

4. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

В предыдущей главе рассматривалась ситуация, когда закон распределения случайной переменной *известен* с точностью до нескольких числовых параметров. Сейчас мы рассмотрим другую задачу: имеется выборка, порожденная случайной переменной (случайным вектором), закон распределения которой *неизвестен*. Формулируется некоторая гипотеза об этом законе распределения, и ставится вопрос: согласуется ли полученная выборка с этой гипотезой? Точнее: дает ли эта выборка основание отвергнуть гипотезу?

Наиболее простая ситуация возникает, если появление выборки *невозможно* при справедливости проверяемой гипотезы. Так, гипотеза о равномерном на $[0, 1]$ распределении наблюдаемой с.п. очевидно *отвергается* при наличии в выборке, например, числа 2 или числа -1 . Подобные тривиальные случаи в дальнейшем не рассматриваются.

Если выборка *не противоречит* гипотезе, то эта гипотеза не может быть отвергнута безоговорочно. Речь может идти лишь о признании полученной выборки *маловероятной* при справедливости данной гипотезы.

Если установлено, что полученная выборка *плохо согласуется* с гипотезой, то гипотезу следует *отклонить*. В противном случае гипотеза *сохраняется* (для дальнейшей проверки).

При этом следует помнить, что в обоих случаях мы можем ошибиться: отклонить верную гипотезу (так называемая *ошибка первого рода*) или сохранить неверную гипотезу (*ошибка второго рода*).

В предположении справедливости проверяемой гипотезы можно оценить вероятность появления наблюдаемой выборки, а, следовательно, и вероятность ошибки первого рода. В то же время оценить вероятность ошибки второго рода невозможно, так как для ее оценки потребовалось бы исследовать все альтернативные гипотезы. Именно поэтому обычно употребляют выражение "гипотеза сохраняется", а не "гипотеза принимается".

Вероятность сохранения верной гипотезы (доверительную вероятность) принято обозначать β . Постановщик задачи обычно назначает допустимую (с его точки зрения) вероятность ошибки первого рода $(1 - \beta)$.

Серьезное предупреждение. Доверять *доверительной вероятности* в задаче проверки гипотезы, как и в задаче оценивания параметров, можно только при *достаточно большом* количестве проверок. При *однократном* применении любого критерия мы либо примем правильное решение, либо ошибемся. И ни о каких вероятностях в этом случае говорить нельзя!

4.1. Критерий хи-квадрат

Изобретенный К. Пирсоном¹ алгоритм проверки статистических гипотез (так называемый *критерий хи-квадрат*) состоит в следующем:

- 1) пространство элементарных событий, т.е. гипотетическое множество значений случайной переменной (случайного вектора), разбивают на m непересекающихся частей (попарно несовместных событий) J_1, \dots, J_m ;
- 2) в предположении справедливости проверяемой гипотезы находят вероятности этих событий p_1, \dots, p_m ($p_1 + \dots + p_m = 1$) и "ожидаемые частоты" $\nu_1 = p_1 n, \dots, \nu_m = p_m n$ (здесь n – объем выборки), т.е. количества элементов выборки, которые *должны попасть* в эти части;
- 3) находят "наблюдаемые частоты" n_1, \dots, n_m ($n_1 + \dots + n_m = n$), т.е. количества элементов выборки, *фактически попавших* в эти части.
- 4) Если гипотеза верна, наблюдаемые частоты должны *мало* отличаться от ожидаемых. Расстояние между векторами наблюдаемых и ожидаемых частот определяется формулой

$$\widehat{\chi^2} = \sum_{k=1}^m \frac{(n_k - \nu_k)^2}{\nu_k}.$$

Число $\widehat{\chi^2}$ может рассматриваться как значение с.п. (статистики) χ^2 . К. Пирсон доказал, что для *больших* объемов выборок плотность распределения этой с.п. (независимо от распределения породившего выборку случайного вектора) хорошо аппроксимируется функцией

$$f_{\chi^2}(t, \kappa) = \begin{cases} 0 & \text{при } t < 0, \\ \frac{1}{2^{\kappa/2} \Gamma(\kappa/2)} \cdot t^{\kappa/2-1} \cdot \exp\left(-\frac{t}{2}\right) & \text{при } t \geq 0, \end{cases} \quad (4.1.1)$$

где $\kappa = m - 1$.

