должно составлять не менее 1000 часов со среднеквадратичным отклонением (с.к.о.) 100 часов. Выборочное среднее времени безотказной работы для случайно отобранных 25 приборов оказалось равным 970 часам. Предположим, что с.к.о. времени безотказной работы для приборов в выборке совпадает с с.к.о. во всей партии. Можно ли считать, что вся партия приборов не удовлетворяет техническим условиям, если: а) α=0,1; б) α=0,01?

Решение:

$m_0 = 1000;\ \ n = 25;\ \ \sigma = 100;\ \ \overline{x} = 970;$

$\alpha_1 = 0,10 \qquad u_{0,9} = 1,282$

$\alpha_2 = 0,01 \qquad u_{0,99} = 2,326$

Альтернативная гипотеза и область принятия гипотезы H_0 *для левостороннего критерия* $H_1 : m < m_0$, *т.е.* $1000 < m_0$

$$-u_{1-\alpha} < \frac{\overline{x} - m_0}{\frac{\sigma}{\sqrt{n}}}$$

$$m_0 = \overline{x} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}$$

1) $m_0 = 970 + \frac{100}{5} 1,28 = 995,64$

$m < 996$

$1000 < 996 \ \Rightarrow\ H_0 - принимаем$

2) $m_0 = 970 + \frac{100}{5} 2,326 = 1016,52$

$m = 1000 < 1016,52 \ \Rightarrow\ H_0 - отклоняем$

Ответ: 1) $H_0 - принимаем$

2) $H_0 - отклоняем$;

Задача 50

Утверждается, что шарики, изготовленные станком-автоматом, имеют средний диаметр $d_0 = 10мм$. Используя односторонний критерий при α=0,05, проверить эту гипотезу, если в выборке из n=16 шариков средний диаметр оказался равным 10,3мм, считая, что:

а) Дисперсия известна и равна $\sigma^2 = 1мм^2$;

б) Оценка дисперсии, определенная по выборке, $S^2 = 1,21мм^2$.

Решение:

$\overline{x}_0 = 10;\ \ \overline{x} = 10,3;\ \ \alpha = 0,05;\ \ n = 16;$

1) $\sigma^2 = 1мм^2$

$H_0 : \overline{x} = \overline{x}_0$

$u_{набл} = \dfrac{\overline{x} - \overline{x}_0}{\sigma}\sqrt{n} = \dfrac{10,3-10}{1}4 = 1,2$

$u_{1-\alpha} = u_{0,95} = 1,645$

$u_{набл} < u_{1-\alpha} \Rightarrow H_0 - принимаем.$

2) $S^2 = 1,21мм^2$

$\dfrac{\overline{x} - \overline{x}}{S\sqrt{n}} < t_{1-\alpha}(n-1)$

$\dfrac{10,3-10}{1,21}4 = 0,991$

$t_{0,95}(15) = 1,753$

$0,991 < 1,753 \Rightarrow принимаемгипотезу$

Ответ: 1)*принимаем*

2)*принимаем.*

Задача 51

Технология производства некоторого вещества дает в среднем 1000кг вещества в сутки с среднеквадратичным отклонением (с.к.о.) среднего, равным 80кг. Новая технология производства в среднем дает 1100кг вещества в сутки с тем же с.к.о. Можно ли считать, что новая технология обеспечивает повышение производительности, если: а) α=0,05; б) α=0,1?

Решение:

$\overline{x} = 1000;\ \ \sigma = 80;\ \ \overline{x}_0 = 1100;\ \ n = 24;$

H_0 : *для правостороннего критерия.*

$u_1 : m > m_0 \qquad (m = 1100)$

$$m_0 = \overline{x} + \frac{\sigma}{\sqrt{n}} u_{1-\alpha}$$

1) $\alpha = 0,05 \qquad u_{0,95} = 1,645$

$$m_0 = 1000 + 1,645 \frac{80}{\sqrt{24}} = 1026,91$$

$1100 > 1026,91 \Rightarrow H_0 - отклоняем;$

2) $\alpha = 0,1 \qquad u_{0,9} = 1,282$

$$m_0 = 1000 + 1,282 \frac{80}{\sqrt{24}} = 1020,97$$

$1100 > 1020,97 \Rightarrow H_0 - отклоняем;$

Ответ: 1) $H_0 - отклоняем$

2) $H_0 - отклоняем;$

Задача 52

На двух станках А и В производят одну и ту же продукцию, контролируемую по внутреннему диаметру изделия. Из продукции станка А была взята выборка из 16 изделий, а из продукции станка В выборка из 25 изделий. Выборочные оценки средних и дисперсий контролируемых размеров $\overline{x}_A = 37,5 мм$ при $S_A^2 = 1,21 мм^2$ и $\overline{x}_B = 36,8 мм$ при $S_B^2 = 1,44 мм^2$. Используя двусторонний критерий, проверить гипотезу о равенстве математических ожиданий контролируемых размеров в продукции обоих станков, если: а) α=0,05; б) α=0,10.

Решение:

$n_1 = 16;\ \ n_2 = 25;\ \ S_1^2 = 1,21 мм^2;$

$\overline{x}_1 = 37,5;\ \ \overline{x}_2 = 36,8;\ \ S_2^2 = 1,44 мм^2;$

Область принятия гипотезы H_0 *для двустороннего критерия*

$$\frac{\left|\overline{x}_1 - \overline{x}_2\right|}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{1-\frac{\alpha}{2}}(n_1 + n_2 - 2)$$

1) $\alpha = 0,05$

$t_{1-\frac{0,05}{2}}(16 + 25 - 2) = t_{0,975}(39) = 1,96$

2) $\alpha = 0,10$

$t_{1-\frac{0,1}{2}}(16 + 25 - 2) = t_{0,95}(39) \approx u_{0,95} = 1,645$

Этой формулой можно пользоваться, если

$H_0 : S_1^2 = S_2^2$ *принимается*

Проверим это $T_{экс} = \frac{S_2^2}{S_1^2} \approx 1,19$

1) $T_{кр} = T_{1-\frac{\alpha}{2}}(n_1 - 1; n_2 - 1) = T_{0,975}(15; 24) = 27.$

2) $T_{кр} = T_{0,95}(15; 24) = 2,29$

$T_{хп} < T_{кр} \ \Rightarrow\ H_0$ *принимаем* $\left(H_0 : S_1^2 = S_2^2\right)$

а) $t_{экс} = \dfrac{\left|\overline{x}_1 - \overline{x}_2\right|}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

$$S = \frac{\left(n_1 - 1\right)S_1^2 + \left(n_2 - 1\right)S_2^2}{n_1 + n_2 - 2} = \frac{15 \cdot 1,21 + 24 \cdot 1,44}{39} = 1,35$$

$$t_{экс} = \frac{3,75 - 36,8}{1,35\sqrt{\frac{1}{16} + \frac{1}{25}}} = 1,86$$

$t_{экс} = 1,86 < 1,96 \Rightarrow H_0 - принимаем.$

б) $t_{экс} = 1,86 > t_{кр} = 1,645 \Rightarrow H_0 - отклоняем$

Ответ: *а)* $H_0 - принимаем$

б) $H_0 - отклоняем.$

Лабораторная работа № 6. Критерий Стьюдента проверки гипотез в пакете STATISTICA

Цель работы – изучить возможности применения критерия Стьюдента для проверки гипотезы о равенстве средних величин независимых выборок.

Теоретические сведения

Определение. Пусть случайные величины $\xi_0, \xi_1, \ldots, \xi_n$ – независимы, и каждая из них имеет стандартное нормальное распределение *N(0,1)*. Введем случайную величину

$$t = \frac{\xi_0}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} \xi_i^2}}$$

Ее распределение называют распределением Стьюдента. Саму случайную величину часто называют стьюдентовской дробью, стьюдентовым отношением и т.п. Число *n, n = 1,2,...* называют числом степеней свободы распределения Стьюдента.

Плотность распределения Стьюдента в точке *x* равна

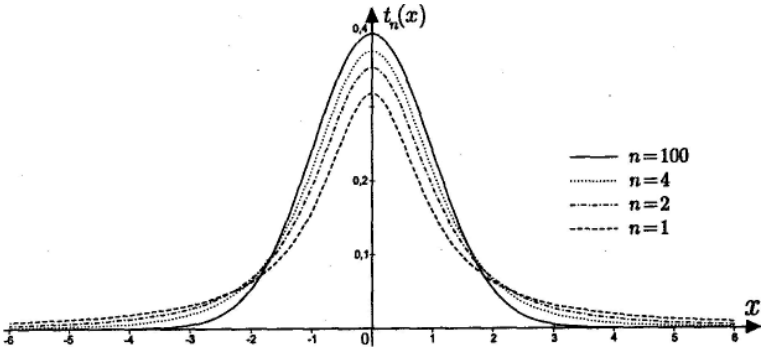
$$\frac{\Gamma((n+1)/2)}{\Gamma(n/2)\sqrt{\pi n}}\left(1+\frac{x^2}{n}\right)^{-(n+1)/2}.$$

Из определения видно, что плотность симметрична относительно $x = 0$. Это обстоятельство используют при составлении таблиц.

На рис. 7.12 изображены функции плотности распределения Стьюдента с различным числом степеней свободы.

Математическое ожидание и дисперсия: распределения Стьюдента имеют вид

$$Mt_n = 0, \qquad Dt_n = \frac{n}{n-2}$$



Рис, 7.12. Функция распределения Стьюдента с различным числом степеней свободы *n.*

Проверка гипотезы о равенстве средних для независимых выборок

Пусть $x_1,...,x_n$, $y_1...,y_m$ - нормальные независимые выборки из законов распределения с параметрами (a_1,σ^2_1) и (a_2, σ^2_2) соответственно. Рассмотрим проверку гипотезы Н: $a_1 = a_2$ против альтернативы $a_1 \neq a_2$.

В самом общем случае, когда обе дисперсии σ^2_1 и σ^2_2 неизвестны, они предполагаются равными. Критерий для проверки гипотезы Н: $a_1 = a_2$ опирается на статистику

$$t = \frac{x - y}{s \cdot \sqrt{\left(\frac{1}{n} + \frac{1}{m}\right)}}$$

Статистика t имеет распределение Стьюдента с $n+m-2$ степенями свободы. Здесь

$$s^2 = \frac{\left(s_1\right)^2(n-1) + \left(s_2\right)^2(n-1)}{(n-1) + (m-1)}$$

Гипотеза Н принимается на уровне значимости α, если $|t| < z_{1-\alpha/2}$. В противном случае гипотеза Н отвергается в

пользу альтернативы.

Задания к лабораторной работе

1. В директории Examples находится файл ***adstudy.sta***, (поставляемый с системой), в который записаны результаты одного социологического опроса: ПОЛ, ПРЕДПОЧТЕНИЕ: PEPSI или COKE, и другие. Определить, зависит ли предпочтение напитка от пола человека.

Шаг 1. Необходимо найти и открыть этот файл в пакете STATISTICA 6.0. Теперь необходимо в Меню выбора основных модулей обработки информации выбрать Статистика(Statistics)►Basic Statistics/Tables.

Шаг 2. В появившемся окне выбрать пункт t-test, independent, by groups.

Шаг 3. В окне *T-Test for independent samples by groups*, нажав на кнопку Variables, установить следующие значения:

- Dependent variables (зависимая): ADVERT (напиток);
- Grouping variables (группирующая): GENDER (пол)
- Нажать OK.

Шаг 4. В возвратившемся окне набрать в поле Code for Group 1: Male (мужчины); в поле Code for Group 2: Female (женщины) и нажать на кнопку Summary.

Наблюдаемая таблица: в столбце **df** указано число степеней свободы - 48, в столбце **t-value** указано значение t статистики равное 1,205214, в столбце ***p*** **–** уровень, на котором можно отвергнуть гипотезу. В данном случае p = 0,234029. Это достаточно большое число. На основании его нельзя отвергнуть гипотезу о равенстве предпочтений в выборе напитка у разных полов. Собранные данные не дают оснований считать, что пол человека влияет на выбор напитка.

2. Проверить гипотезу о значимости ремонта для качества деталей, считая значения переменной *d1* из файла ***3_1.sta*** контролируемым размером диаметра до ремонта, а переменной *d2* - после ремонта станка.

3. Из данных (прил. 1) создать две переменные и проверить гипотезу о равенстве средних для независимых выборок.

Составить отчет по выполненной работе

Отчет по **выполненной** работе должен содержать:

- Постановку задачи.
- Сохраненные на переносном носителе информации, созданные в процессе выполнения лабораторной работы файлы.
- Для наглядности в процессе выполнения работы необходимо сделать несколько Screen Capture, которые в дальнейшем будут размещены в отчете.
- Значения опорных статистик, уровней значимости и статистические выводы.
- Вывод о проделанной работе.

8. КРИТЕРИИ СОГЛАСИЯ И ОДНОРОДНОСТИ

Проверка гипотез о виде функции распределения. Критерии согласия

Постановка задачи. Пусть $\vec{X}=(X_1,...,X_n)$ выборка из распределения $\boldsymbol{L}(\xi)$ с неизвестной функцией распределения $F_\xi(x)$. Выдвигается гипотеза

$$H_0: F_\xi(x)=F_0(x),$$
$$H_1: F_\xi(x)\neq F_0(x)$$

против альтернативы, где $F_0(x)$ выбирается из физических соображений как некоторая гипотетическая функция распределения. Необходимо построить правило, позволяющее на основе значений выборки принять или отвергнуть гипотезу H_0.

Решение: H_0 – простая гипотеза; H_1 – сложная. Формируется некоторая статистика $G=g(\vec{X})$, которая обладает следующими свойствами:

1. Обычно $G\geq 0$;
2. При альтернативе G имеет значение большее, чем при основной гипотезе;
3. Закон распределения статистики G известен точно или асимптотически точно ($n\to\infty$) при основной гипотезе.

Тогда критическая область определяется некоторым значением $g_{кр}$, которое находится из условия:

$$P(G\geq g_{кр}|H_0)=\alpha.$$

Правило принятия решения таково. Если

$$g_{экс}\geq g_{кр}\to\gamma_1 \quad g_{экс}<g_{кр}\to\gamma_0. \tag{8.1}$$

Это правило называют **критерием согласия.**

8.1. Критерий согласия хи - квадрат Пирсона

Этот критерий можно использовать для любых распределений, в том числе и для многомерных. В соответствии с этим критерием, область возможных значений случайной величины ξ разбивается на подобласти с помощью точек $z_0<z_1<...z_m$.

$$P(z_{i-1}\le\xi\le z_i|H_0)=F_0(z_i)-F_0(z_{i-1})=p_i,$$

ν_i - количество элементов выборки, которые попали в интервал (z_{i-1},z_i).

Формируется статистика

$$G=\sum_{i=1}^{m}\left((\nu_i-np_i)^2/np_i\right),$$

где np_i - теоретическое число элементов, попавших в i-тый интервал.

При достаточно большом n эта статистика стремится к χ^2-распределению с (m-1) степенями свобода: $G\xrightarrow[n\to\infty]{F}\chi^2_{m-1}$.

Таким образом, на основе статистики G можно построить следующее правило:

$$g_{экс}\ge g_{кр}\to\gamma_1,$$
$$g_{экс}<g_{кр}\to\gamma_0$$

соответствующее формуле (8.1) из постановки задачи; $g_{кр}$ ищем из условия:

$$P(\chi^2_{m-1}\ge g_{кр})=\alpha$$

$P(\chi^2_{m-1}\ge g_{кр})=1-P(\chi^2_{m-1}<g_{кр})=\alpha$, отсюда имеем:

$$g_{кр}=\chi^2_{m-1,1-\alpha} \qquad (8.2)$$

Формула (8.2) – квантиль распределения χ^2_{m-1} порядка $1-\alpha$.

Замечания.

1. В используемой статистике $G=\sum_{i=1}^{m}\left((\nu_i-np_i)^2/np_i\right)$ число подинтервалов определяется из условия $np_i\ge10$ или $\nu_i\ge10$. При этом длина подинтервалов может быть

разной. Значение $z_0=-\infty$; $z_m=+\infty$ может быть, например, при нормальном законе распределения.

2. При $n\geq 50$ можно считать, что статистика G распределена по закону χ^2.
3. Если случайная величина ξ - дискретная, то разбиение на подинтервалы осуществляется таким образом, чтобы в каждый подинтервал попало значение дискретной случайной величины.
4. Критерий согласия χ^2 можно использовать и тогда, когда распределение $F_0(x)$ известно с точностью до параметра $F_0(x,\theta)$. Если $F_0(x,\theta)$ и $\theta=(\theta_1,...,\theta s)$, то эти параметры можно оценить по той же выборке и подставить в функцию распределения. Тогда $p_i= F_0(z_i,\theta^*)-F_0(z_{i-1},\theta^*)$, а статистика $G\rightarrow \chi^2_{m-s-1}$, где s - число неизвестных оцениваемых по выборке параметров.

Пример 1. От аппаратуры, применяемой при проведении тиража лотереи, требуется, чтобы для 90 возможных значений имелось равномерное распределение. Для проверки в аппарат было положено 5 шаров и проведено n=100 проверочных вынимаий по одному шару. Распределение $F_0(x)$ определяется в соответствии с предположением равномерного распределения пяти возможных значений X (X - номер вынутого шара) следующей функцией вероятности: $p_i=1/5$ (для i=1,...,5). Здесь интервалы - сами возможные значения.

Таблица содержит результаты вынимаий ν_i и данные, нужные для вычислений $\chi^2_{кр}$

i	ν_i	np_i	$\nu_i - np_i$	$\frac{(\nu_i - np_i)^2}{np_i}$
1	18	20	-2	0.20
2	19	20	-1	0.05
3	21	20	1	0.05
4	26	20	6	1.80
5	16	20	-4	0.80
Сумма	100	100	0	2.90

Для α=0.05 и m=5-1=4 степеней свободы $\chi^2_{кр}$=9.5 - найдено по таблице "Критические точки распределения χ^2"[7]. Так как вычисленное значение $\chi^2_{экс}$=2.9 меньше, чем $\chi^2_{кр}$, то результаты тиража не дают повода сомневаться в равномерном распределении.

Пример 2. Случай дополнительных параметров.
Результаты исследования прочности 200 образцов бетона на сжатие представлены в таблице.

Интервал прочности, кг/см^2	Частота, n_i
190-200	10
200-210	26
210-220	56
220-230	64
230-240	30
240-250	14

Проверить гипотезу о законе распределения прочности на сжатие. Уровень значимости принять равным α=0.05.

Решение. Так как точные значения параметров нормального распределения не известны, а объём выборки n=200 достаточно большой, то за их оценки можно принять:

$a=\bar{x}$, $D=S^2$, т.е. будем проверять гипотезу о том, что распределение $F_0(x)$ - нормальное с параметрами $\bar{x}$, S^2.

Определим значения x_i^* середины каждого из шести интервалов $x_1^*=1/2\cdot(200+190)=195$; $x_2^*=205$ и т.д. и вычислим

$\bar{x}=\frac{1}{n}\sum_{i=1}^{6}x_i^*=220$ кг/см^2 - выборочное среднее

$S^2=\frac{1}{n}\sum_{i=1}^{6}(x_i^*-\bar{x})^2\approx 152$ (кг/см^2)^2, - выборочная дисперсия,

$S\approx 12.33$ кг/см^2 – выборочное среднее квадратическое отклонение.

Вычислим теоретические вероятности P_i попадания случайной величины ξ в каждый из 6 интервалов (x_i,x_{i+1}):

$$P_i=p(x_i<\xi\le x_{i+1})=1/2\cdot[\Phi(U_{i+1})-\Phi(U_i)],\ i=1,...,6$$

$$U_i=\frac{x_i-\bar{x}}{S};$$

$$\Phi(U_i)=\frac{2}{\sqrt{2\pi}}\int_0^{U_i}e^{-t^2/2}\,dt$$ - табулировано.

Результаты вычислений сведены в таблицу, где наименьшее значение интервала (-2.1) заменено на (-∞), наибольшее (+2.35) на (+∞)

Интервал изменения ξ	Частота n_i	Нормир интервал U_i U_{i+1}	Вероятность p_i	np_i	$(n_i-np_i)^2$	$\frac{(n_i-np_i)^2}{np_i}$
190-200	10	-∞ -1.70	0.045	9.0	1	0.11
200-210	26	-1.70 -0.89	0.142	28.4	5.76	0.20
210-220	56	-0.89 -0.08	0.281	56.2	0.04	0.001
220-230	64	-0.08 0.73	0.299	59.8	17.64	0.29
230-240	30	0.73 1.54	0.171	34.2	17.64	0.52
240-250	14	1.54 +∞	0.062	12.4	2.56	0.21
Сумма	n=200		1.00	200	$\chi^2_{экс}=1.33$	

По таблице [7] для закона χ^2 по заданному уровню значимости α=0.05 и числу степеней свободы: m=r-s-1=6-2-1=3 находим $\chi^2_{кр}$=7.815. Так как. $\chi^2_{экс}<\chi^2_{кр}$, то нет причины отклонять гипотезу о нормальном законе распределения с параметрами а-220 кг/см^2; σ=12.33 кг/см^2.

8.2. Критерий согласия Колмогорова

Этот критерий применяют в тех случаях, когда функция F(x) непрерывна. Статистикой критерия является величина:

$$G = D_n = \sup_{-\infty<x<\infty} \left|F_n(x) - F(x)\right|. \qquad (8.3)$$

Она представляет собой максимальное отклонение эмпирической функции распределения $F_n(x)$ от гипотетической функции распределения F(x). Это является следствием следующей теоремы.

Теорема.8.1. Относительная частота произвольного события в n независимых испытаниях является оптимальной оценкой для вероятности этого события.

С увеличением объема выборки n происходит сближение $F_n(x)$ с F(x). Поэтому при больших n ($n\to\infty$), когда гипотеза H_0 истинна, значение D_n не должно существенно отклоняться от нуля.

Особенностью статистики D_n является то, что ее распределение при гипотезе H_0 не зависит от вида функции F(x).

Теорема.8.2. Если F(x) - непрерывная функция, то при справедливости гипотезы H_0 закон распределения статистики D_n не зависит от вида функции распределения F(x).

Доказательство. Действительно, полагая в формуле (8.3) $x=F^{-1}(u)$, $0\le u\le 1$, где $F^{-1}(u)$ - функция, обратная к F(x), получаем:

$$D_n = \sup_{0\le u\le 1} \left|F_n(F^{-1}(u)) - u\right|.$$

Перейдем к новым случайным величинам, используя формулу $U_i=F(X_i)$, i=1,...,n; пусть $U_{(1)}\le...\le U_{(n)}$ - их вариационный ряд. Функция F(x) монотонна, поэтому $U_{(k)}=F(X_{(k)})$, k=1,...,n и неравенства $F^{-1}(u)\ge X_{(k)}$ эквивалентны неравенствам $u\ge U_{(k)}$. Используя представление эмпирической функции распределения:

$$F_n(x)=\frac{1}{n}\sum_{k=1}^{n} I\left(X_{(k)}\le x\right)==\frac{1}{n}\sum_{k=1}^{n}\left(X_i\le x\right)$$

имеем:

$$F_n(F^{-1}(u))=\frac{1}{n}\sum_{k=1}^{n} I\left(X_{(k)}\le F^{-1}(u)\right)=\frac{1}{n}\sum_{k=1}^{n} I\left(U_{(k)}\le u\right)=\Phi_n(u)\,.$$

Независимо от вида функции F(x) $\boldsymbol{L}(U_i)=R(0,1)$ и $\Phi_n(u)$ - эмпирическая функция распределения для выборки из равномерного распределения. ▓

Эта теорема позволяет вычислить и протабулировать распределение D_n только один раз (для выборки из равномерного R(0,1) распределения), и использовать ее для проверки гипотезы относительно произвольной непрерывной функции распределения F(x). Функция распределения статистики D_n табулирована при конечных значениях n. При $n\to\infty$ статистика D_n имеет закон распределения Колмогорова

$$F_{D_n}(x)=K(t)=\sum_{i=-\infty}^{\infty}(-1)^i e^{-2i^2t^2}\,.$$

Правило проверки гипотезы на основе критерия Колмогорова: подсчитывается значение статистики $D_n\sim d_{эксс}$

$$d_{эксс}\ge k_{кр}\to\gamma_1,\quad d_{эксс}<k_{кр}\to\gamma_0$$

$k_{кр}$ находится из условия

$$P\left(\sqrt{n}D_n\ge k_{kp}\middle|H_0\right)=1-K(k_{кр})=\alpha \text{ или}$$

$$1-P\left(\sqrt{n}D_n<k_{kp}\middle|H_0\right)=\alpha.\ \text{Отсюда}$$

$$k_{кр}=K_{1-\alpha}\,.$$

$K_{1-\alpha}$ - квантиль распределения Колмогорова порядка $1-\alpha$.

Замечания.

1. В отличие от критерия χ^2, критерий Колмогорова требует точного задания функции F(x).

2. Критерий согласия Колмогорова теоретически обоснован для непрерывных случайных величин.

3. В отличие от критерия χ^2 Пирсона, критерий Колмогорова можно использовать и при n<50 (даже при n≥20).

Пример. Проверить гипотезу о равномерном законе распределения случайных величин, представленных в таблице, при уровне значимости α=0.05.

x	F(x)	$F_n(x)$	$\lvert F_n(x)-F(x)\rvert$
0	0	0	0
0.1	0.1	0.11	0.01
0.2	0.2	0.21	0.01
0.3	0.3	0.334	0.034
0.4	0.4	0.432	0.032
0.5	0.5	0.522	0.022
0.6	0.6	0.641	0.041
0.7	0.7	0.736	0.036
0.8	0.8	0.823	0.023
0.9	0.9	0.899	0.001
1	1	1	0

$D_n = \sqrt{n}\sup_x\left|F_n(x) - F(x)\right|$, n=1000; $F_n(x)=\frac{m}{n}$;

$d_{экс}=\sqrt{1000}\cdot 0.041 \approx 1.3$.

По таблицам критических точек распределения Колмогорова табл.П3.6 находим $K_{kp}=K_{0.95}=1.358$

$d_{экс}<k_{кр}\rightarrow H_0$.

Закон распределения действительно равномерный.

Проверка гипотез о равенстве распределений. Критерии однородности

Постановка задачи. Пусть $\vec{X}=(X_1,...,X_n)$ - выборка из распределения $\boldsymbol{L}(\xi)$ с некоторой неизвестной функцией распределения $F_1(x)$, а $\vec{Y}=(Y_1,...,Y_m)$ - выборка из распределения $\boldsymbol{L}(\eta)$ с неизвестной функцией распределения $F_2(x)$. Требуется проверить гипотезу однородности H_0: $F_1(x)\equiv F_2(x)$. Рассмотрим несколько способов построения критерия однородности для этой гипотезы.

8.3. Критерий однородности Колмогорова - Смирнова

Этот критерий применяют в случае непрерывных распределений. Он использует статистику

$$D_{nm}=\sqrt{\frac{nm}{n+m}}\sup_{-\infty<x<\infty}\left|F_{1n}(x)-F_{2n}(x)\right|, \qquad (8.4)$$

где $F_{1n}(x)$ и $F_{2n}(x)$ - эмпирические функции распределения, построенные по выборкам $\vec{X}$ и $\vec{Y}$ соответственно.

Эмпирическая функция распределения является оптимальной оценкой для теоретической функции распределения и с увеличением объема выборки они сближаются, поэтому, когда справедлива гипотеза H_0 функции $F_{1n}(x)$ и $F_{2ь}(x)$ оценивают одну и ту же неизвестную функцию распределения. В этих случаях статистика D_{nm} не должна отклоняться существенно от нуля. Смирнов Н.В. доказал теорему.

Теорема.8.3. Если $F_1(x)$ и $F_2(x)$ непрерывные функции, то при справедливости гипотезы H_0, статистика D_{nm} не зависит от вида распределения и при $n\to\infty$ и $m\to\infty$, $\frac{n}{m}\to\tau$, $0<\tau<\infty$, случайная величина $\sqrt{\frac{nm}{n+m}}D_{nm}$ распределена по закону Колмогорова. ▩

Отсюда следует критерий проверки гипотезы H_0:

1) вычисляется значение статистики (8.4) ~ $d_{экс}$;
2) с заданным уровнем значимости α находится квантиль распределения Колмогорова $K_{1-\alpha}$;
3) решение принимается следующим образом.

$$d_{экс} \geq K_{1-\alpha} \rightarrow \gamma_1 \quad d_{экс} < K_{1-\alpha} \rightarrow \gamma_0$$

$$P(D_{nm} \geq k_{кр} | H_0) = \alpha, \text{ отсюда}$$

$$1\text{-}P(D_{nm} < k_{кр} | H_0) = \alpha.$$

Другими словами: $F_{D_{nm}}(k_{kp}) = 1 - \alpha \qquad k_{kp} = K_{1-\alpha}.$

8.4. Критерий однородности хи - квадрат

Критерий однородности χ^2 используют для проверки однородности дискретных данных, т.е. когда в опытах наблюдается некоторый переменный признак, принимающий конечное число различных значений Известно, что к такой схеме можно свести любую другую модель, применяя предварительно метод группировки данных. Поэтому критерий χ^2 применим к анализу любых данных, т.е. является универсальным.

