МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Лифшиц М.А.

22 января 2010 г.

1. Базисные понятия теории вероятности. Случайная величина, математическое ожидание, дисперсия, функция распределения. Основные типы распределений (дискретные, непрерывные). Распределение Бернулли. Распределение Пуассона. Равномерное распределение. Нормальное распределение. Независимость величин. Коэффициент корреляции как мера зависимости. Закон больших чисел. Центральная предельная теорема.

2. Базисные понятия математической статистики. Статистическая модель - выборка наблюдений, класс возможных распределений. Типы решаемых задач: оценка параметров распределения данных, проверка гипотез о типе распределения данных.

3. Оценка параметров распределения. Статистика как функция наблюдений. Состоятельность. Несмещенность. Примеры : - оценивание математического ожидания эмпирическим средним;

- оценивание параметра распределения Бернулли;

- оценивание границ интервала по наблюдениям, равномерно распределенным в интервале с неизвестными границами.

- эмпирическая дисперсия как оценка дисперсии распределения наблюдений.

Эмпирическое распределение и его характеристики (функция распределения, математическое ожидание и дисперсия).

Оценки максимального правдоподобия

а) дискретные распределения (Бернулли, Пуассона);

б) непрерывные распределения (нормальный закон).

Интервальное оценивание параметров распределения. Доверительный интервал.

Построение доверительного интервала для математического ожидания нормальной выборки.

а) при известной дисперсии, по таблицам нормального распределения.

б) при неизвестной дисперсии, по таблицам распределения Стьюдента.

Построение доверительного интервала для дисперсии нормальной выборки.

Построение доверительного интервала для параметра распределения Бернулли.

4. Проверка гипотез. Общие понятия, связанные с проверкой гипотез - гипотеза, альтернатива, принятие и отклонение гипотезы, решающее правило, ошибки первого и второго рода.

Проверка гипотезы о соответствии распределения элементов выборки заданному распределению:

а) Дискретное распределение - критерий хи-квадрат. Числовой пример. Таблица распределения критерия хи-квадрат.

б) Непрерывное распределение - критерий Колмогорова, основанный на сравнении функций распределения числовой пример. Таблица распределения критерия Колмогорова.

Применение критерия хи-квадрат к проверке однородности двух выборок.

Применение критерия Вилкоксона к проверке однородности двух выборок.

Применение критерия хи-квадрат к проверке независимости двух величин.

5. Проверка наличия связи между двумя величинами.

Проверка гипотезы о наличии линейной связи между двумя нормальными выборками (выборочный коэффициент корреляции).

Проверка гипотезы о наличии связи между двумя нечисловыми выборками (ранговый коэффициент корреляции Спирмана).

6. Линейная регрессия. Постановка задачи линейной регрессии. Метод наименьших квадратов. Решение задачи линейной регрессии в матричной форме. Решение простейших задач регрессии с одной и двумя объясняющими величинами.

Вероятностная трактовка задачи линейной регрессии. Эффективные оценки скалярных и векторных параметров. Теорема Гаусса–Маркова об

эффективности оценки метода наименьших квадратов. Оценка точности решения задачи регрессии.

1 Базисные понятия математической статистики

Энциклопедия теории вероятностей и математической статистики определяет математическую статистику как раздел математики, посвященный математическим методам сбора, систематизации, обработки и интерпретации статистических данных, а также их использованию для научных и практических выводов.

1.1 Основная задача математической статистики

Имеется некоторое неизвестное распределение, числовые параметры или качественные свойства которого требуется определить. Исходными данными для решения этой задачи служат

1. *выборка*, т.е. набор значений независимых с.в., имеющих данное (неизвестное нам) распределение; количество имеющихся значений называется *объемом выборки*;

2. информация о принадлежности изучаемого распределения к определенному семейству распределений (ее может и не быть).

Пример 1 . Задача о контролерах. При многократных поездках на каком-либо виде транспорта, например, электричке, практический интерес представляет оценка вероятности появления контролеров. В этом случае простейшая модель выглядит следующим образом. Считаем появления контролеров в различных поездках независимыми событиями, имеющими одну и ту же вероятность p. Свяжем с каждым из этих событий случайную величину X_j, равную единице, если контроль во время j-й поездки произошел, и нулю в противном случае. Тогда выборка может иметь следующий вид.

j	1	2	3	4	5	6	7	8	9	10
X_j	0	0	1	0	0	0	1	0	0	1

Мы предполагаем, что наблюдаемые величины следуют распределению Бернулли $B(p)$ с неизвестным параметром p.

Пример 2 . Задача о наработке на отказ. Требуется оценить характеристики наработки на отказ (времени работы до первого отказа) подшипников по данным испытаний 9 образцов. Выборка может иметь следующий вид

Номер испытания(j)	1	2	3	4	5	6	7	8	9
Наработка на отказ (X_j), час	375	377	438	472	480	580	638	690	728

Предполагается, что распределение времени наработки на отказ имеет плотность

$$p(x) = a^{-1}e^{-(x-b)/a}, \qquad x > b.$$

Здесь b – гарантированное время наработки. По данным выборки требуется оценить параметры a и b. Например, проверить гипотезу $b > 0$, т.е. наличие гарантированного времени наработки на отказ.

Пример 3 . Дорожно-транспортные происшествия в Приморском районе. Количество зарегистрированных происшествий может быть представлено следующей выборкой

День (j)	1	2	3	4	5	6	7	8	9	10	11	12
Кол-во происшествий (X_j)	3	6	5	4	3	0	2	1	3	4	7	2

Требуется проверить, принадлежит ли распределение числа ДТП к семейству распределений Пуассона $\mathcal{P}(a)$ и если принадлежит, то оценить параметр a.

Пример 4 . Задача Грегора Менделя. Рассматривается цвет горошин двухцветного горошка (он может быть желтым или зеленым). Согласно закону Менделя при определенных условиях наблюдения доля (вероятность) желтых горошин должна составить 1/4. По наблюдениям цвета горошин требуется проверить справедливость закона Менделя.

1.2 Два типа решаемых задач

Мы будем рассматривать два основных типа задач математической статистики:

1. оценка параметров распределения;
2. проверка гипотез о виде распределения.

К первому типу относится задача о контролерах, задача о наработке на отказ, вопрос об оценке параметра в задаче о ДТП. Более общая задача первого типа - по имеющейся выборке $X_1, \dots X_n$ оценить $\mathbf{E}X$ и $\mathbf{D}X$ (без предположений о виде распределения).

Ко второму типу задач относится задача о ДТП и задача Менделя. Более общие примеры задач второго типа:

1. Проверить гипотезу о том, что неизвестное распределение элементов выборки совпадает с заданным распределением;

2. В выборке представлены одновременные наблюдения двух величин $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$. Проверить гипотезу о том, что величины X и Y независимы;

3. По имеющимся двум выборкам $X_1, \dots X_n$ и $Y_1, \dots, Y_m$ проверить гипотезу о том, что они отвечают одному и тому же распределению;

4. Проверить, что параметры неизвестного распределения принадлежат определенной области.

Пример 5 . Игральный автомат. Пусть плата за одну партию с автоматом составляет A, размер выигрыша V, вероятность выигрыша p. Игра будет в среднем прибыльна для игрока, если $pV > A$. Если же $pV < A$, то в среднем прибыль будет получать владелец автомата. Задача: по результатам сыгранных партий (их можно закодировать единицами и нулями как в задаче о контролерах) оценить, какое из двух неравенств будет иметь место.

1.3 Асимптотический характер статистических решений

Располагая выборкой конечного объема, задачу математической статистики никогда нельзя решить точно. Можно лишь строить приближенные оценки и решения, которые при большом объеме выборки будут давать хорошее решение. Поэтому важной частью статистического исследования является оценка точности его выводов и результатов.

2 Оценка параметров

В задаче оценки параметров исходными данными является выборка наблюдений. Предполагается, что они независимы и имеют одинаковое (неизвестное) распределение, принадлежащее к определенному семейству, зависящему от параметра (или нескольких параметров). Значение параметра требуется уточнить.

В задаче о контролерах это было семейство распределений Бернулли с параметром p. В задаче о наработке на отказ - семейство сдвинутых показательных распределений, зависящее от параметров a, b.

В зависимости от способа представления решения задачи оценки параметров выделяют *точечные оценки* и *интервальные оценки*.

2.1 Точечные оценки

2.1.1 Общие положения

В случае построения точечной оценки параметра решением задачи является *статистическая оценка* (для краткости часто называемая просто *статистикой*). Статистическая оценка - это функция, которая сопоставляет выборочным значениям оценочное значение параметра

$$T_n : (X_1, X_2, \ldots, X_n) \to T_n(X_1, X_2, \ldots, X_n).$$

Оценка опирается исключительно на данные выборки. Она всегда дает лишь приближенное значение, поэтому нужно уметь отличать хорошие оценки от плохих. Кроме того оценка должна быть достаточно хорошей при всех возможных значениях параметра.

Хорошие оценки обладают двумя важными достоинствами – *несмещенностью* и *состоятельностью*.

Оценка T_n называется *несмещенной*, если ее математическое ожидание равно значению оцениваемого параметра при всех возможных значениях этого параметра. Математическая запись несмещенности

$$\mathbf{E}_\theta T_n(X_1, \ldots, X_n) = \theta.$$

Здесь запись $\mathbf{E}_\theta$ означает, что математическое ожидание вычисляется в предположении, что истинное значение параметра равно θ.

Последовательность оценок $T_n,\ n \to \infty$, называется *состоятельной*, если вероятность отклонения T_n от истинного значения оцениваемого параметра, превышающего любую заданную величину, стремится к нулю с ростом числа наблюдений при всех возможных значениях этого параметра. Математическая запись состоятельности

$$\mathbf{P}_\theta \{|T_n(X_1, \ldots, X_n) - \theta| > r\} \to 0, \qquad n \to \infty.$$

Запись $\mathbf{P}_\theta$ означает, что вероятность вычисляется в предположении, что истинное значение параметра равно θ.