Функция $f_{\chi^2}(t, \kappa)$ называется *плотностью распределения хи-квадрат с κ степенями свободы*. Ее график при $\kappa = 4$ изображен на рис.4.1.

¹Карл ПИРСОН (1857-1936) – английский математик и биолог, профессор Лондонского университета, основатель журнала "Биометрика".

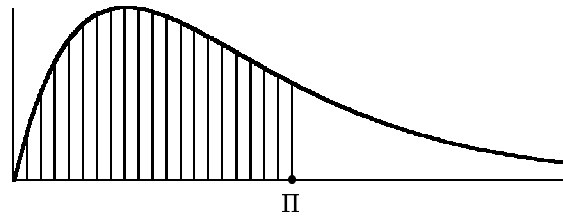


Рис. 1:

5) Назначив допустимую вероятность ошибки первого рода $(1 - \beta)$, можно определить *пороговое значение* статистики χ^2 (точка Π на рис.4.1) как решение уравнения

$$P(\chi^2 < \Pi) = \int_{-\infty}^{\Pi} f_{\chi^2}(t, \kappa) dt = \beta,$$

т.е. как значение в точке β функции, *обратной*² к функции распределения с.п. χ^2 .

6) Проверяемая гипотеза отклоняется, если вычисленное по выборке значение статистики *больше* порогового ($\chi^2 > \Pi$), и сохраняется – в противном случае.

При многократном повторении процедуры проверки гипотезы по критерию хи-квадрат можно ожидать, что относительная частота ошибки первого рода будет близка к заданной вероятности $(1 - \beta)$.

Как видно из рис.4.1, статистика хи-квадрат может принимать на выборках как угодно большие значения, но чем больше значение статистики, тем реже оно будет встречаться. Площадь заштрихованной части фигуры на рис.4.1 равна доверительной вероятности β . Площадь незаштрихованной части $(1 - \beta)$ – это вероятность ошибки первого рода.

Замечания. 1. Каково бы ни было гипотетическое распределение с.п. X , применение критерия хи-квадрат начинается с замены этого распределения на дискретное распределение с вероятностями элементарных событий $p_k = P(X \in J_k)$, $k = 1, \dots, m$.

Если m мало, то будут потеряны характерные особенности гипотетического распределения, и сделанные выводы будут ненадежными. Если же m слишком велико, то наблюдаемые частоты могут оказаться очень малыми, что тоже снижает достоверность выводов.

Обсуждение оптимального значения m выходит за рамки нашего курса. Однако очевидно, что объем выборки должен быть достаточно велик, чтобы обеспечить и достаточно большое число m , и достаточно большую среднюю наблюдаемую частоту $\frac{n}{m}$.

2. Если m – количество частей, на которые разбивается гипотетическое пространство элементарных событий, – задано, то возникает вопрос: а как выбирать эти части? По-видимому, не будет хуже, если брать их *равновероятными*:

$$p_k = P(X \in J_k) \equiv \frac{1}{m}; \quad k = 1, \dots, m.$$

3. Если некоторые параметры гипотетического распределения оцениваются по выборке, то алгоритм применения критерия хи-квадрат сохраняется, но число степеней свободы уменьшается на число этих параметров, т.е. в формуле (4.1.1) следует положить $\kappa = m - 1 - r$, где r – количество оцениваемых параметров.

4.2. Проверка гипотезы о независимости в совокупности координат случайного вектора (критерий хи-квадрат)

Рассмотрим *двухмерный* случайный вектор с координатами X и Y . Имея выборку объема N , разделим координатную ось OX на непересекающиеся промежутки J_{x1}, \dots, J_{xK} , а координатную ось OY – на непересекающиеся промежутки J_{y1}, \dots, J_{yR} . Если верна гипотеза о независимости координат, то вероятность события $(X \in J_{xk}) \cap (Y \in J_{yr})$ равна произведению вероятностей событий $X \in J_{xk}$ и $Y \in J_{yr}$.