Пусть имеется k серий опытов, состоящих из наблюдений за случайной величиной ξ, которая может принимать одно из m возможных значений (ξ - дискретная случайная величина). Серии наблюдений характеризуются выборками

$$\vec{X} = (X_1, ..., X_{n1}) \sim F_1(u)$$

$$\vec{Y} = (Y_1, ..., Y_{n2}) \sim F_2(u)$$

$$.....$$

$$\vec{Z} = (Z_1, ..., Z_{nk}) \sim F_k(u).$$

$F_1(u)$, $F_2(u)$,...,$F_k(u)$ - предполагаемые функции распределения. Требуется проверить гипотезу H_0 о том, что все наблюдения производились над одной и той же случайной величиной

$$H_0: F_1(u) = F_2(u) = ... = F_k(u)$$

против альтернативы H_1: хотя бы одно распределение не равно остальным.

Необходимо построить правило, позволяющее на основе значений выборок принять или отвергнуть H_0.

Для решения используется статистика

$$G = n\left(\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\nu_{ij}^2}{\nu_i n_j} - 1\right), \qquad (8.5)$$

где ν_{ij} - число появлений i-го значения в j-ой серии,

$\nu_{ij} = \sum_{j=1}^{k}\nu_{ij}$ - число появлений i-го значения во всех сериях;

n_j - объем j-ой серии;

$n=\sum_{j=1}^{k} n_j$ - общий объем всех выборок.

При $n\to\infty$ (при увеличении объема выборок) статистика (8.5) будет иметь распределение χ^2 с (m-1)(k-1) степенями свободы, т.е. $G \xrightarrow[n\to\infty]{F} \chi^2_{(m-1)(k-1)}$ при справедливости основной гипотезы. Отсюда следует критерий проверки гипотезы:

1) вычисляется $g_{экс}$;

2) при заданном уровне значимости ищется g_{kp} по таблицам распределения χ^2:

$$g_{kp}=\chi^2_{1-\alpha,(m-1)(k-1)};$$

3) решение принимается следующим образом:

$$g_{экс}\geq g_{kp}\to\gamma_1 \qquad g_{экс}<g_{kp}\to\gamma_0.$$

Замечание.

Критерий χ^2 применяется для непрерывной случайной величин ξ, тогда область возможных значений ξ можно разбить на подинтервалы и ν_{ij} будет числом значений в j-той серии, попавших в i-ый подинтервал.

Пример. Два игрока бросали монету по 100 раз. У первого - герб выпал 57 раз; у второго - 48. Проверить

гипотезу о том, что монеты идентичны, при уровне значимости 0.05.

Решение. Предполагаем, что исследуется случайная величина ξ, показывающая число выпавших гербов при одном подбрасывании.

0	1
1-p	p

У первого игрока $p=p_1$; у второго - $p=p_2$.

$$H_0: \ p_1=p_2$$

$$H_1: \ p_1\neq p_2$$

k=2 - число игроков; m=2 - число возможных значений св. ξ; $n_1=n_2=100$; $n=n_1+n_2=200$

$\nu_{11}=57$ $\nu_{12}=48$ - выпадение герба;

$\nu_{21}=43$ $\nu_{22}=52$ - выпадение решки;

$\nu_1=57+48$ - число появлений герба во всех опытах;

$\nu_2=43+52$ - число появлений решки во всех опытах.

1) Находим $g_{экс}=200\left(\frac{\nu_{11}^2}{n_1\nu_1}+\frac{\nu_{12}^2}{n_2\nu_1}+\frac{\nu_{21}^2}{n_1\nu_2}+\frac{\nu_{22}^2}{n_2\nu_2}-1\right)\approx 16.$

2) По таблицам [7]определяем $g_{kp}=\chi^2_{0.95,1}=3.8.$

3) Решение принимается по следующей схеме:

$g_{экс}>g_{kp}\rightarrow\gamma_1$ - справедлива альтернатива, т.е. монеты не идентичны.

8.5. Непараметрические критерии проверки гипотез

Любой критерий, служащий для проверки гипотез относительно распределения случайной величины, является функцией от наблюдаемых значений этой случайной величины. Такие критерии называются параметрическими. Существуют критерии, вид распределения которых не зависит от распределения генеральной совокупности. Такие критерии называются непараметрическими.

Предположим, что имеются две генеральные совокупности ξ и η, соответствующие непрерывным случайным величинам X и Y. Функции распределения этих совокупностей обозначим так:

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

Необходимо проверить гипотезу: X и Y имеют одно и то же распределение $H_0 : F(x) = G(x)$ для всех x. Альтернативные гипотезы таковы:

$H_1 : F(x) \ne G(x)$; или $H_1 : F(x) < G(x)$; или $H_1 : F(x) > G(x)$.

Проверка производится по 2-м независимым выборкам разного объема из этих совокупностей.

Рассмотрим непараметрические критерии Вилкоксона, Манна-Уитни и знаков. Именно эти критерии реализованы в пакете STATISTICA в модуле “Nonparametric Statistics”. Эти критерии используются для проверки и других гипотез.

Критерий Вилкоксона

Это ранговый критерий. Рассмотри что это такое. Элементы выборки ранжируются, т.е. располагаются в порядке неубывания.

Определение. Рангом элемента выборки называется ее порядковый номер в полученной ранжированной последовательности. ▩

Если встречаются одинаковые элементы, то каждому из них приписывается средний ранг. Критерии, основанные на рангах элементов (а не на самих значениях элементов) называются ранговыми.

Итак, имеются две выборки: $X_1, X_2, ..., X_n$ и $Y_1, Y_2, ..., Y_m$. Объединим их в одну последовательность из n+m элементов и ранжируем ее. Подсчитаем сумму рангов порядковых номеров элементов первой выборки

$X_1, X_2, ..., X_n$ в объединенной последовательности и обозначим ее W. Это и есть критерий Вилкоксона.

Пусть $r_1, r_2, ..r_n$ - ранги (порядковые номера) элементов выборки $X_1, X_2, ..., X_n$ в общем вариационном ряду. Обычно в качестве статистики рангового критерия используют сумму

$$f(r_1) + f(r_2) + ... + f(r_n),$$

где $f(r)$ – некоторая функция, определенная для всех $r = 1,2,..,n+m$.

Пусть $(s_1, s_2, ..., s_{n+m})$ одна из $(n+m)!$ возможных перестановок чисел 1,2,..,n+m. Положим $f(r) = s_r$, тогда статистика Вилкоксона задается формулой

$$W = s_{r_1} + s_{r_2} + .. + s_{r_n}.$$

Односторонний критерий Вилкоксона позволяет принять гипотезу H_0, если W>C, и отвергнуть, если $W \le C$, где С – критическое значение одностороннего критерия Вилкоксона.

Односторонний критерий.

Если $W > C \rightarrow$ принимается H_0

$W \le C \rightarrow$ отвергается H_0, принимается H_1.

Двусторонний критерий.

$C_1 < W < C_2$ $(C_1 < C_2) \rightarrow$ принимается H_0

$W \ge C_2$, либо $W \le C_1 \rightarrow$ отвергается H_0, принимается H_1.

Нижнее C_1 и верхнее C_2 критические значения двустороннего критерия связаны между собой отношениями

$C_1 + C_2 + N$,

где N=m (m+n+1).

Критические значения находят по таблицам [7].

Если объемы выборок велики, то можно воспользоваться асимптотической нормальностью статистики Вилкоксона с математически ожиданием M[W]=m(m+n+1)/2 и дисперсией

D[W]=nm(n+m+1)/12. В этом случае при заданном уровне значимости α для одностороннего критерия имеем

$$C_1 = \frac{N}{2} + \frac{mn(m+n+1)}{12} u_\alpha ,$$

а для двустороннего –

$$C_{1,2} = \frac{N}{2} \pm \frac{mn(m+n+1)}{12} u_{\alpha/2},$$ где $u_\alpha, u_{\alpha/2}$ - квантили стандартного нормального закона.

Критерий Манна-Уитни

Этот критерий проверяет гипотезу об одинаковом распределении двух генеральных совокупностей, как и критерий Вилкоксона.

Рассмотрим всевозможные пары (X_i, Y_j), $i = 1,2,..,m$; $j = 1,2,..,n$. Здесь X_i - i-ый элемент первой выборки, Y_j - j-ый элемент второй выборки. Подсчитаем число пар, для которых $X_i < Y_j$, и обозначим его U. Это и есть критерий Манна-Уитни. Между обоими критериями существует соотношение

$$W+U=mn+m(m+1)/2$$

Это говорит о том, что оба критерия эквивалентны. Математическое ожидание критерия Манна-Уитни находится как M[U]=mn+m(m+1)/2-M[W]=mn/2. Дисперсии в силу линейной зависимости совпадают. При $m \to \infty$ и $n \to \infty$ распределение критерия U стремится к нормальному распределению. В связи с этим все остальное совпадает с процедурой использования критерия Вилкоксона, за исключением того, что M[W] заменяется на M[U]=mn/2.

Критерий знаков

Этот критерий применяется, когда обе выборки имеют одинаковый объем (n=m). Происходит попарное сравнение

элементов выборок: (X_1, Y_1), (X_2, Y_2),.., (X_n, Y_m). Фактически здесь имеется выборка для двумерной случайной величины (X,Y). Пары (X_i, Y_i) i=1,2,..,n являются взаимно независимыми, но компоненты внутри пары могут зависеть друг от друга.

Такая ситуация возникает при анализе эффективности некоторого мероприятия. Для группы объектов измеряется некоторый показатель до и после внедрения данного мероприятия. Результаты измерений для различных объектов можно считать независимыми, но для одного объекта измерения могут быть зависимыми случайными величинами.

Итак, проверяемая гипотеза утверждает, что X_i и Y_i распределены одинаково, т.е.

$$P(X<Y)=P(X>Y)=1/2.$$

Рассмотрим для каждого i разницу $Z_i = X_i - Y_i$, и в общем случае Z=X-Y. Тогда рассматриваемая гипотеза эквивалентна гипотезе

$$P(Z<0)=P(Z>0)=1/2.$$

Следовательно, с равной вероятностью 1/2 знак любого элемента последовательности $Z_1, Z_2, ..., Z_n$ может быть положительным и отрицательным. Критерием R здесь является число положительных знаков. Если гипотеза верна, то R имеет биномиальное распределение с параметрами n и p=1/2. Следовательно, можно применить процедуру проверки гипотезы о вероятности события $p_0 = 1/2$.

Критерий однородности Вилкоксона

Критерий Вилкоксона служит для проверки однородности двух независимых выборок: $x_1, x_2, ..., x_{n1}$ и $y_1, y_2, ..., y_{n2}$. Достоинство этого критерия состоит в том, что он применим к случайным величинам, распределения которых неизвестны.

Если выборки однородны, то они извлечены из одной генеральной совокупности, т.е. $F_1(x) = F_2(x)$.

Нулевая гипотеза утверждает, что функции распределения равны $H_0 : F_1(x) = F_2(x)$.

Альтернативными являются следующие гипотезы:

$$H_1 : F_1(x) \neq F_2(x);\ F_1(x) < F_2(x);\ F_1(x) > F_2(x).$$

Предполагается, что объем первой выборки меньше объема второй: $n_1 \leq n_2$. Если это не так, то выборки можно перенумеровать (поменять местами).

А. Проверка нулевой гипотезы в случае, если объём обеих выборок не превосходит 25

Правило 1. Для того, чтобы при заданном уровне значимости $\alpha = 2Q$ проверить нулевую гипотезу $H_0 : F_1(x) = F_2(x)$ об однородности двух выборок ($n_1 \leq n_2$) при альтернативе $H_1 : F_1(x) \neq F_2(x)$, необходимо

1. Расположить обе выборки в возрастающем порядке, т.е. в виде одного вариационного ряда и найти наблюдаемое значение критерия $W_{набл.}$ – сумму порядковых номеров элементов первой выборки;
2. найти по таблице нижнюю критическую точку $w_{нижс.кр.}(Q, n_1, n_2)$, где $Q = \alpha / 2$;
3. найти верхнюю критическую точку по формуле
   i. $w_{верх.кр.} = (n_1 + n_2 + 1)n_1 - w_{нижн.кр.}$

Если $W_{набл.} < w_{нижн.кр.}$ или $W_{набл.} > w_{верх.кр.}$ - нулевую гипотезу отвергают

Если $w_{нижн.кр.} < W_{набл.} < w_{верхн.кр.}$ - нулевую гипотезу принимают.

Пример. При уровне значимости $\alpha = 0{,}05$ проверить гипотезу $H_0 : F_1(x) = F_2(x)$ об однородности двух выборок объемов $n_1 = 6,\ n_2 = 8$

x_i	15	23	25	26	28	29		
y_i	12	14	18	20	22	24	27	30

при альтернативе $H_1 : F_1(x) \neq F_2(x)$

Решение. Их двух выборок построим один вариационный ряд, элементы пронумеруем (перенумеруем).

Порядковый номер	1	2	3	4	5	6	7
Вариационный ряд	12	14	15	18	20	22	23

Порядковый номер	8	9	10	11	12	13	14
Вариационный ряд	24	25	26	27	28	29	30

Найдем наблюдаемое значение критерия Вилкоксона – сумму порядковых номеров (они подчеркнуты) элементов первой выборки.

$W_{набл.}$=3+7+9+10+12+13=54

Найдем по таблицам [7] нижнюю критическую точку, учитывая, что $Q = \alpha / 2 = 0{,}05/2 = 0{,}025$, $n_1 = 6$, $n_2 = 8$

$w_{нижс.кр.}(0{,}025;6;8) = 29$

Найдем верхнюю критическую точку

$w_{верх.кр.} = (n_1 + n_2 + 1)n_1 - w_{нижн.кр.}$=(6+8+1)*6-29=61

Так как 29<54<61, т.е. $w_{нижн.кр.} < W_{набл.} < w_{верхн.кр.}$ - гипотеза об однородности выборок принимается.

Правило 2. При альтернативе – гипотезе $H_1 : F_1(x) > F_2(x)$ надо найти по таблице [7] нижнюю критическую точку $w_{нижс.кр.}(Q, n_1, n_2)$, где $Q = \alpha$.

Если $W_{набл.} > w_{нижн.кр.}$ - нулевая гипотеза принимается.

Если $W_{набл.} < w_{нижн.кр.}$ - нулевую гипотезу отвергают.

Правило 3. При конкурирующей (альтернативной) гипотезе $H_1 : F_1(x) < F_2(x)$ надо найти верхнюю критическую точку

$$w_{верх.кр.}(Q; n_1; n_2) = (n_1 + n_2 + 1)n_1 - w_{нижн.кр.}(Q; n_1; n_2),$$

где $Q = \alpha$.

Если $W_{набл.} < w_{верх.кр.}$ - нулевую гипотезу принимают.

Если $W_{набл.} > w_{верх.кр.}$ - нулевую гипотезу отвергают.

Замечание.

Если несколько элементов только одной выборки одинаковы, то в общем вариационном ряду их нумеруют как различные числа. Если совпадают элементы разных выборок, то им присваивают один и тот же порядковый номер, равный среднему арифметическому порядковых номеров, которые имели бы эти элементы до совпадения.

Б. Проверка нулевой гипотезы в случае, если объём хотя бы одной из выборок больше 25.

Правило 1. При альтернативе $H_1 : F_1(x) \neq F_2(x)$ нижняя критическая точка определяется по формуле

$$w_{нижс.кр.}(Q, n_1, n_2) = \left[\frac{(n_1 + n_2 + 1)n_1 - 1}{2} - z_{кр} \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \right], \quad (*)$$

где $Q = \alpha / 2$;

$z_{кр}$ находят по табл. П3.2. функции нормального распределения по равенству:

$$\Phi(z_{кр}) = \frac{(1-\alpha)}{2},$$

знак [ ] означает целую часть числа. В остальном правило 1, приведенное п. А, сохраняется.

Правило 2. При альтернативах $F_1(x) < F_2(x)$ и $F_1(x) > F_2(x)$ нижнюю критическую точку находят по формуле (*), положив $Q = \alpha$;

$z_{кр}$ находят по табл. П3.2. функции нормального распределения по равенству

$$\Phi(z_{кр}) = \frac{(1-2\alpha)}{2}.$$

В остальном правила 2-3, приведенные в п. А, сохраняются.

Задачи и решения

Задача 53

При 50 подбрасываниях монеты герб появился 20 раз. Можно ли считать монету симметричной? Принять α=0,10.

Решение:

При решении этой задачи используется статистика

$$G = \sum_{i=1}^{m} \left((\nu_i - np_i)^2 / np_i \right).$$

В этом случае число интервалов разбиения определяется из условия $np_i \geq 10$ или $\nu_i \geq 10$.

При этом длина интервалов может быть разной.

Следует заметить, что при $n \geq 50$ можно считать, что статистика G распределена по закону χ^2 в соответствии с теоремой Пирсона.

$n = 50;\ \ \nu_1 = 20;\ \ \nu_2 = 30;\ \ \alpha = 0{,}1;$

$$\chi^2_{экс} = \sum \frac{(\nu_i - np_i)^2}{np_i}$$

$$p_i = \frac{1}{2} \quad (т.к.\ подбрасываем\ монету)$$

$$\chi^2_{экс} = \frac{(20-25)^2}{25} + \frac{(30-25)^2}{25} = 2$$

$r - l - 1 = 2 - 0 - 1 = 1$ (r – *число разрядов, на разбиен.опытные значения*; l – *число параметров*)

$\chi^2_{1-0,1,1} = 2,706$

$\chi^2_{экс} < \chi^2_{кр} \ \Rightarrow\ H_0$ – *принимаем.*

Ответ: H_0 – *принимаем.*

Задача 54

Число выпадений герба при 20 подбрасываниях двух монет распределились следующим образом:

Количество гербов	0	1	2
Число подбрасываний	4	8	8

Согласуются ли эти результаты с предположениями о симметричности монет и независимости результатов подбрасываний? Принять α=0,05.

Решение:

При решении этой задачи, как и задачи 52, также используется статистика вида

$$G = \sum_{i=1}^{m} \left((\nu_i - np_i)^2 / np_i \right).$$

число гербов 0 1 2

число подбрасываний 4 8 8

α=0,05

$$\chi^2_{экс} = \frac{\left(4 - 20\frac{1}{2}\right)^2}{10} + \frac{(8-10)^2}{10} + \frac{(8-10)^2}{10} = 4,4;$$

$$\chi^2_{кр} = \chi^2_{1-0,05,2-l-1} = \chi^2_{0,95,2} = 5,99$$

$$r - l - 1 = 3 - 0 - 1 = 2$$

$$\chi^2_{экс} < \chi^2_{кр} \Rightarrow H_0 - принимаем.$$

Ответ: $H_0 - принимаем.$

Задача 55

Для определения зависимости цвета волос жителей от их местожительства были обследованы три группы людей из районов A, B и C. Свидетельствуют ли приводимые ниже результаты обследования о зависимости цвета волос жителей от их местожительства? Принять α=0,05.

Район	Цвет волос		
	Рыжий	Светлый	Темный
A	2	9	9
B	3	6	21
C	15	15	20

Решение:

Для решения данной задачи следует применить критерий однородности хи-квадрат. Статистика критерия такова:

$$G = n\left(\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\nu_{ij}^2}{\nu_i n_j} - 1\right),$$

$\alpha = 0,05$

$$\chi^2_{экс} = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{\left(v_{ij} - \frac{v_i v_j}{n}\right)^2}{\frac{v_i v_j}{n}} = n\left(\sum_{i=1}^{3}\sum_{j=1}^{3} \frac{v_{ij}}{v_i v_j} - 1\right)$$

$$\chi^2_{экс} = 100(\frac{2^2}{20\cdot 20} + \frac{9^2}{30\cdot 20} + \frac{9^2}{30\cdot 20} + \frac{3^2}{20\cdot 30} + \frac{3^2}{20\cdot 30} + \frac{6^2}{50\cdot 30} + \frac{21}{50\cdot 50} +$$

$$+ \frac{15^2}{30\cdot 50} + \frac{20}{50\cdot 50} - 1) = 100\cdot 0,11 = 11$$

$$\chi^2_{кр} = \chi^2_{1-\alpha,(m-1)(k-1)} = \chi^2_{0,95,4} = 9,49$$

$$\chi^2_{экс} > \chi^2_{кр} \quad \Rightarrow \quad H_0 - \textit{отклоняется}$$

Ответ: H_0 – *отклоняется*.

Задача 56

Отдел технического контроля проверил n=200 партий одинаковых изделий и получил распределение:

x_k	0	1	2	3	4
n_k	116	56	22	4	2

(где в первой строке стоит количество нестандартных изделий в одной партии, во второй количество партий содержащих это количество нестандартных изделий). Требуется при уровне значимости 0,05 проверить гипотезу о том, что число нестандартных изделий X распределено по закону Пуасона.

Решение:

1.Найдем выборочную среднюю

$$\overline{x} = \frac{\sum_{k=0}^{4} n_k x_k}{n} = (116*0+56*1+22*2+4*3+2*4)/200=0.6$$

2. Примем в качестве оценки параметра λ распределения Пуассона выборочную среднюю λ=0.6 , то закон распределения будет иметь вид:

$$P_{200}(k) = \frac{(0.6)^k}{k!} e^{-0.6} \equiv p_k$$

3. По табл. П3.1 находим вероятность $p_к$ появления k нестандартных изделий в 200 партиях, положив k=0, 1, 2, 3, 4

$p_0=P_{200}(0)=0.5488$

$p_1=P_{200}(1)=0.3293$

$p_2=P_{200}(2)=0.0988$

$p_3=P_{200}(3)=0.0198$

$p_4=P_{200}(4)=0.0030$

4. Найдем теоретические частоты по формуле

$$n'_k = np_k = 200p_k .$$

Подставив в эту формулу найденные в пункте 3 значения вероятностей $p_к$, получим:

$$n'_0 = 200*0,5488=109,76$$

$$n'_1 = 200*0,3293=65,86$$

$$n'_2 =200*0,0988=19,76$$

$$n'_3 =200*0,0198=3,96$$

$$n'_4 =200*0,0030=0,6$$

5. Сравним эмпирические и теоретические частоты с помощью критерия χ^2 Пирсона. Для этого составим расчетную таблицу, учитывая значения, объединим малочисленные частоты (4+2=6) и соответствующие им теоретические частоты (3,96+0,60=4,56). Результаты занесем в таблицу

k	n_k	n'_k	$n_k - n'_k$	$(n_k n'_k)^2$-	$(n_k - n'_k)^2/n'_k$
0	116	109,76	6,24	38,9376	0,3548
1	56	65,86	-986	97,2196	1,4762
2	22	19,76	2,24	5,0176	0,2539
3	6	4,56	1,44	2,0736	0,4547
$\sum$	200				$\chi^2_{набл}$=2,54

Из расчетной таблицы находим наблюдаемые значения критерия Пирсона: $\chi^2_{набл}$=2,54. По выборке оценивался один параметр λ, то число степеней свободы L=4-1-1=2, где 4-число групп выборок после сокращения. По таблице [7] находим $\chi^2_{,.95}(2)=6,0$, т.е. $\chi^2_{набл}<\chi^2_{кр}$, и гипотеза о распределении случайной величины X по закону Пуассона принимается.

Задача 57

Измерения 1000 деталей представлены в виде группированной выборки в таблице.

i	x_i	$n_i m_i$	i	x_i	$n_i m_i$
1	98,0	21	6	100,5	201
2	98,5	47	7	101,0	142
3	99,0	87	8	101,5	97
4	99,5	158	9	102,0	41
5	100,0	181	10	1025	25

Проверить, пользуясь критерием Колмогорова, согласие полученных наблюдений с предположением, что величина X подчиняется нормальному закону с математическим. ожиданием $\bar{x}$=100,25 мм и СКО δ=1 мм при уровне значимости α=0,005.

Решение:

Теоретическая функция распределения определяется формулой

$$F(x) = \frac{1}{2} + \frac{1}{2}\Phi(x - \bar{x})$$

Эмпирическая функция распределения $F_n(x)$ определяется по формуле

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{k} m_i n_i \quad F_k(x) = \frac{1}{n}\sum_{i=1}^{k} m_i$$

Расчеты заносим в таблицу

i	$x_i - \bar{x}$	$\frac{1}{2}\Phi(x_i - \bar{x})$	$F(x_i)$	$F_n(x_i)$	$\lvert F_n(x_i) - F(x_i)\rvert$
1	-2,25	-0,4877	0,0123	0,0210	0,0087
2	-1,75	-0,4599	0,0401	0,0680	0,0279
3	-1,25	-0,3944	0,1056	0,1550	0,0494
4	-0,75	-0,2734	0,2266	0,3130	0,0864
5	-0,25	-0,0987	0,4013	0,4940	0,0927
6	0,25	0,0987	0,5987	0,6950	0,0963
7	0,75	0,2734	0,7734	0,8370	0,0636
8	1,25	0,3944	0,8944	0,9340	0,0396
9	1,75	0,4599	0,9599	0,9750	0,0151
10	2,25	0,4877	0,9877	1,000	0,0123

$\frac{1}{2}\Phi(x_i - \bar{x})$ - ищется в табл. П3.2.

Из таблицы следует, что

$$d_{экс} = \sqrt{n}\sup_x |F_n(x) - F(x)| = \sqrt{1000} * 00963 = 31.6 * 0.0963 = 3.04 .$$

Для α=0,005 из табл. П3.6 находим $\lambda_a = k_{кр} = 1{,}358$

$$d_{экс} > k_{кр} \qquad 3{,}04 > 1{,}224 ,$$

следовательно, принимается альтернатива H_1, т.е. гипотеза H_0 о нормальности распределения выборки наблюдений отвергается.

Лабораторная работа № 7. Критерии хи-квадрат проверки гипотез в пакете STATISTICA

Цель работы – изучить применение критерия хи-квадрат для анализа законов распределения случайных величин, полученных в результате эксперимента.

Теоретические сведения

Проверка простой гипотезы о вероятностях

Обозначим: A_1, ..., A_m - m возможных исходов некоторого опыта;

p_1, ..., p_m - вероятности соответствующих исходов, $\sum_{i=1}^{m} p_i = 1$;

n - число независимых повторений опыта;

ν_1, ..., ν_m - число появлений соответствующих исходов в n опытах, $\sum_{i=1}^{m} \nu_i = n$;

p_1^0, ..., p_m^0 - гипотетические значения вероятностей, $p_i^0 > 0$, $\sum_{i=1}^{m} p_i^0 = 1$.

Требуется по наблюдениям $\nu_1,...,\nu_m$ проверить гипотезу H о том , что вероятности p_1, ..., p_m имеют значения p_1^0, ..., p_m^0, т.е. H: $p_i = p_i^0$, i=1, ...,m.

Оценками для p_1, ..., p_m являются $\hat{p}_1 = \nu_1 /n$, ..., $\hat{p}_m = \nu_m/n$. Мерой расхождения между гипотетическими и эмпирическими вероятностями принимается величина

$$X^2 = n\sum_{i=1}^{m} p_i^0 \left(\frac{\widehat{p}_i - p_i^0}{p_i^0} \right)^2,$$

которая с точностью до множителя n есть усредненное с весами p_i^0 значение квадрата относительного отклонения значений $\widehat{p}_i$ от p_i^0.

Статистика X^2 называется статистикой хи-квадрат Пирсона. Для ее вычисления используются две формулы:

$$X^2 = \sum_{i=1}^{m} \frac{(\nu_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^{m} \frac{\nu_i^2}{np_i^0} - n.$$

Условно статистику можно записать так:

$$X^2 = \sum \frac{(\mathrm{H} - \mathrm{T})^2}{\mathrm{T}},$$

где Н - наблюдаемые частоты ν_i, Т - теоретические (ожидаемые) частоты np_i^0.

Поскольку по закону больших чисел $\widehat{p}_i \to p_i$ при $n \to \infty$, то

$$\sum_{i=1}^{m} p_i^0 \left(\frac{\widehat{p}_i - p_i^0}{p_i^0} \right)^2 \to \sum_{i=1}^{m} \frac{(p_i - p_i^0)^2}{p_i^0}.$$

Последняя величина равна 0, если верна H; если же H не верна, то $X^2 \to \infty$.

Процедура проверки гипотезы состоит в том, что если величина X^2 приняла "слишком большое" значение, т.е. если

$$X^2 \geq h, \qquad (8.6)$$

то гипотеза H отклоняется; если это не так, будем говорить, что наблюдения не противоречат гипотезе. На вопрос, что означает "слишком большое" значение, отвечает теорема Пирсона.

Теорема К. Пирсона. Если гипотеза H верна и $p_i^0 > 0$, i=1,...,m, то при $n \to \infty$ распределение статистики X^2 асимптотически подчиняется распределению хи-квадрат с m-1 степенями свободы, т.е. $P\{ X^2 < x / H \} \to F_{m-1}(x) \equiv P\{ \chi^2_{m-1} < x \}$.

Порог h выберем из условия: вероятность ошибки первого рода должна быть малой, равной выбираемому значению α - уровню значимости:

$P\{$ отклонить H / H верна$\} = P\{ X^2 \geq h / H \} \cong P\{\chi^2_{m-1} \geq h\} = \alpha$, откуда

$$h = Q(1-\alpha, n-1), \tag{8.7}$$

квантиль уровня $1-\alpha$ распределения хи-квадрат с m -1 степенями свободы. Процедура (8.6) - (8.7) проверки H может быть записана иначе: гипотеза H отклоняется, если:

$$P\{\chi^2_{m-1} \geq X^2\} \leq \alpha, \tag{8.8}$$

т.е. если мала вероятность получения (при справедливости H) такого же расхождения, как в опыте (т.е. X^2), или ещё большего. Вероятность слева в (8.8) называется минимальным уровнем значимости (при любом значении α, большем $P\{X^2_{m-1} \geq X^2\}$, гипотеза, очевидно, отклоняется).

Замечание.