2.1.2 Оценка математического ожидания

Постановка задачи - дана выборка $X_1, \ldots, X_n$ величин, имеющих одинаковое, но неизвестное распределение. Требуется оценить математическое ожидание величин с таким распределением. Решением задачи служит статистическая оценка

$$T_n(X_1, \ldots, X_n) = \frac{X_1 + \cdots + X_n}{n} = \bar{X}.$$

Она называется *эмпирическим средним* или *выборочным средним*. В отличие от более специальных постановок из примеров, здесь нет никакого предположения о виде распределения с.в. X_j.

Утверждение. *Оценка $\bar{X}$ является несмещенной и состоятельной.*

Несмещенность проверяется прямым вычислением. Состоятельность следует из закона больших чисел.

Примеры применения к конкретным ситуациям - оценка параметра p распределения Бернулли, оценка параметра a показательного распределения, оценка параметра a пуассоновского распределения, оценка параметра a нормального распределения.

Пример состоятельной, но смещенной оценки. Пусть $\{X_j\}$ выборка из независимых наблюдений с.в., равномерно распределенных на интервале $[a, a+1]$ с неизвестным параметром a. Для оценки параметра a можно выбрать оценку

$$T_n = \min\{X_1, X_2, \ldots, X_n\}.$$

Она будет смещенной, т.к. $T_n > a$ с вероятностью 1. С другой стороны, она будет состоятельной, т.к.

$$P\{|T_n - a| > r\} = \mathbf{P}\{T_n > a + r\} = \mathbf{P}\{X_1 > a + r, X_2 > a + r \dots X_n > a + r\} =$$

$$(1 - \mathbf{P}(X_1 \le a + r))^n = (1 - r)^n \to 0, \qquad r < 1.$$

Эта оценка все же хорошая, т.к. она *асимптотически* несмещенная. Можно проверить, что

$$\mathbf{E}_a T_n(X_1, \dots, X_n) = a + \frac{1}{n+1} \ .$$

т.е. оценка становится "почти" несмещенной с ростом n. Пример несмещенной оценки в той же задаче $T_n(X_1, \dots, X_n) = \bar{X} - 1/2$.

Преимущество первой оценки состоит в том, что она не использует информацию о длине интервала.

Упражнение. Для выборки из равномерного распределения $U(a, b)$ с неизвестными a, b длину интервала можно оценить через

$$T_n = \max\{X_1, \dots, X_n\} - \min\{X_1, \dots, X_n\}.$$

Проверить состоятельность и асимптотическую несмещенность этой оценки.

2.1.3 Оценка дисперсии

Постановка задачи - дана выборка $X_1, \dots, X_n$ величин, имеющих одинаковое, но неизвестное распределение. Требуется оценить дисперсию величин с таким распределением.

По определению, $\mathbf{D}X = \mathbf{E}X^2 - (\mathbf{E}X)^2 = \mathbf{E}(X - \mathbf{E}X)^2$. Заменяя знак $\mathbf{E}$ на выборочное среднее, получим оценку

$$s_n^2(X_1, \dots, X_n) = \frac{1}{n}\sum_{j=1}^{n} X_j^2 - \left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)^2 ,$$

называемую *эмпирической дисперсией*. Посмотрим, будет ли она несмещенной. Имеем

$$\begin{aligned}
\mathbf{E}s_n^2 &= \frac{1}{n}\sum_{j=1}^{n}\mathbf{E}X_j^2 - \frac{1}{n^2}\sum_{j=1}^{n}\sum_{i=1}^{n}\mathbf{E}X_iX_j \\
&= \mathbf{E}X^2 - \frac{1}{n^2}\left(\sum_{j=1}^{n}\mathbf{E}X_j^2 + \sum_{i\neq j}\mathbf{E}X_iX_j\right) \\
&= (1-\frac{1}{n})\mathbf{E}X^2 - \frac{n(n-1)}{n^2}(\mathbf{E}X)^2 \\
&= (1-\frac{1}{n})(\mathbf{E}X^2 - (\mathbf{E}X)^2) = (1-\frac{1}{n})\mathbf{D}X.
\end{aligned}$$

Поэтому оценка s_n будет несколько смещенной, а несмещенной будет оценка

$$\begin{aligned}
\bar{s}_n^2 &= \frac{n}{n-1}s_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}X_j^2 - \frac{1}{n(n-1)}\left(\sum_{j=1}^{n}X_j\right) \qquad (1)\\
&= \frac{1}{n-1}\left(\sum_{j=1}^{n}X_j^2 - n\bar{X}^2\right) = \frac{1}{n-1}\sum_{j=1}^{n}\left(X_j-\bar{X}\right)^2 \\
&= \frac{1}{n-1}\sum_{j=1}^{n}X_j^2 - \frac{n}{n-1}\bar{X}^2.
\end{aligned}$$

Обе эти оценки будут состоятельными.

2.2 Интервальные оценки

2.2.1 Общие положения

Пытаясь оценить параметр a, мы даем не приблизительное значение, а указываем целый интервал $[a_1, a_2]$, в котором с большой вероятностью находится значение нашего параметра. Здесь a_1, a_2 - статистические оценки, определяемые по выборке. Интервал $[a_1, a_2]$ называется *доверительным интервалом*, а вероятность попадания в него параметра $\mathbf{P}\{a \in [a_1, a_2]\}$ – *доверительной вероятностью*.

Чем больше интервал, тем выше доверительная вероятность, но менее точна оценка. Таким образом, требования точности и надежности

взаимно противоречивы. Поэтому сначала задают требуемый уровень (порог) доверительной вероятности γ, выбирая его из набора чисел 0.95, 0.99, 0.999. Затем строят как можно более узкий доверительный интервал $[a_1, a_2]$, удовлетворяющий условию

$$\mathbf{P}\{a \in [a_1, a_2]\} \geq \gamma$$

при всех допустимых распределениях элементов выборки.

2.2.2 Нормальное распределение - оценка a при известном σ

Пусть допустимыми распределениями элементов выборки являются нормальные $\{\mathcal{N}(a, \sigma^2)\}$. Предположим, что σ известно, и построим доверительный интервал для неизвестного параметра a. Будем строить его в виде $[a_1(X), a_2(X)] = [\bar{X} - r, \bar{X} + r]$. Тогда нужно решить уравнение

$$\begin{aligned} \gamma &= \mathbf{P}\{a \in [\bar{X} - r, \bar{X} + r]\} = \mathbf{P}\{\bar{X} - r \leq a \leq \bar{X} + r]\} \\ &= \mathbf{P}\{-r \leq a - \bar{X} \leq r\} = \mathbf{P}\{|a - \bar{X}| \leq r\}. \end{aligned}$$

Заметим, что $\bar{X}$ имеет распределение $\mathcal{N}(a, \frac{\sigma^2}{n})$, соответственно $\bar{X} - a$ – $\mathcal{N}(0, \frac{\sigma^2}{n})$. По свойствам нормального закона, в том числе используя формулу $\mathbf{P}\{|Y| \leq t\} = 2\Phi(t) - 1$, получаем

$$\mathbf{P}\{|a - \bar{X}| \leq r\} = \mathbf{P}\{\frac{\sigma}{\sqrt{n}}|Y| \leq r\} = \mathbf{P}\{|Y| \leq \frac{\sqrt{n}r}{\sigma}\} = 2\Phi\left(\frac{\sqrt{n}r}{\sigma}\right) - 1,$$

где Y величина со стандартным нормальным распределением. Обозначим $t = \frac{\sqrt{n}r}{\sigma}$. Уравнение примет вид

$$2\Phi(t) - 1 = \gamma \quad \Leftrightarrow \quad \Phi(t) = \frac{1 + \gamma}{2}.$$

Величину t, удовлетворяющую уравнению такого типа, называют квантилью нормального закона. Ее можно найти по таблице

Таблица 1 *Квантили нормального закона*

γ	$\frac{1+\gamma}{2}$	t
0.9	*0.95*	*1.645*
0.95	*0.975*	*1.96*
0.99	*0.995*	*2.58*
0.999	*0.9995*	*3.30*

Окончательно находим

$$r = \frac{\sigma t}{\sqrt{n}} \tag{2}$$

и доверительный интервал имеет вид

$$[a_1, a_2] = [\bar{X} - \frac{\sigma t}{\sqrt{n}}, \bar{X} + \frac{\sigma t}{\sqrt{n}}].$$

Таким образом, ширина интервала обратно пропорциональна *корню* из числа наблюдений и прямо пропорциональна среднеквадратическому отклонению.

Замечание. Можно, наоборот, найти минимальный объем выборки, при котором достигается заданная ширина доверительного интервала. Пусть заданы r, γ. Из уравнения (2) найдем

$$n = \frac{\sigma^2 t^2}{r^2}.$$

Пример 6 . Построить доверительный интервал для параметра a при известном $\sigma = 1$ и доверительной вероятности $\gamma = 0.95$, исходя из следующей выборки.

j	1	2	3	4	5	6	7
X_j	0.464	0.137	2.45	-0.323	-0.068	0.296	-0.288

Решение. По таблице находим $t = 1.96$. Из выборки находим $\bar{X} = \frac{2.668}{7} = 0.38$. Вычисляем $r = \frac{t\sigma}{\sqrt{n}} = \frac{1.96 \cdot 1}{2.64} = 0.74$. Доверительный интервал:

$$[\bar{X} - r, \bar{X} + r] = [-0.36, 1.12].$$

Пример 7 . Многолетние наблюдения за осадками в г.Страсбург показали, что количество осадков выпадающих за год - нормальная случайная величина со среднеквадратическим отклонением 100 мм/год. По данным следующих наблюдений построить доверительный интервал для среднего количества осадков, отвечающий доверительному уровню $\gamma = 0.95$.

