ПРОВЕРКА НЕПАРАМЕТРИЧЕСКИХ ГИПОТЕЗ

Статистические методы, изложенные в **2–4**, опираются на различные априорные допущения о виде исследуемой *статистической модели.* Например, *метод максимального правдоподобия* применяют при известном (с точностью до вектора параметров) законе распределения *генеральной совокупности.* Основные методы построения *доверительных интервалов* и проверки *статистических гипотез* основаны на предположении о нормальном законе распределения генеральной совокупности. Все эти методы предполагают, что результаты наблюдений являются реализациями независимых случайных величин.

Оказывается, что многие предположения о виде статистической модели, в том числе все перечисленные выше, можно сформулировать как статистические гипотезы и проверить при помощи *статистических критериев* на основании *статистических данных.* Наиболее важные из этих критериев рассмотрены в этой главе.

1. Критерии согласия. Простая гипотеза

Критериями согласия называют статистические критерии, предназначенные для обнаружения расхождений между гипотетической статистической моделью и реальными данными, которые эта модель призвана описать. Другими словами, они выясняют, насколько предположения о распределении случайных величин соответствуют экспериментальным данным, т.е. не вступает ли принятая статистическая модель в противоречие с имеющимися данными.

Критерий Колмогорова. Пусть $\vec{X}_n$ — *случайная выборка объема* n *из генеральной совокупности* X. Рассмотрим задачу проверки *простой статистической гипотезы* H_0 о том, что функция распределения $F(t)$ случайной величины X совпадает с некоторой известной функцией $F_0(t)$:

$$H_0\colon\ F(t) = F_0(t), \quad t \in \mathbb{R}. \tag{5.1}$$

Предположим, что случайная величина X непрерывна. Проверка *основной гипотезы* H_0 против *альтернативной гипотезы*

$$H_1\colon\ F(t) \neq F_0(t) \ \text{для некоторых } t \in \mathbb{R} \tag{5.2}$$

основана на *статистике* $D(\vec{X}_n)$, реализации $D(\vec{x}_n)$ которой определяют по формуле

$$D(\vec{x}_n) = \sup_t |F_n(t) - F_0(t)|, \tag{5.3}$$

где $F_n(t)$ — *эмпирическая функция распределения*, построенная по *реализации* $\vec{x}_n$ *случайной выборки* $\vec{X}_n$.

При заданной вероятности α совершения *ошибки первого рода* ***критерий Колмогорова**** отклоняет гипотезу H_0 в пользу H_1 на *уровне значимости* α, если

$$D(\vec{x}_n) > D_{1-\alpha},$$

где $D_{1-\alpha}$ — квантиль уровня $1-\alpha$ распределения случайной величины $D(\vec{X}_n)$ при условии истинности основной гипотезы H_0.

Если же

$$D(\vec{x}_n) \leqslant D_{1-\alpha},$$

то делается вывод о непротиворечивости (согласии) статистических данных гипотезе H_0.

Разобраться в сути этого формального определения можно при помощи следующего нестрогого рассуждения. Согласно теореме 1.1, случайная величина $\widehat{F}(t;\vec{X}_n) - F_0(t)$, где $\widehat{F}(t;\vec{X}_n)$ — *выборочная функция распределения*, для любого t в случае истинности основной гипотезы H_0 стремится к нулю при $n \to \infty$, а в случае истинности альтернативной гипотезы H_1 — к величине $F(t) - F_0(t)$, которая для некоторых значений t может быть отлична от нуля. Поэтому при $n \to \infty$ случайная величина $D(\vec{X}_n)$ стремится к неслучайной величине $\sup|F(t) - F_0(t)|$, которая в случае истинности основной гипотезы H_0 равна нулю, а в случае истинности альтернативной гипотезы H_1 является

положительной величиной. Следовательно, если для *статистических данных*, представленных *выборкой* $\vec{x}_n$, случайная величина $D(\vec{X}_n)$ приняла „ достаточно большое“ значение, то гипотезу H_0 естественно отклонить в пользу гипотезы H_1, а если $D(\vec{X}_n)$ приняла значение, „близкое к нулю“, то гипотезу H_0 следует принять.

Оказывается, что при истинности основной гипотезы H_0 распределение случайной величины $D(\vec{X}_n)$ не зависит от $F_0(t)$ (хотя зависит от *объема выборки* n), что чрезвычайно важ-но для вычисления квантилей случайной величины $D(\vec{X}_n)$, по-скольку не нужно составлять отдельные таблицы значений функции распределения статистики $D(\vec{X}_n)$ для каждой функ-ции $F_0(t)$, а можно обойтись всего лишь одной таблицей. Это свойство вытекает из приводимой без доказательства следую-щей теоремы*.