Теорему Пирсона можно применять, если все ожидаемые частоты удовлетворяют условию: $np_i^0 \geq 10, \quad i=1, ...,m.$

Если m порядка десяти и более, то достаточно выполнения данного: условия: $np_i^0 \geq 4, \quad i=1, ...,m.$

Если рассмотренные условия не выполняется, то необходимо некоторые исходы A_i объединять.

Проверка сложной гипотезы о вероятностях

Пусть $A_1, ..., A_m$ - m исходов некоторого опыта,
n - число независимых повторений опыта,

$\nu_1,...,\nu_m$ - числа появлений исходов.

Проверяемая гипотеза H предполагает, что вероятности исходов $P(A_i)$ являются известными функциями $p_i(a)$ k-мерного параметра $a = (a_1,...,a_k)$, т.е. $H: P(A_i) = p_i(a)$, $i = 1, ..., m$, но значение a неизвестно. Для проверки гипотезы H определим статистику

$$\tilde{X}^2 = \min_a \sum_{i=1}^{m} \frac{(\nu_i - np_i(a))^2}{np_i(a)} \tag{8.9}$$

По теореме Фишера, если H верна, то при $n \to \infty$ распределение статистики X^2 асимптотически подчиняется распределению *хи-квадрат* с числом степеней свободы $f = m - 1 - k$, и потому ***отклоняем*** H, если

$$\tilde{X}^2 \geq h, \tag{8.10}$$

где $h = Q(1-\alpha, f)$ - квантиль уровня $1-\alpha$ распределения *хи-квадрат* с числом степеней свободы f; такой порог обеспечивает выбранный уровень α вероятности P (отклонить H / H) ошибки 1-го рода. Если (8.10) не выполняется, делаем вывод, что ***наблюдения не противоречат гипотезе***. Распределению *хи-квадрат* с $f = m–1–k$ степенями свободы асимптотически подчиняется также статистика

$$\tilde{X}^2 = \sum_{i=1}^{m} \frac{(\nu_i - np_i(\hat{a}))^2}{np_i(\hat{a})}, \tag{8.11}$$

где $\hat{a}$ - оценка максимального правдоподобия для a, и потому в (8.10) может быть использована статистика (8.11) вместо (8.9). Процедура (8.10) может быть записана иначе: если

$$P\{\chi_f^2 \geq X^2\} \leq \alpha,$$

то гипотеза H отклоняется.

Проверка гипотезы о типе распределения

Пусть требуется проверить гипотезу о том, что выборка $x_1, ..., x_n$ извлечена из совокупности, распределенной по некоторому закону, известному с точностью до k-мерного параметра $a=(a_1,...,a_k)$. Оказываются теоретически обоснованными следующие действия: разобьем весь диапазон наблюдений на m интервалов, определим значения ν_i -число наблюдений в i-м интервале, получим значение оценки $\hat{a}$ минимизацией (8.9) или методом максимального правдоподобия, определим вероятности $p_i(\hat{a})$ попадания в i-й интервал, вычислим (8.9) или (8.11) и примем решение по (8.10).

Проверка гипотезы о независимости признаков (таблица сопряженности признаков)

Предположим, имеется большая совокупность объектов, каждый из которых обладает двумя признаками A и B; признак A имеет m уровней: $A_1, ..., A_m$, а признак B – k уровней: $B_1, ..., B_k$. Пусть уровень A_i встречается с вероятностью $P(A_i)$, а уровень B_j - с вероятностью $P(B_j)$. Признаки A и B независимы, если

$$P(A_i B_j) = P(A_i)\cdot P(B_j), \quad i = 1, ..., m, j = 1, ..., k$$

т.е. вероятность встретить комбинацию A_i B_j равна произведению вероятностей. Пусть признаки определены на n объектах, случайно извлеченных из совокупности; ν_{ij} - число объектов, имеющих комбинацию A_i B_j, $\sum_{i=1}^{m}\sum_{j=1}^{k}\nu_{ij}=n$. По совокупности наблюдений $\{\nu_{ij}\}$ (таблица $m \times k$) требуется проверить гипотезу H о независимости признаков A и B. Задача сводится к случаю с неизвестными параметрами; ими являются вероятности

$$P(A_i),\ i = 1, ..., m;\ P(B_j),\ j = 1, ..., k,$$

всего *(m*-1*)* + *(k*-1*)*; их оценки:

$$\widehat{P}(A_i) = \frac{\sum_{j=1}^{k} \nu_{ij}}{n} \equiv \frac{\nu_{i\bullet}}{n} \quad , \quad \widehat{P}(B_j) = \frac{\sum_{i=1}^{m} \nu_{ij}}{n} \equiv \frac{\nu_{\bullet j}}{n}$$

(в обозначениях точка означает суммирование по соответствующему индексу), и статистика (8.9) принимает вид:

$$\widetilde{X}^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{\nu_{ij}^2}{n\widehat{P}(A_i)\widehat{P}(B_j)} - n = n\left(\sum_{i=1}^{m}\sum_{j=1}^{k} \frac{\nu_{ij}^2}{\nu_{i\bullet}\nu_{\bullet j}} - 1\right) \quad (8.12)$$

Если гипотеза *H* верна, то по теореме Фишера $\widetilde{X}^2$ асимптотически распределена по закону хи-квадрат с числом степеней свободы

$$f = mk - 1 - (m - 1) - (k - 1) = (m - 1)(k - 1),$$

и потому, если

$$P\{\chi_f^2 \geq \widetilde{X}^2\} \leq \alpha, \quad (8.13)$$

то гипотезу о независимости признаков следует отклонить.

Ясно, что по (8.12) - (8.13) можно проверять независимость двух случайных величин, разбив диапазоны их значений на *m* и *k* частей.

Проверка гипотезы об однородности выборок

Пусть имеется *m* выборок объемами $n_1,..., n_m$, извлеченных из различных совокупностей. Измеряемая величина в каждой из выборок может иметь *k* уровней $B_1, ..., B_k$. Требуется проверить гипотезу о том, что исходные совокупности распределены одинаково. Обозначим ν_{ij} - число наблюдений в *i*-й выборке, имеющих уровень B_j, $\sum_j \nu_{ij} \equiv \nu_{i\bullet} = n_i$. Имеем таблицу $m \times k$ наблюдений аналогично предыдущему пункту. Можно показать, что для проверки гипотезы справедлива процедура (8.12) - (8.13).

Задания к лабораторной работе

1. Необходимо проверить гипотезу о нормальном законе распределения.

Проверим гипотезу о нормальном законе распределения диаметров валов, выточенных на одном станке, по выборке объема $n = 200$; измерения приведены в прил. 2. Оценками для a (среднего) и σ (стандартного отклонения) являются:

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

Шаг 1. Для начала результаты измерения диаметров валов, взятых из ***приложения 2,*** занесем в таблицу с одним столбцом (d) и 200 строками; соответствующий файл назовем, ***8_1.sta***

Шаг 2. Теперь необходимо в Меню выбора основных модулей обработки информации в программном обеспечении STATISTICA6.0 выбрать Статистика(Statistics)►Distribution Fitting (подбор распределения). В появившемся окне выбрать поле Continuous Distributions: Normal и нажать ОК.

Шаг 3. В окне Fitting Continuous Distributions выбрать Variable: *d* и перейти на вкладку Options. Далее в поле Plot distribution: *Frequency distribution* (частоты распределения) и отказываемся от теста Колмогорова-Смирнова. Для получения конечного результата нажать кнопку Summary. Перед вами сформированная таблица частот, в которой нам нужны столбцы *observed frequency* (наблюдаемые частоты) и *expected frequency* (ожидаемые частоты). Сравним графически наблюдаемые и ожидаемые частоты построением 2D Histogram. Наблюдаем некоторое различие.

В таблице приведено значение статистики (8.11) Chi-Square: 12.55864, количество степеней свободы d.f. = 3, которое получилось при объединении интервалов для выполнения условий (4.5): $f = 6 - 1 - 2 = 3$. Приведено

значение вероятности:

$$P\{\chi^2_3 \geq 12.55864\} = p = 0.00570.$$

Последнее означает, что если гипотеза верна, вероятность получить 12.55864 или больше равна 0.00570 - слишком мала, чтобы поверить в нормальность. Гипотезу о нормальности отклоняем.

Если посмотреть гистограмму наблюдений, видно, что в выборке имеется одно аномальное значение 14.56 (№ 188), которое могло появиться в результате какой-либо ошибки (при записи наблюдений, при перепечатке или попалась деталь с другого станка и т.д.). Удалим его и снова проверим гипотезу. Удаление одного наблюдения, если оно типично, не может изменить характеристики совокупности из 200 элементов; если же изменение происходит, следовательно, это наблюдение типичным не является и должно быть удалено.
Чтобы не портить исходные данные, продублируем их в новый столбец, например, *dc*, и удалим аномальное наблюдение. Повторим проверку гипотезы для "цензурированной" выборки и убедимся в том, что наблюдения не противоречат гипотезе о нормальности.

2. Необходимо проверить простую гипотезу о распределении.

Проверим генератор случайных чисел. Сгенерируем выборку заданного объема с заданным в табл. 8.1 законом распределения и по полученным результатам проверим гипотезу о согласии данных с этим распределением. Созданный файл с выборкой назовем ***8_2.sta***. В таблице приняты обозначения для распределений: *R* - равномерное, *N* - нормальное, *E* - показательное, *Bi* - биномиальное, *P* - Пуассона.

Таблица 8.1

Исходные данные для построения выборки в зависимости от варианта

№ варианта	Распределение	Объем выборки
1.	*R*[0, 5]	130
2.	*N*(10, 2^2=4)	140
3.	*E*(3)	140
4.	*B*(10, 0.5)	160
5.	*P(15)*	130
6.	*beta*(1, 1)	140
7.	*R*[0, 10]	130
8.	*N*(15, 3^2=9)	160
9.	*E(5)*	130
10.	*B*(15, 0.3)	140
11.	*P(20)*	150
12.	*beta*(2, 2)	160
13.	*R*[-1, 1]	130
14.	*N*(0, 1)	140
15.	*E*(1)	150

Задание выполняется аналогично предыдущему с некоторыми отличиями:

- в окне Fitting Continuous Distribution нужно ввести значения параметров распределения (вместо их оценок) и, возможно, поправить параметры группировки;
- приводимый результат для уровня значимости p не соответствует рассматриваемому случаю, так как число степеней свободы *d.f.* должно быть равным m -1; пакет же указывает с учетом числа оцениваемых параметров. Нужное значение для p получим в модуле Basic Statistics and Tables *в* Probability calculator.

3. Необходимо проверить гипотезу о независимости признаков.

Группа данных, собранных по ряду школ, относительно физических недостатков школьников (P_1, P_2, P_3 - признак A) и дефектов речи (S_1, S_2, S_3 - признак B) приведены в табл. 8.2. Помимо этого в табл. 8.3 даны частоты.

Для проверки гипотезы о независимости этих двух признаков вычислим статистику (8.11): $\tilde{X}^2 = 34.88$; число степеней свободы $f = (3-1)\times(3-1) = 4$; минимальный уровень значимости

$$P\{\chi_4^2 \geq 34.88\} \leq 0.001;$$

это значит, что при независимых признаках вероятность получить значение такое же, как в опыте или большее, меньше 0.001, и потому гипотезу о независимости следует отклонить.

Шаг 1. Образуем таблицу с двумя столбцами (P и S) и 217 строками и назовем ее ***8_3.sta*** в соответствии с табл. 8.2.

Таблица 8.2

Дефекты речи (S) и физические недостатки (P) школьников

	P S	P S	P S	P S	P S	P S	P S	P S	P S
1	P1 S1	P1 S1	P3 S2	P2 S2	P1 S3	P1 S1	P1 S1	P2 S1	P3 S3
2	P2 S3	P2 S2	P1 S3	P1 S1	P2 S2	P2 S1	P2 S2	P3 S3	P1 S1
3	P1 S1	P2 S3	P1 S2	P1 S1	P2 S2	P2 S2	P1 S3	P3 S2	P2 S3
4	P1 S2	P2 S3	P3 S1	P2 S1	P2 S2	P3 S3	P1 S1	P2 S1	P1 S3

Продолжение табл. 8.2

5	P1 S1	P2 S1	P2 S1	P1 S1	P1 S1	P2 S1	P2 S2	P2 S3	P2 S2
6	P3 S3	P1 S2	P3 S3	P2 S2	P1 S3	P1 S1	P2 S3	P1 S1	P2 S1
7	P1 S1	P2 S3	P1 S2	P2 S2	P2 S1	P2 S2	P1 S3	P2 S3	P1 S1
8	P1 S2	P1 S1	P2 S3	P1 S2	P2 S2	P1 S3	P2 S2	P2 S2	P3 S3
9	P2 S2	P2 S1	P1 S2	P1 S1	P2 S2	P2 S3	P2 S3	P1 S2	P2 S1
10	P2 S2	P2 S1	P2 S2	P1 S3	P3 S3	P1 S1	P1 S3	P2 S2	P2 S2
11	P2 S2	P2 S1	P1 S2	P1 S2	P2 S1	P1 S1	P1 S3	P1 S2	P1 S1
12	P1 S2	P2 S1	P1 S2	P2 S2	P1 S1	P1 S1	P1 S1	P2 S3	P2 S1
13	P1 S1	P3 S3	P2 S2	P2 S2	P2 S2	P2 S1	P2 S3	P2 S2	P2 S2
14	P2 S3	P1 S1	P2 S3	P2 S1	P2 S1	P1 S2	P2 S1	P1 S2	P3 S3
15	P2 S1	P1 S1	P3 S2	P2 S2	P1 S1	P2 S2	P3 S2	P2 S2	P1 S2
16	P2 S1	P2 S1	P1 S2	P2 S1	P2 S2	P3 S3	P2 S2	P2 S3	P3 S3
17	P3 S2	P1 S1	P2 S2	P3 S3	P1 S1	P2 S1	P2 S2	P1 S1	P1 S2
18	P1 S1	P2 S2	P1 S1	P3 S2	P3 S3	P2 S2	P1 S2	P1 S2	

Продолжение табл. 8.2

19	P1 S2	P3 S3	P2 S1	P1 S1	P1 S1	P2 S2	P1 S1	P1 S1	
20	P3 S3	P3 S3	P1 S1	P1 S1	P3 S2	P1 S1	P1 S1	P2 S1	
21	P2 S2	P2 S1	P2 S3	P3 S2	P2 S2	P1 S2	P2 S1	P2 S2	
22	P1 S3	P1 S1	P2 S2	P2 S2	P3 S1	P2 S2	P2 S3	P1 S1	
23	P2 S3	P2 S2	P3 S3	P3 S3	P1 S1	P2 S1	P1 S1	P2 S1	
24	P3 S2	P3 S2	P2 S3	P1 S3	P2 S2	P3 S2	P2 S2	P1 S2	
25	P3 S1	P2 S3	P2 S1	P1 S2	P2 S2	P1 S2	P2 S1	P2 S2	

Таблица 8.3

Таблица частот

	S_1	S_2	S_3	Сумма
P_1	45	26	12	83
P_2	32	50	21	103
P_3	4	10	17	31
Сумма	81	86	50	217

Шаг 2. Теперь необходимо в Меню выбора основных модулей обработки информации в программном обеспечении STATISTICA6.0 выбрать Статистика(Statistics)►Basic Statistics and Tables►Tables and banners►OK. В появившемся окне нажать на кнопку *Specify Table* и в появившемся окне

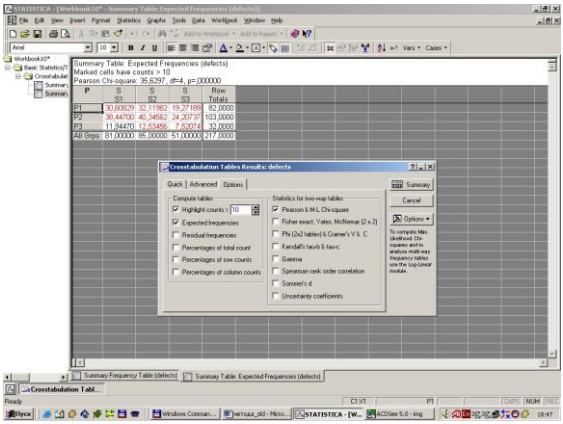
отобрать признаки: list 1: *P*, list 2: *S*. Нажать ОК в форме *Specify Table* и в форме *Crosstabulation tables.*

Шаг 3. В следующем окне (*Crosstabulation Tables Results*) необходимо перейти на вкладку OPTIONS и отметить следующее:

- Highlight counts > = 10;
- Expected frequencies (ожидаемые или теоретические частоты);
- *Pearson & M-L Chi-square.*

После выполнения вышеописанных действий нажать на кнопку Summary (см. рисунок).

Наблюдаем две таблицы: таблицу частот *Summary Frequency Table* и *Expected Frequencies*; в верхней части последней указано значение статистики (8.12) (*Chi-square*), число степеней свободы *df* и уровень значимости *p* (вероятность в (8.13)). Поскольку значение *p* мало, гипотеза о независимости речевых и физических дефектов отклоняется.



Установки в форме Crosstabulation Tables Results

Замечание.

Если бы исходные признаки *X*, *Y*,... были не символьными, а числовыми, нужно было бы сначала их классифицировать: разбить диапазон значений на части, и для

каждой ввести свой символ (например, *x*1, *x*2, ..., *y*1, *y*2, ...) введением дополнительных столбцов и использованием операции *Recode...* (кнопка *Vars* или *Edit - Variables*).

4. Необходимо проверить гипотезу об однородности выборок.

Имеются данные о наличии примесей серы в углеродистой стали, выплавляемой двумя заводами (см. табл. 8.4). Проверим гипотезу о том, что распределения содержания серы (нежелательный фактор) одинаковы на этих заводах.

Таблица 8.4

Число плавок

	Содержание серы, 10^{-2} %				
	0÷2	2÷4	4÷6	6÷8	Сумма
Завод 1	82	535	1173	1714	3504
Завод 2	63	429	995	1307	2794
Сумма	145	964	2168	3021	

По (8.12) находим: $\tilde{X}^2 = 3.39$. Число степеней свободы f = (2-1)×(4-1) = 3; квантиль уровня 0.95 *h* = *Q*(0.95, 3) = 7.8. Полученное нами из опыта значение 3.39 лежит в области допустимых значений, и потому у нас нет оснований считать, что содержание серы в стали этих заводов имеют различные распределения.

Шаг 1. Образуем таблицу *4v* × *2c*, в которую занесем данные; столбцы назовем, например, S1 ÷ S4 (сера), а строки - Z1, Z2 (заводы).

Шаг 2. Теперь необходимо в Меню выбора основных модулей обработки информации в программном обеспечении STATISTICA6.0 выбрать Статистика(Statistics)►Advanced linear / Nonlinear Models►Log - Linear Analysis of frequency Tables.

Шаг 3. В появившемся окне необходимо установить следующие значения:

- Input file: Frequencies w/out coding variables (частоты без кодирующих переменных);
- *Variables: Select All.*

Далее нажать OK. В следующем открывшемся окне необходимо установить:

- *Factor Name: S;*
- *No. of levels:* 4;
- *Factor Name:* Z;
- *No. of level*s: 2

Далее нажать OK.

Шаг 4. В открывшемся окне (*Log - Linear Model Specification*) перейти на вкладку Advanced и нажать Test all marginal & partial association models. В полученной таблице *Results of Fitting,* в последней строке столбца *Person Chi-Squ* получаем $X^2 = 3.59$, число степеней свободы *Degrs of Freedom f* = 3, и уровень значимости *Probab. p* = 0.31. Поскольку эта вероятность не мала, гипотезу об одинаковом распределении содержания серы в металле на двух заводах можно принять (вернее, наблюдения этому не противоречат).

5. Проверить 3 гипотезы: о нормальном, равномерном и показательном распределении выборки из прил. 1 в соответствии с вашим вариантом.

6. Проверить генератор случайных чисел на основе сгенерированной выборки по закону, заданному в табл. 8.1. Сравнить гипотетические и вычисленные характеристики.

Замечание.

Выполнение этого задания отличается от предыдущего следующим:

- в окне Fitting Continuous Distribution нужно ввести значения параметров распределения, а не их оценки и, возможно, поправить параметры группировки;
- приводимый результат для уровня значимости *p* не соответствует рассматриваемому случаю, так как число степеней свободы *df* должно быть равным *m-1*; пакет же указывает с учетом числа оцениваемых параметров. Нужное значение для *p* можно вычислить в Probability Calculator.

7. Генерировать три выборки (варианты *n, n+1, n+2*, где *n* – ваш вариант) из табл. 8.1. Провести их группирование на 8-10 интервалах. Проверить гипотезу об однородности трех выборок.

Составить отчет по выполненной работе

Отчет **по выполненной работе** должен содержать:

- Постановку задачи.
- Сохраненные на переносном носителе информации, созданные в процессе выполнения лабораторной работы файлы.
- Краткое описание критерия *хи-квадрат.*
- Двумерные гистограммы, диаграммы рассеивания и таблицы частот.
- Для наглядности в процессе выполнения работы необходимо сделать несколько Screen Capture, которые в дальнейшем будут размещены в отчете.
- Значения опорных статистик, уровней значимости и статистические выводы.
- Вывод о проделанной работе.

9. ЭЛЕМЕНТЫ РЕГРЕССИОННОГО И ДИСПЕРСИОННОГО АНАЛИЗА

Регрессионный анализ является основным методом современной математической статистики. Идея регрессионного анализа заключается в том, что все доступные нам ресурсы необходимо использовать полно и эффективно, особенно если это требуется для анализа и обработки экспериментальных данных.

Родоначальником регрессионного анализа принято считать К. Гаусса. К. Гаусс (и независимо от него А. Лежандр) на рубеже XVIII - XIX столетий заложили основы метода наименьших квадратов. Этот метод составляет математическую основу регрессионного анализа. Поводом для создания метода наименьших квадратов послужили потребности астрономии и геодезии. Усилиями ученых многих стран была развита и теория, которая стала теперь классической. Примерно 150 лет, до середины XX века, длился период классического регрессионного анализа. За это время к алгебраической процедуре метода наименьших квадратов прибавилось система статистических положений, задающих математическую модель. Были отработаны методы проверки статистических гипотез о значимости коэффициентов уравнения, полученного методом наименьших квадратов. Сочетание метода наименьших квадратов с указанными статистическими процедурами и привело к созданию того, что стало называться регрессионным анализом. Постепенно расширялись и области приложений. Так, например, Д. И. Менделеев начал применять регрессию для описания температурных и иных зависимостей свойств химических веществ. Однако до конца первой мировой войны метод не нашел широкого применения. Появлялись лишь отдельные работы. Следует отметить, что после классических работ К. Пирсона в самом начале XX века теория была хорошо и подробно изложена, а практическое приложение не наблюдалось и резко отставало от теории.

В 20-е годы сложилось новое направление в экономике – эконометрия. Она взяла на вооружение регрессионные методы, что способствовало их распространению. Другой толчок произошел в связи с развитием способов измерения психических свойств личности, имевших большое значение не только для психологии, но и для тесно связанных с нею педагогики, социологии и медицины. Лишь вторая мировая война и особенно послевоенное время привели к широчайшему внедрению регрессии во все области научных исследований, экономического анализа и промышленного производства.

В данном случае решающую роль сыграла вычислительная техника. Появление в 50-е годы массового производства ЭВМ привело к регрессионному буму. Сейчас наступил новый этап развития вычислительной техники. Появились персональные компьютеры. Повышение быстродействия, увеличение памяти и удешевление компьютеров, а также значительный прогресс в сервисных устройствах вызвали к жизни новые подходы к анализу данных, основанные на применении вычислительной техники. Это прежде всего относится к имитационному моделированию, предложенному Т. Нейлором и Р.Шенноном. Все эти методы обогатили регрессионный анализ. С другой стороны, сама регрессионная модель выступает теперь в качестве инструмента, связывающего эти методы в нечто целостное.

В анализе экспериментальных данных используется дисперсионный анализ. Дисперсионный анализ – статистический метод, предназначенный для выявления влияния отдельных факторов на результат эксперимента, а также для последующего планирования аналогичных экспериментов. Дисперсионный анализ первоначально был предложен Р. Фишером в 1925 году. Он сделал обработку результатов агрономических опытов, чтобы определить условия, при которых испытываемый сорт сельскохозяйственной культуры даст максимальный урожай.

9.1 Модель линейной регрессии. Метод наименьших квадратов

Мы рассматривали до сих пор статистические выводы для моделей, которые соответствовали повторным независимым наблюдениям над некоторой случайной величиной ξ. Исходные статистические данные в этих случаях представляют собой реализацию случайного вектора $\vec{X}=(X_1,...,X_n)$, компоненты которого независимы и одинаково распределены, а именно $F_{X_i} = F_{\xi}$, i=1,...,n. Однако, на практике, предположение о независимости и одинаковой распределенности компонент X_i не всегда выполняется. В этих случаях используют линейную регрессионную модель. В этой модели предполагается, что математические ожидания наблюдений X_i являются линейными функциями $\varphi_i(\vec{\beta})$ от неизвестных параметров $\beta=(\beta_1,...,\beta_k)$ и делаются предположения о вторых моментах.

Пусть производится n опытов, на результат которых оказывают влияние неслучайные переменные - факторы - $\vec{z}=(z_1,...,z_k)$. Значения этих факторов меняются от опыта к опыту. Результат i-го опыта можно представить в виде:

$$X_i = \vec{z}^{(i)T}\vec{\beta} + \varepsilon_i,\ i=1,...,n$$

где ε_i - погрешность измерения некоторой случайной величины или ошибка; $\vec{z}^{(i)}$ - вектор-столбец факторов в i-ом опыте.

Предполагаем, что математическое ожидание $M\varepsilon_i=0$, т.е. отсутствуют систематические ошибки и распределение «ошибок» ε_i от параметров β не зависит. Введем матрицу плана

$$Z = \left\| \vec{z}^{(1)} ... \vec{z}^{(n)} \right\|$$

размером k×n, составленную из вектор-столбцов $\vec{z}^{(1)} ... \vec{z}^{(n)}$, и вектор ошибок $\vec{\varepsilon}=(\varepsilon_1,...,\varepsilon_n)$. В матричных обозначениях предыдущее равенство принимает вид:

$$\vec{X} = Z^T\vec{\beta} + \vec{\varepsilon}, \ M(\varepsilon)=0. \qquad (9.1)$$

Предполагают, что случайные величины $\varepsilon_1,...,\varepsilon_n$ (или, что то же самое, $X_1,...,X_n$) не коррелированы и имеют одинаковые дисперсии: $Dx_i = D\varepsilon_i = \sigma^2>0$, i=1,...,n, где σ^2, обычно неизвестно. В этом случае матрица вторых моментов вектора наблюдений $\vec{X}$ имеет вид:

$$D(\vec{X})=D(\vec{\varepsilon})=M(\vec{\varepsilon}\,\vec{\varepsilon}^{\,T})= K_{\varepsilon_i\varepsilon_j} = \begin{cases}\sigma^2, & i=j\\ 0, & i\neq j\end{cases} \quad i,j=1,...,n. \qquad (9.2)$$

Если выполняются условия (9.1)-(9.2), то имеет место модель **линейной регрессии**. Параметры $\beta_1,...,\beta_k$ называют **коэффициентами регрессии**, а σ^2 - **остаточной дисперсией.** Важную роль играет матрица

$$A=ZZ^T \qquad (9.3)$$

Предполагается, что rank Z=k, т.е. строки матрицы A линейно независимы. Предполагается также, что матрица A невырождена (det A≠0 или |A|≠0).

Общим методом оценивания неизвестных коэффициентов регрессии $\beta_1,...,\beta_k$ является **метод наименьших квадратов**, разработанный К. Гауссом и, независимо от него, Лежандром примерно в 1795—1803 гг.

В соответствии с этим методом оценки этих параметров находят из условия обращения в минимум квадратичной формы

$$S(\vec{\beta}) = S(X;\vec{\beta}) = (X-Z^T\vec{\beta})^T(X-Z^T\vec{\beta}), \qquad (9.4)$$

представляющей собой сумму квадратов разностей между наблюдениями и их математическими ожиданиями.

Точку $\vec{b}=(b_1,...,b_k)$, удовлетворяющую равенству $S(\vec{b}) = \min_{\beta} S(\vec{\beta})$, называют, по определению, **оценкой наименьших квадратов** (о.н.к.) параметра $\vec{\beta}=(\beta_1,...,\beta_k)$.

Пусть Y=ZX; тогда с помощью непосредственных вычислений можно убедиться, что система уравнений $\frac{\partial S(\vec{\beta})}{\partial \beta_i} = 0$, i=1,...,k, в матричной форме записывается в виде:

$$A\vec{\beta} = Y, \qquad (9.5)$$

где матрица A задана в (9.3).

Это уравнение для экстремальных точек $\vec{\beta}$ называют **нормальным уравнением** метода наименьших квадратов. Справедлива следующая теорема.

Теорема. Пусть $\vec{\beta}^*$ - любое решение нормального уравнения. Тогда

$$\min_{\beta} S(\vec{\beta}) = S(\vec{\beta}^*)$$

и, следовательно, этот минимум одинаков для всех $\vec{\beta}^*$. Если $\det A \neq 0$, то оценка наименьших квадратов (о.н.к.) единственна и определяется равенством:

$$b = \vec{\beta}^* = A^{-1}Y = A^{-1}ZX. \qquad (9.6)$$

Примем без доказательства.