Построим так называемую *таблицу сопряженности признаков*, т.е. $(K \times R)$ -матрицу T , элемент t_{kr} которой – количество элементов выборки, попавших в прямоугольник $J_{xk} \times J_{yr}$.

²Функция распределения хи-квадрат и обратная ей функция табулированы. Их значения "умеют" вычислять среды конечного пользователя и библиотеки ФОРТРАНа.

Вычислим частоты попадания с.п. X в промежутки J_{xk} :

$$n_{xk} = \sum_{r=1}^R t_{kr}, \quad k = 1, \dots, K$$

и частоты попадания с.п. Y в промежутки J_{yr} :

$$n_{yr} = \sum_{k=1}^K t_{kr}, \quad r = 1, \dots, R.$$

Поскольку *вероятности* событий $X \in J_{xk}$ и $Y \in J_{yr}$ неизвестны, заменим их точечными оценками – *относительными частотами* этих событий. Получим, что при *независимых* координатах ожидаемая частота попадания случайного вектора в прямоугольник $J_{xk} \times J_{yr}$ равна

$$\nu_{kr} = \frac{n_{xk} \cdot n_{yr}}{N}.$$

Вычислим значение статистики хи-квадрат:

$$\widehat{\chi^2} = \sum_{k=1}^K \sum_{r=1}^R \frac{(t_{kr} - \nu_{kr})^2}{\nu_{kr}} = N \cdot \left(\sum_{k=1}^K \sum_{r=1}^R \frac{t_{kr}^2}{n_{xk} \cdot n_{yr}} - 1 \right).$$

Определим число степеней свободы. Гипотетическое пространство элементарных событий мы разбили на $K \cdot R$ частей. При этом мы оценивали по выборке параметры дискретного распределения – вероятности

$$p_{xk} = \mathbf{P}(X \in J_{xk}), \quad k = 1, \dots, K; \quad p_{yr} = \mathbf{P}(Y \in J_{yr}), \quad r = 1, \dots, R.$$

Но из этих $K + R$ параметров лишь $K + R - 2$ свободных, так как $\sum_{k=1}^K p_{xk} = 1$ и $\sum_{r=1}^R p_{yr} = 1$. Согласно замечанию 3 из п.4.1,

$$\kappa = K \cdot R - 1 - (K + R - 2) = (K - 1) \cdot (R - 1). \quad (4.2.1)$$

Пример. Имея выборку объема $N = 20$,

X	62	72	88	3	51	84	84	41	46	17
Y	73	99	23	35	59	41	26	53	28	15
X	57	30	49	74	89	84	21	13	27	41
Y	80	53	95	55	62	15	71	9	0	2

разделим координатные оси на непересекающиеся части (мы взяли на каждой оси по четыре *равночастотных* промежутка):

$$J_{x1} =] - \infty, 28[, \quad J_{x2} = [28, 50[, \quad J_{x3} = [50, 80[, \quad J_{x4} = [80, +\infty[;$$

$$J_{y1} =] - \infty, 20[, \quad J_{y2} = [20, 50[, \quad J_{y3} = [50, 70[, \quad J_{y4} = [70, +\infty[.$$

Построим таблицу сопряженности признаков – (4×4) -матрицу T :

	J_{y1}	J_{y2}	J_{y3}	J_{y4}
J_{x1}	3	1	0	1
J_{x2}	1	1	2	1
J_{x3}	0	0	2	3
J_{x4}	1	3	1	0

Так как промежутки на обеих осях взяты равночастотные, имеем $n_{xk} = n_{yr} \equiv 5$.

Находим значение статистики хи-квадрат на имеющейся выборке:

$$\widehat{\chi^2} = 20 \cdot \left(\sum_{k=1}^4 \sum_{r=1}^4 \frac{t_{kr}^2}{5 \cdot 5} - 1 \right) = 13.6.$$

Число степеней свободы, согласно формуле (4.2.1), равно 9. В таблице 4.1 приведены пороговые значения статистики хи-квадрат с 9 степенями свободы для некоторых часто используемых значений доверительной вероятности β .