Год	1991	1992	1993	1994	1995	1996	1997	1998
Осадки (мм)	510	614	780	512	501	534	603	788

Решение. По таблице находим $t = 1.96$. Из выборки находим $\bar{X} = \frac{4842}{8} = 605$. Вычисляем $r = \frac{t\sigma}{\sqrt{n}} = \frac{1.96 \cdot 100}{2.828} = 69.3$. Доверительный интервал:

$$[\bar{X} - r, \bar{X} + r] = [535.7, 674.3].$$

2.2.3 Нормальное распределение - оценка a при неизвестном σ

Решается та же задача, что и раньше, по-прежнему предполагается, что выборка взята из нормального распределения, но теперь мы считаем, что оба параметра a, σ неизвестны. Снова ищем доверительный интервал в виде $[\bar{X}-r, \bar{X}+r]$. Но если раньше мы использовали формулу $r = \frac{\sigma t}{\sqrt{n}}$, где σ было известно, то теперь мы вынуждены пользоваться оценкой параметра σ, взятой из формулы (1). Соответственно, и квантиль t приходится находить по новой таблице, с помощью т.н. *распределения Стьюдента*[1] . В отличие от предыдущего случая, квантиль t зависит от объема выборки n. Впрочем, при больших n квантили распределения Стьюдента близки к квантилям нормального закона из таблицы 1 , поскольку в этом случае $\bar{s}$ близко к σ.

Таблица 2 *Квантили распределения Стьюдента*

γ	$\frac{1+\gamma}{2}$	t	t	t	t
		$n=5$	$n=10$	$n=15$	$n=100$
0.95	*0.975*	*2.78*	*2.26*	*2.15*	*1.98*
0.99	*0.995*	*4.60*	*3.25*	*2.98*	*2.63*
0.999	*0.9995*	*8.61*	*4.78*	*4.14*	*3.39*

Таким образом, доверительный интервал имеет вид

$$[\bar{X} - \frac{\bar{s}t}{\sqrt{n}}, \bar{X} + \frac{\bar{s}t}{\sqrt{n}}].$$

Пример 8 .Рост стоимости акций нефтяных компаний составил (в %)

Недели	1	2	3	4	5
Рост(%)	2.3	4.1	0.8	6.4	-2.3

Считая распределение данного показателя нормальным, оценить среднее значение и дисперсию. Построить доверительный интервал с доверительной вероятностью 0.95.

Решение. Находим оценки параметров

$$\bar{X} = \frac{11.3}{5} = 2.26,$$

$$\bar{s}^2 = \frac{1}{n-1}\sum_{j=1}^{n} X_j^2 - \frac{n}{n-1}\bar{X}^2 = \frac{68.99}{4} - \frac{5}{4}2.26^2 = 17.24 - 6.38 = 10.85,$$

[1] W.Gosset

$$\bar{s} = \sqrt{10.85} = 3.29.$$

По таблице $t = 2.78$. Отсюда

$$\frac{\bar{s}t}{\sqrt{n}} = \frac{3.29 \cdot 2.78}{\sqrt{5}} = \frac{9.1462}{2.236} = 4.09.$$

Доверительный интервал

$$[\bar{X} - \frac{\bar{s}t}{\sqrt{n}}, \bar{X} + \frac{\bar{s}t}{\sqrt{n}}] = [2.26 - 4.09, 2.26 + 4.09] = [-1.83, 6.35].$$

Видно, что доверительный интервал слишком большой (включает положительные и отрицательные значения) из-за того, что мало данных.

2.2.4 Распределение Бернулли - оценка p

Рассматривается выборка, состоящая из нулей и единиц (обычно это связано с наступлением или ненаступлением определенного события). Можно считать, что величины в такой выборке следуют распределению Бернулли и оценивать по выборке неизвестный параметр. При построении доверительного интервала будем исходить из центральной предельной теоремы, которая утверждает асимптотическую нормальность сумм бернуллиевских величин. Поэтому (при достаточно большом числе наблюдений) можно строить доверительный интервал вида $[\bar{X} - r, \bar{X} + r]$, а в качестве ширины интервала использовать

$$r = \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\ t.$$

Здесь t квантиль нормального закона (из таблицы 1), а величина $\bar{X}(1-\bar{X})$ является оценкой дисперсии, истинное значение которой $p(1-p)$. Таким образом доверительный интервал имеет вид

$$I = \left[\bar{X} - \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\ t, \bar{X} + \sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\ t\right].$$

Пример 9 . Предвыборный опрос. В ходе предвыборного опроса в поддержку определенного кандидата высказались 56% опрошенных. Оценить шансы этого кандидата с помощью построения доверительного интервала с доверительным уровнем вероятности 0.95 при объеме выборки 50, 500, 5000 чел.

Решение. Имеем $\bar{X} = 0.56$, по таблице $t = 1.96$. Поэтому

$$r = \frac{\sqrt{0.56 \cdot 0.44}\ 1.96}{\sqrt{n}} = \frac{0.973}{\sqrt{n}}.$$

Результаты дальнейшего расчета можно представить в виде таблицы

n	r	I
50	0.138	[0.422,0.698]
500	0.043	[0.516,0.603]
5000	0.014	[0.546,0.574]

Видим, что в первом случае трудно даже предсказать исход выборов, во втором кандидат побеждает, но разброс возможных результатов значителен, в третьем случае можно довольно точно предсказать результат выборов.

2.2.5 Нормальное распределение - оценка σ

В этой задаче допустимыми считаются всевозможные нормальные распределения $\mathcal{N}(a, \sigma^2)$. Имея точечную оценку $\bar{s}$ для σ, целесообразно искать доверительный интервал в виде $I = [\bar{s}(1-t), \bar{s}(1+t)]$. Можно показать, что t является квантилью т.н. χ^2-распределения ("хи-квадрат") и его значение, зависящее от выбранного доверительного уровня γ и объема имеющейся выборки n, но не от неизвестных a, σ, может быть найдено в следующей таблице

Таблица 3 *Квантили для оценки дисперсии в нормальной выборке*

γ			t	
	$n = 5$	$n = 10$	$n = 20$	$n = 100$
0.95	*1.37*	*0.65*	*0.37*	*0.143*
0.99	*2.67*	*1.08*	*0.58*	*0.198*
0.999	*5.64*	*1.8*	*0.88*	*0.27*

Если $q > 1$, то в качестве нижней границы доверительного интервала следует взять ноль.

Пример 10 .Оценить среднеквадратическое отклонение по данным об осадках из примера 7.

Решение. Находим $\bar{X} = 605$ и

$$\bar{s}^2 = \frac{1}{n-1}\sum_{j=1}^{n} X_j^2 - \frac{n}{n-1}\bar{X}^2 = \frac{3028350}{7} - \frac{8}{7}605^2 = 432621 - 418314 = 14306,$$

откуда $\bar{s} = 119.6$. По таблице при $n = 8$ имеем $q \approx 0.8$. Соответственно доверительный интервал

$$I = [119.6 \cdot 0.2, 119.6 \cdot 1.8] = [23.9, 215.3].$$

Он содержит величину $\sigma = 100$, использованную в примере 7. Можно отметить, что доверительный интервал слишком большой. Вообще, для точной оценки σ нужно (при одинаковых требованиях к точности) гораздо больше наблюдений по сравнению с оценкой параметра a.

3 Проверка гипотез

3.1 Общие понятия, связанные с проверкой гипотез

Проверить гипотезу - значит определить, обладает ли распределение данных выборки определенным свойством. Гипотеза называется *простой*, если ей удовлетворяет единственное распределение и *сложной* в остальных случаях. Проверяемую гипотезу часто называют *нулевой* и обозначают H_0. Распределения, допустимые в данной задаче, но не удовлетворяющие гипотезе, называют *альтернативами*.

На основании данных выборки мы можем *принять* или *отклонить* гипотезу. Правило, указывающее в каком случае следует принять, а в каком - отклонить гипотезу, называется *решающим правилом*.

$$R = R(X_1, \ldots, X_n) = \begin{cases} \text{принять гипотезу} \\ \text{отклонить гипотезу} \end{cases}$$

В каждой конкретной ситуации возможны четыре варианта

	Гипотеза верна	Гипотеза неверна
Гипотеза принята	верно	ошибка II рода
Гипотеза отклонена	ошибка I рода	верно

Ошибки первого и второго рода являются конкурирующими - желая уменьшить вероятность одной ошибки, всегда увеличиваем вероятность другой (при постоянном числе наблюдений). Уменьшить их одновременно можно только увеличивая число наблюдений.

Между принятием и отклонением гипотезы имеется асимметрия. Принятие гипотезы всего лишь означает, что она не противоречит данным выборки. Возможно, гипотезу следует при этом проверять и другими способами. Отклонение же гипотезы означает, что она противоречит имеющимся данным, и является окончательным (другие проверки не нужны). В связи с этим различием более опасной можно считать ошибку первого рода - как ведущую к непоправимым последствиям.

Многокритериальную задачу минимизации вероятностей ошибок первого и второго рода принято решать следующим образом. Выбирается более опасный тип ошибки (в дальнейшем - всегда ошибка первого рода) и определенный *уровень значимости*, например, $\alpha = 0.05$ или

$\alpha = 0.01$. Далее рассматриваются только те решающие правила, для которых во всех случаях вероятность избранного типа ошибки не превосходит уровня значимости. Среди таких правил выбирают те, у которых меньше вероятность ошибки противоположного рода. Поскольку вероятность ошибки второго рода зависит от альтернативы, то при сравнении двух правил может оказаться, что одно из них лучше при одних альтернативах, а другое - при других. Но бывают правила, оказывающиеся наилучшими при всех альтернативах.

Решающие правила строят на основе критериев. *Критерий* - это обычная статистическая оценка $T = T(X_1, \ldots, X_n)$ (см. раздел 2.1). На основе критерия можно построить *одностороннее* и *двустороннее* решающее правило. При одностороннем правиле по подходящей специальной таблице выбирается *критический уровень* $t = t(\alpha)$. Если $T > t$, то гипотеза отклоняется, иначе она принимается. При двустороннем критерии по таблице находятся два критических уровня t_1 и t_2, гипотеза отклоняется в каждом из случаев $T < t_1$ и $T > t_2$ и принимается, если $t_1 \leq T \leq t_2$.