Интерес представляют не сами параметры $\beta_1,...,\beta_k$, а их некоторые линейные комбинации, т.е. новый параметрический вектор $\vec{t} = (t_1,...,t_m)$, $m \leq k$, связанный с $\vec{\beta}$ соотношением $\vec{t} = T\vec{\beta}$, где T - заданная матрица размером m×n. В этом случае о.н.к. $\vec{t}^*$ для $\vec{t}$ определяется равенством $\vec{t}^* = T\vec{\beta}^*$, где $\vec{\beta}^*$ - любое решение нормального уравнения (9.5). Если $\det A \neq 0$, то из (9.6) следует, что $\vec{t}^*$ определяется однозначно и имеет вид

$$\vec{t}^* = TA^{-1}Y = TA^{-1}ZX. \qquad (9.7)$$

9.2. Свойства оценок наименьших квадратов

Свойства оценок наименьших квадратов определяется следующей теоремой.

Теорема. Пусть матрица A не вырождена. Тогда для произвольного вектора $\vec{t}=T\vec{\beta}$ о.н.к. $\vec{t}^{\,*}$, определенная равенством (9.7) является несмещенной оценкой с минимальной дисперсией в классе всех линейных несмещенных оценок t; при этом матрица вторых моментов случайного вектора $\vec{t}^{\,*}$ имеет вид:

$$D(\vec{t}^{\,*})=\sigma^2 TA^{-1}T^T\equiv\sigma^2 D. \qquad (9.8)$$

Из этой теоремы видно, что

1. Оценка несмещенная, т.е. подставляя (9.1) в (9.7) имеем:

$$\vec{t}^{\,*}=TA^{-1}Z(Z^T\vec{\beta}+\varepsilon)=TA^{-1}(ZZ^T)\vec{\beta}+TA^{-1}Z\varepsilon=T\vec{\beta}+TA^{-1}Z\varepsilon.$$

Находим математическое ожидание:

$$M[\vec{t}^{\,*}]=M[T\vec{\beta}]+M[TA^{-1}Z\varepsilon],$$

но $M\varepsilon=0$, отсюда $M[\vec{t}^{\,*}]=t$, т.е. $\vec{t}^{\,*}$ - линейная несмещенная оценка t.

Пусть $I=LX$ - произвольная линейная несмещенная оценка t, т.е.

$$M[I]=LM[X]=LZ^T\vec{\beta}=T\vec{\beta}.$$

Это равенство должно выполняться для всех $\vec{\beta}$, поэтому отсюда следует, что

$$LZ^T=T. \qquad (**)$$

Из (9.2) находим

$$D(I)=LD(X)L^T=\sigma^2 LL^T. \qquad (*)$$

Наша цель - минимизировать диагональные элементы матрицы LL^T, т.е. дисперсии оценок $l_1,...,l_m$. Для этого запишем тождество

$$LL^T=(TA^{-1}Z)(TA^{-1}Z)^T+(L-TA^{-1}Z)(L-TA^{-1}Z)^T,$$

которое непосредственно следует из равенства (**). Каждое слагаемое правой части тождества имеет вид HH^T, откуда следует неотрицательность диагональных элементов. Но от L зависит только второе слагаемое, поэтому диагональные элементы D(I) одновременно достигают минимума тогда и только тогда, когда $L=TA^{-1}Z$. Соответствующая оптимальная

оценка имеет вид $I^*=TA^{-1}ZX=\vec{t}^{\,*}$, т.е. совпадает с о.н.к. (9.7). Наконец, формула (9.8) следует из соотношения (*), если подставить вместо *L* найденное оптимальное решение. ▓

2. Поскольку оценка наименьших квадратов имеет минимальную дисперсию, она является эффективной оценкой. В качестве <u>следствия</u> теоремы получаем, что $D(\vec{\beta})=\sigma^2A^{-1}$ или

$$\operatorname{cov}\left(\beta_i^*,\beta_j^*\right)=\sigma^2\left\|a^{ij}\right\|,\quad i,j=1,\ldots,k, \tag{9.9}$$

где $$\left\|a^{ij}\right\|=A^{-1}=\left\|a_{ij}\right\|^{-1}.$$

Теорема позволяет решить задачу о построении оптимальных оценок для произвольных линейных функций от коэффициентов регрессии - это оценки наименьших квадратов.

<u>Оценивание остаточной дисперсии.</u> Из равенства (9.4) имеем $MS(\vec{\beta})=\sum_{i=1}^{n}DX_i=n\sigma^2$. Учитывая (9.9), найдем

$$M\left[(\vec{\beta}^*-\vec{\beta})^T A(\vec{\beta}^*-\vec{\beta})\right]=\sum_{i,j=1}^{k}a_{ij}M\left[(\vec{\beta}_i^*-\vec{\beta}_i)(\vec{\beta}_j^*-\vec{\beta}_j)\right]=$$

$$=\sigma^2\sum_{i,j=1}^{k}a_{ij}a^{ij}=\sigma^2tr(AA^{-1})=\sigma^2tr(E_k)=k\sigma^2,\quad AA^{-1}=E_k.$$

Отсюда следует, что $M[S(\vec{\beta}^*)]=(n-k)\sigma^2$, т.е. несмещенной оценкой для остаточной дисперсии σ^2 является статистика

$$\tilde{\sigma}^2=\frac{1}{n-k}S\left(\vec{\beta}^*\right)=\frac{1}{n-k}\left(X-Z^T\vec{\beta}^*\right)^T\left(X-Z^T\vec{\beta}^*\right).$$

Вектор $U=X-Z^T\vec{\beta}^*$ называют **остаточным** вектором, а его компоненты - **остатками**. Таким образом, оценка $\tilde{\sigma}^2$ равна сумме квадратов остатков, поделенной на (n-k) – это есть разность между числом наблюдений и числом параметров β_i.

Рассмотрим пример применения метода наименьших квадратов

Пример.(Простая регрессия) Рассмотренную теорию применим для оценивания параметров в случае простой регрессии. Пусть число параметров k=2, т.е. $\vec{\beta}=(\beta_1,\beta_2)$, а векторы $\vec{z}^{(i)}$ имеют вид $\vec{z}^{(i)}=(1,t_i)$, i=1,...,n. Тогда

$$M[X_i]=\beta_1+\beta_2 t,\ i=1,...,n. \qquad (9.10)$$

т.е. среднее значение наблюдений является линейной функцией одного фактора t. Так, t может быть температурой, при которой производится эксперимент, дозой лечебного препарата, возрастом обследуемых лиц и т.д.

Следует выявить связь между исходом эксперимента и фактором t на основании выборки. Поэтому регистрируют n пар значений t_i фактора t. Прямую $\varphi(t)=\beta_1+\beta_2 t$, соответствующую (9.10) называют **линией регрессии**, а коэффициент β_2 - ее **наклоном.**

В данном случае матрицы Z,A и столбец Y=ZX равны

$$Z\left\|\begin{matrix}1 & \dots & 1\\ t_1 & \dots & t_n\end{matrix}\right\|,\quad A=\left\|\begin{matrix}n & \sum t_i\\ \sum t_i & \sum t_i^2\end{matrix}\right\|,\quad Y=\begin{pmatrix}\sum X_i\\ \sum t_i X_i\end{pmatrix}.$$

Будем предполагать, что не все t_i одинаковы (чтобы rank Z=2), тогда

$$\det A\equiv|A|=n\sum_{i=1}^{n}t_i^2-\left(\sum_{i=1}^{n}\bar{t}_i\right)^2=n\sum_{i=1}^{n}(t_i-\bar{t})^2>0.$$

(черта сверху означает арифметическое среднее).

$$A^{-1}=\frac{1}{|A|}\left\|\begin{matrix}\sum t_i^2 & -nt\\ -nt & n\end{matrix}\right\|,\quad A^{-1}Y=\frac{n}{|A|}\begin{pmatrix}\bar{X}\sum t_i^2-\bar{t}\sum t_i X_i\\ \sum t_i X_i-n\bar{X}\bar{t}\end{pmatrix}=\begin{pmatrix}\beta_1^*\\ \beta_2^*\end{pmatrix}.$$

В результате несложных преобразований запишем оценки β_1^* и β_2^* в следующем удобном для вычислений виде:

$$\beta_2^*=\frac{\sum(t_i-\bar{t})(X_i-\bar{X})}{\sum(t_i-\bar{t})^2},\ \beta_1^*=\bar{X}-\bar{t}\beta_2^*$$

Вторые моменты этих оценок образуют матрицу σ^2A^{-1}, поэтому

$$D(\beta_1^*) = \frac{\sum t_i^2}{n\sum (t_i - \bar{t})^2}\sigma^2;\ D(\beta_2^*) = \frac{\sigma^2}{\sum (t_i - \bar{t})^2};$$

$$K_{\beta_1^*\beta_2^*} \equiv \operatorname{cov}(\beta_1^*, \beta_2^*) = \frac{-\bar{t}\sigma^2}{\sum (t_i - \bar{t})^2}.$$

Величина $S(\vec{\beta}^*)$ равна

$$S\left(\vec{\beta}^*\right) = \sum \left(X_i - \overline{X}\right)^2 - \beta_2^{*2} \sum \left(t_i - \bar{t}\right)^2$$

и несмещенная оценка для остаточной дисперсии имеет вид

$$\tilde{\sigma}^2 = S(\vec{\beta}^*) \Big/ (n-2).$$

9.3. Подбор прямой методом наименьших квадратов

Одна из наиболее общих задач статистики состоит в оценивании связи между двумя случайными величинами (если такая связь существует). Такими парами случайных величин могут быть, например, рост и вес, зарплата и уровень интеллекта, возраст мужа и жены в момент вступления в брак, длина металлического стержня и его температура, давление и температура некоторого объема газа и т.д.

Если имеется n пар наблюдений (x_i, y_i) $i = 1,2,..n$ над такими случайными величинами, то наблюдения можно представить точками на плоскости. Затем можно попытаться подобрать по этим точкам некоторую гладкую кривую таким образом, чтобы они располагались как можно ближе к этой кривой. Эту кривую будем называть эмпирической или аппроксимирующей. Ясно, сто все точки не лягут на соответствующую кривую, поскольку каждая из случайных величин в рассмотренных примерах подвержена случайным флуктуациям в результате воздействия факторов, которыми мы не в состоянии управлять.

Мы можем различать здесь два основных типа переменных. Назовем их предсказывающими (предикторами), или независимыми переменными (факторами) и зависимыми

переменными (откликами). Обычно зависимая переменная – отклик является случайной величиной, а независимые переменные – детерминированные величины. При этом предикторы и отклики связываются уравнением аппроксимирующей кривой.

В общем случае мы будем интересоваться тем, какие изменения предикторов влияют на значения откликов. Это можно осуществить, применяя метод анализа, называемый методом наименьших квадратов (МНК). Его можно применять для обработки данных эксперимента и для получения заключений о свойствах выбранного уравнения. Этот метод часто называют регрессионным анализом. МНК является составной частью регрессионного анализа.

Слово «регрессия» (движение вспять, обращение) впервые употребил английский антрополог и генетик Френсис Гальтон в своем труде «Основные законы наследственности человека».

Будем использовать метод наименьших квадратов в связи с простым приложением – подбором «наилучшей» прямой по данным для двух переменных X и Y, а затем рассмотрим случаи, когда рассматривается большее число факторов

Рассмотрим прямолинейную зависимость между переменными. Во многих экспериментальных работах мы хотим исследовать, как изменения одной переменной влияют на другую. Иногда случается так, что две исследуемые переменные оказываются связаны между собой точным уравнением прямой.

Пример 1. Закон Ома утверждает, что если I – ток (в амперах), протекающий через сопротивление R (в омах), а V – напряжение (в вольтах) на этом сопротивлении, то эти три величины связаны соотношением $V = R \cdot I$. В прямоугольных координатах (X, Y) закон Ома выражается прямой линией, проходящей через начало координат. Если бы мы не знали закон Ома, мы могли бы найти зависимость эмпирически, поддерживая R фиксированным и изменяя X и Y. Иногда линейная зависимость не точна, но она имеет смысл.

Пример 2. Рассматривается рост и вес взрослых мужчин из данной выборки, или популяции. Если мы нанесем на график пары чисел (X,Y)=(рост, вес), то результат будет соответствовать рис. 9.1.

Такое изображение обычно называют диаграммой рассеяния или полем рассеяния . Для любого заданного роста встречаются различные веса и наоборот. Такая вариация может получиться из-за ошибки измерения и, самое главное, это следствие разброса между индивидами. Однако можно обнаружить, что средний наблюденный вес при заданном росте растет с увеличением роста. Геометрическое место точек средних наблюденных весов при данных ростах (при изменении роста) назовем регрессионной кривой веса от роста.

Обозначим это следующим образом: $Y = f(X)$. Существует также и регрессионная кривая роста от веса, которую обозначим так: $X = g(Y)$. Пусть эти «кривые» будут прямыми, как на рис. 9.1., хотя в общем случае они могут быть и не прямыми.

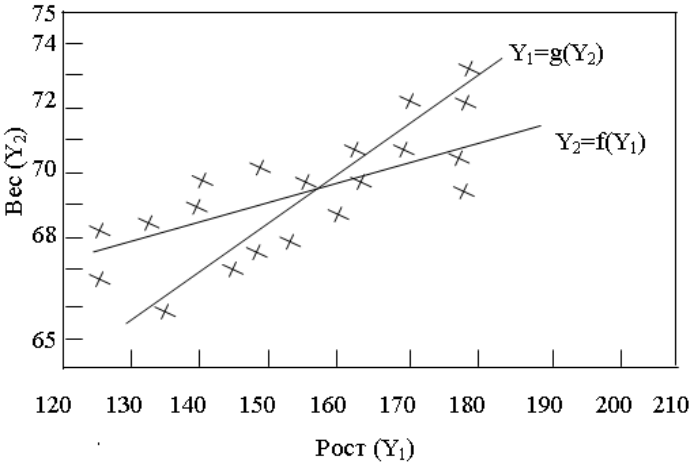


Рис. 9.1. Рост и вес 30 мужчин

Если мы установим связь между зависимой случайной величиной Y и величиной X, которая является переменной,

но не является случайной переменной, то уравнение Y относительно X будет называться уравнением регрессии**.** Будем далее предполагать, что переменные - предикторы не подвержены случайным вариациям (изменениям), а отклики – подвержены. Если же это не так, то тогда требуются более сложные методы построения зависимостей.

Линейная регрессия: подбор прямой

Предположим, что линия регрессии переменной, которую обозначим Y, от переменной X имеет вид $\beta_0 + \beta_1 X$. Тогда можно записать линейную модель

$$Y = \beta_0 + \beta_1 X + \varepsilon \ . \qquad (9.11)$$

Уравнение (9.11) – это модель, которой мы задаемся, или которую мы постулируем. Постулирование модели есть предварительное допущение об ее правильности. Модель надо критически исследовать в разных аспектах. Мнение о модели может измениться на более поздней стадии исследования. При этом величины β_0 и β_1 называют параметрами модели.

Замечание.

Когда мы говорим, что модель линейная или нелинейная, то имеется в виду линейность по параметрам.

Величина наивысшей степени предиктора в модели называется порядком модели. Например,

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

есть регрессионная модель 2-го порядка (по X) и линейная (по β).

Итак, в уравнении (9.11) величины β_0, β_1 и ε неизвестны, причем, ε будет трудно исследовать, поскольку она меняется от наблюдения к наблюдению. Величины β_0 и β_1 остаются постоянными, и мы можем получить для них оценки b_0 и b_1. Запишем это в виде:

$$\widehat{Y} = b_0 + b_1 X \ . \qquad (9.12)$$

Уравнение (9.12) можно использовать как предсказывающее: подстановка в него некоторого значения X позволяет предсказать среднее «истинное» значение Y для этого X. Процедурой оценивания будет метод наименьших квадратов (МНК), разработанный Гауссом и, независимо от него, Лежандром примерно в 1795—1803 гг.

Пусть мы имеем множество из n наблюдений $(X_1, Y_1),...(X_n, Y_n)$. Тогда уравнение (9.11) перепишется в виде:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

где $i = 1,2,..n$.

Следовательно, сумма квадратов отклонений от «истинной» линии есть

$$S = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 . \qquad (9.13)$$

Будем подбирать значения оценок b_0 и b_1 так, чтобы их подстановка вместо β_0 и β_1 в уравнение (9.13) давала минимальное (наименьшее возможное) значение S (см. рис. 9.2).

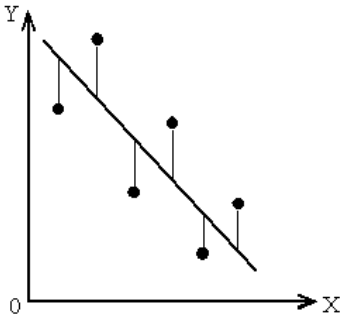


Рис. 9.2

Здесь линия, подобранная методом наименьших квадратов такова, что делает сумму всех вертикальных расхождений настолько малой, насколько это возможно. Заметим, что X_i, Y_i – фиксированные числа, которые нам

известны.

Мы можем определить b_0 и b_1, дифференцируя уравнение (9.13) сначала по β_0, затем по β_1, и приравнивая результаты к нулю. Тогда

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i),$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i \cdot (Y_i - \beta_0 - \beta_1 X_i). \qquad (9.14).$$

Так что для оценок b_0 и b_1 имеем:

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0,$$

$$\sum_{i=1}^{n} X_i \cdot (Y_i - b_0 - b_1 X_i) = 0, \qquad (9.15)$$

где мы приравняли выражения (9.14) к нулю и подставили (b_0, b_1) вместо (β_0, β_1). Из (9.15) имеем:

$$\sum_{i=1}^{n} Y_i - b_0 n - b_1 \sum_{i=1}^{n} X_i = 0,$$

$$\sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 = 0,$$

или

$$b_0 n + b_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i,$$

$$b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i. \qquad (9.16)$$

Эти уравнения называются **нормальными**. Решение уравнений (9.16) относительно угла наклона прямой b_1 дает

$$b_1 = \frac{\sum X_i Y_i - \sum X_i \sum Y_i / n}{\sum X_i^2 - (\sum X_i)^2 / n} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}, \qquad (9.17)$$

где суммирование ведется по $i = 1,2,..n$.

Два выражения для b_1 – это обе правильные, но несколько различные формы одной и той же величины. Так как по определению

$$\overline{X} = (X_1 + ... + X_n)/n = \sum X_i / n,$$
$$\overline{Y} = (Y_1 + ... + Y_n)/n = \sum Y_i / n,$$

имеем:

$$\sum (X_i - \overline{X})(Y_i - \overline{Y}) = \sum X_i Y_i - \overline{X} \sum Y_i - \overline{Y} \sum X_i + n\overline{X}\overline{Y} =$$
$$\sum X_i Y_i - n\overline{X}\overline{Y} = \sum X_i Y_i - \sum X_i \sum Y_i / n.$$

Отсюда следует эквивалентность числителей в (9.17), а также и знаменателей при замене Y на X.

Величина $\sum X_i^2$ называется нескорректированной суммой квадратов X-в,

$\sum X_i^2 / n$ – коррекцией на среднее значение X-в.

Разность между ними $\sum X_i^2 - \sum X_i^2 / n$ называется скорректированной суммой квадратов X-в.

Аналогично $\sum X_i Y_i$ называется нескорректированной суммой смешанных (парных) произведений X и Y,

$\sum X_i \sum Y_i / n$ – коррекцией на среднее значение произведений.

Разность между ними $\sum X_i Y_i - \sum X_i \sum Y_i / n$ называется скорректированной суммой произведений X и Y.

Введем удобные обозначения:

$$S_{XY} = \sum (X_i - \overline{X})(Y_i - \overline{Y}) = \sum X_i Y_i - \sum X_i \sum Y_i / n =$$
$$\sum X_i Y_i - n\overline{X}\overline{Y} = \sum (X_i - \overline{X}) Y_i = \sum (Y_i - \overline{Y}) X_i.$$

Заметим, что все эти выражения эквивалентны. Запишем далее по аналогии с предыдущим:

$$S_{XX} = \sum (X_i - \overline{X})^2 = \sum X_i^2 - (\sum X_i)^2 / n = \sum (X_i - \overline{X}) X_i =$$
$$\sum X_i^2 - n\overline{X}^2.$$

$$S_{YY} = \sum (Y_i - \overline{Y})^2 = \sum Y_i^2 - (\sum Y_i)^2 / n = \sum (Y_i - \overline{Y}) Y_i = \sum Y_i^2 - n\overline{Y}^2 .$$

Для b_1 получается легко запоминающаяся формула:

$$b_1 = S_{XY} / S_{XX} . \qquad (9.17.а)$$

Решение уравнения (9.16) относительно свободного члена (отрезка на оси ординат при $X = 0$) дает:

$$b_0 = \overline{Y} - b_1 \overline{X} . \qquad (9.18)$$

Подстановка (9.18) в уравнение (9.12) дает оцениваемое уравнение регрессии

$$\widehat{Y}_i = \overline{Y} + b_1 (X_i - \overline{X}), \qquad (9.19)$$

где b_1 определяется уравнением (9.17).

Если в (9.19) положить $X_i = \overline{X}$, то окажется, что $\widehat{Y}_i = \overline{Y}$, а это означает, что точка (X, Y) лежит на подобранной прямой.

Разность между наблюдаемым (истинным) значением Y_i и оценкой прогнозируемой величины $\widehat{Y}_i$ называется остатком $Y_i - \widehat{Y}_i$. Остатков получается столько же, сколько исходных данных.

Так как $\widehat{Y}_i = \overline{Y} + b_1 (X_i - \overline{X})$, то

$$Y_i - \widehat{Y}_i = (Y_i - \overline{Y}) - b_1 (X_i - \overline{X}),$$

$$\sum_{i=1}^{n} (Y_i - \widehat{Y}_i) = \sum_{i=1}^{n} (Y_i - \overline{Y}) - b_1 \sum_{i=1}^{n} (X_i - \overline{X}) = 0.$$

Следовательно, и сумма остатков будет равна нулю. На практике из-за ошибок округления она может оказаться не точно равной нулю.

В любой регрессионной задаче сумма остатков всегда равна нулю, если член β_0 входит в модель. Это следствие первого из нормальных уравнений. Исключение β_0 из модели приводит к тому, что отклик обращается в нуль, когда все предикторы равны нулю. Такое предположение слишком

сильно и потому обычно не справедливо.

В линейной модели $Y = \beta_0 + \beta_1 X + \varepsilon$ исключение β_0 означает, что линия проходит через точку $X = 0$, $Y = 0$, т.е. она отсекает нулевой отрезок $\beta_0 = 0$ при $X = 0$. Исключение β_0 из модели возможно с помощью «центрирования» данных, но это не то же самое, что приравнивание $\beta_0 = 0$. Если мы запишем уравнение (9.11) в виде:

$$Y - \overline{Y} = (\beta_0 + \beta_1 \overline{X} - \overline{Y}) + \beta_1 (X - \overline{X}) + \varepsilon,$$

или

$$y = \beta_0' + \beta_1 x + \varepsilon,$$

где $y = Y - \overline{Y}$, $x = X - \overline{X}$ $\beta_0' = \beta_0 + \beta_1 \overline{X} - \overline{Y}$,

то оценки для β_0' и β_1 будут такими:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2},$$

в соответствии с уравнением (9.17) и

$$b_0' = \bar{y} - b_1 x = 0$$

так как $\bar{x} = \bar{y} = 0$ при любом значении b_1.

Поэтому можно записать центрированную модель, совсем опуская свободный член (отрезок)

$$\beta_0' : Y - \overline{Y} = \beta_1 (X - \overline{X}) + \varepsilon.$$

Таким образом, мы потеряли один параметр, что соответствует потере данных, а это влечет за собой потерю части информации. Потерянная часть информации эффективно используется для корректировки модели, позволяющей исключить свободный член

9.4. Точность оценки регрессии

Рассмотрим вопрос о том, какая точность может быть приписана оценке линии регрессии. Рассмотрим тождество:

$$Y_i - \hat{Y}_i = Y_i - \overline{Y} - (\hat{Y}_i - \overline{Y}). \qquad (9.20)$$

Что это означает геометрически, показано на рис. 9.3.

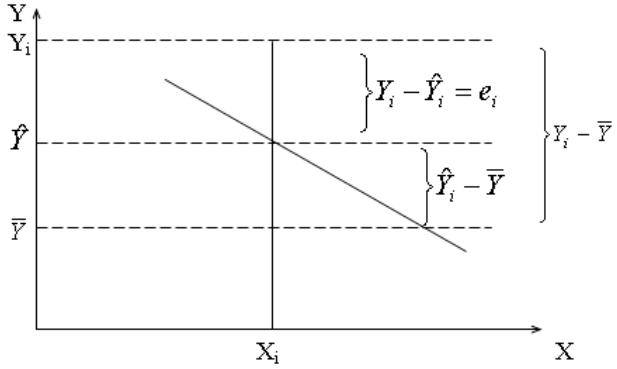


Рис. 9.3

Остаток $e_i = Y_i - \hat{Y}_i$ представляет собой разность между двумя величинами:

1) отклонением наблюдаемого значения отклика Y_i от общего среднего откликов $\overline{Y}$ и

2) отклонение предсказанного значения отклика $\hat{Y}_i$ от того же общего среднего $\overline{Y}$.

Заметим, что среднее арифметическое предсказанных значений $\hat{Y}_i$ равно

$$\sum \hat{Y}_i / n = \sum (b_0 + b_1 X_i)/n = (nb_0 + nb_1 \overline{X})/n = b_0 + b_1 \overline{X} = \overline{Y}.$$

Иными словами, среднее арифметическое предсказанных значений $\hat{Y}_i$ то же, что и наблюдаемых откликов Y_i.Отсюда, как было установлено ранее,

$$\sum e_i = \sum (Y_i - \hat{Y}_i) = n\overline{Y} - n\overline{Y} = 0.$$

Уравнение (9.20) можно переписать еще и так:

$$Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i).$$

Если мы возведем обе части этого выражения в квадрат и просуммируем по $i = 1,2,..n$, то получим:

$$\sum (Y_i - \overline{Y})^2 = \sum (\hat{Y}_i - \overline{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \quad . \qquad (9.21)$$

В уравнении (9.21) величина $Y_i - \overline{Y}$ – это отклонение $\boldsymbol{i}$-го наблюдения от общего среднего, следовательно, левая часть уравнения (9.21) – это сумма квадратов отклонений относительно среднего наблюдений (сокращенно SS *относительно среднего*), а также скорректированная сумма квадратов Y-в. Так как $Y_i - \widehat{Y}_i$ есть отклонение i-го наблюдения от его предсказанного или вычисленного значения (i-й остаток), а $\widehat{Y}_i - \overline{Y}_i$ – отклонение предсказанного значения i-го наблюдения от среднего, то мы можем выразить уравнение (9.21) словесно следующим образом: «сумма квадратов относительно среднего» = «сумма квадратов относительно регрессии» + «сумма квадратов, обусловленная регрессией».

Пригодность линии регрессии для целей предсказания зависит от того, какая именно часть SS относительно среднего приходится на SS, обусловленную регрессией, а какая – соответствует SS относительно регрессии.

Удовлетворительные результаты получаются, если SS, обусловленная регрессией, будет много больше, чем SS относительно регрессии или, то же самое, если отношение

$$R^2 = \frac{SS, обусловленная}{SS, относительная} \approx 1.$$

будет не слишком отличаться от 1.

Всякая сумма квадратов связана с числом, называемым ее степенями свободы. В статистике числом степеней свободы некоторой величины часто называют разность между числом различных опытов и числом констант, найденных по этим опытам независимо друг от друга.

Это понятие можно применить к сумме квадратов. Мы получим число, которое показывает, как много независимых элементов информации, получающихся из n независимых чисел $Y_1,...Y_n$, требуется для образования данной суммы квадратов. Например, для SS относительно среднего

требуется $n-1$ независимый элемент (из чисел $Y_1-\bar{Y},...Y_n-\bar{Y}$ независимы только $n-1$, так как сумма всех n чисел при определении среднего приравнивалась к нулю).

Мы можем вычислить сумму квадратов SS, обусловленную регрессией, используя единственную функцию от $Y_1,...Y_n$, а именно b_1, [т.к. $\sum(\hat{Y}_i-\bar{Y})^2=b_1^2\sum(X_i-\bar{X})^2$], и поэтому данная сумма квадратов имеет одну степень свободы.

По разности SS относительно регрессии имеет $(n-2)$ степени свободы. Это отражает тот факт, что рассматриваемые остатки получены для модели прямой линии, которая требует оценивания двух параметров. Вообще, остаточная сумма квадратов основывается на числе степеней свободы, равном числу наблюдений минус число оцениваемых параметров. Следовательно, в соответствии с уравнением (9.21) мы можем разложить степени свободы таким образом:

$$n-1=1+(n-2)\,. \qquad (9.22)$$

Пользуясь уравнениями (9.21) и (9.22), мы можем построить таблицу дисперсионного анализа. (ANOVA). Обозначение ANOVA произошло от английских слов « Analysis of variance». «Средний квадрат» MS получается при делении каждой суммы квадратов SS на соответствующее ей число степеней свободы.

Таблица 9.1.