Таблица 1:

β	0.9	0.95	0.99	0.995	0.999
Π	14.7	16.9	21.7	23.6	27.9

Назначив $\beta = 0.9$, т.е. допуская ошибочное отклонение гипотезы в одном случае из десяти, мы видим, что $\widehat{\chi^2} = 13.6 < 14.7 = \Pi$, и гипотеза о независимости сохраняется.

Замечания. 1. Рассмотренный пример носит методический характер: имея выборку столь малого объема, конечно, не следует строить таблицу сопряженности признаков из 16 клеток.

2. Нетрудно распространить описанный алгоритм на случай проверки гипотезы о независимости в совокупности координат случайного вектора, размерность которого больше двух.

4.3. Проверка гипотезы о том, что выборки порождены одной и той же случайной переменной (критерий хи-квадрат)

Имеются s выборок:

$$x^{(1)}, \dots, x^{(s)},$$

где $x^{(j)} = \{x_1^{(j)}, \dots, x_{n_j}^{(j)}\}$. Проверяется гипотеза: все выборки представляют одну и ту же случайную переменную. Опишем алгоритм применения критерия хи-квадрат в этой задаче.

- 1) Выборки объединяются, образуя новую выборку z объема $\sum_{j=1}^s n_j$.
- 2) Выборка z упорядочивается по возрастанию.
- 3) Назначается число m , и числовая ось делится на m непересекающихся промежутков J_1, \dots, J_m , в каждый из которых попадает одинаковое число элементов выборки z . Границы этих промежутков служат точечными оценками границ *равновероятных* промежутков, вероятность попадания с.п. в каждый из которых равна $\frac{1}{m}$.
В предположении, что гипотеза верна, относительная частота попадания с.п. в любой из этих промежутков должна быть близка к $\frac{1}{m}$ для *каждой* выборки $x^{(j)}$, $j = 1, \dots, s$. Поэтому ожидаемое количество элементов этой выборки в любом промежутке равно $\frac{n_j}{m}$.
- 4) Находятся t_{jk} – наблюдаемые количества элементов выборки $x^{(j)}$ в промежутке J_k .
- 5) Вычисляется значение статистики хи-квадрат:

$$\widehat{\chi^2} = \sum_{j=1}^s \sum_{k=1}^m \frac{\left(t_{jk} - \frac{n_j}{m}\right)^2}{\frac{n_j}{m}}.$$

- 6) Можно показать, что число степеней свободы в этой задаче

$$\varkappa = (s - 1) \cdot (m - 1)$$

(по $m - 1$ степени свободы на каждую из s выборок, причем по выборке оценивается $m - 1$ параметров гипотетического распределения – границы равновероятных промежутков).

- 7) Задается доверительная вероятность β и находится Π – соответствующее пороговое значение статистики хи-квадрат.

- 8) Если $\widehat{\chi^2} > \Pi$, гипотеза отвергается.

4.4. Проверка гипотезы о законе распределения случайной переменной

По утверждению разработчика программного датчика псевдослучайных чисел (п.2.7), этот датчик генерирует случайную переменную с заданным законом распределения. Имеется выборка, порожденная этим датчиком. Рассмотрим два критерия, используемых для ответа на вопрос, согласуется ли полученная выборка с декларированным законом распределения: уже известный критерий хи-квадрат и критерий Колмогорова³. Алгоритмы применения этих критериев будут проиллюстрированы на примере.

Пример. Имеется выборка, содержащая $n = 100$ псевдослучайных чисел, полученных от датчика стандартного нормального распределения ($M(X) = 0$, $\sigma(X) = 1$):

³Андрей Николаевич КОЛМОГОРОВ (1903-1987) – один из крупнейших математиков XX века. Действительный член АН СССР, член практически всех наиболее авторитетных научных сообществ мира. Создатель одной из крупнейших в стране научных школ. Автор основополагающих работ по теории функций, теории вероятностей, теории информации, теории алгоритмов...