Теорема 5.1. Пусть $\widehat{R}(t, \vec{X}_n)$ — выборочная функция рас-пределения, построенная по случайной выборке $\vec{X}_n$ объема n из генеральной совокупности с равномерным законом распре-деления на отрезке $[0, 1]$. Тогда при истинности H_0 функция распределения случайной величины $D(\vec{X}_n)$ совпадает с функцией распределения случайной величины

$$\sup_{0 \leqslant t \leqslant 1} \left| \widehat{R}(t, \vec{X}_n) - t \right|. \quad \#$$

Из теоремы 5.1 следует, что для проверки гипотезы о виде распределения достаточно составить таблицы значений функ-ции распределения статистики $D(\vec{X}_n)$ только для случайной выборки $\vec{X}_n$ из генеральной совокупности X с равномерным законом распределения. Для $n \leqslant 100$ такие таблицы существу-ют*. При больших n для вычисления квантилей $D_{1-\alpha}$ уровня $1 - \alpha$ следует использовать приближенную формулу, которая основана на доказанном А.Н. Колмогоровым предельном соот-ношении

$$\lim_{n \to \infty} \mathbf{P}\left\{ \sqrt{n} D(\vec{X}_n) < t \right\} = K(t), \quad t > 0,$$

где

$$K(t)=\sum_{k=-\infty}^{\infty}(-1)^k e^{-2k^2t^2}. \tag{5.4}$$

Это соотношение справедливо при истинности основной гипотезы H_0. Из него следует, что если n достаточно велико, то

$$D_{1-\alpha}\approx\frac{t_{1-\alpha}}{\sqrt{n}},$$

где величина $t_{1-\alpha}$ определяется уравнением

$$K(t_{1-\alpha})=1-\alpha.$$

Подробные таблицы значений функции $K(t)$ приведены в литературе**. Как показывает практика, приближением с помощью функции $K(t)$ можно пользоваться уже при $n\geqslant 20$. Для вычисления значений $D(\vec{x}_n)$ статистики $D(\vec{X}_n)$ удобна формула

$$D(\vec{x}_n)=\max_{1\leqslant i\leqslant n}\left\{\frac{i}{n}-F_0(x_{(i)}),F_0(x_{(i)})-\frac{i-1}{n}\right\}, \tag{5.5}$$

которую можно также записать в виде

$$D(\vec{x}_n)=\max_{1\leqslant i\leqslant n}\left(\left|F_0(x_{(i)})-\frac{2i-1}{2n}\right|+\frac{1}{2n}\right).$$

Здесь $x_{(i)}$, $i=\overline{1,n}$, — *члены вариационного ряда*, построенного по выборке $x_1, \ldots, x_n$.

Пример 5.1. Для выборки $\vec{x}_{10}$ объема 10 с элементами

−0,29;	1,06;	0,16;	−0,12;	−1,20;
1,09;	−0,91;	1,22;	−1,15;	1,29

на уровне значимости $\alpha=0{,}1$ проверим гипотезу H_0 о том, что эта выборка является реализацией случайной выборки $\vec{X}_n$ из генеральной совокупности X, имеющей стандартное нормальное распределение. Это распределение, согласно (5.1), имеет функцию распределения

$$F_0(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{-\frac{s^2}{2}}\, ds.$$

В качестве альтернативной возьмем гипотезу (5.2).

Вариационный ряд $x_{(1)}, \ldots, x_{(10)}$ выборки $\vec{x}_{10}$ будет иметь вид

−1,20;	−1,15;	−0,91;	−0,29;	−0,12;
0,16;	1,06;	1,09;	1,22;	1,29.

Значения функции распределения $F_0(t)$ в этих точках равны

0,115;	0,125;	0,181;	0,386;	0,452;
0,564;	0,855;	0,862;	0,899;	0,901.

Вычисляем значения функции $\frac{i}{n} - F_0(x_{(i)})$ при $i = \overline{1, 10}$ и $n = 10$:

−0,015;	0,075;	0,119;	0,014;	0,048;
0,036;	−0,155;	−0,062;	0,001;	0,099,

и значения $F_0(x_{(i)}) - \frac{i-1}{n}$ при тех же i и n:

0,115;	0,025;	−0,019;	0,086;	0,052;
0,064;	0,255;	0,162;	0,089;	0,001.