Таблица дисперсионного анализа (ANOVA). Основное разложение

Источник вариации	Число степеней свободы	Суммы квадратов SS	Средние квадраты MS
Обусловленный регрессией	1	$\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$	MS

Продолжение табл. 9.1

Относительно регрессии (остаток)	$n-2$	$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	$s^2=\frac{SS}{n-2}$
Общий, скорректированный на среднее Y	$n-1$	$\sum_{i=1}^{n}(Y_i-\bar{Y})^2$	

Более общая форма таблицы дисперсионного анализа получается при добавлении в таблицу корректирующего фактора для среднего Y-в, который называется $SS(b_0)$.Это название будет пояснено позже

Таблица 9.2

Таблица дисперсионного анализа (ANOVA), включающая $SS(b_0)$

Источник вариации	Число степеней свободы	Суммы квадратов SS	Средние квадраты MS
Обусловленный b_0/b_1	1	$SS(b_1/b_0)=$ $=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$	MS
Остаток	$n-2$	$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	s^2

Общий, скорректированный	$n-1$	$\sum_{i=1}^{n}(Y_i-\bar{Y})^2$	
Корректирующий фактор, обусловленный b_0	1	$SS(b_0)=\frac{(\sum_{i=1}^{n}Y_i)^2}{n}$	
Общий	n	$\sum_{i=1}^{n}Y_i^2$	

Сумма SS редко подсчитывается так, как показано в таблице, а обычно получается делением $SS(b_1/b_0)$ на общую скорректированную SS. Сумму квадратов, обусловленную регрессией, можно вычислять множеством способов (суммирование везде идет по $i=1,2,..n$):

$$SS(b_1/b_0)=\sum(\hat{Y}_i-\bar{Y})^2=b_1[\sum(X_i-\bar{X})(Y_i-\bar{Y})]=b_1S_{XY}. \quad (9.23)$$

$$\frac{S_{XY}^2}{S_{XX}}=\frac{[\sum(X_i-\bar{X})(Y_i-\bar{Y})]^2}{\sum(X_i-\bar{X})^2}.$$

$$\frac{S_{XY}^2}{S_{XX}}=\frac{[\sum X_iY_i-\sum X_i\sum Y_i/n]^2}{\sum X_i^2-(\sum X_i)^2/n}. \quad (9.24)$$

$$\frac{S_{XY}^2}{S_{XX}}=\frac{[\sum(X_i-\bar{X})Y_i]^2}{\sum(X_i-\bar{X})^2}.$$

Уравнение (9.23) проще всего использовать, поскольку оба сомножителя уже получены при подборе уравнения прямой. Округление может внести неточности, лучше использовать формулу (9.24), где деление производится в

последний момент.

Общую скорректированную сумму квадратов можно записать и вычислить следующим образом:

$$S_{YY} = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - (\sum Y_i)^2 / n = .$$
$$= \sum Y_i^2 - n\bar{Y}^2 \ .$$

Обозначение $SS(b_1/b_0)$ читается так: «сумма квадратов для b_1 с учетом поправки на b_0». Средний квадрат относительно регрессии s^2 дает оценку дисперсии относительно регрессии, основанную на (n-2) степенях свободы. Будем обозначать эту величину σ_{YX}^2. Если уравнение регрессии будет оцениваться из большого числа наблюдений, то дисперсия относительно регрессии будет представлять ошибку измерения, с которой любое измеренное Y предсказывается для данного значения X по известному уравнению:

Исследуем уравнение регрессии. Пока не были использованы предположения о распределении вероятностей. Теперь введем основные предположения о том, что в модели

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \ i = 1,2,..n:$$

1) остаток ε_i есть случайная величина со средним, равным нулю, и неизвестной дисперсией σ^2, т.е.

$E(\varepsilon_i) = 0$, $V(\varepsilon_i) = \sigma^2$ – англоязычное

$M(\varepsilon_i) = 0$, $D(\varepsilon_i) = \sigma^2$ – русское обозначение;

2) остатки ε_i и ε_j некоррелированы при $i \neq j$, так что $cov(\varepsilon_i, \varepsilon_j) = 0$. Поэтому $M(Y_i) = \beta_0 + \beta_1 X_i$, $D(Y_i) = \sigma^2$ и значения Y_i и Y_j некоррелированы при $i \neq j$;

3) остаток ε_i есть нормально распределенная случайная величина со средним нуль и дисперсией σ^2, т.е. $\varepsilon_i \sim N(0, \sigma^2)$. При добавлении этого предположения остатки ε_i и ε_j становятся не только некоррелированными, но даже

независимыми

Замечание 1.

Дисперсия σ^2 может быть равной или не равной σ_{YX}^2 – дисперсии относительно регрессии. Если постулированная модель не соответствует «истинной», то $\sigma^2 < \sigma_{YX}^2$. Из этого следует, что s^2 – остаточный средний квадрат, который в любом случае оценивает σ_{YX}^2 – служит оценкой σ^2, если только модель корректна. Если $\sigma^2 < \sigma_{YX}^2$, то постулируемая модель некорректна или страдает неадекватностью.

Замечание 2.

Во многих реальных ситуациях ошибки, в соответствии с центральной предельной теоремой, подчиняются нормальному распределению. Если ε оказывается суммой ошибок, то, независимо от того, как распределены отдельные ошибки, их сумма ε имеет тенденцию к нормальному распределению в соответствии с центральной предельной теоремой.

9.5. Интервальное оценивание параметров регрессии

Рассмотрим **стандартное отклонение** b_1 **и доверительный интервал для** β_1**.**

Мы знаем, что

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum (X_i - \overline{X})Y_i}{\sum (X_i - \overline{X})^2} = \frac{(X_1 - \overline{X})Y_1 + \ldots + (X_n - \overline{X})Y_n}{\sum (X_i - \overline{X})^2}.$$

Далее, дисперсия некоторой функции

$$F = a_1 Y_1 + \ldots a_n Y_n$$

равна

$$D(F) = a_1^2 D(Y_1) + \ldots + a_n^2 D(Y_n),$$

если Y_i попарно некоррелированы и a_i – константы.

Кроме того, если $D(Y_i) = \sigma^2$, то

$$D(F) = (a_1^2 + ... + a_n^2)\sigma^2 = \sigma^2 \sum a_i^2 .$$

В выражении для b_1: $a_i = \dfrac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2}$,

так как X_i можно рассматривать как константы. Отсюда, после преобразований получаем

$$D(b_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} = \frac{\sigma^2}{S_{XX}} .$$

Стандартное отклонение b_1 есть квадратный корень из дисперсии, т.е. $\dfrac{\sigma}{\sqrt{\sum (X_i - \bar{X})^2}}$.

Если σ неизвестна и мы используем вместо нее оценку s, предполагая, что модель корректна, то оценка стандартного отклонения b_1 есть

$$\frac{s}{\sqrt{\sum (X_i - \bar{X})^2}} .$$

Вместо термина «оцениваемое стандартное отклонение» обычно используют термин «стандартная ошибка». Если мы предполагаем, что разброс наблюдений относительно линии нормален, т.е. что ошибки ε_j все принадлежат некоторому нормальному распределению $N(0, \sigma^2)$, то можно показать, что $100 \cdot (1 - \alpha)\%$ доверительные интервалы для β_1 получаются, если вычислить

$$b_1 \pm \frac{t(n - 2{,}1 - \frac{\alpha}{2})s}{\sqrt{\sum (X_i - \bar{X})^2}}, \qquad (9.25)$$

где $t(n - 2{,}1 - \dfrac{\alpha}{2})$ – это $100 \cdot (1 - \dfrac{\alpha}{2})\%$ точка t-распределения Стьюдента с $n - 2$ степенями свободы.

С другой стороны, мы можем проверить нуль - гипотезу о том, что $\beta_1 = \beta_{10}$,

где β_{10} – частное значение, которое может быть нулем против альтернативы, что β_1 отлично от β_{10}.

Обычно пишут: $H_0 : \beta_1 = \beta_{10}$ против $H_1 : \beta_1 \neq \beta_{10}$.

Для этого надо вычислить

$$t = (\beta_1 - \beta_{10})\sqrt{\sum (X_i - \bar{X})^2} \Big/ s \qquad (9.26)$$

и сравнить $|t|$ с $t(n-2,1-\frac{\alpha}{2})$ из таблицы t-критерия с $(n-2)$ степенями свободы – числом, на котором основана оценка s^2 для σ^2. В таком виде критерий будет двусторонним с $100 \cdot \alpha\,\%$ процентным уровнем значимости.

После того как мы получили доверительный интервал для β_1, уже нет необходимости находить величину $|t|$ для проверки гипотезы с помощью t-критерия. Достаточно исследовать доверительный интервал для β_1 и посмотреть, содержит ли он значение β_{10}. Если это так, то гипотезу $H_0 : \beta_1 = \beta_{10}$ нельзя отвергнуть, а если не так, то она отвергается. Это можно увидеть из уравнений (9.26), $H_0 : \beta_1 = \beta_{10}$ отвергается при α-уровне, если $|t| > t(n-2,1-\frac{\alpha}{2})$, откуда следует, что

$$|\beta_1 - \beta_{10}| > \frac{t(n-2,1-\frac{\alpha}{2})s}{\sqrt{\sum (X_i - \bar{X})^2}},$$

т.е. что β_{10} лежит за пределами, соответствующими уравнению (9.25).

Определим стандартное отклонение свободного члена b_0 и доверительный интервал для β_0.

Доверительный интервал для β_0 и проверку гипотезы о том, что β_0 равно или не равно некоторому заданному числу, удается построить аналогично тому, как было получено для β_1.

Можем показать, что стандартное отклонение b_0 есть

$$\sigma\sqrt{\frac{\sum X_i^2}{n\sum(X_i-\bar{X})^2}} \ .$$

Более детально эта формула будет выводиться позже.

Замена σ на s дает оценку стандартному отклонению b_0. Отсюда получаем $100\cdot(1-\alpha)\%$ доверительные пределы для β_0:

$$b_0 \pm t(n-2,1-\frac{\alpha}{2})s\sqrt{\frac{\sum X_i^2}{n\sum(X_i-\bar{X})^2}} \ .$$

Критерий t для нулевой гипотезы $H_0:\beta_0=\beta_{00}$ против альтернативы $H_1:\beta_0\neq\beta_{00}$,

где β_{00} – заданное значение,

будет отвергать ее с $100\cdot\alpha\%$ уровнем значимости, если β_{00} попадет за доверительные границы, или не будет ее отвергать, если β_{00} попадет внутрь интервала.

Проверку гипотезы H_0 можно выполнить и иначе, находя величину

$$t=\frac{(b_0-\beta_{00})}{s\sqrt{\frac{\sum X_i^2}{n\sum(X_i-\bar{X})^2}}}$$

и сравнивая ее с процентной точкой $t(n-2,1-\frac{\alpha}{2})$, так как

$(n-2)$ – это число степеней свободы, на котором основана оценка s^2 для σ^2.

Возможно построение совместной доверительной области для β_0 и β_1 одновременно, если применить формулу, которая будет получена для многомерных величин.

Стандартное отклонение $\hat{Y}$

Ранее было показано, что подобранное уравнение регрессии имеет вид

$$\hat{Y} = \bar{Y} + b_1(X_i - \bar{X}),$$

где как $\bar{Y}$, так и b_1 подвержены ошибкам, которые будут влиять на $\hat{Y}$.

Далее, если a_i и c_i – константы и

$$a = a_1Y_1 + \ldots + a_nY_n,$$
$$c = c_1Y_1 + \ldots + c_nY_n,$$

то, в случае некоррелированности Y_i и Y_j при $i \neq j$ и при условии $D(Y_i) = \sigma^2 \ \forall i$, имеем:

$$\operatorname{cov}(a,c) = (a_1c_1 + \ldots + a_nc_n)\sigma^2 .$$

Если заменим a на $\hat{Y}$, т.е. $a = \hat{Y}$, то это влечет $a_i = \frac{1}{n}$,

замена $c = b_1$ влечет $c_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}$, так что

$$\operatorname{cov}(\bar{Y}, b_1) = 0,$$

т.е. $\bar{Y}$ и b_1 – некоррелированные случайные величины.

Поэтому дисперсия предсказываемого среднего значения Y (или $\hat{Y}_0$ при заданном X_0) в зависимости от X есть

$$D(\hat{Y}_0) = D(\bar{Y}) + (X_0 - \bar{X})^2 V(b_1) = \frac{\sigma^2}{n} + \frac{(X_0 - \bar{X})^2 \sigma^2}{\sum (X_i - \bar{X})^2}.$$

Отсюда оценка стандартного отклонения

$$\hat{Y}_0 = s\sqrt{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X_i - \overline{X})^2}}\ .$$

Следовательно, эта величина достигает минимума, когда $X_0 = \overline{X}$ и возрастает, по мере того как мы "удаляем" X_0 от $\overline{X}$ в любом направлении.

Другими словами, чем больше разность между X_0 и средним значением $\overline{X}$, тем больше ошибка, с которой мы будем предсказывать среднее значение Y для данного X_0.

Следовательно, мы можем ожидать наилучшее предсказание в центре области наблюдений $\overline{X}$ и не должны ожидать хорошего предсказания при удалении от центра. Дисперсия и оценка стандартного отклонения, которые мы рассмотрели, относятся к предсказываемому среднему значению Y при данном X_0. Так как фактические значения Y варьируют около "истинного" среднего значения с дисперсией σ^2 (не зависимой от $D(Y)$), то предсказанное значение индивидуального наблюдения будет определяться величиной $\hat{Y}$, но с дисперсией

$$\sigma^2[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X_i - \overline{X})^2}].$$

Доверительные пределы можно найти уже рассмотренным способом, Мы вычисляем 95% доверительный интервал для нового наблюдения, который будет симметричен относительно $\hat{Y}_0$ и длина которого будет зависеть от оценки этой новой дисперсии

$$\hat{Y}_0 \pm t(\nu,0{,}975)s\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum (X_i - \overline{X})^2}}\ ,$$

где ν – число степеней свободы, на котором основана оценка s^2 (здесь это число равно $(n-2)$).

Доверительный интервал для среднего из q новых

наблюдений $\hat{Y}_0$ находится аналогично, исходя из следующего. Пусть $\hat{Y}_0$ есть средне из q новых наблюдений при X_0. Тогда

$$\bar{Y}_0 \sim N(\beta_0 + \beta_1 X_0, \sigma_0^2),$$

$$\hat{Y}_0 \sim N(\beta_0 + \beta_1 X_0, D(\hat{Y}_0)),$$

так что $\bar{Y}_0 - \hat{Y}_0 \sim N(0, \sigma_0^2 - D(\hat{Y}_0))$ и $(Y_{0\text{-}} - Y_{0\text{-}})/s^2$ распределено как $t(v)$,

где v – число степеней свободы, на котором основана s^2, оценка σ^2.

Поэтому вероятность

$$[|\bar{Y}_0 - \hat{Y}_0| \le t(v,0{,}975)s^2 \sqrt{\frac{1}{n} + \frac{1}{q} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}}] = 0{,}95.$$

Так что мы можем построить доверительный интервал для $\bar{Y}_0$ относительно $\hat{Y}_0$:

$$\hat{Y}_0 \pm t(v,0{,}975)s \sqrt{\frac{1}{n} + \frac{1}{q} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2}}.$$

Эти пределы, конечно, шире, чем для среднего значения Y при данном X_0 так как ожидается, что 95% будущих наблюдений при X_0 (для $q = 1$) или будущих средних из q наблюдений (для $q > 1$) лежат внутри них.

F-критерий значимости регрессии

F-критерий известен из математической статистики и может быть использован для исследования моделей регрессии. Так как Y_i – случайные величины, то любая функция от них тоже будет случайной величиной. Например, две функции:

MS_R – средний квадрат, обусловленный регрессией, и s^2 – средний квадрат, обусловленный остаточной вариацией, тоже будут случайными. Они представлены в таблице дисперсионного анализа. Эти функции имеют свои

собственные распределения, средние, дисперсии и моменты. Их средние значения будут:

$$E(MS_R) = \sigma^2 + \beta_1 \sum (X_i - \overline{X})^2 \ , \ E(s^2) = \sigma^2 ,$$

где E означает среднее или математическое ожидание случайной величины.

Положим, что ошибки ε_i – независимые случайные величины с распределением $N(0, \sigma^2)$. Можно показать, что если $\beta_1 = 0$, то величина MS_R, умноженная на свое число степеней свободы (в данном случае на 1), подчиняется χ^2-распределению с тем же самым числом степеней свободы. Более того, $(n-2)s^2/\sigma^2$ тоже имеет χ^2-распределение с $(n-2)$ степенями свободы.

Так как эти две случайные величины независимы, то из статистической теории вытекает, что отношение

$$F = \frac{MS_R}{s^2}$$

подчиняется $F-$ распределению с 1 и $(n-2)$ степенями свободы при условии, что $\beta_1 = 0$. Этот факт можно использовать как критерий выполнимости равенства $\beta_1 = 0$.

Мы должны сравнить отношение $F = \frac{MS_R}{s^2}$ с $100 \cdot (1-\alpha)\%$ табличной точкой $F(1, n-2)$, чтобы посмотреть, можно ли на основе имеющихся данных рассматривать β_1 как число, отличное от нуля.

Для объяснения доли разброса мы определили, что

$$R^2 = \frac{\sum (\hat{Y}_i - \overline{Y})^2}{\sum (Y_i - \overline{Y})^2},$$

где суммирование ведется по $i = 1,2,..n$. Тогда R^2 измеряет долю общего разброса относительно среднего $\overline{Y}$, объясняемую регрессией. Ее часто выражают в процентах, умножая на 100. Фактически R – это корреляция между Y и

$\hat{Y}$, и ее обычно называют множественным коэффициентом корреляции. Коэффициент R^2 самое большее может достигнуть величины 1 (или 100%), когда все значения X различны. Если в данных имеются повторяющиеся опыты, то величина R^2 не может достигнуть 1, как бы хороша ни была модель. Это объясняется вариацией в данных из-за "чистой" ошибки опыта (ошибки воспроизводимости).

9.6. Проверка адекватности модели линейной регрессии

Обсудим методы анализа точности описания данных предложенной моделью. Рассмотрим, что такое **неадекватность и «чистая» ошибка**.

Построенная линия регрессии – это расчетная линия, основанная на предположениях. Эти предположения мы должны рассматривать как предварительные. Мы можем при некоторых обстоятельствах (условиях) проверить, корректна ли наша модель. Мы будем изучать проявления предполагаемой некорректности модели.

Вспомним, что $e_i = Y_i - \hat{Y}_i$ – остатки при $X = X_i$. Это величины, на которые действительные наблюдаемые значения Y_i отличаются от $\hat{Y}_i$, вычисленных по уравнению. Было показано, что $\sum e_i = 0$. Остатки содержат информацию о том, почему построенная модель недостаточно правильно объясняет наблюдаемый разброс зависимой переменной Y.

Пусть $\eta_i = E(Y_i)$ обозначает величину среднего для «истинной» модели при $X = X_i$. Тогда можем записать

$$Y - \hat{Y}_i = (Y_i - \hat{Y}_i) - E(Y_i - \hat{Y}_i) + E(Y_i - \hat{Y}_i) =$$

$$= [(Y_i - \hat{Y}_i) - (\eta_i - E(\hat{Y}_i)) + (\eta_i - E(\hat{Y}_i))] = q_i + B_i,$$

где $q_i = (Y_i - \hat{Y}_i) - (\eta_i - E(\hat{Y}_i))$, $B_i = \eta_i - E(\hat{Y}_i)$.

Величина B_i – это ошибка смещения при $X = X_i$. Если модель верна, то $E(\hat{Y}_i) = \eta_i$ и $B_i = 0$. Если же модель не верна,

то $E(\hat{Y}_i) \neq \eta_i$ и $B_i \neq 0$, и его значение зависит от «истинной» модели и значения X_i.

Переменная q_i – это случайная величина, имеющая нулевое среднее, т.к.

$$E(q_i) = E(Y_i - \hat{Y}_i) - (\eta_i - E(\hat{Y}_i)) =$$
$$= \eta_i - E(\hat{Y}_i) - (\eta_i - E(\hat{Y}_i)) = 0,$$

и это верно независимо от того, будет ли модель правильна, т.е. $E(\hat{Y}_i) = \eta_i$.

Можно показать, что q_i коррелированны, и величина $q_1^2 + ... + q_n^2$ имеет математическое ожидание, или среднее значение, $(n-2)\sigma^2$,

где $\sigma^2 = V(Y_i)$ – дисперсия ошибки.

Исходя из этого, можно показать, что остаточный средний квадрат, т.е. величина

$$\frac{1}{n-2}\sqrt{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}$$

имеет математическое ожидание, или среднее значение, σ^2, если предложенная модель корректна, и

$$\sigma^2 + \sum_{i=1}^{n} B_i^2 \Big/ (n-2),$$

если модель не корректна.

Если модель корректна, т.е. $B_i = 0$, то остатки будут коррелированными случайными отклонениями q_i, и остаточный средний квадрат можно использовать как оценку дисперсии ошибки σ^2.

Если модель не корректна, т.е. $B_i \neq 0$, то остатки содержат оба компонента: случайный q_i и систематический B_i. Мы можем рассматривать их как случайную ошибку разброса и систематическую ошибку смещения. В простейшем

случае подбора прямой, как правило, можно определить ошибку смещения, непосредственно, исследуя график с данными (см. рис. 9.4 а, б, в, г).

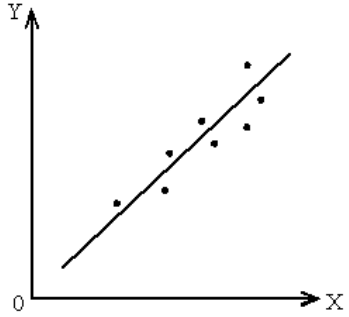


Рис.9.4 а

На рис.9.4 а представлен случай 1, когда проверяется модель $Y = \beta_0 + \beta_1 X + \varepsilon$. В этом случае нет неадекватности, линейная регрессия значима, используется модель $\widehat{Y} = b_0 + b_1 X$

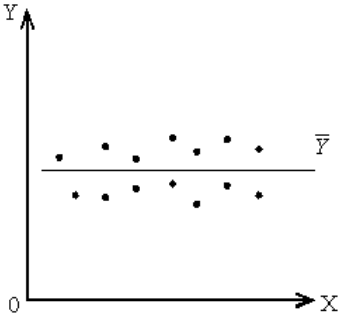


Рис.9.4 б

На рис.9.4 б представлен случай 2, когда проверяется модель $Y = \beta_0 + \beta_1 X + \varepsilon$. В этом случае нет неадекватности, линейная регрессия незначима, используется модель $\widehat{Y}_i = \overline{Y} + b_1 (X_i - \overline{X})$

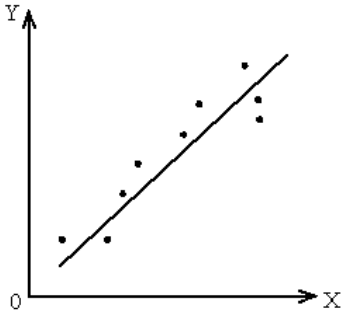


Рис.9.4 в

На рис.9.4 в представлен случай 3, когда проверяется модель $Y = \beta_0 + \beta_1 X + \varepsilon$. В этом случае неадекватность значима, линейная регрессия незначима, следует проверить модель $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$.

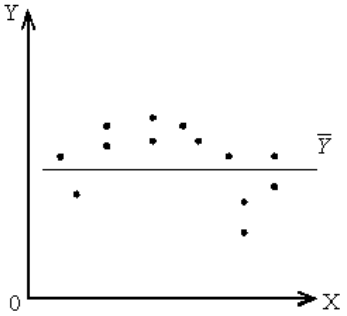


Рис.9.4 г

На рис.9.4 г представлен случай 4, когда проверяется модель $Y = \beta_0 + \beta_1 X + \varepsilon$. В этом случае неадекватность значима, следует проверить модель $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$.

Если модель более сложна или включает больше переменных, то это невозможно (т.е. невозможно определить ошибку смещения из данных). Если существует априорная оценка σ^2 (под «априорной оценкой» мы понимаем оценку, полученную на основе ранее выполненных опытов), то можно увидеть (или проверить по F-критерию), значимо ли

остаточная сумма квадратов превышает нашу априорную оценку. Если это так, то говорят, что имеет место неадекватность и следует пересмотреть модель.

Если априорной оценки нет σ^2, но измерения Y повторялись (два или более раза) при одинаковых значениях X, то мы можем использовать эти повторения для получения оценки σ^2.

Такую оценку называют «чистой» ошибкой, потому что если сделать X одинаковыми для двух наблюдений, то только случайные вариации могут влиять на результаты и создавать разброс между ними. Эти различия обеспечивают получение оценки σ^2, которая более надежна, чем оценки, получаемые из других источников. По этой причине имеет смысл ставить опыты с повторением.

Когда в данных содержатся повторные опыты, нужны дополнительные обозначения для множества наблюдений Y при одном и том же значении X.

Пусть мы имеем m различных значений X и к j-му из этих значений X_i, где $i = 1,2,..m$, относятся n_j наблюдений. Тогда мы говорим, что

$Y_{11}, Y_{12},...Y_{1n} - n_1$ повторных наблюдений при X_1,

$Y_{21}, Y_{22},...Y_{2n} - n_2$ повторных наблюдений при X_2,

$Y_{ju} - u$-е наблюдение при X_j, $u = 1,2,...n_j$,

$Y_{m1}, Y_{m2},...Y_{mn} - n_m$ повторных наблюдений при X_m.

Всего получается $n = \sum_{j=1}^{m}\sum_{u=1}^{n} 1 = \sum_{j=1}^{m} n_j$ наблюдений.

Вклад суммы квадратов, связанной с «чистой» ошибкой для n_1 наблюдений при X_1, будет равен внутренней сумме квадратов Y_{1u} относительно их среднего $\bar{Y}_1$, т.е.

$$\sum_{u=1}^{n}(Y_{1u} - \bar{Y}_1)^2 = \sum_{u=1}^{n} Y_{1u}^2 - n_1\bar{Y}_1^2 = \sum_{u=1}^{n} Y_{1u}^2 - (\sum_{u=1}^{n} Y_{1u})^2 \Big/ n_1 . \quad (9.27)$$

Объединяя внутренние суммы квадратов для всех серий повторных опытов, мы получим общую сумму квадратов «чистых» ошибок в виде

$$\sum_{j=1}^{m}\sum_{u=1}^{n}(Y_{ju}-\overline{Y}_j)^2$$

со степенями свободы

$$n_e=\sum_{j=1}^{m}(n_j-1)=\sum_{j=1}^{m}n_j-m\,.$$

Отсюда средний квадрат «чистых» ошибок равен

$$s_e^2=\frac{\sum_{j=1}^{m}\sum_{u=1}^{n}(Y_{ju}-\overline{Y}_j)^2}{\sum_{j=1}^{m}n_j-m}$$

и он служит оценкой σ^2 независимо от того, корректна ли подобранная модель. Словом, эта величина – полная SS между повторениями, деленная на общее число степеней свободы.

Замечание.

Если имеются два наблюдения Y_{j1} и Y_{j2} в точке X_j, то

$$\sum_{u=1}^{2}(Y_{ju}-\overline{Y}_j)^2=\frac{1}{2}(Y_{j1}-Y_{j2})^2\,. \qquad (9.28)$$

Это удобная форма для вычислений. Такая SS имеет только одну степень свободы.

Таким образом, сумма квадратов «чистых» ошибок фактически оказывается частью остаточной суммы квадратов, что мы и покажем.

Остаток для u-го наблюдения при X_j можно записать в виде:

$$Y_{ju}-\widehat{Y}_j=(Y_{ju}-\overline{Y}_j)-(\widehat{Y}_j-\overline{Y}_j)\,.$$

Здесь мы пользуемся тем обстоятельством, что все повторные точки при любом X_j имеют одно и то же предсказанное

значение $\hat{Y}_j$. Если мы возведем в квадрат обе части этого выражения, а затем просуммируем их по u и по j, то получим:

$$\sum_{j=1}^{m}\sum_{u=1}^{n}(Y_{ju}-\hat{Y}_j)^2=\sum_{j=1}^{m}\sum_{u=1}^{n}(Y_{ju}-\bar{Y}_j)^2+\sum_{j=1}^{m}n_j(\hat{Y}_j-\bar{Y}_j)^2, \quad (9.29)$$

причем парные произведения исчезают при суммировании по u для каждого j. Слева в уравнении (9.29) стоит остаточная сумма квадратов. Первый член в правой части – это сумма квадратов «чистых» ошибок. Последний член мы называем суммой квадратов неадекватности. Отсюда следует, что сумму квадратов, обусловленную «чистой» ошибкой, можно ввести в таблицу дисперсионного анализа, как это показано на рис. 9.5.

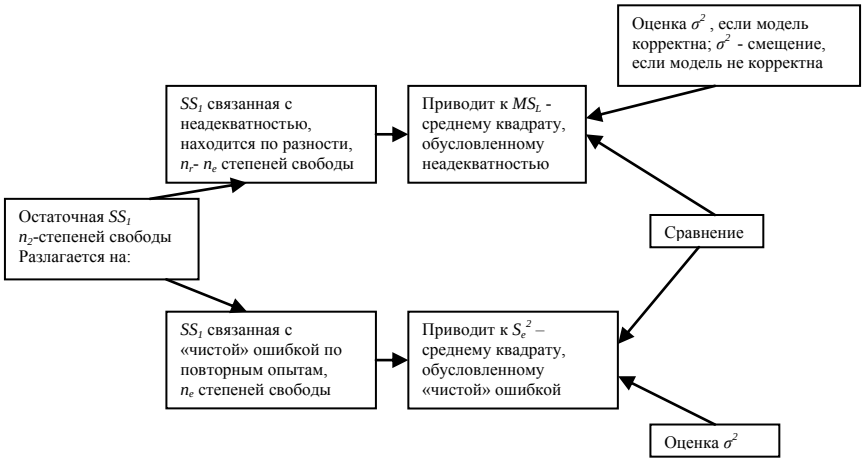


Рис. 9.5

Затем производится сравнение отношения $F=MS_L/s_e^2$ со $100\cdot(1-\alpha)\%$ точкой F-распределения при $(n_\eta-n_e)$ и n_e степенях свободы. Если это отношение является:

1) значимым, то это показывает, что модель, по-видимому, неадекватна и следует изучить, когда и как встречается неадекватность.