3.2 Проверка гипотезы о соответствии распределения элементов выборки заданному распределению (критерий хи-квадрат)

Задача состоит в том, чтобы по заданной выборке определить, имеют ли ее элементы заданное распределение. Методом хи-квадрат задача решается следующим образом. Область возможных значений элементов выборки разбивается на m классов. Подсчитывается количество попаданий элементов выборки в каждый класс n_j. Общее число наблюдений $n = \sum_j n_j$. Подсчитывается вероятность попадания одного элемента выборки в каждый класс согласно предполагаемому распределению p_j и соответственно ожидаемое количество наблюдений $n'_j = np_j$. Значением критерия хи-квадрат будет

$$T = \sum_{j=1}^{m} \frac{(n_j - n'_j)^2}{n'_j} = \left[\sum_{j=1}^{m} \frac{n_j^2}{n'_j} \right] - n \ .$$

Затем подсчитывается число степеней свободы

$$k = m - 1,$$

а критический уровень t находится по следующей таблице.

Таблица 4 *Критические уровни критерия хи-квадрат.*

$k\backslash\alpha$	*0.05*	*0.01*
1	*3.8*	*6.6*
2	*6.0*	*9.2*
3	*7.8*	*11.3*
4	*9.5*	*13.3*
5	*11.1*	*15.1*
...	...	...
10	*18.3*	*23.2*
...	...	...
20	*31.4*	*37.6*
...	...	...
30	*43.8*	*50.9*

Полную таблицу см. например [6], с.470, [3], с.465.

Если $T > t$, гипотеза отвергается, в противном случае - принимается.

Группы рекомендуется строить так, чтобы в одну группу попадали не менее 5-8 наблюдений, и $n'_j \geq 10$.

При проверке сложных гипотез, например, пуассоновости или нормальности, следует оценить значение параметра (или нескольких параметров) по выборке, а затем применить указанную процедуру. При этом число степеней свободы следует уменьшить на число оцененных параметров.

Литература: [6], с.267.

Пример 11 . *Анализ числа отказов при испытании двигателя. Имея статистику числа отказов при испытании 87 образцов двигателя*

Число отказов	0	1	2	3	4	5	6	7	8	9
Количество двигателей	27	24	9	8	7	6	3	2	0	1

проверить гипотезу о том, что это число имеет геометрическое распределение с параметром $q = 0.6$, *т.е.* $P\{N = k\} = (1 - q)q^k$, $k = 0, 1, \ldots$.

Решение. Разобьем данные на 5 групп (0,1,2,3 и более трех отказов). Данные расчета можно представить следующим образом.

отказы	n_j	p_j	n'_j	слагаемое n_j^2/n'_j
0	27	0.4	34.8	20.94
1	24	0.24	20.9	27.58
2	9	0.144	12.5	6.48
3	8	0.086	7.5	8.55
>3	19	0.13	11.3	31.94
всего	87	1	87	95.49

Имеем $T = 95.49 - 87 = 8.49$. Поскольку $k = 4$, принимая $\alpha = 0.05$, по таблице находим $t = 9.5$. Соответственно $T < t$, и данные не противоречат гипотезе.

Пример 12 . *При тех же данных проверить сложную гипотезу об их соответствии закону Пуассона.*

Решение. Сначала оценим параметр $a = \bar{x} = 1.896$. Затем посчитаем теоретические вероятности по формуле

$$P(X = k) = e^{-a}\frac{a^k}{k!}.$$

(имеем $e^{-a} \approx 0.15$). Данные всего расчета можно представить следующим образом.

отказы	n_j	p_j	n'_j	слагаемое n_j^2/n'_j
0	27	0.15	13.05	55.87
1	24	0.285	24.79	23.23
2	9	0.27	23.49	3.45
3	8	0.171	14.88	4.30
>3	19	0.124	10.79	33.46
всего	87	1	87	120.31

Число степеней свободы $k = 5 - 1 - 1 = 3$, значение критерия $T = 120.31 - 87 = 33.31$, критический уровень из таблицы 4 равен 7.8. Поскольку $T > t$, гипотеза отклоняется.

Пример 13 . *При анализе металла на наличие микро-трещин получены данные*

Число трещин	0	1	2	3	4	5	6	7
Число образцов	112	168	130	68	32	5	1	1

Проверить теоретическое предположение о соответствии закону Пуассона с параметром $a = 1.5$.

Решение. Данные всего расчета можно представить следующим образом.

отказы	n_j	p_j	n'_j	слагаемое n_j^2/n'_j
0	112	0.223	115.3	108.79
1	168	0.335	173.2	162.95
2	130	0.251	129.7	130.3
3	68	0.126	65.1	71.03
4	32	0.047	24.3	42.14
>4	7	0.0186	9.6	5.1
всего	517	1	517	520.31

Число степеней свободы $k = 6 - 1 = 5$, критический уровень из таблицы 4 равен 11.1. Поскольку $T = 3.31 < t$, гипотеза принимается.

Пример 14 . *Проведена выборка из 2020 семей, имеющих двух детей.*

Группы	2 мальчика	2 девочки	мальчик и девочка
Число семей	527	476	1017

а) Можно ли считать, что мальчики и девочки появляются независимо с равной вероятностью ?

б) Можно ли считать, что мальчики и девочки появляются независимо, но вероятности их появления оцениваются из выборки ?

а) Решение:

группы	n_j	p_j	n'_j	слагаемое n_j^2/n'_j
2 мальчика	527	0.25	505	549.96
2 девочки	476	0.25	505	448.66
мальчик и девочка	1017	0.5	1010	1024.05
всего	2020	1	2020	2022.67

Число степеней свободы $k = 2$, критический уровень из таблицы $t = 6$. Поскольку $T = 2.67 < t$, гипотеза принимается.

б) Решение: оценка вероятностей имеет вид

$$p_M = (1017 + 2 \cdot 527)/2 \cdot 2020 = 0.513, \qquad p_D = 1 - p_M = 0.487.$$

Далее

группы	n_j	p_j	n'_j	слагаемое n_j^2/n'_j
2 мальчика	527	0.2632	531.6	522.44
2 девочки	476	0.2371	479.1	472.92
мальчик и девочка	1017	0.4997	1009.3	1024.76
всего	2020	1	2020	2020.12

Число степеней свободы $k = 3 - 1 - 1 = 1$, критический уровень из таблицы $t = 3.8$. Поскольку $T = 0.12 < t$, гипотеза принимается.

3.3 Проверка гипотезы о соответствии распределения элементов выборки заданному распределению (критерий Колмогорова)

Здесь рассматривается та же задача, что и в предыдущем пункте, но решается она с помощью другого критерия, который применим при несколько других условиях - а именно, им можно пользоваться при небольшом числе наблюдений (когда хи-квадрат не работает), но важно, чтобы наблюдения были числовыми и имели непрерывное распределение.

Для того, чтобы сформулировать критерий Колмогорова, нам потребуется понятие предполагаемой *функции распределения* элементов выборки $F(x) = P\{X \le x\}$, а также ее выборочного аналога, *эмпирической функции распределения*,

$$F_n(x) = \frac{1}{n}\#\{j : X_j \le x\} = \frac{1}{n}\sum_{j=1}^{n} 1_{X_j \le x}.$$

По закону больших чисел имеем при каждом x,

$$F_n(x) \to F(x), \qquad n \to \infty,$$

и можно показать даже, что эта сходимость равномерна по x. Критерий Колмогорова основан на статистике Колмогорова

$$D_n = \sup_x |F_n(x) - F(x)|,$$

которую удобнее записать с помощью *вариационного ряда*

$$X^{(1)}, \ldots, X^{(k)}, \ldots, X^{(n)},$$

т.е. последовательности элементов выборки, записанных в порядке возрастания. Тогда

$$D_n = \max_{1 \le k \le n} \max \left\{ \frac{k}{n} - F(X^{(k)}); F(X^{(k)}) - \frac{k-1}{n} \right\}. \qquad (3)$$

Критерий Колмогорова работает следующим образом. Вычисляются величины $X^{(k)}$ и $F(X^{(k)})$. Затем по формуле (3) находится D_n. Задавшись уровнем значимости α и учитывая число наблюдений (объем выборки) n находим критический уровень t в следующей таблице.

Таблица 5 *Критические уровни критерия Колмогорова.*

$n \backslash \alpha$	*0.05*
5	*1.26*
7	*1.27*
10	*1.29*
20	*1.31*
30	*1.32*
60	*1.33*
90	*1,34*
...	...
∞	*1.36*

Более подробную таблицу можно найти в [2], с.407.

Полагаем $T = \sqrt{n} D_n$. Гипотезу принимаем, если $T < t$ и отвергаем, если $T > t$.

Пример 15 . *Соответствуют ли данные датчика случайных чисел из следующей таблицы закону* $N(0,1)$ *?*

X_1	X_2	X_3	X_4	X_5	X_6	X_7
0.464	0.137	2.45	-0.323	-0.068	0.296	-0.288

Решение можно представить в такой таблице.

$X^{(k)}$	-0.323	-0.288	-0.068	0.137	0.296	0.464	2.45
$\Phi(X^{(k)})$	0.37	0.385	0.47	0.55	0.62	0.68	0.974
$\frac{k}{n}$	0.143	0.286	0.429	0.571	0.714	0.857	1
$\frac{k-1}{n}$	0	0.142	0.286	0.429	0.571	0.714	0.857

Далее находим $T = \sqrt{7}\ 0.37 = 0.98$, а по таблице $t = 1.27$, так что гипотеза принимается.

Литература: см. [5],с.605; [2],с.92.

3.4 Проверка однородности двух выборок (критерий хи-квадрат)

Пример 16 . *Пытаемся понять, одинаково ли распределение уровней зарплаты в Москве и Санкт-Петербурге по следующим данным опроса 1000 семей в каждом из городов ("АиФ", 1996 г.).*

Доход (тыс руб/чел)	<50	50-100	100-150	150-200	>200	Всего
Москва	225	285	220	150	120	1000
Санкт-Петербург	240	305	225	155	75	1000

В общем случае мы наблюдаем две выборки X и Y. Задача состоит в том, чтобы понять, относятся ли обе выборки к одному и тому же распределению (это называется однородностью выборок). Для этого наблюдения группируются в классы и подсчитываются количества наблюдений n_i^X, n_i^Y, попавших в класс i. Для проверки гипотезы используем критерий хи квадрат (см, [6], с.275)

$$T = n^X n^Y \sum_{i=1}^{m} \frac{\left(\frac{n_i^X}{n^X} - \frac{n_i^Y}{n^Y}\right)^2}{n_i^X + n_i^Y},$$

где $n^X = \sum_{i=1}^m n_i^X, n^Y = \sum_{i=1}^m n_i^Y$, общее число элементов в каждой из выборок. Если $n^X = n^Y$, то формула упрощается до

$$T = \sum_{i=1}^{m} \frac{(n_i^X - n_i^Y)^2}{n_i^X + n_i^Y}.$$

Полученное значение сравнивается со значением критического уровня $t = t(\alpha, k)$ из таблицы 4. При этом α - уровень значимости, а число степеней свободы

$$k = m - 1,$$

где m - число групп. Если $T < t$, то гипотеза об однородности выборок принимается, если же $T > t$, то гипотеза отклоняется, т.е. мы заключаем, что распределения, из которых сделаны выборки, различны.