Наибольшим из этих чисел будет 0,255. Значит, $D(\vec{x}_{(10)}) = 0{,}255$. По таблице квантилей статистики* $D(\vec{X}_n)$ для $n = 10$ и $\alpha = 0{,}1$ находим $D_{1-\alpha} = D_{0,9} = 0{,}369$. Так как $D(\vec{x}_{10}) < D_{0,9}$, то оснований отклонить гипотезу H_0 нет.

Критерий ω^2. Из равенства (5.3), задающего статистику $D(\vec{X}_n)$, следует, что критерий Колмогорова „хорошо различает" функции распределения $F(t)$ и $F_0(t)$, отличающиеся друг от друга достаточно сильно пусть даже на небольшом интервале. Если же число $\sup\limits_t |F(t) - F_0(t)|$ невелико, но $F(t) \neq F_0(t)$ на достаточно большом промежутке, то можно показать, что для

проверки гипотезы (5.1) при альтернативе (5.2) целесообразно использовать так называемый ***критерий*** ω^2 (омега-квадрат), использующий статистику

$$\omega^2(\vec{X}_n) = \frac{1}{12n^2} + \frac{1}{n}\sum_{i=1}^{n}\left(F_0(X_{(i)}) - \frac{2i-1}{2n}\right)^2, \qquad (5.6)$$

где $X_{(i)}$, $i = \overline{1, n}$, — элементы вариационного ряда случайной выборки $X_1, \ldots, X_n$. Основная гипотеза (5.1) отклоняется в пользу альтернативной гипотезы (5.2) на уровне значимости α, если

$$\omega^2(\vec{x}_n) > \omega^2_{1-\alpha},$$

где $\omega^2_{1-\alpha}$ — квантиль уровня $1-\alpha$ распределения статистики $\omega^2(\vec{X}_n)$ при условии истинности гипотезы H_0.

Так же как и для критерия Колмогорова, можно доказать, что распределение статистики $\omega^2(\vec{X}_n)$ при истинности основной гипотезы H_0 не зависит от F_0. Для малых n существуют таблицы* квантилей статистики $\omega^2(\vec{X}_n)$. При больших n нужно пользоваться предельным распределением статистики $n\omega^2(\vec{X}_n)$, для которого также составлены таблицы.

Пример 5.2. Вернемся к задаче, рассмотренной в примере 5.1, но для ее решения используем критерий ω^2.

С помощью формулы (5.6) найдем значение $\omega^2(\vec{x}_{10})$ статистики $\omega^2(\vec{X}_n)$, используя значения $F_0(x_{(i)})$, $i = 1, 10$, вычисленные в примере 5.1:

$$\omega^2(\vec{x}_{10}) = \frac{1}{12\cdot 10^2} + \frac{1}{10}\Big(0{,}065^2 + 0{,}125^2 + 0{,}169^2 + 0{,}064^2 + 0{,}098^2 +$$

$$+ 0{,}086^2 + 0{,}105^2 + 0{,}012^2 + 0{,}061^2 + 0{,}149^2\Big) \approx 0{,}0114.$$

По таблицам распределения статистики $\omega^2(\vec{X}_n)$ для $n = 10$ находим

$$\omega^2_{1-\alpha} = \omega^2_{0{,}95} \approx 0{,}046.$$

Так как $\omega^2(\vec{x}_{10}) < \omega^2_{0{,}95}$, то гипотеза H_0 на уровне значимости $\alpha = 0{,}05$ не отклоняется.

Критерий согласия χ^2. При анализе критериев Колмогорова и ω^2 предполагалось, что $\vec{X}_n$ — случайная выборка объема n из генеральной совокупности непрерывной случайной величины X. Пусть теперь наблюдается дискретная случайная величина X, принимающая r различных значений $u_1, \ldots, u_r$ с положительными вероятностями $p_1, \ldots, p_r$:

$$\mathbf{P}\{X = u_k\} = p_k, \quad k = \overline{1, r}, \quad \sum_{k=1}^{r} p_k = 1.$$

Допустим, что в выборке $\vec{x}_n = (x_1, \ldots, x_n)$ число u_k встретилось $n_k(\vec{x}_n)$ раз, $k = \overline{1, r}$. Отметим, что $\sum\limits_{k=1}^{r} n_k(\vec{x}_n) = n$, т.е. случайные величины $n_1(\vec{X}_n), \ldots, n_r(\vec{X}_n)$ зависимы. При этих условиях справедлива следующая теорема*.