2) не значимым, то это показывает, что, по-видимому, оснований сомневаться в адекватности модели нет, и что как средний квадрат, связанный с «чистой» ошибкой, так и средний квадрат, обусловленный неадекватностью, могут использоваться как оценки σ^2.

Объединенная оценка σ^2 может быть получена из суммы квадратов, связанной с «чистой» ошибкой, и суммы квадратов, связанной с неадекватностью, путем объединения их в остаточную сумму квадратов и деления ее на остаточное число степеней свободы n_r, что дает $s^2 = SS/n_r$.

Если используются не повторные опыты в качестве повторных, то s^2 будет проявлять склонность к переоценке σ^2, а F-критерий для проверки неадекватности будет иметь тенденцию к ошибочному определению ее отсутствия

Пример на иллюстрацию неадекватности и «чистой» ошибки

В табл. 9.3 представлены результаты 24 наблюдений.

Таблица 9.3

Номер наблюдения	Y	X	Номер наблюдения	Y	X	Номер наблюдения	Y	X
1	2,3	1,3	9	1,7	3,7	17	3,5	5,3
2	1,8	1,3	10	2,8	4,0	18	2,8	5,3
3	2,8	2,0	11	2,8	4,0	19	2,1	5,3
4	1,5	2,0	12	2,2	4,0	20	3,4	5,7
5	2,2	2,7	13	5,4	4,7	21	3,2	6,0
6	3,8	3,3	14	3,2	4,7	22	3,0	6,0
7	1,8	3,3	15	1,9	4,7	23	3,0	6,3
8	3,7	3,7	16	1,8	5,0	24	5,9	6,7

По этим данным была оценена линия регрессии

$$\hat{Y}_i = 1{,}436 + 0{,}338 X_i .$$

Рассчитаем данные для таблицы дисперсионного анализа. Сумма квадратов, обусловленная регрессией,

$$SS(b_1/b_0) = \sum (\hat{Y}_i - \bar{Y})^2 = \frac{[\sum X_i Y_i - \sum X_i \sum Y_i / n]^2}{\sum X_i^2 - (\sum X_i)^2 / n} = 6{,}326,$$

(число степеней свободы 1).

Общая (полная) скорректированная сумма квадратов есть

$$s_{YY} = \sum Y_i^2 - (\sum Y_i)^2 / n = 27{,}518,$$

(число степеней свободы $n-1=23$);

$$s_{YY} = \sum (Y_i - \bar{Y})^2 = 27{,}518;$$

$$s_{YY} = \sum Y_i^2 - n\bar{Y}^2 .$$

Остаток считается по формуле $e_i = Y_i - \hat{Y}_i$, для каждого наблюдения

$\hat{Y}_i$ считается по предложенной прямой $\hat{Y}_i = 1{,}436 + 0{,}338$,

а сумма квадратов для остатков считается по формуле

$e_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ со степенями свободы $n-2$.

Для примера $SS = \sum (Y_i - \hat{Y}_i)^2 = 21{,}192$ $(n-2=22)$;

$s^2 = SS/(n-2) = 0{,}963.$

Таблица 9.4 является таблицей дисперсионного анализа. Сумма квадратов:

1) связанная с "чистой" ошибкой, из повторений при $X = 1{,}3$ есть $\frac{1}{2}(2{,}3 - 1{,}8)^2 = 0{,}125$ с 1 степенью свободы (использована формула (9.28);

2) связанная с "чистой" ошибкой, из повторений при $X = 4{,}7$ (считается по формуле (9.27)

$(5{,}4)^2 + (3{,}2)^2 + (1{,}9)^2 - (5{,}4 + 3{,}2 + 1{,}9)^2 / 3 = 43{,}01 - 36{,}75 = 6{,}26$с 2 степенями свободы.

Таблица 9.4

Источник	Число степеней свободы	Суммы квадратов SS	Средние квадраты MS	F -отношение
Регрессия	1	6,326	6,326	MS – средний. квадрат, обусловл. регресс-сией
Остаток	22	21,192	0,963	s^2 – средний квадрат, обусловленный остаточной вариацией
Общий, скорректированный	23	27,518		$F = MS/s^2$ =6,569 значимо при α =0,05, если нет неадекватности

Аналогичные вычисления дают следующие величины.

Таблица 9.5

Уровень X	$\sum_{u=1}^{n}(Y_{ju}-\bar{Y}_j)^2$	Число степеней свободы	Уровень X	$\sum_{u=1}^{n}(Y_{ju}-\bar{Y}_j)^2$	Число степеней свободы
1,3	0,125	1	4,0	0,240	2
2,0	0,845	1	4,7	0,260	2
3,3	2,000	1	5,3	0,980	2
3,7	2,000	1	6,0	0,020	1
			Итого	12,470	11

Теперь полученные данные можно переписать в таблицу 9.6 дисперсионного анализа.

Таблица .9.6

Дисперсионный анализ (демонстрация неадекватности)

Источник	Число степеней свободы	Суммы квадратов SS	Средние квадраты MS	F - отношение
Регрессия	1	6,326	6,326	
Остаток	22	21,192	$s^2=0{,}963$	
Неадекват-ность	11	8,722	$MS_L=0{,}793$	

Продолжение табл.9.6

«Чистая» ошибка	11	12,470	s_e^2 =1,134	$F = MS/s^2$ =6,569 значимо α =0,05
Общий, скоррек-тирован-ный	23	27,518		$F = MS_L/s_e^2$ =0,699 не значимо

Неадекватность находится как разность

$$SS_{ОСТАТОК} - SS_{ЧИТСАЯ\ ОШИБКА}.$$

Отношение

$$F = MS_L/s_e^2 = 0{,}699$$

не значимо, так как оно меньше 1. Поэтому на основе такого критерия нет оснований сомневаться в адекватности нашей модели и можно использовать s^2=0,963 как оценку для σ^2, чтобы иметь возможность воспользоваться F - критерием для проверки значимости всей регрессии. F - критерий состоятелен, только если нет неадекватности представления результатов нашей моделью

Итак, в итоге рассмотрим все необходимые действия, когда наши данные содержат повторные наблюдения.

1) Подобрать модель, составить простую таблицу дисперсионного анализа с двумя входами: регрессией и остатком. Но для общей регрессии пока не использовать F - критерий.

2) Вычислить сумму квадратов, связанную с “чистой” ошибкой, и разложить остаточную сумму квадратов как на рис. 9.5.

3) Применить F - критерий для неадекватности. Если критерий неадекватности не значим, т.е. нет смысла сомневаться в адекватности модели, то перейти к пункту 4.б.

4.а) Значимая неадекватность. Прекратить анализ подобранной модели и искать пути ее улучшения методами анализа остатков. Не применять F - критерий для общей регрессии и не пытаться строить доверительные интервалы. Если нет адекватности подобранной модели, то не верны предпосылки, которые лежат в основе этих операций.

4.б) Неадекватность не значима. Снова объединить суммы квадратов для "чистых" ошибок и неадекватности в остаточную сумму квадратов. Использовать остаточный средний квадрат s^2 в качестве оценки для $D(Y) = \sigma^2$, применить F - критерий для общей регрессии, получить доверительные пределы для истинного среднего значения Y, вычислить R^2 и т.д.

Заметим, что если модель проходит все барьеры, это еще не означает, что она правильна, просто нет оснований считать ее неадекватной имеющимся данным. Если неадекватность обнаружена, то может понадобиться другая модель, возможно, квадратичная, вида

$$Y = \alpha + \beta X + \gamma X^2 + \varepsilon .$$

На рис. 9.4 показаны некоторые ситуации, которые могут возникнуть, когда прямая строится по данным шаг за шагом.

Влияние повторных опытов на R^2

Мы уже отмечали, что величина R^2 не может достичь 1, если есть повторные опыты. Никакая модель не может изменить вариацию, обусловленную "чистой" ошибкой. В нашем последнем примере: сумма квадратов, обусловленная "чистой" ошибкой, равна 12,470 при 11 степенях свободы. То, что модель подогнана к этим данным, не имеет значения, все равно величина 12,470 остается неизменяемой и

необъясняемой. Следовательно, максимум R^2, достижимый при этих данных, есть

$$\max R^2 = \frac{SS_{общая} - SS_{обусловл.}}{SS_{общая}} = \frac{27{,}518 - 12{,}470}{27{,}518} = 0{,}5468,$$

или 54,68 %.

То значение R^2, которое фактически достигнуто для подобранной модели, равно:

R^2 = SS $_{РЕГР.}$ / общаяSS $_{скор}$ = 6,326/27,518=0,2299, или 22,99 %,

Иными словами, мы можем объяснить 0,2299/0,5468=0,4202, или 42,02 % того, что возможно объяснить.

«Чистая» ошибка в многофакторном случае

Полученные формулы для одной переменной применимы в общем случае для n предикторов $X_1, X_2, \ldots$. Но у повторных опытов должны совпадать все координаты, т.е., например, следующие четыре отклика для четырех точек

$(X_1, X_2, X_3, X_4) = (4,2,17,1), (4,2,17,1), (4,2,17,1), (4,2,17,1)$

дают повторные опыты. Однако четыре точки

$(X_1, X_2, X_3, X_4) = (4,2,16,1), (4,2,17,1), (4,2,18,1), (4,2,19,1)$

уже не дают повторных опытов, поскольку координаты X_3 во всех этих случаях различны.

Корреляция между переменными X и Y и регрессия

Когда мы выдвигали постулат о линейности модели

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

то мы предварительно полагали, что Y можно выразить как функцию 1-го порядка от X без учета ошибок.

В такой зависимости X обычно предполагается фиксированным (неслучайным), т.е. не имеющим вероятностного распределения, Y предполагается случайной величиной, имеющей распределение вероятностей со средним $\beta_0 + \beta_1 X$ и дисперсией $D(\varepsilon)$.

Рассмотрим две случайные величины U и W с некоторым непрерывным совместным двумерным распределением вероятностей $f(U,W)$. Тогда мы определяем коэффициент корреляции между ними как

$$\rho_{UW} = \frac{\mathrm{cov}(U,W)}{\sqrt{D(U)D(W)}},$$

где $\mathrm{cov}(U,W) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(U - M(U))(W - M(W))\, f(U,W)\,dU dW$,

$$D(U) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(U - M(U))^2\, f(U,W)\,dU dW ,$$

$$M(U) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} U f(U,W)\,dU dW .$$

Значения $D(W)$ и $M(W)$ определяются аналогично в терминах W. Известно, что $-1 \le \rho_{UW} \le 1$. Величина ρ_{UW} служит мерой линейной зависимости между случайными величинами U и W. Если имеется выборка объема n из величин $(U_1,W_1),...(U_n,W_n)$ с совместным распределением, то величина

$$r_{UW} = \frac{\sum_{i=1}^{n}(U_i - \overline{U})(W_i - \overline{W})}{\sqrt{\sum_{i=1}^{n}(U_i - \overline{U})^2}\sqrt{\sum_{i=1}^{n}(W_i - \overline{W})^2}} \qquad (9.30)$$

называется выборочным коэффициентом корреляции между U и W, оценивает ρ_{UW} и представляет собой эмпирическую меру линейной зависимости между U и W. r_{UW} лежит между -1 и +1.

Для нашей регрессионной задачи будем рассматривать r_{XY}. Если корреляция r_{XY} не равна нулю, это значит, что в нашем множестве данных существует некоторая линейная зависимость между конкретными значениями X_i и Y_i при $i = 1,2,..n$. (Мы предполагаем, что X_i не подвержены

воздействию случайных ошибок, а значения Y_i имеют случайный разброс относительно среднего, зависящего от модели.) Допустим, что имеются данные $(X_1,Y_1),...(X_n,Y_n)$. Применяя уравнение (9.30), мы можем получить $r_{XY} = r_{YX}$, а если постулировать модель $Y = \beta_0 + \beta_1 X + \varepsilon$, то можно получить оценку коэффициента регрессии b_1 по уравнению:

$$b_1 = \frac{\sum X_i Y_i - (\sum X_i \sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2/n} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

Рассмотрим, как связаны между собой r_{XY} и b_1. Сравнивая уравнение (9.30) при замене U и W на X и Y с уравнением для b_1, видим, что

$$b_1 = r_{XY} \sqrt{\frac{\sum (Y_i - \overline{Y})^2}{\sum (X_i - \overline{X})^2}},$$

где суммирование ведется по $i = 1,2,..n$.

Иными словами, b_1 – это «взвешенный» вариант величины r_{XY}, причем взвешивание происходит с помощью отношения разброса Y_i к разбросу X_i. Если мы запишем, что

$$(n-1)s_Y^2 = \sum (Y_i - \overline{Y})^2,$$

$$(n-1)s_X^2 = \sum (X_i - \overline{X})^2, \text{ то}$$

$$b_1 = r_{XY} \frac{s_Y}{s_X}.$$

Таким образом, b_1 и r_{XY} весьма близки, но интерпретируются по-разному. Коэффициент r_{XY} измеряет связь между X и Y, в то время как b_1 измеряет величину изменения переменной Y, которую можно предсказать, если изменение переменной $X = 1$.

Множественный коэффициент корреляции, который уже был рассмотрен, равен

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(Y_i - \bar{Y})^2}.$$

Кроме того $$r_{Y\hat{Y}} = R, \qquad (9.31)$$

т.е. R равно корреляции между имеющимися наблюдениями Y_i и предсказанными значениями $\hat{Y}_i$. Уравнение (9.31) справедливо для любой линейной регрессии с любым числом предикторов.

Обратная регрессия (случай прямой линии)

Допустим, что мы подобрали уравнение прямой $\hat{Y} = b_0 + b_1 X$ по множеству данных (X_i, Y_i), $i = 1,2,..n$. И теперь хотим для определенного значения Y, например Y_0, получить предсказанное значение $\hat{X}_0$, соответствующее значению X. А еще хотим получить доверительный интервал, устанавливаемый для X вокруг $\hat{X}_0$. Это задача обратной регрессии.

Есть несколько способов решения задач такого типа. Допустим, что Y_0 есть среднее арифметическое q наблюдений. Нарисуем полученную прямую и доверительные интервалы для $\bar{Y}$ при данном X (рис.9.6).

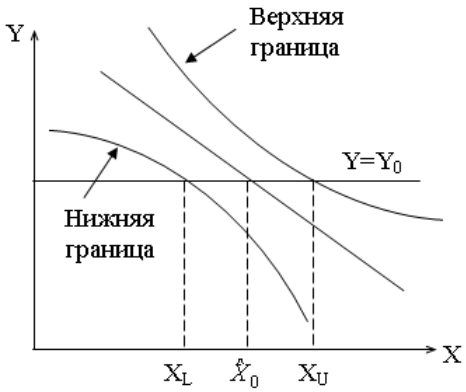


Рис.9.6

На высоте Y_0 проведем горизонтальную линию, параллельную оси X. Там, где эта линия пересечет кривые доверительных интервалов, опустим перпендикуляры на ось X и получим точки: X_L – нижний $100\cdot(1-\alpha)\%$ и X_U – верхний $100\cdot(1-\alpha)\%$ пределы. Перпендикуляр, опущенный на ось X из точки пересечения двух прямых, дает обратную оценку X, определяемую как решение уравнения $Y_0 = b_b + b_1\widehat{X}_0$ относительно X_0, а именно:

$$\widehat{X}_0 = (Y_0 - b_0)/b_1 .$$

Для получения значений X_L и X_U можно поступить так. На рис. 9.6 X_L – это координата точки пересечения прямой

$$Y = Y_0 \text{ (т.е. } Y = b_0 + b_1\widehat{X}_0) \qquad (9.32)$$

и кривой

$$Y = Y_{XL} - ts\sqrt{\frac{1}{n} + \frac{(X_L - \overline{X})^2}{s_{XX}}} \ , \qquad (9.33)$$

,

где

$s_{XX} = \sum (X_i - \overline{X})^2$, $Y_{XL} = b_0 + b_1 X_L$, $t = t(v, 1 - \frac{\alpha}{2})$ – обычная процентная точка для t-критерия, а v – число степеней свободы для s^2.

Приравнивание уравнений (9.32) и (9.33), сокращение b_0, перенесение квадратного корня из левой части уравнения в правую, возведение обеих частей в квадрат для избавления от корня приводит к следующему уравнению относительно X_L:

$$PX_L^2 + 2QX_L + R = 0, \qquad (9.34)$$

где

$$P = b_1^2 - \frac{t^2 s^2}{s_{XX}},$$

$$Q = -b_1^2 \hat{X}_0 + \frac{t^2 s^2 \overline{X}}{s_{XX}},$$

$$R = b_1^2 \hat{X}_0^2 - \frac{t^2 s^2}{n} - \frac{t^2 s^2 \overline{X}^2}{s_{XX}}.$$

Мы получим то же самое уравнение для X_U.

Таким образом, X_L и X_U – оказываются корнями уравнения (9.34), т.е.

$$\overline{X} + \frac{b_1 (Y_0 - \overline{Y}) \pm ts\sqrt{[(Y_0 - \overline{Y})^2 / s_{XX}] + (b_1^2 / n) - (t^2 s^2 / n s_{XX})}}{b_1^2 - (t^2 s^2 / s_{XX})}.$$

Обратное оценивание не имеет большого практического значения, если регрессия не достаточно хорошо определена, т.е. если b_1 – не значим. При этом может случиться так, что корни X_L и X_U могут, вообще говоря, оказаться комплексными

Решение о стратегии эксперимента

Пусть экспериментатор хочет собрать данные об отклике Y при n выбранных значениях предиктора для определения

эмпирической зависимости между Y и этим предиктором. Пусть предиктор не подвержен действию случайной ошибки, а Y-отклик – подвержен. Будем считать, что допускаются повторные опыты.

Перед экспериментатором стоит масса вопросов.

1) Какой диапазон значений предиктора выбрать?

Диапазон должен быть достаточно широк, чтобы сделать полезные выводы. Вместе с тем, он должен быть достаточно узок, чтобы результаты представить простейшей моделью. Когда решение принято, диапазон, или интервал, $(-1,1)$ кодируется без нарушения общности.

Допустим, что если время Т. изменяется в диапазоне $140 \cdot c < T < 200 \cdot c$, то кодирование $X = (T - 170)/30$ даст интервал $(-1,1)$. Преобразование здесь имеет вид

X= (натур. величина – середина натур. интервала)/половина диапазона

2) Какого рода зависимость окажется правильной?

3) А если предложенная зависимость ошибочна? Какую альтернативу выбрать? Если была прямая линия, то альтернатива представляет квадратичную зависимость?

4) Каков разброс, присущий отклику, т.е. чему равна $D(Y) = \sigma^2$. В данном случае экспериментатор, возможно, пожелает для оценки σ^2 присоединить повторные опыты.

5) Сколько опытов может понадобиться?

6) Сколько мест (т.е. различных значений X) стоит выбрать? Сколько повторных опытов имеет смысл проводить в каждом месте?

Рассмотрим конкретный пример. Допустим, наш экспериментатор решил, что во всем диапазоне $-1 \le X \le 1$ кодированного предиктора наиболее правдоподобна линейная зависимость, возможна квадратичная альтернатива, дисперсии σ^2, всего возможны 14 опытов.

Так при каких же значениях X (т.е. в каких местах) стоит проводить опыты, сколько в каждом из этих мест и на каком основании? Каждый план с самого начала имеет 14

степеней свободы. Две из них идут на оценки параметров b_0 и b_1. Остается 12 степеней свободы, которые надо разделить между неадекватностью и «чистой» ошибкой.

Рассмотрим таблицу. Строки (1) и (2) в таблице показывают, как эти остаточные степени свободы разбиваются в различных планах. В строке (3) приведены значения $\sqrt{\sum (X_i - \overline{X})^2}$, которые пропорциональны стандартному отклонению коэффициента b_1 подобранной прямой. В строке (4) показано число параметров, которые можно найти по данным соответствующего плана. Заметим, что число степеней свободы для неадекватности равно числу различных мест для X в данных минус число параметров в постулированной модели. Так как в нашем примере есть два параметра, подлежащих оценке (β_0 и β_1), то разность между числами, стоящими в строках (4) и (1) таблицы всюду равна 2.

Поскольку в примере требуется, чтобы σ^2 оценивалась через «чистую» ошибку, стратегия (а) оказывается в данном случае плохой. Поскольку мы не в состоянии проверить адекватность, то вариант (ж) автоматически исключается. Случай (б) исключается, т.к. этот план из оставшихся имеет наибольшее стандартное отклонение b_1/σ, а также в нем предлагается использовать 7 разных уровней, когда главной альтернативой служит квадратичная модель. Семь уровней слишком много! Ясно, что наилучший выбор заключается в вариантах (в), (г), (д) или (е). Какой из них выбрать – зависит от предпочтений экспериментатора. С точки зрения стандартного отклонения b_1/σ лучше взять вариант (е). Варианты (в) и (г) отклоняются, т.к. 3-х и 2-х степеней свободы для неадекватности много, особенно когда альтернатива всего лишь квадратичная зависимость.

Таблица 9.7

Характеристики различных стратегий

		(а)	(б)	(в)	(г)	(д)	(е)	(ж)
1	число степеней свободы для неадекватности	12	5	3	2	1	1	0
2	число степеней свободы для чистой ошибки	0	7	9	10	11	11	12
3	Стандартное отклонение b_1/σ	0,43	0,40	0,33	0,31	0,32	0,29	0,27
4	число мест	14	7	5	4	3	3	2

Использование табл. 9.7, в которой представлены характеристики различных стратегий, позволит экспериментатору принять правильное решение о проведении экспериментов по оценке моделей регрессии.

10. КЛАСТЕРНЫЙ АНАЛИЗ

10.1. Основные понятия кластерного анализа

В задачах обработки результатов экспериментов группировка первичных данных является основным приемом решения задачи классификации. При наличии нескольких признаков (исходных или обобщенных) задача классификации может быть решена методами кластерного анализа. Основное отличие этих методов заключается в том, что отсутствуют обучающие выборки, т.е. априорная информация о распределении генеральной совокупности, которая представляет собой вектор X.

Рассмотрим следующую задачу. Пусть исследуется совокупность n объектов, каждый из которых характеризуется X признаками, измеренными k раз. Требуется разбить эту совокупность на однородные группы (классы). При этом отсутствует априорная информация о характере распределения измерений X внутри классов.

Полученные в результате разбиения группы называют кластерами (от англ. cluster – группа элементов, характеризуемых каким-либо общим свойством). Методы нахождения кластеров называются кластер – анализом или распознаванием образов с самообучением.

Рассмотрим три различных подхода к проблеме кластерного анализа: эвристический, экстремальный и статистический.

Эвристический подход характеризуется отсутствием формальной модели изучаемого явления и критерия для сравнения различных решений. Основой подхода является алгоритм, построенный исходя из интуитивных соображений.

При экстремальном подходе также не формулируется исходная модель, а задается критерий, определяющий качество разбиения на кластеры. Этот подход полезен, когда цель исследования четко определена. Качество разбиения в этом случае может измеряться эффективностью выполнения цели.

Основой статистического подхода решения задач кластерного анализа является вероятностная модель исследуемого процесса. Данный подход дает возможность решать задачи, связанные с воспроизводимостью результатов кластерного анализа.

Рассмотрим формы представления исходных данных и определение мер близости. В кластерном анализе формой представления исходных данных служит прямоугольная матрица, каждая строка которой представляет результат измерения k признаков на одном из обследованных объектов.

$$\vec{X} = \begin{Bmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{Bmatrix}.$$

В конкретных ситуациях может представлять интерес, как группировка объектов, так и группировка признаков.

Числовые значения, входящие в матрицу X, могут соответствовать трем типам переменных – количественным, ранговым и качественным. Количественные переменные обладают свойством упорядоченности и над ними можно производить арифметические операции. Значения ранговых переменных тоже упорядочены, и их можно пронумеровать натуральными числами. Однако использование этих чисел в арифметических операциях будет некорректным. Качественными называются переменные, принимающие два (дихотомные) или более значений. Этим значениям также можно поставить в соответствие некоторые числа, но они не будут отражать упорядоченности значений качественной переменной. Упорядоченности подвергаются дихотомные переменные, два значения которых (как правило, они обозначаются числами 0 и 1) можно считать упорядоченными. Желательно, чтобы таблица исходных данных содержала один

тип переменных. В противном случае разные типы переменных стараются свести к одному типу.

Матрица X не является единственным способом представления исходных данных. Исходная информация может быть задана в виде квадратной матрицы

$$A = \left(a_{ij} \right),\ i,j=1,2,\ldots,k,$$

где элемент a_{ij} который определяет степень близости i – го объекта к j – му, т.е. сходство этих объектов.

Большинство алгоритмов кластерного анализа исходят из матрицы расстояний (или сходства), либо требуют вычисления отдельных ее элементов. Если данные представлены в форме X, то первым этапом решения задачи поиска кластеров будет выбор способа вычисления расстояний или близости (сходства) между объектами или признаками.

Достаточно просто определяется близость между признаками. Чаще всего мерами близости служат различные статистические коэффициенты связи. Если признаки количественные, то можно использовать оценки обычных парных выборочных коэффициентов корреляции R_{ij}, $i, j = 1,2,\ldots,k$. Однако коэффициент корреляции измеряет только линейную связь. Если связь нелинейная, то следует произвести подходящее преобразование шкалы признаков.

Рассмотрим наиболее распространенные типы нормировок, переводящих признаки в безразмерные величины. Пусть имеются одномерные наблюдения $x_1, x_2, \ldots, x_n$. Нормировки:

$$x_i^{'} = \left(x_i - x_{\min}\right) / \left(x_{\max} - x_{\min}\right),$$

$$x_i^{'} = \left(x_i - \bar{x}\right) / S,$$

где $\bar{x} = \frac{1}{n}\sum x_i$ - среднее арифметическое,

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

позволяют изменять масштабы шкал признаков для использования корреляционных статистических связей. Существуют различные коэффициенты связи, определенные для ранговых, качественных и дихотомных переменных.

10.2. Расстояние между объектами и мера близости

Наиболее трудным и наименее формализованным в задаче классификации является определение понятия однородности объектов.

В общем случае понятие однородности объектов задается либо введением правила вычисления расстояния $\rho\left(X_i, X_j\right)$ между любой парой исследуемых объектов $(X_1, X_2, ..., X_n)$, либо заданием некоторой функции $L\left(X_i, X_j\right)$, характеризующей степень близости (сходства) i – го и j – го объектов. Если задана функция $\rho\left(X_i, X_j\right)$, то близкие с точки зрения этой метрики объекты считаются однородными, принадлежащими одному классу. При этом необходимо сопоставлять $\rho\left(X_i, X_j\right)$ с некоторым пороговым значением, в каждом конкретном случае определяемом по-своему.

Аналогично используется и мера близости $L\left(X_i, X_j\right)$. При классификации изображений эту меру рассматривают как меру сходства изображений. Общие требования к мере близости сводятся к следующему:

1) мера близости должна быть всегда положительной величиной, т.е.

$$L\left(X_i, X_j\right) \geq 0,$$

2) должна обладать свойством симметрии

$$L\left(X_i, X_j\right) = L\left(X_j, X_i\right),$$

3) мера близости объекта с самими собой должна быть максимальной

$$L\left(X_i, X_j\right) = \max_j L\left(X_i, X_j\right),$$

4) мера близости должна обладать свойством монотонности убывания $L\left(X_i, X_j\right)$ по $\rho\left(X_i, X_j\right)$, т.е. из $\rho\left(X_k, X_e\right) \geq \rho\left(X_i, X_j\right)$ должно следовать неравенство $L\left(X_k, X_e\right) \leq L\left(X_i, X_j\right)$.

Выбор метрики или меры близости является узловым моментом исследования, от которого зависит окончательный вариант разбиения объектов на классы при заданном алгоритме разбиения. В каждом конкретном случае это выбор должен производиться по-своему в зависимости от целей исследования, физической и статистической природы вектора наблюдений X, априорных сведений о характере вероятностного распределения X.

Рассмотрим наиболее часто используемые расстояния и меры близости в задачах кластерного анализа. В случае зависимых компонент вектора наблюдений X и их различной значимости при решении задач классификации используют обобщенное расстояние Махаланобиса, задаваемое формулой

$$\rho_0\left(X_i, X_j\right) = \sqrt{\left(X_i - X_j\right)^T \Lambda^T \sum^{-1} \Lambda \left(X_i - X_j\right)}, \quad (10.1)$$

где $\sum$ – ковариационная матрица генеральной совокупности, из которой извлекаются наблюдения; Λ – некоторая симметрическая неотрицательно – определенная матрица весовых коэффициентов. Обычно она выбирается диагональной.

Обычное евклидово расстояние

$$\rho_E\left(X_i, X_j\right) = \sqrt{\sum_{l=1}^{k}\left(x_{il} - x_{jl}\right)^2}, \qquad (10.2)$$

где x_{il}, x_{jl} – величина l – й компоненты i – го (j – го) объекта ($l = 1,2,\ldots,k;\ i, j = 1,2,\ldots,n$).