Данный критерий можно применять только при достаточно большом объеме выборок.

В нашем примере

$$T = 0.484 + 0.678 + 0.056 + 0.082 + 10.38 = 11.68.$$

Число степеней свободы $k = 4$ и $\alpha = 0.05$, по таблице находим $t = 9.5$. Поскольку $T > t$, гипотеза о равенстве распределений отклоняется (главным образом, из-за последнего класса).

Пример 17 . *Оценить влияние рекламы на ход продаж по следующим данным опроса менеджеров 100 магазинов.*

Ход продаж	Хороший	Плохой	Посредственный	Всего
До рекламы	39	21	40	100
После рекламы	51	20	29	100

Решение. Имеем $k = 2$, $T = 1.6 + 0.02 + 1.75 = 3.37$. Критический уровень из таблицы (при $\alpha = 0.05$) $t = 6$. Гипотеза об однородности (отсутствии влияния рекламы на ход продаж) принимается, так как $T < t$. Но она была бы отклонена при уровне значимости $\alpha = 0.2$.

Замечание. Ответ изменится, если все числа выборки удвоить.

Литература: см. Смирнов, [6],с.272; Крамер, [4], с.485.

3.5 Проверка однородности двух выборок (критерий Вилкоксона)

Решается задача об однородности двух выборок, описанная в предыдущем пункте, но по малому числу наблюдений случайных величин с непрерывным распределением.

Пример 18 . *В следующих выборках приведены данные о выручке магазинов, расположенных в двух различных районах города (тыс руб на кв м). Для определения арендной платы (имеющей районный коэффициент) важно понять, одинаково ли распределение в этих двух выборках.*

	1	2	3	4	5	6	7
X	2.3	3.2	2.8	3.1	2.6	2.4	2.25
Y	3.11	2.5	2.7	3.4	2.0	2.05	2.31

Решение. Для проверки гипотезы используем совместный вариационный ряд двух выборок. В данном примере это будет

$$Y_5Y_6X_7X_1Y_7X_6Y_2X_5Y_3X_3X_4Y_1X_2Y_4.$$

Затем вычисляем ранги наблюдений элементов первой выборки в этом ряду (3,4,6,8,10,11,13) и подсчитываем суммарный ранг T. Найденное значение сравнивается с критическими уровнями t_1, t_2. При этом t_1 находится из следующей Таблицы 6, а t_2 рассчитывается из соотношения

$$\frac{t_1+t_2}{2} = \frac{(n^X+n^Y+1)n^X}{2} \ ,$$

т.е.

$$t_2 = (n^X+n^Y+1)n^X - t_1.$$

Таблица 6 . *Критические уровни критерия Вилкоксона*[2]

[2]F.Wilcoxon

n^X	n^Y	$\alpha = 0.05$
6	*6*	*26*
7	*7*	*36*
8	*8*	*49*
9	*9*	*62*
10	*10*	*78*
11	*11*	*96*
12	*12*	*115*
13	*13*	*136*
14	*14*	*160*
15	*15*	*184*
20	*20*	*337*

Если $t_1 < T < t_2$, то гипотеза об однородности выборок принимается, если же $T < t_1$ или $T > t_2$, то гипотеза отклоняется, т.е. мы заключаем, что распределения, из которых сделаны выборки, различны.

В нашем примере $T = 3{+}4{+}\cdots{+}13 = 55$, $n^X = n^Y = 7$, следовательно, при $\alpha = 0.05$ по таблице находим $t_1 = 36$, $t_2 = 15\dot{7} - 36 = 69$. Гипотеза о равенстве распределений принимается.

Заметим, что если $T = r_1 + \cdots + r_{n_x}$ - суммарный ранг, то

$$ET = \frac{n^X(n^X + n^Y + 1)}{2} \ .$$

Можно показать, что

$$DT = \frac{n^X \ n^Y}{12}\left(n^X + n^Y + 1\right) \ .$$

и что при большом (≥ 20) числе наблюдений в обеих выборках имеет приближенно нормальное распределение. Поэтому критические уровни в этом случае можно находить в виде

$$t_{1,2} = \frac{n^X(n^X + n^Y + 1)}{2} \pm \zeta\sqrt{\frac{n^X \ n^Y}{12}\left(n^X + n^Y + 1\right)} \ .$$

Критическая область имеет вид $\{T < t_1\} \cup \{T > t_2\}$. Здесь ζ – квантиль нормального закона.

Пример построения интервалов. Пусть $n^X = n^Y = 20$.

Точный способ: по таблице находим $t_1 = 337$. Тогда $t_2 = (20 + 20 + 1)20 - 337 = 41 \cdot 20 - 337 = 483$. Критическая область $\{T < 337\} \cup \{T > 483\}$.

Приближенный способ: $t_{1,2} = \frac{20(20+20+1)}{2} \pm 1.96\sqrt{\frac{20 \cdot 20 \cdot 41}{12}} = 410 \pm 72.5 = [337.5; 482.5]$. Таким образом, разницы между двумя результатами практически нет.

Альтернативный подход к критерию Вилкоксона. Подсчитываются инверсии $I = |\{y_j < x_i\}|$. Тогда имеем $I = T - \frac{n^X(n^X+1)}{2}$ и можно строить проверку на базе критерия I (с поправкой на сдвиг). Пояснение: число пар $(\cdot, x_i)$ есть $\sum_i (r_i - 1) = T - n^X$; среди них не являются инверсиями $0 + 1 + \cdots + (n^X - 1) = \frac{n^X(n^X-1)}{2}$.

Литература: о критерии Вилкоксона см. также [6],с.283, [5],с.661.

3.6 Проверка независимости двух величин (критерий хи-квадрат)

Пример 19 . *Проверить, имеется ли зависимость между распределением числа преступлений и типом населенного пункта, в котором они совершаются.*

Вид преступления	Тяжкие преступления	Преступления средней тяжести	Мелкие нарушения	Всего (A_i)
Большие города	62	120	310	492
Малые города	31	91	270	392
Деревни, поселки	7	35	52	94
Всего (B_j)	100	246	632	978

Для проверки независимости двух признаков используется критерий χ^2. Пусть $n_{i,j}$ - числа из клеток таблицы (т.е. число совместных появлений i-значения первого признака и j-значения второго признака). Подсчитаем суммарные количества по строкам A_i и по столбцам B_j. Пусть $n = \sum_{i,j} n_{i,j}$ суммарное число наблюдений. Тогда критерий имеет вид

$$T = \sum_{i=1}^{r} \sum_{j=1}^{s} \left(n_{i,j} - \frac{A_i B_j}{n} \right)^2 \frac{n}{A_i B_j} = n \left(\left[\sum_{i=1}^{r} \sum_{j=1}^{s} \frac{n_{i,j}^2}{A_i B_j} \right] - 1 \right)$$

(см. [7], с.295). Полученное значение сравнивается со значением критического уровня $t = t(\alpha, k)$ из таблицы 4. При этом α - уровень значимости, а

$$k = (r-1)(s-1)$$

число степеней свободы, где r, s число строк и столбцов таблицы (т.е. число значений каждого признака). Если $T < t$, то гипотеза о независимости признаков принимается, если же $T > t$, то гипотеза отклоняется, т.е. мы заключаем, что между двумя признаками имеется связь.

В нашем примере при расчете элементов суммы находим по клеткам,

$$\begin{pmatrix} 0.078 & 0.11898 & 0.309 \\ 0.0245 & 0.0859 & 0.02942 \\ 0.005 & 0.053 & 0.045 \end{pmatrix},$$

а после суммирования находим $n = 978$, $T = 978(1.0135 - 1) = 13.2$. Число степеней свободы $k = 4$. Принимая $\alpha = 0.05$, находим по таблице $t = 9.5$. Поскольку $T > t$, гипотеза о независимости признаков отклоняется. Имеющиеся данные указывают на определенную связь между двумя признаками.

Пример 20 . *(из промышленной психологии). Определить наличие зависимости между нервным типом испытуемого оператора установки и видом инструкции, который он предпочитает по следующим данным.*

Тип оператора / Вид инструкции	Высокореактивный	Низкореактивный	Всего
Детальная	63	42	105
Краткая	34	56	90
Всего	97	98	195

Решение. При расчете элементов суммы находим по клеткам,

$$\begin{pmatrix} 0.39 & 0.171 \\ 0.132 & 0.356 \end{pmatrix},$$

а после суммирования находим $n = 195$, $T = 195 \cdot 0.049 = 9.55$. Число степеней свободы $k = 1$. Принимая $\alpha = 0.05$, находим по таблице $t = 3.8$. Поскольку $T > t$, гипотеза о независимости признаков отклоняется. Имеющиеся данные указывают на определенную связь между двумя признаками.

4 Проверка наличия связи между величинами

4.1 Проверка наличия связи между числовыми признаками с помощью выборочного коэффициента корреляции

Пример 21 . *Имеется ли связь между умственными способностями родителей и детей по следующим данным о коэффициенте интеллектуального развития родителей (X) и детей (Y) (в среднем по каждой семье).*

Семья	1	2	3	4	5	6	7	8	9	10
Родители	125	120	110	105	105	95	95	90	80	75
Дети	100	105	95	125	120	105	75	95	90	80

Мы будем проверять гипотезу об *отсутствии* линейной связи (незначимости коэффициента корреляции). При этом принятие гипотезы говорит о незначимости связи (т.е. нет данных, свидетельствующих о связи), а отклонение гипотезы означает, что связь между признаками есть.