Теорема 5.2 (теорема Пирсона). Распределение случайной величины

$$\sum_{k=1}^{r} \frac{(n_k(\vec{X}_n) - np_k)^2}{np_k}$$

при $n \to \infty$ слабо сходится к *χ^2-распределению* с $r - 1$ степенями свободы. #

Этой теоремой можно воспользоваться для проверки *простой гипотезы*

$$H_0\colon\ p_1 = p_{10},\ \ldots,\ p_r = p_{r0}, \tag{5.7}$$

где $p_{10}, \ldots, p_{r0}$ — известные величины, против альтернативной гипотезы

$$H_1\colon \text{ существуют такие } k, \text{ что } p_k \neq p_{k0},\ k = \overline{1, r}. \tag{5.8}$$

Если истинной является гипотеза H_0, то по теореме 5.2 при $n \to \infty$ распределение случайной величины

$$\chi^2(\vec{X}_n) = \sum_{k=1}^{r} \frac{(n_k(\vec{X}_n) - np_{k0})^2}{np_{k0}} = n \sum_{k=1}^{r} \frac{\left(\frac{n_k(\vec{X}_n)}{n} - p_{k0}\right)^2}{p_{k0}} \qquad (5.9)$$

стремится к распределению χ^2 с $r-1$ степенями свободы.

Если основная гипотеза H_0 не является истинной, то в этом случае по закону больших чисел при $n \to \infty$

$$\frac{n_k(\vec{X}_n)}{n} \to p_k, \quad k = \overline{1, r}.$$

Поэтому при $n \to \infty$

$$\frac{n_k(\vec{X}_n)}{n} - p_{k0} = \left(\frac{n_k(\vec{X}_n)}{n} - p_k\right) + (p_k - p_{k0}) \to p_k - p_{k0}.$$

Следовательно, если $p_k - p_{k0} \neq 0$ для некоторых $k = \overline{1, r}$, то статистика $\chi^2(\vec{X}_n)$ принимает бóльшие значения, чем в случае истинности основной гипотезы H_0.

Таким образом, становится естественным следующее определение ***критерия согласия*** χ^2 (хи-квадрат). Этот критерий при больших n на уровне значимости α отклоняет гипотезу H_0 в пользу альтернативной гипотезы H_1, если

$$\chi^2(\vec{x}_n) > \chi^2_{1-\alpha}(r-1),$$

где $\chi^2_{1-\alpha}(r-1)$ — квантиль уровня $1-\alpha$ χ^2-распределения с $r-1$ степенями свободы, а $\chi^2(\vec{x}_n)$ — реализация случайной величины (5.9).

Если же

$$\chi^2(\vec{x}_n) \leqslant \chi^2_{1-\alpha}(r-1),$$

то делается вывод о том, что гипотеза H_0 не противоречит статистическим данным и ее следует принять.

В отличие от критериев Колмогорова и ω^2 критерием χ^2 при небольших объемах выборки n пользоваться нельзя. Более того, для удовлетворительной аппроксимации распределения случайной величины $\chi^2(\vec{X}_n)$ распределением χ^2 необходимо, чтобы

не только n было велико, но и все величины np_k, $k = 1, r$, так-же были немалыми. На практике при небольших r необходимо, чтобы выполнялись условия $np_k \geqslant 10$, $k = 1, r$, а если r велико ($r \geqslant 20$), достаточно, чтобы было $np_k \geqslant 5$, $k = 1, r$. Поскольку теорема Пирсона носит асимптотический характер, то ***крите-рий*** χ^2 является ***асимптотически непараметрическим.***

Критерий χ^2 можно использовать и тогда, когда случайная величина X непрерывна или дискретна, но принимает счет-ное множество значений с положительными вероятностями.

В этом случае множество M возможных значений X разбивают на r непересекающихся подмножеств M_k, $k = \overline{1, r}$, таким образом, чтобы вероятность p_k, $k = \overline{1, r}$, попадания случайной величины X в k-е подмножество M_k удовлетворяла условию $np_k \geqslant 5$ или $np_k \geqslant 10$, $k = \overline{1, r}$. Если X — непрерывная случайная величина, то в качестве M_k, $k = \overline{1, r}$, обычно берут множества вида

$$(-\infty, s_1), \quad [s_1, s_2), \quad \ldots, \quad [s_{r-2}, s_{r-1}), \quad [s_{r-1}, \infty),$$

где $s_1 < s_2 < \ldots < s_{r-1}$, $s_k \in \mathbb{R}$, $k = 1, r-1$.