Использование этого расстояния оправдано в следующих случаях, если:

а) наблюдения берутся из генеральных совокупностей, имеющих многомерное нормальное распределение, причем наблюдения независимы и имеют одну и ту дисперсию;

б) компоненты вектора наблюдений X однородны по физическому смыслу и одинаково важны для классификации;

в) признаковое пространство совпадает с геометрическим пространством.

С геометрической точки зрения евклидово расстояние может оказаться бессмысленным, если признаки имеют разные единицы измерения. Для приведения признаков к одинаковым единицам используют нормировку каждого признака путем деления центрированной величины на среднее квадратическое отклонение и переходят от матрицы $\vec{X}$ к нормированной матрице с элементами

$$x_{il}^{H} = \frac{x_{il} - \bar{x}_l}{S_l},$$

где x_{il} - значение l – го признака у i – го объекта;

$\bar{x}_l$ - среднее арифметическое значение l – го признака;

$S_l = \sqrt{\frac{1}{n-1}\sum\left(x_{il} - \bar{x}_l\right)^2}$ – среднее квадратическое отклонение l – го признака.

В результате этой операции могут быть нежелательные последствия. Если кластеры, например, хорошо разделены по одному признаку и не разделены по другому, то после

нормировки разделительные свойства первого признака будут уменьшены с увеличением «шумового» эффекта второго.

«Взвешанное» евклидово расстояние

$$\rho_{BE}\left(X_i, X_j\right) = \sqrt{\sum_{l=1}^{k} \mu_l \left(x_{il} - x_{jl}\right)^2} \qquad (10.3)$$

применяется в случаях, когда каждой компоненте x_l вектора наблюдений X удается приписать некоторый «вес» μ_l, пропорциональный степени важности признака в задаче классификации. Обычно принимают $0 \le \mu_l \le 1$, где $l = 1,2,...,k$.

Хеммингово расстояние

$$\rho_{H}\left(X_i, X_j\right) = \sum_{l=1}^{k} \left| x_{il} - x_{jl} \right| \qquad (10.4)$$

используется как мера различия объектов, задаваемых дихотомическими признаками.

Решение задачи классификации многомерных объектов предусматривает в качестве предварительного этапа исследования реализацию методов выделения наиболее существенных информативных признаков, т.е. уменьшения размерности наблюдаемого пространства. С этой целью каждую из компонент $x_1, x_2, \ldots, x_k$ рассматривают как объект, подлежащий классификации. После разбиения на небольшое число однородных групп, для дальнейшего исследования оставляют по одному представителю от каждой группы. Предполагается, что признаки, попавшие в группу, связаны друг с другом и несут информацию о каком-то одном свойстве объекта.

В качестве близости между отдельными признаками обычно используют различные характеристики степени их коррелированности, в первую очередь коэффициенты корреляции. Другие расстояния (метрики) также

используются. Выбор метрики определяется структурой признакового пространства и целью классификации. Формализовать этот этап задачи классификации пока не представляется возможным.

10.3. Расстояние между кластерами

В ряде процедур классификации (кластер – процедур) используют понятия расстояния между группами объектов и меры близости двух групп объектов.

Пусть S_i - i – я группа (класс, кластер), состоящая из n_i объектов;

$\bar{x}_i$ - среднее арифметическое векторных наблюдений S_i группы, т.е. «центр тяжести» i – й группы;

$\rho(S_l, S_m)$ - расстояние между группами S_l и S_m.

Наиболее употребительными расстояниями и мерами близости между классами объектов является:

- - расстояние, измеряемое по принципу «ближайшего соседа»

$$\rho_{\min}(S_l, S_m) = \min_{\substack{x_i \in S_l \\ x_j \in S_m}} \rho(x_i, x_j); \qquad (10.5)$$

- - расстояние, измеряемое по принципу «дальнего соседа»

$$\rho_{\max}(S_l, S_m) = \max_{\substack{x_i \in S_l \\ x_j \in S_m}} \rho(x_i, x_j); \qquad (10.6)$$

- - расстояние, измеряемое по «центрам тяжести» групп

$$\rho(S_l, S_m) = \rho(\bar{x}_l, \bar{x}_m) \qquad (10.7)$$

- - расстояние, измеряемое по принципу «средней связи».

Это расстояние определяется как среднее арифметическое всех попарных расстояний между представителями рассматриваемых групп.

$$\rho_{ср}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{x_i \in S_l} \sum_{x_j \in S_m} \rho(x_i, x_j) \quad (10.8)$$

Здесь n_l и n_m - количество объектов в классах S_l и S_m.

Колмогоров А.Н. предложил обобщенное расстояние между классами. Оно в качестве частных случаев включает в себя все расстояния, рассмотренные ранее. Обобщенное расстояние – это степенное среднее, которое определяется формулой

$$\rho_{об}(S_l, S_m) = \left[\frac{1}{n_l n_m} \sum_{x_i \in S_l} \sum_{x_j \in S_m} \rho^r(x_i, x_j) \right]^{1/r} \quad (10.9)$$

Можно показать, что

при $r \to \infty$ $\rho_{об}(S_l, S_m) = \rho_{\max}(S_l, S_m)$;

при $r \to -\infty$ $\rho_{об}(S_l, S_m) = \rho_{\min}(S_l, S_m)$;

при $r = 1$ $\rho_{об}(S_l, S_m) = \rho(S_l, S_m)$.

Из формулы (10.9) следует, что если $S_{(m,q)} = S_m \cup S_q$ – группа элементов, полученная путем объединения кластеров S_m и S_q, то обобщенное расстояние между кластерами S_l и $S_{(mq)}$ определяется по формуле

$$\rho_{об}(S_l, S_{(mq)}) = \left\{ \frac{n_m \left[\rho_{об}(S_l, S_m) \right]^r + n_q \left[\rho_{об}(S_l, S_q) \right]^r}{n_m + n_q} \right\}^{1/r} \quad (10.10)$$

Расстояние между группами элементов важно в агломеративных иерархических кластер - процедурах. Принцип работы таких алгоритмов состоит в последовательном объединении сначала самых близких

элементов, а затем и целых групп все более и более отдаленных друг от друга элементов. При этом расстояние между классами S_l и $S_{(mq)}$, являющимися объединением двух других классов S_m и S_q, можно определить по формуле:

$$\rho_{l,(mq)} = \rho\left(S_l, S_{(mq)}\right) = \alpha\rho_{lm} + \beta\rho_{lq} + \gamma\rho_{mq} + \delta\left|\rho_{lm} - \rho_{lq}\right|, \qquad (10.11)$$

где $\rho_{lm} = \rho\left(S_l, S_m\right);\ \rho_{lq} = \rho\left(S_l, S_q\right);\ \rho_{mq} = \rho\left(S_m, S_q\right)$ - расстояния между классами S_l, S_m и S_q;

α, β, γ и δ – числовые коэффициенты, значение которых определяет специфику процедуры, ее алгоритм.

Например, при $\alpha = \beta = -\delta = 1/2$ и $\gamma = 0$ приходим к расстоянию, построенному по принципу «ближайшего соседа». При $\alpha = \beta = \delta = 1/2$ и $\gamma = 0$ расстояние между классами определяется по принципу «дальнего соседа», как расстояние между двумя самыми дальними элементами этих классов.

При $\alpha = \frac{n_m}{n_m + n_q}$, $\beta = \frac{n_q}{n_m + n_q}$, $\gamma = \delta = 0$ соотношение (10.11) приводит к расстоянию между классами, вычисленному как среднее из расстояний между всеми парами элементов, один из которых берется из одного класса, а другой – из другого класса.

10.4. Функционалы качества разбиения

Существует много способов разбиения на классы заданной совокупности элементов. Возникает задача сравнительного анализа качества этих способов разбиения. С этой целью вводится понятие функционала качества разбиения $Q(S)$, определенного на множестве всех возможных

разбиений. Наилучшее разбиение представляет собой такое разбиение, при котором достигается экстремум выбранного функционала качества. Выбор того или иного функционала качества разбиения, как правило, опирается на эмпирические соображения.

Рассмотрим некоторые наиболее распространенные функционалы качества разбиения. Пусть при исследовании выбрана метрика ρ в пространстве X и $S = (S_1, S_2, ..., S_\rho)$ некоторое фиксированное разбиение наблюдений $X_1, X_2, ..., X_n$ на заданное число ρ классов $S_1, S_2, ..., S_\rho$.

Существуют следующие характеристики функционала качества

- сумма внутриклассовых дисперсий

$$Q_1(S) = \sum_{l=1}^{p} \sum_{x_i \in S_l} \rho^2(x_i, \bar{x}_l);$$

- сумма попарных внутриклассовых расстояний между элементами

$$Q_2(S) = \sum_{l=1}^{p} \sum_{x_j \in S_l} \rho^2(x_i, x_j).$$

$Q_1(S)$ и $Q_2(S)$ широко используются в задачах кластерного анализа для сравнения качества процедур разбиения;

- обобщенная внутриклассовая дисперсия

$$Q_3(S) = \det\left(\sum_{l=1}^{p} n_l C_l \right),$$

где detA – определитель матрицы A;

C_l – выборочная ковариационная матрица класса S_l, элементы которой определяется по формуле

$$c_{qm}(l) = \frac{1}{n_l} \sum_{x_i \in S_l} \left(x_{iq} - \bar{x}_q \right)\left(x_{im} - \bar{x}_m \right),\ q, m = \overline{1, k},$$

где x_{iq} - q – я компонента многомерного наблюдения x_i;

$\bar{x}_q$ - среднее значение q – й компоненты, вычисленное по наблюдениям l – го класса.

Качество разбиения характеризуют и другим видом обобщенной дисперсии, в которой операция суммирования C_l заменена операцией умножения

$$Q_4(S) = \prod_{l=1}^{p} (\det C_l)^{n_l} .$$

Функционалы $Q_3(S)$ и $Q_4(S)$ обычно используют при решении вопроса: не сосредоточены ли наблюдения, разбитые на классы, в пространстве размерности, меньшей, чем k.

10.5. Иерархические процедуры

Иерархические (деревообразные) процедуры бывают двух типов: агломеративные и дивизимные. В агломеративных процедурах начальным является разбиение, состоящее из n одноэлементных классов, а конечным – из одного класса, в дивизимных наоборот. Принцип работы иерархических агломеративных (дивизимных) процедур состоит в последовательном объединении (разделении) групп элементов сначала самых близких (далеких), а затем все более отдаленных (близких) друг от друга. Большинство этих алгоритмов исходит из матрицы расстояний (сходства).
Громоздкость вычислительной реализации является недостатком иерархических процедур.

Рассмотрим пример агломеративного иерархического алгоритма. На первом шаге каждое наблюдение $X_i (i = 1,2,\ldots,n)$ рассматривается как отдельный кластер. В дальнейшем на каждом шаге работы алгоритма происходит объединение двух самых близких кластеров, и, с учетом принятого расстояния, по формуле пересчитывается матрица расстояний. Размерность матрицы, очевидно, снижается на единицу. Работа алгоритма заканчивается, когда все наблюдения объединены в один класс. Иерархическую

классификацию представляют в виде дендрограммы (dendron (греч.) – дерево). Дивизимные иерархические процедуры используются для распознавания образов.

Пример. Провести классификацию n=6 объектов, каждый из которых характеризуются двумя признаками:

№ объекта i	1	2	3	4	5	6
x_{i1}	5	6	5	10	11	10
x_{i2}	10	12	13	9	9	7

Расположение объектов в виде точек на плоскости показано на рис. 10.1.

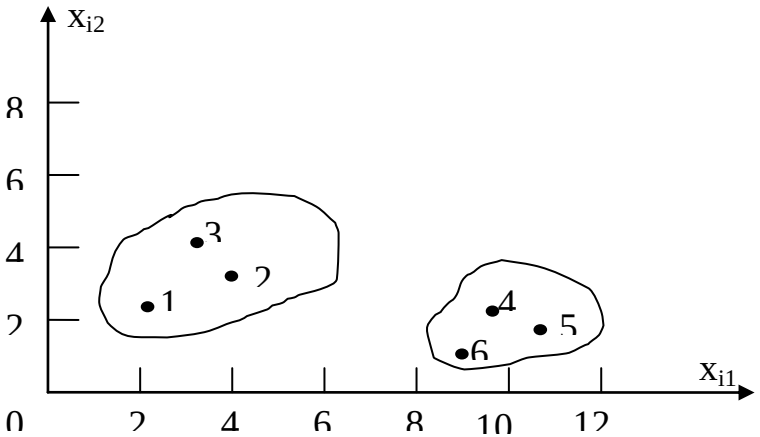


Рис. 10.1. Классификация объектов

Решение

Воспользуемся агломеративным иерархическим алгоритмом классификации. В качестве расстояния между объектами возьмем обычное евклидово расстояние. Тогда согласно формуле (10.2) расстояние между первым и вторым объектами

$$\rho_{12} = \sqrt{(5-6)^2 + (10-12)^2} = 2{,}24,$$

а между первым и третьим объектами

$$\rho_{13} = \sqrt{(5-5)^2 + (10-13)^2} = 3.$$

Очевидно, что $\rho_{11} = 0$.

Аналогично находим расстояние между шестью объектами и строим матрицу расстояний

$$R_1 = \left\{\rho\left(x_i, x_j\right)\right\} = \begin{bmatrix} 0 & 2.24 & 3 & 5.10 & 6.08 & 5.83 \\ 2.24 & 0 & 1.41 & 5 & 5.83 & 6.40 \\ 3 & 1.41 & 0 & 6.40 & 7.21 & 7.81 \\ 5.10 & 5 & 6.40 & 0 & 1 & 2 \\ 6.08 & 5.83 & 7.21 & 1 & 0 & 2.24 \\ 5.83 & 6.40 & 7.81 & 2 & 2.24 & 0 \end{bmatrix}.$$

Из матрицы расстояний следует, что четвертый и пятый объекты наиболее близки $\rho_{4,5} = 1{,}00$ и поэтому объединяются в один кластер. После объединения объектов имеем пять кластеров:

Номер кластера	1	2	3	4	5
Состав кластера	(1)	(2)	(3)	(4,5)	(6)

Расстояние между кластерами определим по принципу «ближайшего соседа», воспользовавшись формулой пересчета (10.11). Расстояние между объектом S_1 и кластером $S_{(4,5)}$ будет

$$\rho_{1,(4.5)} = \rho\left(S_1, S_{(4.5)}\right) = \frac{1}{2}\rho_{14} + \frac{1}{2}\rho_{15} - \frac{1}{2}\left|\rho_{14} - \rho_{15}\right| =$$

$$= \frac{1}{2}\left[5.10 + 6.08\right] - \frac{1}{2}\left[\left|5.10 - 6.08\right|\right] = 5.10.$$

Таким образом, расстояние $\rho_{1,(4.5)}$ равно расстоянию от объекта 1 до ближайшего к нему объекта, входящего в кластер $S_{(4,5)}$, т.е. $\rho_{1,(4.5)} = \rho_{1,4} = 5{,}10$. Тогда матрица расстояний примет вид

$$R_2 = \begin{bmatrix} 0 & 2.24 & 3 & 5.10 & 5.83 \\ 2.24 & 0 & 1.41 & 5 & 6.40 \\ 3 & 1.41 & 0 & 6.40 & 7.81 \\ 5.10 & 5 & 6.40 & 0 & 2 \\ 5.83 & 6.40 & 7.81 & 2 & 0 \end{bmatrix}.$$

Объединим второй и третий объекты, имеющие наименьшее расстояние $\rho_{2,3} = 1.41$. После объединения объектов имеем четыре кластера:

$$S_{(1)}; S_{(2,3)}; S_{(4,5)}; S_{(6)}.$$

Вновь найдем матрицу расстояний. Для того чтобы рассчитать расстояние до кластера $S_{(2,3)}$ воспользуемся матрицей расстояний R_2. Например, расстояние между кластерами $S_{(4,5)}$ и $S_{(2,3)}$ равно

$$\rho_{(4,5),(2,3)} = \frac{1}{2}\rho_{(4,5),2} + \frac{1}{2}\rho_{(4,5),3} - \frac{1}{2}\left|\rho_{(4,5),2} - \rho_{(4,5),3}\right| = \frac{5}{2} + \frac{6{,}40}{2} - \frac{1{,}40}{2} = 5.$$

Проведя аналогичные расчеты, получим

$$R_3 = \begin{bmatrix} 0 & 2.24 & 5.10 & 5.83 \\ 2.24 & 0 & 5 & 6.40 \\ 5.10 & 5 & 0 & 2 \\ 5.83 & 6.40 & 2 & 0 \end{bmatrix}.$$

Объединим кластеры $S_{(4,5)}$ и S_6, расстояние между которыми, согласно матрице R_3, наименьшее $\rho_{(4,5),6} = 2$. В результате получим три кластера

$$S_{(1)}; S_{(2,3)} \text{ и } S_{(4,5,6)}.$$

Матрица расстояний будет иметь вид:

$$R_4 = \begin{bmatrix} 0 & 2.24 & 5.10 \\ 2.24 & 0 & 5 \\ 5.10 & 5 & 0 \end{bmatrix}.$$

Объединим теперь кластеры $S_{(1)}$ и $S_{(2,3)}$, расстояние между которыми $\rho_{1,(2,3)} = 2{,}24$. В результате получим два кластера: $S_{(1,2,3)}$ и. $S_{(4,5,6)}$ Расстояние между ними, найденное по принципу «ближайшего соседа», будет $\rho_{(1,2,3)(4,5,6)} = 5$.

Результаты иерархической классификации объектов представлены на рис. 10.2 в виде дендрограммы: по горизонтали откладываются номера объектов, а по вертикали – значения мер близости, при которых происходили соединения классов.

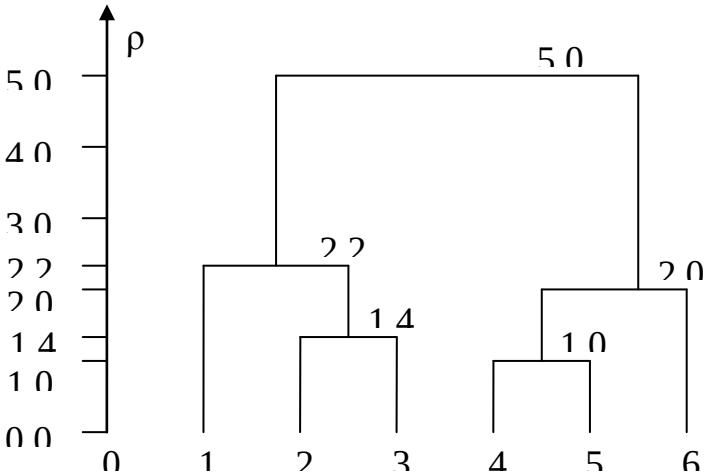


Рис. 10.2. Дендрограмма

На рис. 10.2 приводятся расстояния между кластерами, которые объединяются на одном этапе. В этом примере предпочтение следует отдать предпоследнему этапу

классификации, когда все объекты объединены в два кластера $S_{(1,2,3)}$ и $S_{(4,5,6)}$.

10.6. Эвристические методы и алгоритмы

Подавляющая часть классификаций на практике производится эвристическими методами. Это объясняется относительной простотой и содержательной ясностью таких алгоритмов. Эти алгоритмы допускают вмешательство в их работу путем изменения одного или нескольких параметров, кроме того, они обладают невысокой трудоемкостью. Наиболее распространенные эвристические алгоритмы используют для своей реализации различные типы классов. Разделение классов на типы впервые осуществил советский ученый Миркин Б.Г. Какие множества называются группами однородных объектов или классов можно уточнить на основе заданного расстояния между объектами. Выделим некоторые типы классов.

1. Класс типа ядра. Такой класс называется *сгущением*. Все расстояния между объектами внутри класса меньше любого из расстояний между объектами класса и остальной частью множества объектов. На рис. 10.3 сгущениями являются A и B. Остальные пары множеств не разделяются с помощью этого определения.

2. Кластер (*сгущение в среднем*). Среднее расстояние внутри класса меньше среднего расстояния от объектов класса до всех остальных. Множества C и Д теперь разделяются, но у E(G) среднее внутреннее расстояние больше, чем среднее расстояние между E и F (G и H).

3. Класс типа ленты (*слабое сгущение*). Существует пороговое значение $T>0$ такое, что для любого x_i из класса S найдется такой объект $x_j \in S$, что $\rho_{ij} \le T$, а для всех $x_k \notin S$ справедливо неравенство $\rho_{ik} > T$. В смысле этого

определения на рис. 10.3 разделяются все пары множеств кроме I и J, K и L.

4. Класс с центром. Существует порог R>0 и некоторая точка x^* в пространстве, занимаемом объектами класса *S*, такие, что все объекты из S и только они содержатся в шаре радиуса R с центром в x^*. Часто в качестве x^* выступает центр масс класса *S*, т.е. координаты центра определяются как средние значения признаков у объектов класса. Множества I и J являются классами с центром, а E, F и G – нет.

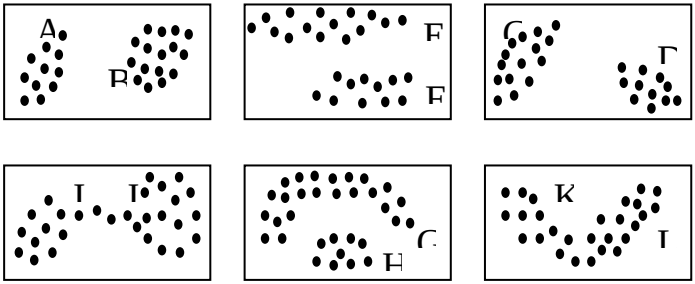


Рис. 10.3. Типы классов

Рассмотрим наиболее часто используемые эвристические алгоритмы.

1. Метод К-средних предназначен для выделения и распознавания классов типа 4 («класс с центром»). Приведем два варианта.

а) Алгоритм Г. Бола и Д. Холла был предложен в 1965 году. Суть алгоритма состоит в следующем. Случайно выбираются *K* объектов (эталонов). Каждый объект присоединяется к ближайшему эталону (тем самым образуются *K* классов), в качестве новых эталонов принимаются центры масс классов. Считается, что каждому объекту приписана масса 1. После пересчета объекты снова распределяются по ближайшим эталонам и т.д. Критерием окончания алгоритма служит стабилизация центров масс всех классов.

б) Алгоритм Дж. Мак-Кина предложен в 1967 году. Он отличается от метода Бола и Холла тем, что при просмотре списка объектов пересчет центра масс класса происходит после присоединения к нему каждого очередного объекта. Этот алгоритм связан с функционалом качества разбиения $Q_1(S)$ - сумма внутриклассовых дисперсий.

2. Алгоритм «Форель». Первоначальное название этого алгоритма было: ФОРЭЛ – ФОР малый ЭЛемент. Этот алгоритм предложен в 1966 году Елкиной В.Н. и Загоруйко Н.Г. Этот алгоритм выполняется следующим образом. Случайный объект объявляется центром класса; все объекты, находящиеся от него на расстоянии не большем R, входят в первый класс. В нем определяется центр масс, который объявляется новым центром класса и т.д. до стабилизации центра. Затем все объекты, попавшие в первый класс, изымаются, и процедура повторяется с новым случайным центром.

После проведения классификации важно в удобной форме представить ее результаты. Для этого необходимы следующие характеристики классификации.

1. Распределение номеров объектов по номерам классов.
2. Гистограмма межобъектных расстояний.
3. Средние внутриклассовые расстояния.
4. Матрица средних межклассовых расстояний.
5. Визуальное представление данных на плоскости двух (в пространстве трех) «наиболее информативных» признаков.
6. Дендрограмма для иерархических процедур.
7. Средние значения и размахи во всех классах для каждого признака.

Последний пункт очень важен, так как в большинстве случаев описание классов происходит по средним значениям признаков в них. Сопоставление средних значений для заданного признака наиболее просто осуществляется, если

классы не имеют наложения проекций. Степень разделенности классов по каждой оси можно охарактеризовать с помощью коэффициента

$$\gamma = -\sum L_j / \sum R_i ,$$

где R_i – размах по заданному признаку l – го класса, а $L_1, L_2, \ldots$ - длины наложений проекций классов на ось признака. Если γ=1, то классы полностью разделены. Чем ближе γ к 0, тем больше наложение проекций классов друг на друга.

10.7. Алгоритм К – внутригрупповых средних

Рассмотрим компьютерную реализацию алгоритма К – внутригрупповых средних. Этот алгоритм наиболее часто используется в задачах классификации объектов. Он также используется для организации входных данных для нейронных сетей. Алгоритм, представленный ниже, минимизирует показатель качества, определенный как сумма квадратов расстояний всех точек, входящих в кластерную область, до центра кластера. Эта процедура, которую часто называют алгоритмом, основанным на вычислении К внутригрупповых средних, состоит из следующих шагов.

Шаг 1. Выбираются К исходных центров кластеров $z_1(1)$, $z_2(1), \ldots, z_k(1)$. Этот выбор производится произвольно, и обычно в качестве исходных центров используются первые К результатов выборки из заданного множества объектов.

Шаг 2. На k-м шаге итерации заданное множество объектов $\{x\}$ распределяется по К кластерам по правилу:

$$x \in S_j(k), \quad \text{если} \quad \|x - z_j(k)\| < \|x - z_i(k)\| \qquad (10.12)$$

для всех $i=1, 2, \ldots, K$, $i \neq j$, где $S_j(k)$ – множество объектов, входящих в кластер с центром $z_j(k)$. В случае равенства в (10.12) решение принимается произвольным образом.

Шаг 3. На основе результатов шага 2 определяются новые центры кластеров $z_j(k+1)$, $j=1, 2, \ldots, K$, исходя из условия, что сумма квадратов расстояний между всеми объектами,

принадлежащими множеству $S_j(k)$, и новым центром кластера должна быть минимальной. Другими словами, новые центры кластеров $z_j(k+1)$ выбираются таким образом, чтобы минимизировать показатель качества

$$Q_j = \sum_{x \in S_j(k)} \left\| x - z_j(k+1) \right\|^2, \quad j = 1,2,...,K. \tag{10.13}$$

Центр $z_j(k+1)$, обеспечивающий минимизацию показателя качества, является в сущности, выборочным средним, определенным по множеству $S_j(k)$. Следовательно, новые центры кластеров определяются как

$$z_j(k+1) = \frac{1}{N_j} \sum_{x \in S_j(k)} x, \quad j = 1,2,..., K, \tag{10.14}$$

где N_j – число выборочных объектов, входящих в множество $S_j(k)$. Очевидно, что название алгоритма «К внутригрупповых средних» определяется способом, принятым для последовательной коррекции назначения центров кластеров.

Шаг 4. Равенство $z_j(k+1) = z_j(k)$ *при* $j = 1, 2, ..., K$ является условием сходимости алгоритма, и при его достижении выполнение алгоритма заканчивается. В противном случае алгоритм повторяется от шага 2.

Качество работы алгоритмов, основанных на вычислении К внутригрупповых средних, зависит от числа выбираемых центров кластеров, от выбора исходных центров кластеров, от последовательности осмотра объектов и от геометрических особенностей данных. Хотя для этого алгоритма общее доказательство сходимости не известно, получение приемлемых результатов можно ожидать в тех случаях, когда данные образуют характерные гроздья, отстоящие друг от друга достаточно далеко. В большинстве случаев практическое применение этого алгоритма потребует проведения экспериментов, связанных с выбором различных значений параметра К и исходного расположения центров кластеров.

Лабораторная работа № 8. Реализация алгоритма К – внутригрупповых средних в пакете MATHCAD

Целью лабораторной работы является программная реализация алгоритма К – внутригрупповых средних для корректирования центров исходных классов. Классы составляют случайные величины с различными законами распределения.

Задачами лабораторной работы являются:

- формирование случайных величин с нормальным, равномерным и экспоненциальным законом распределения с использованием математического пакета MATHCAD;
- описание эталонных образов классов в виде векторов, характеристиками которых являются: математическое ожидание, дисперсия, среднеквадратическое отклонение, эксцесс и асимметрия законов распределения;
- применение алгоритма К– внутригрупповых средних для корректирования центров исходных классов случайных величин с рассмотренными законами распределения.

Пример выполнения задания

В качестве примера представлен графический интерфейс программы, выполненной в среде Borland Delphi 7. Форма, предназначенная для реализации алгоритма К - внутригрупповых средних, показана на рис. 10.4.

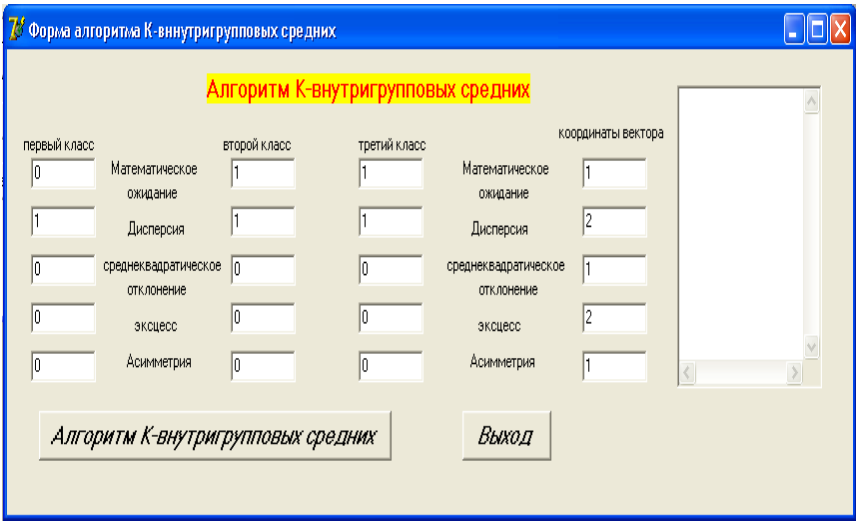


Рис. 10.4. Форма для алгоритма К– внутригрупповых средних

При запуске алгоритма выдается сообщение о том, что идет обработка введенного вектора (рис. 10.5).