-0.64	1.78	0.24	1.07	-1.74	0.18	0.09	1.57
0.08	-1.76	-0.83	-0.15	-0.43	0.79	-1.60	-0.69
0.17	0.86	1.60	-0.80	-0.15	-1.99	0.97	1.32
0.85	-0.49	1.00	0.44	0.81	-0.35	-0.43	-0.34
1.04	0.67	-0.72	-0.76	1.23	-0.10	-0.85	-0.84
0.29	0.41	1.08	1.82	0.81	0.59	0.28	-0.21
-0.35	0.15	-0.84	1.84	-1.71	0.56	0.90	-1.69
0.93	-0.32	0.07	-1.39	1.68	-1.80	0.45	0.82
-1.09	1.73	-0.08	-0.25	0.06	0.15	-0.14	-0.04
1.22	0.01	0.20	-0.58	-1.72	2.25	1.93	-1.82
-0.35	0.29	-2.15	-1.21	-1.20	-0.98	-1.25	-1.25
0.93	-1.48	-0.89	2.10	0.76	-0.25	-0.93	0.06
-0.38	-0.34	-2.39	0.27				

Критерий хи-квадрат

Согласно алгоритму, изложенному в п.4.1, разделим множество значений с.п. (\mathbb{R}) на $m = 10$ равновероятных (для стандартного нормального распределения) промежутков.

Решая уравнения

$$\frac{1}{\sqrt{2\pi}} \cdot \int_{-\infty}^{x_k} \exp\left(-\frac{t^2}{2}\right) dt = \frac{1}{2} \cdot \left(1 + \operatorname{erf}\left(\frac{x_k}{\sqrt{2}}\right)\right) = \frac{k}{10}; \quad k = 1, \dots, 9,$$

найдем границы этих промежутков:

k	1	2	2	4	5	6	7	8	9
x_k	-1.28	-0.84	-0.52	-0.25	0.00	0.25	0.52	0.84	1.28

Очевидно, *ожидаемое* количество чисел в каждом из равновероятных промежутков

$$]-\infty, x_1[, [x_1, x_2[, \dots, [x_9, +\infty[$$

равно 10. Количества фактически наблюдаемых чисел в промежутках приведены в следующей таблице:

k	1	2	3	4	5	6	7	8	9	10
n_k	12	10	9	11	8	12	7	8	12	11

Вычислим значение статистики хи-квадрат на выборке:

$$\widehat{\chi^2} = 0.1 \sum_{k=1}^{10} (n_k - 10)^2 = 3.2$$

Число степеней свободы $\varkappa = 9$. Выбрав $\beta = 0.95$, т.е. допуская ошибочное отклонение гипотезы в пяти случаях из ста, из таблицы 4.1 находим $\Pi = 16.9$. Поскольку $\widehat{\chi^2} = 3.2 < 16.9$, мы не имеем основания предъявить претензии разработчику датчика.

Критерий Колмогорова

Пусть F_ξ – гипотетическая функция распределения наблюдаемой с.п., а $x = [x_1, \dots, x_n]$ – *вариационный ряд*, т.е. *упорядоченная по возрастанию* выборка. Тогда гипотетическая вероятность события $(\xi < t)$ равна $F_\xi(t)$, а наблюдаемая относительная частота этого события равна количеству элементов вариационного ряда, лежащих левее точки t . Пусть Δ – наибольшее отклонение гипотетической вероятности от наблюдаемой относительной частоты.

А.Н. Колмогоров показал, что при большом объеме выборки n функция распределения статистики $\Lambda = \sqrt{n} \cdot \Delta$ хорошо аппроксимируется функцией

$$K(\lambda) = \begin{cases} 0 & \text{при } \lambda \leq 0, \\ 1 - 2 \sum_{m=1}^{+\infty} (-1)^{m-1} \exp(-2m^2\lambda^2) & \text{при } \lambda > 0. \end{cases}$$

Эта функция *не зависит* от гипотетической функции распределения F_ξ . Ее называют *функцией распределения Колмогорова*. Функция $K(\lambda)$ и обратная ей функция табулированы, их "умеют" вычислять среды конечного пользователя и библиотеки ФОРТРАНа.

Сформулируем алгоритм критерия Колмогорова.