Заметим, что если величины нормально распределенные, то говорят просто о проверке наличия связи, а если нет - то о наличии *линейной* связи.

По аналогии с формулой для ковариации

$$cov(X,Y) = E(XY) - EXEY$$

определим

$$\overline{cov}(X,Y) = \overline{XY} - \bar{X}\bar{Y} = \frac{\sum_j X_j Y_j}{n} - \frac{\sum_j X_j}{n}\ \frac{\sum_j Y_j}{n}.$$

Далее можно определить эмпирический коэффициент ковариации

$$\bar{\rho} = \frac{\overline{cov}(X,Y)}{\bar{s}_X \bar{s}_Y}\ ,$$

где в знаменателе стоят выборочные оценки среднего квадратического, вычисляемые по формуле (1). Расчет значения критерия ведется по формуле

$$T = \bar{\rho}\sqrt{n-2}/\sqrt{1-\bar{\rho}^2}.$$

Критический уровень t находится по таблице критических уровней критерия Стьюдента (см. Табл. 8) с числом степеней свободы $k = n - 2$ и произвольно выбираемым уровнем значимости α. Гипотеза об отсутствии связи отклоняется, если $|T| > t$ и принимается, если $|T| < t$.

В нашем примере имеем $\bar{X} = 100$, $\bar{Y} = 99$, $\overline{XY} = 10012.5$, $\overline{cov}(X, Y) = 112.5$. Далее $\bar{s}_X^2 = 261.1$, $\bar{s}_X = 16.16$, $\bar{s}_Y^2 = 248.89$,$\bar{s}_Y = 15.78$. Отсюда $\bar{\rho} = 0.44$, $T = 0.44 \cdot 2.83/0.898 = 1.386$. С другой стороны, число степеней свободы $k = n-2 = 8$. Выбирая уровень значимости $\alpha = 0.05$, по таблице находим $t = t(\gamma, k) = 2.31$. Получается $|T| < t$. Вывод: линейной связи двух факторов не усматривается.

Литература [3], гл.19, §22.

4.2 Проверка наличия связи между числовыми признаками с помощью коэффициента ранговой корреляции Спирмана

Рассматривается та же задача, что и в предыдущем пункте, т.е. оценка связи между двумя признаками. Однако в отличие от предыдущей постановки не предполагается, что признаки имеют числовую природу. Вместо этого делается более слабое предположение - что их значения сравнимы, т.е. внутри выборки по каждому признаку можно провести упорядочение (ранжирование).

Пример 22 . *Рассмотрим наличие зависимости между местом музыкальной группы в хит-параде X и объемом продаж дисков данной группы Y (тыс шт) по следующим данным.*

X	1	2	3	4	5	6	7	8	9	10
Y	8	3.9	4	3.8	5.2	2.6	2.1	5.7	5	1.2

Идея метода Спирмана состоит в том, чтобы применить расчет корреляций из предыдущего пункта не к самим данным, а к их рангам. В результате получаем оценку коэффициента корреляции

$$\bar{\rho} = 1 - \frac{6 \sum d_j^2}{n^3 - n},$$

где $d_j = R_j^X - R_j^Y$ разницы рангов в первой и второй выборке, n - число наблюдений. Этот коэффициент называется коэффициентом ранговой корреляции Спирмана.

Проверим гипотезу об *отсутствии* связи (незначимости коэффициента ранговой корреляции).

Расчет значения критерия ведется по старой формуле

$$T = \bar{\rho}\sqrt{n-2}/\sqrt{1-\bar{\rho}^2}$$

Как и ранее, критический уровень t находится по таблице критических уровней критерия Стьюдента (см. Табл. 2). Гипотеза об отсутствии связи отклоняется, если $|T| > t$ и принимается, если $|T| < t$. В нашем примере имеем

R^X	1	2	3	4	5	6	7	8	9	10
R^Y	1	6	5	7	3	8	9	2	4	10
d	0	-4	-2	-3	2	-2	-2	6	5	0

Число наблюдений $n = 10$. По таблице рангов $\sum d_j^2 = 102$, откуда по формулам $\bar{\rho} = 0.38$, $T = 1.16$.

С другой стороны, число степеней свободы $k = n - 2 = 8$. Выбирая уровень значимости $\alpha = 0.05$, по таблице находим $t = t(\gamma, k) = 2.31$. Получается $T < t$. Вывод: связи двух факторов не усматривается.

Поясним формулу коэффициента ранговой корреляции Спирмана. Имея две классические формулы

$$1 + 2 + \cdots + n = \frac{n^2+n}{2},$$

$$1 + 4 + \cdots + n^2 = \frac{2n^3+3n^2+n}{6},$$

найдем $\bar{R} = \frac{1+\cdots+n}{n} = \frac{n+1}{2}$ и $\bar{R}^2 = \frac{1+\cdots+n^2}{n} = \frac{2n^2+3n+1}{6}$, находим $s_R^2 = \frac{2n^2+3n+1}{6} - \frac{(n+1)^2}{4} = \frac{n^2-1}{12}$. Соответственно,

$$\bar{\rho} = \frac{\frac{1}{n}\sum_i R_i^X R_i^Y - \bar{R^X}\bar{R^Y}}{\sqrt{s_{R^X}^2 \cdot s_{R^Y}^2}} = \frac{12}{n^3-n}\left[\sum_i R_i^X R_i^Y - n(\bar{R})^2\right]$$

$$= \frac{12}{n^3-n}\left[\frac{\sum_i (R_i^X)^2 + \sum_i (R_i^Y)^2 - \sum_i (R_i^X - R_i^Y)^2}{2} - n(\bar{R})^2\right]$$

$$= \frac{12}{n^3-n}\left[n s_R^2 - \frac{\sum_i (R_i^X - R_i^Y)^2}{2}\right] = 1 - \frac{6\sum_i (R_i^X - R_i^Y)^2}{n^3-n}\ .$$

Литература: [3], гл.19, §25.

4.3 Проверка гипотез о параметрах нормального закона

В этом параграфе рассматривается нормальная выборка величин, следующих закону $N(a, \sigma^2)$ с неизвестным параметром a и проверяется гипотеза $\{a = a_0\}$ для заданного значения a_0.

Следует различать случаи известного и неизвестного σ. При известном σ используется *критерий Гаусса*

$$T = \frac{\sqrt{n}(\bar{x} - a_0)}{\sigma} \ .$$

Критическая область, используемая при проверке гипотезы, зависит от рассматриваемой альтернативы. Если рассматривается простейшая альтернатива $\{a \neq a_0\}$, то критическая область имеет вид $\{|T| > t\}$, где $t = t(\alpha)$ находится по уровню значимости α из уравнения

$$\mathbf{P}\{|Y| > t\} = \alpha \quad \Leftrightarrow \quad \Phi(t) = 1 - \frac{\alpha}{2} \ ,$$

где Y стандартная нормальная величина, а Φ ее функция распределения. (см. таблицу 7) . Иными словами, гипотеза принимается при $\{|T| \leq t\}$ и отклоняется, если $\{|T| > t\}$. Если рассматривается односторонняя альтернатива $\{a > a_0\}$, то критическая область имеет вид $\{T > t'\}$, где t' находится из уравнения

$$\mathbf{P}\{Y > t\} = \alpha \quad \Leftrightarrow \quad \Phi(t) = 1 - \alpha,$$

(см также таблицу 7) по уровню значимости Иными словами, гипотеза принимается при $\{T \leq t'\}$ и отклоняется при $\{T > t'\}$. Обе процедуры имеют одинаковую вероятность ошибки первого рода, но при односторонней альтернативе для любой альтернативы односторонняя критическая область приводит к меньшей вероятности ошибки второго рода (принятю гипотезы, когда она фактически неверна).

Таблица 7 . *Критические уровни критерия Гаусса.*

α	$\alpha/2$	t'	t
0.1	*0.05*	*1.28*	*1.645*
0.05	*0.025*	*1.645*	*1.96*
0.01	*0.005*	*2.33*	*2.58*

При неизвестном σ приходится аппроксимировать среднеквадратическое отклонение по данным выборки и используется критерий Стьюдента

$$T = \frac{\sqrt{n-1}(\bar{x} - a_0)}{\bar{s}} \ .$$

Построение критических областей и технология проверки гипотез аналогичны предыдущему случаю, стой лишь разницей, что квантили приходится находить по распределению Стьюдента, и в этом случае они зависят не только от α, но и от числа степеней свободы $k = n - 1$, где n – объем выборки. Квантили распределения Стьюдента приведены в следующей таблице.

Таблица 8 . *Квантили распределения Стьюдента.*

α	$\alpha/2$	$k=5$	6	7	8	9	10
0.05	0.025	2.57	2.45	2.365	2.31	2.26	2.23
		2.015	1.94	1.895	1.86	1.83	1.81
0.01	0.005	4.03	3.71	3.5	3.355	3.25	3.17
		3.365	3.14	3	2.9	2.82	2.76

α	$\alpha/2$	$k=15$	20	60	100	120	∞
0.05	0.025	2.13	2.08	2.00	1.99	1.98	1.96
		1.75	1.725	1.67	1.665	1.66	1.645
0.01	0.005	2.94	2.845	2.66	2.63	2.61	2.58
		2.6	2.53	2.39	2.37	2.36	2.33

Пример 23 . *Мониторинг цен. Систематический мониторинг цен на туристские велосипеды среднего класса в течение лета позволил установить их нормальное распределение с параметрами* $a_0 = 9000р.$, $\sigma_0 = 400р.$

В сентябре была сделана новая выборка по данным в 15 торговых точках и зафиксирована средняя цена 8700 р. при оценке среднего квадратического $\bar{s} = 480р.$ *Можно ли говорить о систематическом сезонном снижении цен?*

Решение. Возможны два подхода к проверке гипотезы.