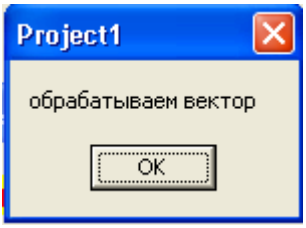


Рис. 10.5. Сообщения, выдаваемые при обработке вектора

После выполнения алгоритма в окне результатов будут выведены скорректированные центры исходных классов (рис. 10.6).

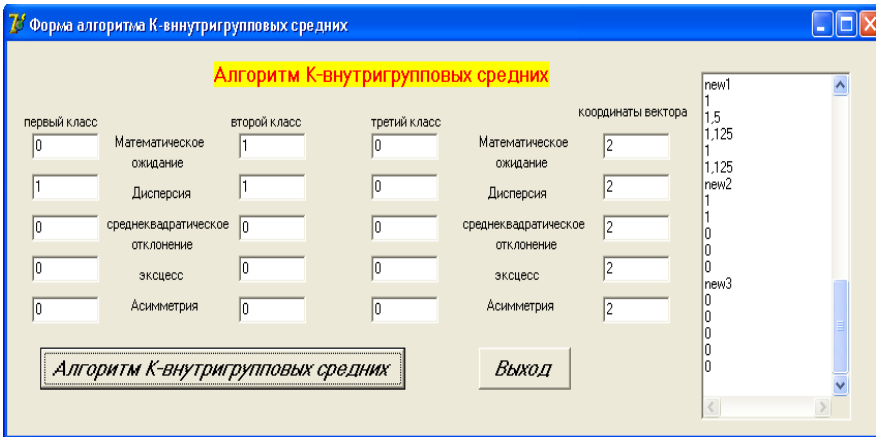


Рис. 10.6. Результат работы критерия

Задание для самостоятельной работы

Применить алгоритм К – внутригрупповых средних для корректирования центров исходных классов случайных величин с законами распределения: χ^2, Стьюдента, Фишера-Снедекора.

ЗАКЛЮЧЕНИЕ

Современный мир насыщен информацией. Математические технологии анализа данных позволяют человеку существовать в информационной среде и принимать правильные решения. К этим математическим технологиям относится прикладная математическая статистика, с помощью которой можно компактно описать данные, выявить их структуру, провести классификацию и найти закономерности в случайных явлениях.

Данное учебное пособие представляет собой учебник по прикладной математической статистике, ориентированный на постоянное использование компьютеров, и состоит из двух частей. В первой части были рассмотрены основные понятия выборочной теории и методы обработки статистических данных, оценивание неизвестных параметров распределения, а также методы максимального правдоподобия и моментов нахождения точечных оценок, интервальное оценивание параметров распределений. В первой части были представлены основные сведения о распределениях случайных величин, которые используются в задачах прикладной математической статистики.

Вторая часть учебного пособия состоит из пяти глав. В шестой главе рассмотрены основные понятия и теоремы проверки статистических гипотез. Седьмая глава посвящена задачам проверки статистических гипотез о параметрах нормального распределения. В восьмой главе рассмотрены критерии согласия и однородности: хи - квадрат Пирсона, Колмогорова, Колмогорова – Смирнова, непараметрические критерии. Девятая глава посвящена рассмотрению элементов регрессионного и дисперсионного анализа. В десятой главе приводятся основные сведения по кластерному анализу, который очень важен для обработки экспериментальных данных.

Отличительной особенностью данного учебного пособия является наличие задач с решениями, что значительно

облегчает понимание методов прикладной математической статистики. В конце глав имеются описания лабораторных работ. Эти работы ставят своей целью реализацию на компьютере соответствующих теоретических положений и получение практических навыков в решении задач. Два компьютерных пакета программ: MATHCAD и STATISTICA. используются в лабораторных работах. Первый представляет собой язык программирования, который ориентирован на математические вычисления. Пакет STATISTICA представляет собой мощную интегрированную систему статистического анализа и обработки данных.

Следует отметить, что данное учебное пособие представляет собой совокупность теории, задач с решениями и лабораторных работ, что значительно отличает его от многих учебников.

Учебное пособие предназначено для бакалавров и магистров

ПРИЛОЖЕНИЕ 1

ВАРИАНТЫ ВЫБОРОК

Вариант 1

65,3	69,1	56,1	57,1	73,1	74,4	99,9	57,1	97,6	73,4
25,1	36,1	60,1	36,6	61,1	59,1	57,6	70,1	64,1	63,8
90,1	78,1	30,4	70,1	61,6	68,1	77,9	93,6	82.1	49,6
64,9	63,8	64,8	84,1	45,1	47,1	47,5	67,1	85,5	72,3
38,9	65,1	86,3	38,1	62,9	62,5	65,9	93,4	42,1	71,7
66,4	73,7	43,5	57,5	69,5	66,8	63,9	64,5	72,6	62,5
48,6	65,1	42,1	54,8	72,6	62,9	33,5	50,7	43,5	52,7
64,6	41,8	69,1	83,7	44,4	77,8	46,4	58,9	59,1	84,2
61,4	91,8	52,2	84,3	54,4	45,1	58,1	61,2	105,1	78,1
45,1	58,9	79,9	37,1	54,2	47,4	76,9	56,1	32,9	45,1

Вариант 2

54,2	58,0	45,0	46,0	62,2	63,3	88,8	46,0	80,5	62,3
14,0	25,0	49,0	25,5	50,0	48,0	46,5	59,0	53,0	52,7
79,0	67,0	19,3	59,0	50,5	57,0	66,8	82,5	71,0	38,5
53,9	52,8	53,7	73,0	34,0	36,0	26,4	56,0	74,4	61,2
27,8	54,0	75,2	27,0	51,8	51,4	54,8	82,3	31,0	60,6
55,3	62,6	32,4	46,4	58,4	55,7	52,8	53,4	61,5	51,4
37,5	54,0	31,0	43,7	61,5	51,8	22,4	39,6	32,4	41,6
53,5	30,7	58,0	72,6	33,3	66,7	35,2	47,8	48,0	73,1
50,3	80,7	41,1	73,2	43,3	34,0	47,0	50,1	94,0	67,0
34,0	47,8	68,8	26,0	42,8	46,3	68,8	45,0	21,8	34,7

Вариант 3

43,9	55,5	59,3	46,3	47,3	63,5	64,6	90,1	47,3	81,8
63,6	15,3	26,3	50,3	26,8	51,3	49,3	47,8	60,3	54,3
54,0	80,3	68,3	20,6	60,3	20,6	51,8	59,3	68,1	83,8
39,8	55,2	54,1	55,0	74,3	35,3	37,3	72,3	57,3	75,6
62,5	29,1	55,2	76,5	28,3	53,1	52,7	56,1	83,6	92,3
61,9	56,6	63,9	33.7	47,7	59,7	56,0	54,1	54,7	62,8
52,7	38,8	55,3	32,3	45,0	62,8	53,1	23,7	40,9	33,7
42,9	54.8	32,0	59,3	79,9	34,6	68,0	36,5	49,1	74,4
51,6	82,0	42,4	74,5	44,6	35,3	48,3	51,4	95,3	68,3
35,3	49,1	70,1	27,3	44,1	47,6	70,1	46,3	23,1	35,3

Вариант 4

63,2	67,0	54,0	55,0	71,2	72,3	97,8	55,0	89,5	71,3
23,0	34,0	58,0	34,5	59,0	57,0	55,5	68,0	62,0	61,7
88,0	76,0	28,3	68,0	75,8	91,5	80,0	59,5	66,0	47,5
62,9	61,8	62,7	82,0	43,0	45,0	35,4	65,0	83,4	70,2
36,8	63,0	84,2	36,0	60,8	60,4	63,8	91,3	40,0	69,6
64,3	71,6	41,4	55,4	67,4	64,7	61,8	62,4	70,5	60,4
46,5	63,0	40,0	52,7	70,5	60,8	31,4	48,6	41,4	50,6
62,5	39,7	67,0	81,6	42,3	75,7	44,2	56,8	47,0	82,1
59,3	89,7	50,1	82,2	52,3	43,0	56,0	59,1	103,0	76,0
43,0	56,8	77,8	35,0	52,1	55,3	77,8	54,0	30,8	43,0

Вариант 5

49,6	53,4	40,4	41,4	57,6	58,2	84,2	41,4	75,9	57,7
35,6	20,4	44,4	20,9	45,4	43,4	41,9	54,4	48,4	48,1
74,4	62,4	14,7	54,4	52,9	62,2	77,9	66,4	33,9	45,9
49,3	48,2	49,1	68,4	29,4	31,4	21,8	51,4	69,8	56,6

23,2	49,4	70,6	9,4	47,2	46,8	50,2	77,7	26,4	56,0
50,7	58,0	27,8	41,8	53,8	51,1	48,2	48,8	56,9	46,8
32,9	49,4	26,4	39,1	56,9	47.2	17,8	35,0	27,8	37,0
48,9	46,6	53,4	68,0	28,7	62,1	30,6	43,2	43,4	68,5
45,7	76,1	36,5	68,6	38,7	29,4	42,4	45.5	89,4	62,4
29,4	43,2	64,2	21,4	38,2	41,7	64,2	40,4	17,2	29,4

Вариант 6

48,9	52,7	39,7	40,7	56,9	58,0	83,5	40,7	75,2	57,0
8,7	19,7	43,7	20,2	44,7	42,7	41,2	53,7	47,7	47,4
73,7	61,7	14,0	53,8	45,6	51.7	61,5	77,2	65,7	33,2
48,6	47,5	48,4	67,7	28,7	30,7	21,1	50,7	69,1	55,9
22,5	48,7	69,9	21,7	46,5	49,5	77,0	25,7	55.3	46,1
50,0	57,3	27,1	41,1	53,1	50,4	47,5	48,1	56,2	46,1
32.2	48,7	25,7	38,4	56,2	46,5	17,1	34,3	27,1	36,3
48,2	25,4	52,7	67,3	28,0	61,4	29,9	42,5	42,7	67,8
45,0	75,4	35,8	67,9	38,0	28,7	41,7	44,8	88,7	61,7
28,7	42,7	63,5	20,7	37,5	41,0	63,5	39,7	16,5	28,7

Вариант 7

58,7	62,5	49,5	50,5	66,7	67,8	93,3	50,5	85,0	66,8
18,5	29,5	53,5	30,0	54,5	52,5	51,0	63,5	57,5	57,2
83,5	71,5	23,8	63,5	55,0	61,5	71,3	87,0	75,5	43,0
58,4	57,3	58,2	77,5	38,5	40,5	30,9	60,5	78,9	65,7
32,3	58,5	79,7	31,5	56,3	55,9	59,3	86,8	36,5	65,1
59,8	67,1	36,9	62,9	50,9	60,2	57,3	64,7	66,0	55,9
42,0	58,5	35,5	48,2	66,0	56,3	26,9	44,1	36,9	46,1
58,0	35,2	65,2	77,1	37,8	71,2	39,7	52,3	52,5	77,6

54,8	85,2	45,6	77,7	47,8	38,5	51,5	54,6	98,5	71,5
38,5	52,3	73,3	30,5	47,3	50,8	73,3	49,5	26,3	38,5

Вариант 8

58,2	62,0	49,0	50,0	66,2	67,3	92,8	50,0	84,5	66,3
18,0	29,0	53,0	29,5	54,0	51,8	50,5	63,0	57,0	56,7
90,0	71,0	23,3	63,0	54,5	61,0	70,8	86,5	75,0	42,5
57,9	56,8	57,7	77,0	38,0	40,0	30,4	60,0	78,4	65,2
31,8	58,0	79,2	31,0	55,8	55,4	58,8	86,3	35,0	64,6
59,3	66,6	36,4	50,4	62,4	59,7	56,8	57,4	65,5	55,4
41,5	58,0	35,0	47.7	65,5	55,8	26.4	43,6	36,4	45,6
57,5	34,7	62,0	76,6	37,3	70,7	39,2	51,8	52,0	77,1
54,3	84,7	45,1	77,2	47,2	38,0	51,0	54,1	98,0	71,0
38,0	51,8	72,8	30,0	46,8	50,3	72,8	49,0	25,8	38,0

Вариант 9

46,6	50,4	37,4	38,4	54,6	55,7	81,2	38,4	72,9	54,7
6,4	17.4	41,4	17,9	42,4	40,4	38,9	51,4	45,4	45,1
71,4	59,4	11,7	51,4	42,9	49,4	59,2	74,9	63,4	30,9
46,3	45,2	46,1	65,4	26,4	28,4	18,8	48,4	66,8	53,6
20,2	46,4	67,6	19,4	44,2	43,8	47,2	74,7	23,8	53,0
47,7	55,0	24,8	38,8	50,8	48,1	45,2	45,8	53,9	43,8
29,9	46,4	23,4	36,1	53,9	44,2	14,8	32,0	24,8	34,0
45,9	23,1	50,4	65,0	25,7	59,1	27,6	40,2	40,4	65,5
42,7	73,1	33,5	65,6	35,7	26,4	39,4	42,5	86,4	59,4
26,4	40,2	61,2	18,4	35,2	39,0	61,2	37,4	14,2	26,4

Вариант 10

53,8	57,6	44,6	45,6	61,8	62,9	88,4	45,6	80,1	61,9
13,6	24,6	48,6	25,1	49,6	47,6	46,1	58,6	52,6	52,3
78,6	66,6	18,9	58,6	50,1	56.6	66,4	82,1	70,6	38,1
53,5	52,4	53,3	72,6	33,6	35,6	26,0	55,6	74,0	60,8
27,4	53,6	74,8	26,6	51,4	51,0	54,4	81,9	30,6	60,2
54,9	62,2	32,0	46,0	58,0	55,3	52.4	53,0	61,1	51,1
37,1	53,6	30,6	43,3	61,1	51,4	22.0	39,2	32,0	41,2
53,1	30,3	37,6	72,2	32,9	66,3	34,8	47,4	47,6	72,7
49,9	80,3	40,7	72,8	42,9	33,6	46,6	49,7	93,6	66,6
33,6	47,4	68,4	25,6	42,4	45,9	68,4	44,6	21,4	33,6

Вариант 11

58,0	61,8	48,8	49,8	66,0	67,1	92.6	49,8	84,3	66,1
17,8	28,3	52,8	29,3	53,8	51,8	50,3	62,8	56,8	56,5
82,8	70,8	23,1	62,8	54,3	60,8	70,6	86,3	74,8	42,3
57,7	56,6	57,5	76,8	37,8	39,8	30,2	59,8	78,2	65,0
31,6	57,8	79,0	30,8	55,6	55,2	58,2	86,1	34,8	64,1
59,1	66,4	36,2	50,2	62,2	59,5	56,6	57,2	65,3	55,2
41,3	57,8	34,8	47,5	65,3	55,6	26,2	43,4	36,2	45,4
57,3	34,5	61,8	76,4	37,1	70,5	39,0	51,6	51,8	76,9
54,1	84,5	44,9	77,0	47,1	37,8	50,8	53,9	97,8	70,8
37,8	51,6	72,6	29,8	46,6	50,1	72,6	48,8	25,6	37,8

Вариант 12

47,0	50,8	37,8	38,8	55,0	56,1	81,6	38,8	73,3	55,1
6,8	17,8	41,8	18,3	42,8	40,8	39,3	51,8	45,8	45,5
71,8	59,8	12,1	51,8	43,3	49,8	59,6	75,3	63,8	31,3
46,7	45,6	46,5	65,8	26,8	28,8	19,2	48,8	67,2	54,0

20,6	46,8	68,0	19,8	44,6	44,2	47,6	75,1	23,8	53,4
48,1	55,4	25,2	39,2	51,2	48,5	45,6	46,2	54,3	44,2
30,3	46,8	23,8	36,5	44,6	54,3	15.2	32,4	25,2	34,4
46,3	23,5	50,8	65,4	26,1	59,5	28,0	40,6	40,8	65,9
43,1	73,5	33,9	66,0	36,1	26,8	39,8	42,9	86,8	59,8
26,8	40,6	61,6	18,8	35,6	39.1	61,6	37,8	14,6	26,8

Вариант 13

54,7	72,9	38,4	46,6	50,4	37,4	38,4	54,6	55,7	81,2
45,1	45,4	51,4	6,4	17,4	41,4	17,9	42,4	40,4	38,9
30,9	63,4	74,9	71,4	59,4	11,7	51,4	42,9	49,4	59,2
53,6	66,8	48,4	46,3	45,2	46,1	65,4	26,4	28,4	18,8
53,0	23,8	74,7	20,2	46,4	67,6	19,4	44,2	43,8	47,2
43,8	53,9	45,8	47,7	55,0	24,8	38,8	50,8	48,1	45,2
34,0	24,	32,0	29,9	46,4	23,4	36,1	53,9	44,2	14,8
65,5	40,4	40,2	45,9	23,1	50,4	65,0	25,7	59,1	27,6
59,4	86,4	42,5	42,7	73,1	33,5	65,6	35,7	26,4	39,4
26,4	14,2	37,4	26,4	40,2	61,2	18,4	35,2	39,0	61,2

Вариант 14

45,0	62,2	63,3	88,8	54,2	46,0	46,0	80,5	58,0	62,3
49,0	50,0	48,0	46,5	14,0	25,5	59,0	53,0	25,0	52,7
19.3	50,5	57,0	66,8	79,0	59,0	82,5	71,0	67,0	38,5
53,7	34,0	36,0	26,4	53,9	73,0	56,0	74,4	52,8	61,2
75,2	51,8	51,4	,54,8	27,8	27,0	82,3	31,0	54,0	60,6
32,4	58,4	55,7	52,8	55,3	46,4	53,4	61,5	62,6	51,4
31,0	61,5	51,8	22,4	37,5	43,7	39,6	32,4	54,0	41,6
58,0	33,3	66,7	35,2	53,5	72,6	47,8	48,0	30,7	73,1
41,1	43,3	34,0	47,0	50,3	73,2	50,1 ,	94,0	80,7	67,0

ПРИЛОЖЕНИЕ 2

ДИАМЕТРЫ 200 ВАЛОВ ПРОТОЧЕННЫХ НА СТАНКЕ, ММ.

d1	**d2**	**d1**	**d2**	**d1**	**d2**	**d1**	**d2**
13.39	13.33	13.56	13.38	13.43	13.37	13.53	13.40
13.28	13.34	13.50	13.38	13.38	13.45	13.47	13.62
13.53	13.58	13.32	13.27	13.42	13.40	13.57	13.46
13.57	13.36	13.43	13.38	13.26	13.52	13.35	13.29
13.40	13.39	13.50	13.52	13.39	13.39	13.46	13.29
13.29	13.33	13.38	13.61	13.55	13.40	13.20	13.31
13.43	13.51	13.50	13.38	13.44	13.62	13.42	13.54
13.41	13.49	13.42	13.45	13.34	13.47	13.48	13.59
13.55	13.44	13.50	13.40	13.48	13.29	13.31	13.42
13.43	13.26	13.58	13.38	13.48	13.45	13.29	13.32
13.34	13.14	13.31	13.51	13.59	13.32	13.52	13.57
13.23	13.37	13.64	13.30	13.40	13.58	13.24	13.32
13.43	13.58	13.63	13.48	13.34	13.37	13.18	13.50
13.38	13.33	13.57	13.28	13.32	13.40	13.40	13.33
13.34	13.54	13.40	13.47	13.28	13.41	13.39	13.48
13.28	13.46	13.37	13.53	13.43	13.30	13.45	13.40
13.33	13.39	13.56	13.46	13.26	13.35	13.42	13.36
13.43	13.51	13.51	13.24	13.34	13.28	13.37	13.54
13.52	13.23	13.48	13.48	13.54	13.41	13.51	13.44
13.53	13.44	13.69	13.66	13.32	13.26	13.51	13.38

ПРИЛОЖЕНИЕ 3

Таблица П3.1

Распределение Пуассона $P[X = k] = \frac{\lambda^k}{k!} e^{-\lambda}$

.

λ / k	0,1	0,2	0,3	0,4	0,5	0,6	0,7
0	0,90484	0,81873	0,74082	0,67032	0,60653	0,54881	049659
1	0,09048	0,16375	0,22223	0,26813	0,30327	0,32929	0,34761
2	0,00452	0,01638	0,03334	0,05363	0,07582	0,09879	0,12166
3	0,00015	0,00109	0,00333	0,00715	0,01204	0,01976	0,02839
4		0,00006	0,00025	0,00072	0,00158	0,00296	0,00497
5			0,00002	0,00006	0,00016	0,00036	0,00070
6					0,00001	0,00004	0,00008
7							0,00001
	0,8	0,9	1,0	2,0	3,0	4,0	5,0
0	0,44933	0,40657	0,36788	0,13534	0,04979	0,01832	0,00674
1	0,35946	0,36591	0,36788	0,27067	0,14936	0,07326	0,03369
2	0,14379	0,16466	0,18394	0,27067	0,22404	0,14653	0,08422
3	0,03834	0,04940	0,06131	0,18045	0,22404	0,19537	0,14037
4	0,00767	0,01112	0,01533	0,09022	0,16803	0,19537	0,17547
5	0,00123	0,00200	0,00307	0,03609	0,10082	0,15629	0,17547
6	0,00016	0,00030	0,00051	0,01203	0,05041	0,10419	0,14622
7	0,00002	0,00004	0,00007	0,00344	0,02160	0,05954	0,10445
8			0,00001	0,00086	0,00810	0,02977	0,06528
9				0,00019	0,00270	0,01323	0,03627
10				0,00004	0,00081	0,00529	0,01813
11				0,00001	0,00022	0,00193	0,00824
12					0,00006	0,00064	0,00343
13					0,00001	0,00020	0,00132
14						0,00006	0,00047
15						0,00002	0,00016
16							0,00005
17							0,00001

Таблица П3.2

Функция распределения Φ(x) нормального закона N(0,1)

$$\Phi(-x) = 1 - \Phi(x),\ \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$

X	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	05948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	07967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8261
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	09345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	09834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	09971	09972	09973	0,9974
2,8	0,9974	0,9975	09976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3,0	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998

Квантили u_p нормального распределения N(0,1):

p	0.90	0.95	0.975	0.99	0.995	0.999	0.9995
u_p	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Таблица П3.3

Функция Лапласа $\Phi(z) = \frac{2}{\sqrt{2\pi}} \int_0^z e^{-\frac{x^2}{2}} dx$

z	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
0,	0	797	1585	2358	3108	3829	4515	5161	5763	6319
1,,	6827	7287	7699	8064	8385	8664	8904	9109	9231	9426
2,	9545	9643	9722	9786	9836	9876	9907	9931	9949	9963
3,	9973	9981	9986	9990	9993	9995	9997	9998	9999	9999

В таблице приведены значения Ф (z) • 10^4. В первом столбце указаны целые, а в верхней строке — десятые доли аргумента z.

Приведем также некоторые значения z, отвечающие круглым значениям функции Ф(z),

Ф(z)	0,80	0,90	0,95	0,99	0,999
z	1,2816	1,6449	1,9600	2,5758	3,2905

Критические точки распределения Стьюдента Таблица П3.4

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0, 10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,06	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05,	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	0,025	0,01	0,005	0,001	0,0005
	Уровень значимости α (односторонняя критическая область)					

Таблица П3.5

Критические точки распределения F Фишера—Снедекора

k_1 — число степеней свободы большей дисперсии,

k_2 —число степеней свободы меньшей дисперсии

Уровень значимости α =0,01												
k_2	k_1											
	1	2	3	4	5	6	7	8	9	10	11	12
1	405	499	540	562	576	588	592	598	602	605	608	610
2	98,4	99,	99,	99,2	99,3	99,3	99,3	99,3	99,3	99,4	99,4	99,4
3	34,1	30,	29,	28,7	28,2	27,9	27,6	27,4	27,3	27,2	27,1	27,0
4	21,2	18,	16,	15,9	15,5	15,2	14,9	14,8	14,6	14,5	14,4	14,3
5	16,2	13,	12,	11,3	10,9	10,6	10,4	10,2	10,1	10,0	9,96	9,89
6	13,7	10,	9,7	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72
7	12,2	9,5	8,4	7,85	7,46	7,19	7,00	6,84	6,71	6,62	6,54	6,47
8	11,2	8,6	7,5	7,01	6,63	6,37	6,19	6,03	5,91	5,82	5,74	5,67
9	10,5	8,0	6,9	6,42	6,06	5,80	5,62	5,47	5,35	5,26	5,18	5,11
10	10,0	7,56	6,5	5,99	5,64	5,39	5,21	5,06	4,95	4,85	4,78	4,71
11	9,86	7,2	6,2	5,67	5,32	5,07	4,88	4,74	4,63	4,54	4,46	4,40
12	9,33	6,9	5,9	5,41	5,06	4,82	4,65	4,50	4,39	4,30	4,22	4,16
13	9,07	6,7	5,7	5,20	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96
14	8,86	6,51	5,5	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80
15	8,68	6,3	5,4	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67
16	8,5	6,2	5,2	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61	3,55
17	8,4	6,1	5,1	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,45

Продолжение табл.П3. 5

Уровень значимости α =0,05												
k_2	k_1											
		2	3	4	5	6	7	8	9	10	11	12
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18.5	19.0	19.1	19.2	19.3	19.3	19.3	19.3	19.3	19.3	19.4	19.4
3	10.	9.55	9.28	9.12	9.01	8.94	8.88	8.84	8.81	8.78	8.76	8.74
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.0	5.96	5.93	5.91
5	6.6	5.7	5.41	5.19	5.05	4.95	4.8	4.82	4.78	4.74	4.70	4.68
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.03	4.00
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.63	3.60	3.57
8	5.32	4.46	4.07	3.84	3.6	3.58	3.50	3.44	3.39	3.34	3.31	3.28
9	5.12	4.26	3.86	3.63	3.4	3.37	3.29	3.23	3.18	3.13	3.10	3.07
10	4.96	4.10	3.71	3.48	3.3	3.22	3.14	3.07	3.02	2.97	2.94	2.91
11	4.84	3.98	3.59	3.36	3.2	3.09	3.01	2.95	2.90	2.86	2.82	2.79
12	4.75	3.88	3.49	3.26	3.11	3.00	2.92	2.85	2.8	2.76	2.72	2.69
13	4.67	3.80	3.41	3.18	3.0	2.92	2.84	2.77	2.72	2.67	2.63	2.60
14	4.60	3.74	3.34	3.11	2.9	2.85	2.77	2.70	2.6	2.60	2.56	2.53
15	4.54	3.68	3.29	3.06	2.9	2.79	2.7	2.64	2.5	2.55	2.51	2.48
16	4.49	3.6	3.24	3.01	2.8	2.74	2.66	2.59	2.54	2.49	2.45	2.42
17	4.45	3.59	3.20	2.96	2.81	2.70	2.62	2.55	2.5	2.4	2.41	2.38

Таблица П3.6

Критические точки распределения Колмогорова

α	0,10	0,05	0,01	0,001
$\lambda_{кр}$	1,224	1,358	1,627	1,950

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Андронов А.М. Теория вероятностей и математическая статистика: учебник для вузов / А.М.Андронов, Е.А. Копытов, Л.Я Гринглаз. – СПб.: Питер, 2004. – 461 с

2. Айвазян С.А. Прикладная статистика: Исследование зависимостей / С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1985. – 471 с.

3.Боровиков В. STATISTICA: искусство анализа данных на компьютере / В. Боровиков. – СПб.: Питер, 2001. – 656 с.

4. Бочаров П.П. Теория вероятностей. Математическая статистика / П.П. Бочаров, А.В. Печинкин. – М.: Гардарика, 1998. – 328 с.

5. Вентцель Е.С. Теория вероятностей: учебник для вузов / Е.С. Вентцель М.: Высш. шк., 1999. – 576 с.

6. Гнеденко Б.В. Курс теории вероятностей / Б.В. Гнеденко. – М.: Эдиториал УРСС, 2001. – 320 с.

7. Гмурман В.Е.. Теория вероятностей и математическая статистика / В.Е. Гмурман – М.: Высш. шк., 2010. – 479 с.

8. Дрейпер Н. Прикладной регрессионный анализ, кн.1 / Н. Дрейпер, Г. Смит. – М.: Финансы и статистика, 1986. – 366 с.

9. Ивченко Г. И. Математическая статистика / Г. И. Ивченко, Ю. И. Медведев. – М.: Высш. шк., 2003. – 357 с.

10. Лагутин М.Б. Наглядная математическая статистика / М.Б Лагутин. – М.:,БИНОМ. Лаборатория знаний.2007. – 472 с.

11. Математическая статистика: учебник для вузов / под ред. В.С.Зарубина, А.П.Крищенко. – М.: Изд-во МГТУ им. Н.Э. Баумана, 2001. – 424 с.

12. Новикова Н.М. Обработка экспериментальных данных: учеб. пособие / Н.М. Новикова. 2-е изд., ВГТУ, 2010. . – 119 с.

13. Орлов А.И. Прикладная статистика: учебник для вузов / А.И.Орлов. – М.: Экзамен, 2004. – 656 с.

14 Сборник задач по математике для втузов. Ч.4. Теория вероятностей и математическая статистика: учеб. пособие / под ред. А.В.Ефимова, А.С.Поспелова. – М.: Изд-во Физ.-мат. лит-ра, 2003. – 432 с.

ОГЛАВЛЕНИЕ