- 1) Выборка упорядочивается по возрастанию.
- 2) Вычисляется $\hat{\lambda}$ – значение статистики Λ на выборке.
- 3) Задается доверительная вероятность β .
- 4) Вычисляется пороговое значение статистики Колмогорова Π , т.е. решение уравнения $K(\Pi) = \beta$.
- 5) Если $\hat{\lambda} > \Pi$, гипотеза отклоняется.

В таблице 4.2 приведены пороговые значения статистики Колмогорова для некоторых часто используемых значений доверительной вероятности β .

Таблица 2:

β	0.9	0.95	0.99	0.995	0.999
Π	1.22	1.36	1.63	1.73	2.23

Пример. Упорядочив выборку, уже рассмотренную в этом пункте, найдем $\hat{\lambda} = 0.572$.

Зададим (как и для критерия хи-квадрат) $\beta = 0.95$. Этой доверительной вероятности соответствует пороговое значение статистики Колмогорова $\Pi = 1.36$. Поскольку $\hat{\lambda} < \Pi$, критерий Колмогорова также не дает оснований забраковать датчик псевдослучайных чисел.

Замечание. Повторим еще раз: мы не можем утверждать, что гипотеза о нормальности верна. Имеющаяся выборка лишь *не дает оснований отвергнуть* ее.

Серьезное предупреждение. Мы применили два статистических критерия для проверки качества программного датчика псевдослучайных чисел.

Иногда эти критерии применяются в другой ситуации. Пусть требуется оценить вероятность некоторого события. Следовало бы это оценивание провести так: проделать серию экспериментов и найти относительную частоту появления события. Однако если событие редкое, то потребуются очень большой объем выборки (не встретив событие ни разу при трех испытаниях, не стоит говорить, что вероятность его появления равна нулю!). Экспериментатор, желая сэкономить на эксперименте, делит множество значений с.п. на *небольшое* число частей m , выдвигает некоторую гипотезу о законе распределения, а затем, получив удовлетворительное (за счет малости m) согласие выборки с гипотезой, применяет гипотетическое распределение для оценки вероятности редкого события. Нетрудно убедиться в том, что таким "методом" можно получить любое заданное наперед значение искомой вероятности.

ЗАКЛЮЧЕНИЕ

Напомним еще раз, что практическое применение теории вероятностей и математической статистики возможно только при описании *массовых* явлений при условии *статистической устойчивости* относительных частот. Это условие часто постулируется без достаточных на то оснований. Приведем в связи с этим цитату из учебника В.Н. Тутубалина¹ (Теория вероятностей. – М.: МГУ, 1972; второе издание – М.: МГУ, 1993): "Чрезвычайно важно искоренить заблуждение, встречающееся иногда у недостаточно знакомых с теорией вероятностей инженеров и естествоиспытателей, что результат любого эксперимента можно рассматривать как случайную величину. В особо тяжелых случаях к этому присоединяется вера в нормальный закон распределения".

Из-за этого заблуждения теория вероятностей очень часто оказывается объектом вульгаризаций и некорректного применения; встречаются и откровенные спекуляции (достаточно упомянуть псевдоисторические труды академика А.Т. Фоменко).

Мы рекомендуем читателю три брошюры В.Н. Тутубалина (Теория вероятностей в естествознании. – М.: "Знание", 1972; Статистическая обработка рядов наблюдений. – М.: "Знание", 1973; Границы применимости. – М.: "Знание", 1977), в которых обсуждаются "идеологические" проблемы, связанные с применением вероятностно-статистических методов.

Тем, кому необходимо более основательное знакомство с теорией вероятностей и математической статистикой, мы рекомендуем цитированный выше учебник, особенно вторую его часть – "Научные и методические замечания". Дело в том, что учебники по теории вероятностей и математической статистике можно разделить на две группы: первая – учебники для математиков, недоступные прикладнику, и вторая – "учебники", где предлагаются определения типа "случайной величиной называется величина, принимающая случайные значения". Курс В.Н. Тутубалина – единственное известное нам исключение.

¹Валерий Николаевич ТУТУБАЛИН (род. 1936) – российский математик, профессор Московского университета.