а) Считая теоретическую дисперсию неизменной, находим по критерию Гаусса

$$T = \frac{\sqrt{n}(\bar{x} - a_0)}{\sigma_0} = \frac{\sqrt{15}(8700 - 9000)}{400} = -2.9 \ .$$

Если исходить из уровня значимости $\alpha = 0.05$, то односторонняя критическая область ограничена квантилью $-t'(\alpha) = -1.645$. Имеем $T < -t'$. Поэтому отклонение следует признать значимым, отвергнуть гипотезу $a = a_0$ и можно признать наличие сезонного снижения цен.

б) Оценивая дисперсию по выборке, находим по критерию Стьюдента

$$T = \frac{\sqrt{n-1}(\bar{x} - a_0)}{\bar{s}} = \frac{\sqrt{14}(8700 - 9000)}{480} = -2.34 \ .$$

Если исходить из уровня значимости $\alpha = 0.05$, то односторонняя критическая область ограничена квантилью $-t'(\alpha, n-1) \approx -1.75$. Имеем $T < -t'$. Поэтому отклонение следует признать значимым, отвергнуть гипотезу $a = a_0$ и снова можно признать наличие сезонного снижения цен.

Пример 24 . *Средний курс продажи наличных долларов по Петербургу составлял 31.62 р. Ежедневная выборка 20 обменных пунктов дала среднее значение 31.66 р. и среднеквадратическое отклонение* $\bar{s} = 0.12р$. *Можно ли говорить об изменении среднего курса при уровне значимости 0.01?*

Решение. Критическая область будет двусторонней. Используя критерий Стьюдента, находим $t = 2.845$, $T = \frac{\sqrt{19}\ 0.14}{0.12} = 1.49$. Отклонение среднего курса не значимое, т.к. $|T| < t$. Вывод об изменении курса сделать нельзя.

Литература: Гмурман [3], с.308.

5 Линейная регрессия

5.1 Постановка задачи

Мы многократно наблюдаем значения величины Y (объясняемая величина) и величин $X_1, \ldots, X_m$ (объясняющие величины) и пытаемся понять как Y зависит от $X_1, \ldots, X_m$. Будем ограничиваться простейшей – линейной – формой зависимости.

Пример 25 *Пусть* Y *- стоимость квартиры (тыс. долларов),* X_1 *- расстояние до метро (мин),* X_2 *- площадь квартиры(м*^{2}*). Пытаемся понять как зависит стоимость квартиры от расстояния до метро и от площади квартиры, исходя из следующих данных.*

i	X_{i1}	X_{i2}	Y_i
1	10	70	35
2	20	60	26
3	10	65	32
4	20	30	12
5	10	80	41
6	30	35	15
7	20	55	26
8	20	45	20
9	10	70	35
10	15	29	14

Ожидаемая форма зависимости имеет вид

$$Y \approx \theta_1 X_1 + \theta_2 X_2$$

или, для отдельных наблюдений,

$$Y_i = \theta_1 X_{i1} + \theta_2 X_{i2} + \varepsilon_i$$

Здесь θ_1, θ_2*- неизвестные коэффициенты,* ε_i *- ошибки.*

Метод наименьших квадратов для решения задачи линейной регрессии состоит в том, чтобы выбрать коэффициенты θ_1, θ_2 так, чтобы минимизировать суммарную квадратичную ошибку

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \theta_1 X_{i1} - \theta_2 X_{i2})^2 \to \min.$$

Другие примеры задач регрессии:

Оценить как зависит объем продаж детского питания от цены продукта, количества детей в зоне продаж, цен продуктов конкурентов, объема рекламы.

Оценить как зависит цена на золото на бирже драгоценных металлов от индекса инфляции, кредитных ставок, индекса цен акций на бирже, объема спроса на золото в промышленности.

5.2 Решение общей задачи линейной регрессии

Мы будем решать задачу регрессии

$$\sum_{i=1}^{n} (Y_i - \sum_{j=1}^{m} \theta_j X_{ij})^2 \to \min.$$

где n -число наблюдений, m - число объясняющих величин. Введем матричные обозначения. Пусть $\Theta = (\theta_1, \ldots, \theta_m)$ – вектор-столбец неизвестных коэффициентов, $Y = (Y_1, \ldots, Y_n)$ вектор-столбец значений объясняемой величины (иногда называемый вектором отклика), $X = (X_{ij})$ – матрица значений объясняющих величин (называемая матрицей регрессии). $\mathcal{E} = (\varepsilon_i)$ – вектор-столбец ошибок. В матричной форме задача регрессии записывается как

$$Y = X\Theta + \mathcal{E}.$$

Таким образом минимизации подлежит функция

$$\begin{aligned} Q(\Theta) &= \sum_{i=1}^{n} \varepsilon_i^2 = (\mathcal{E}, \mathcal{E}) = (Y - X\Theta, Y - X\Theta) \\ &= (Y, Y) - 2(X\Theta, Y) + (X\Theta, X\Theta). \end{aligned}$$

Мы должны продифференцировать функцию Q по векторной переменной Θ и приравнять производную к нулю. Продифференцируем вдоль некоторого направления v:

$$Q'_v(\Theta) = \lim_{s \to 0} \frac{Q(\Theta + sv) - Q(\Theta)}{s}.$$

В нашем случае

$$\begin{aligned} & Q(\Theta + sv) = (Y, Y) - 2(X(\Theta + sv), Y) + (X(\Theta + sv), X(\Theta + sv)) \\ = \;& (Y, Y) - 2(X\Theta, Y) + (X\Theta, X\Theta) - 2s(Xv, Y) + 2s(X\Theta, Xv) + s^2(Xv, Xv) \\ = \;& Q(\Theta) - 2s(Xv, Y) + 2s(X\Theta, Xv) + s^2(Xv, Xv). \end{aligned}$$

Отсюда следует, что

$$Q'_v(\Theta) = -2(Xv, Y) + 2(X\Theta, Xv) = 2(X\Theta - Y, Xv).$$

Переходя к транспонированным матрицам, пользуясь равенством $(a, Xb) = (X^T a, b)$, получим

$$Q'_v(\Theta) = 2(X^T(X\Theta - Y), v).$$

Приравнивая производную по *всем* направлениям v к нулю, находим

$$X^T(X\Theta - Y) = 0,$$

откуда

$$X^T X\Theta = X^T Y.$$

В конечном счете мы приходим к ответу

$$\Theta = (X^T X)^{-1} X^T Y.$$

5.3 Решение задачи линейной регрессии в простейших случаях

В этом разделе мы посмотрим как общая формула применяется к задачам с одной и двумя объясняющими переменными. Будем пользоваться следующими обозначениями:

$$Z_1 = \sum_{i=1}^{n} X_{i1}, \qquad Z_{1y} = \sum_{i=1}^{n} X_{i1}Y_i,$$
$$Z_2 = \sum_{i=1}^{n} X_{i2}, \qquad Z_{2y} = \sum_{i=1}^{n} X_{i2}Y_i,$$
$$Z_y = \sum_{i=1}^{n} Y_i, \qquad Z_{12} = \sum_{i=1}^{n} X_{i1}X_{i2},$$
$$Z_{11} = \sum_{i=1}^{n} X_{i1}^2, \qquad Z_{22} = \sum_{i=1}^{n} X_{i2}^2.$$

Начнем с наипростейшего случая.

5.3.1 Задача $Y \approx \theta_1 X_1$

Здесь матрица X состоит из единственного столбца, а матрица X^T – из единственной строки

$$X = \begin{pmatrix} X_{11} \\ \dots \\ X_{n1} \end{pmatrix}, \quad X^T = \begin{pmatrix} X_{11} & \dots & X_{n1} \end{pmatrix}.$$

При этом получается, что "матрица" X^TX – это число $X^TX = Z_{11}$. Соответственно, $(X^TX)^{-1} = Z_{11}^{-1}$. Далее, имеем $X^TY = Z_{1y}$ и мы приходим к окончательному ответу в координатной форме

$$\Theta = \theta_1 = \frac{Z_{1y}}{Z_{11}}.$$

Пример 26 *Оценить зависимость объема подоходного налога от объема заработной платы по данным из 12 регионов. Далее Y – объем подоходного налога по региону (млн. рублей), X_1 – объем заработной платы по региону (млн. рублей).*

i	X_{1i}	Y_i
1	120	14
2	160	19
3	130	16
4	140	17
5	290	35
6	230	27
7	180	22
8	500	61
9	410	49
10	210	25
11	300	34
12	310	37

Решение. Получается $Z_{11} = 894200, Z_{1y} = 107010$ и $\theta_1 = 0.119$.

5.3.2 Задача $Y \approx \theta_1 X_1 + \theta_2 X_2$

Здесь матрица X состоит из двух столбцов, а матрица X^T – из двух строк

$$X = \begin{pmatrix} X_{11} & X_{12} \\ \dots & \dots \\ X_{n1} & X_{n2} \end{pmatrix}, \quad X^T = \begin{pmatrix} X_{11} & \dots & X_{n1} \\ X_{12} & \dots & X_{n2} \end{pmatrix}.$$

При этом получается

$$X^T X = \begin{pmatrix} Z_{11} & Z_{12} \\ Z_{12} & Z_{22} \end{pmatrix}.$$

Обратная матрица имеет вид

$$(X^T X)^{-1} = (Z_{11} Z_{22} - Z_{12}^2)^{-1} \begin{pmatrix} Z_{22} & -Z_{12} \\ -Z_{12} & Z_{11} \end{pmatrix}.$$

Далее, имеем

$$X^T Y = \begin{pmatrix} Z_{1y} \\ Z_{2y} \end{pmatrix},$$

и мы приходим к окончательному ответу $\Theta = (X^T X)^{-1} X^T Y$ в координатной форме

$$\theta_1 = \frac{Z_{22} Z_{1y} - Z_{12} Z_{2y}}{Z_{11} Z_{22} - Z_{12}^2}, \qquad \theta_2 = \frac{Z_{11} Z_{2y} - Z_{12} Z_{1y}}{Z_{11} Z_{22} - Z_{12}^2}.$$

Применим полученные формулы к примеру 25. Здесь будет $Z_{11} = 3125$, $Z_{22} = 32041, Z_{12} = 8135, Z_{1y} = 3770, Z_{2y} = 15441$, и получается

$$\theta_1 = -0.142, \qquad \theta_2 = 0.518,$$

то есть каждая лишняя минута удаленности от метро уменьшает стоимость квартиры на 142\$, а стоимость квадратного метра жилья составляет 518\$.

5.3.3 Задача $Y \approx \theta_1 X_1 + \theta_2$

Задача аналогична предыдущей, причем $X_{i2} = 1$ при всех i. Соответственно, $Z_{22} = n$, $Z_{12} = Z_1$, $Z_{2y} = Z_y$ и решение принимает форму

$$\theta_1 = \frac{nZ_{1y} - Z_1 Z_y}{Z_{11} n - Z_1^2}, \qquad \theta_2 = \frac{Z_{11} Z_y - Z_1 Z_{1y}}{Z_{11} n - Z_1^2}.$$

Пример 27 *Определить зависимость цены (в тыс. \$) подержанного автомобиля ВАЗ от его пробега (тыс. км).*

i	X_{i1}	Y_i
1	32	3.7
2	18	3.8
3	65	3.3
4	81	3.2
5	90	3.1
6	50	3.4
7	72	3.3
8	63	3.4
9	104	2.9

Решение. Здесь будет $n = 9$, $Z_1 = 575, Z_{11} = 42703, Z_{1y} = 18629, Z_y = 30.1$ и получается

$$\theta_1 = -0.0101, \qquad \theta_2 = 3.99,$$

то есть новая машина (без пробега) стоит 3990\$, а каждая лишняя тысяча километров уменьшает стоимость автомобиля на 0.0101 тыс.\$= 10.1\$.

5.4 Вероятностная трактовка задачи линейной регрессии

Мы будем решать ту же самую задачу регрессии в матричной форме,

$$Y = X\Theta + \mathcal{E}.$$

Матрицу X мы считаем неслучайной и известной, а вектор коэффициентов Θ – тоже неслучайным, но неизвестным. Ошибки наблюдений $\mathcal{E}$ мы будем интерпретировать как некоррелированные случайные величины с нулевым математическим ожиданием (то есть систематическая ошибка отсутствует) и одинаковой для всех наблюдений дисперсией $D\varepsilon_i = \sigma^2$. Отсюда следует, что вектор Y является случайным вектором с некоррелированными компонентами, причем $EY = X\Theta$, $DY_i = \sigma^2$.

Мы можем строить различные оценки $T = T(X, Y)$ векторного параметра Θ. Одна такая оценка уже построена при помощи метода наименьших квадратов:

$$\hat{\Theta} = \hat{\Theta}(X, Y) = (X^T X)^{-1} X^T Y.$$

Эта оценка линейна по переменной Y, но сложным образом зависит от X.

Отметим, что $\hat{\Theta}$ является несмещенной оценкой. Действительно, в силу линейности математического ожидания,

$$E\hat{\Theta} = (X^T X)^{-1} X^T EY = (X^T X)^{-1} X^T X\Theta = \Theta.$$

Возникает вопрос - можно ли построить оценку параметра Θ, которая была бы лучше, чем $\hat{\Theta}$ или же эта оценка является оптимальной ? Чтобы уточнить содержание этого вопроса, нам потребуется понятие эффективной оценки.

5.5 Эффективные оценки

Начнем с эффективности оценки *числового* (скалярного) параметра.

Эффективной оценкой называется такая оценка параметра, которая имеет минимальную дисперсию среди всех несмещенных оценок, принадлежащих определенному классу. Иными словами, если $\mathcal{T}$ – некоторый класс оценок параметра θ, то эффективная оценка решает задачу

$$\min\{DT, T \in \mathcal{T}, ET = \theta\}.$$

Эффективность оценки означает, что она является наиболее точной, имеет наименьший разброс.

Эффективность оценки зависит от рассматриваемого класса оценок. Некоторая оценка T может быть эффективной (наилучшей) среди "простых"оценок, но уступать в точности какой-то более сложной оценке.

Теперь перейдем к эффективности оценок векторных параметров. Здесь нам потребуется векторный аналог дисперсии – ковариационная матрица. Напомним, что если $T = (T_1, \ldots, T_m)$ – случайный вектор, то ковариационная матица $cov(T) = (cov(T_j, T_j))_{1 \le j,j \le m}$. В частности, на диагонали этой матрицы стоят дисперсии компонент $cov(T_j, T_j) = DT_j$.

С помощью ковариационной матрицы можно находить дисперсии всевозможных линейных выражений вида

$$(a, T) = \sum_{j=1}^{m} a_j T_j.$$

Действительно, в силу линейности ковариации по каждому аргументу,

$$\begin{aligned} D(a,T) &= cov((a,T),(a,T)) = cov\left(\sum_{j=1}^{m} a_j T_j, \sum_{\ell=1}^{m} a_\ell T_\ell\right) \\ &= \sum_{j=1}^{m}\sum_{\ell=1}^{m} a_j a_\ell cov(T_j, T_\ell) = \left(cov(T)a, a\right). \end{aligned}$$

Эффективной называется такая оценка векторного параметра, которая имеет минимальную ковариационную матрицу среди всех несмещенных оценок, принадлежащих определенному классу. Формально это означает следующее: оценка $\hat{\Theta}$ эффективна, если для всех оценок $\tilde{\Theta}$ из класса $\mathcal{T}$ и всех векторов $a \in R^m$ верно

$$D(a, \hat{\Theta}) = \left(cov(\hat{\Theta})a, a\right) \le \left(cov(\tilde{\Theta})a, a\right) = D(a, \tilde{\Theta}).$$

Подставляя сюда единичные векторы a, легко убедиться, что на диагонали верны неравенства

$$D\hat{\Theta}_j = (cov(\hat{\Theta}))_{jj} \le (cov(\tilde{\Theta}))_{jj} = D\tilde{\Theta}_j.$$

Отсюда видно, что эффективная оценка вектора θ дает эффективные оценки каждой его скалярной компоненты θ_j.

Вернемся к оценке параметра в задаче регрессии. Мы докажем такой факт.

Теорема Гаусса–Маркова. *Оценка метода наименьших квадратов* $\hat{\Theta} = (X^T X)^{-1} X^T Y$ *является эффективной оценкой параметра* Θ *среди всех линейных по* Y *оценок этого параметра.*

Доказательство. Мы рассматриваем класс всевозможных оценок $\mathcal{T} = \{T = HY\}$, где $H = H(X)$ – некоторая матрица, определяющая оценку T. Для оценки метода наименьших квадратов $H = \hat{H} = (X^T X)^{-1} X^T$.

Нас интересуют только несмещенные оценки: $ET = \Theta$. В терминах матрицы H это условие примет вид

$$\Theta = ET = EHY = HEY = HX\Theta.$$

Поскольку это должно быть верно при всех $\Theta \in R^m$, то

$$HX = I_m,$$

где I_m – единичная матрица размера m.

Для ковариационных матриц мы имеем

$$cov(T) = cov(HY) = H cov(Y) H^T = H\ \sigma^2 I_m\ H^T = \sigma^2 H H^T.$$

В частности,

$$cov(\hat{\Theta}) = \sigma^2 \hat{H}\hat{H}^T = \sigma^2 (X^T X)^{-1} X^T \cdot X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}.$$

Теперь приступим к доказательству эффективности. Для произвольной оценки T запишем $T = HY = (\hat{H} + B)Y$, причем $B = H - \hat{H}$ удовлетворяет условию

$$BX = HX - \hat{H}X = I_m - I_m = 0.$$

Далее,

$$\begin{aligned} cov(T) &= cov(HY) = \sigma^2 H H^T = \sigma^2 (\hat{H} + B)(\hat{H} + B)^T \\ &= \sigma^2 (\hat{H}\hat{H}^T + B\hat{H}^T + \hat{H}B^T + BB^T). \end{aligned}$$

Поэтому

$$cov(T) - cov(\hat{\Theta}) = \sigma^2 (B\hat{H}^T + \hat{H}B^T + BB^T).$$

Но два первых слагаемых обращаются в ноль, так как

$$B\hat{H}^T = B[(X^TX)^{-1}X^T]^T = BX(X^TX)^{-1} = 0,$$

$$\hat{H}B^T = [B\hat{H}^T]^T = 0.$$

Остается

$$cov(T) - cov(\hat{\Theta}) = \sigma^2 BB^T$$

и этого достаточно, так как

$$(cov(T)a, a) - \left(cov(\hat{\Theta})a, a\right) = \left(BB^Ta, a\right) = (B^Ta, B^Ta) \geq 0.$$

5.6 Оценка точности решения задачи регрессии

Точность оценок, решающих задачу регрессии (в ее вероятностной постановке) зависит от величины ошибок ε_i, характеризующейся значением параметра σ^2. Этот параметр можно оценить величиной

$$S^2 = \frac{1}{n-m}\left(Y - X\hat{\Theta}, Y - X\hat{\Theta}\right).$$

Здесь, как и ранее, n – число наблюдений, m – число объясняющих величин, $\hat{\Theta}$ – оценка метода наименьших квадратов, X – матрица регрессии.

Можно показать, что S^2 – несмещенная оценка, то есть $ES^2 = \sigma^2$ при всех значениях Θ и σ^2.

Литература: см. Бородин, [1],с.178; Смирнов, [6], с.333.

Список литературы

[1] Бородин А. Н. Элементарный курс теории вероятностей и математической статистики. Изд-во "Лань", С-Пб. 1998.

[2] ван дер Варден Б. Л. Математическая статистика, Изд-во иностранной литературы, 1960.

[3] Гмурман В. Е. Теория вероятностей и математическая статистика. Высшая Школа, 1977.

[4] Крамер Г. Математические методы статистики, Мир, 1975.

[5] Кендалл М. Дж., Стьюарт А. Статистические выводы и связи, Мир, 1973.

[6] Смирнов Н. В., Дунин-Барковский И. В. Курс теории вероятностей и математической статистики для технических приложений, Наука, 1965.

[7] Тюрин Ю. Н., Макаров А. А. Статистический анализ данных на компьютере. Изд-во "Инфра", М., 1998.

[8] Юл Дж. Э., Кэндел М. Дж. Теория статистики. Госстатиздат, 1960.