Определим дискретную случайную величину X', принимающую значение k тогда и только тогда, когда $X \in M_k$, $k = \overline{1, r}$. В этом случае исходная задача проверки статистических гипотез сводится к проверке основной гипотезы (5.7) при альтернативной гипотезе (5.8), где в случае непрерывности случайной величины X

$$p_{k0} = \int\limits_{M_k} dF_0(t) = \int\limits_{M_k} p_0(t)\, dt \quad —$$

вероятность попадания случайной величины X в множество M_k в предположении, что функция распределения случайной величины X есть $F_0(t)$, а плотность — $p_0(t)$. Если X — дискретная случайная величина, имеющая счетное множество

возможных значений $z_1, z_2, \ldots$ и $\mathbf{P}\{X = z_j\} = q_j > 0, j = 1, 2, \ldots$, то вместо проверки гипотезы

$$H_0: q_j = q_{j0},\ j = 1, 2, \ldots,$$

где $q_{j0}, j = 1, 2, \ldots$, — известные числа, при альтернативной гипотезе

$$H_1: \text{существуют такие } j, \text{ что } q_j \neq q_{j0},\ j = 1, 2, \ldots,$$

проверяют гипотезу (5.7) при альтернативной гипотезе (5.8), где вероятности p_{k0}, $k = 1, r$, вычисляют по формулам

$$p_{k0} = \sum_{z_j \in M_k} q_{j0}, \quad k = 1, r.$$

Далее для выборки $\vec{x}_n$ находят число $n_k(\vec{x}_n)$ ее элементов, принадлежащих множеству M_k, $k = 1, r$. Затем, подставляя $\vec{x}_n$ вместо $\vec{X}_n$ в формулу (5.9), определяют реализацию $\chi^2(\vec{x}_n)$ случайной величины $\chi^2(\vec{X}_n)$. Гипотеза H_0 отклоняется в пользу гипотезы H_1, если $\chi^2(\vec{x}_n) > \chi^2_{1-\alpha}(r-1)$ и принимается в противном случае.

Недостатком использования критерия χ^2 для случайных величин, принимающих бесконечное множество значений, является некоторая потеря информации при переходе от X к случайной величине X' с конечным числом значений.

Пример 5.3. Среди *элементов выборки* $\vec{x}_{1000}$ дискретной случайной величины X значение 0 встретилось 343 раза, значение 1 — 372 раза, значение 2 — 201 раз, значение 3 — 68 раз, а значения, бóльшие или равные 4, встретились 16 раз. Проверим на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 о том, что наблюдаемая случайная величина имеет распределение Пуассона с параметром $\lambda = 1$, т.е.

$$\mathbf{P}\{X = k\} = \frac{\lambda^k}{k!} e^{-\lambda}, \quad k = 0, 1, 2, \ldots$$

Предполагая истинность основной гипотезы H_0, находим

$$p_{01} = \mathbf{P}\{X = 0\} = 0{,}368, \qquad p_{02} = \mathbf{P}\{X = 1\} = 0{,}368,$$
$$p_{03} = \mathbf{P}\{X = 2\} = 0{,}184, \qquad p_{04} = \mathbf{P}\{X = 3\} = 0{,}061,$$
$$p_{05} = \mathbf{P}\{X \geqslant 4\} = 0{,}019.$$

Заменим случайную величину X, принимающую бесконечное число значений, случайной величиной X', принимающей только пять различных значений 0, 1, 2, 3 и 4 с положительными вероятностями $p_{10} = 0{,}368$, $p_{20} = 0{,}368$, $p_{30} = 0{,}184$, $p_{40} = 0{,}061$ и $p_{50} = 0{,}019$ соответственно. По формуле (5.9) для $r = 5$, $n = 1000$ получаем

$$\chi^2(\vec{x}'_n) = \frac{(343 - 0{,}368 \cdot 1000)^2}{0{,}368 \cdot 1000} + \frac{(372 - 0{,}368 \cdot 1000)^2}{0{,}368 \cdot 1000} +$$
$$+ \frac{(201 - 0{,}184 \cdot 1000)^2}{0{,}184 \cdot 1000} + \frac{(68 - 0{,}061 \cdot 1000)^2}{0{,}061 \cdot 1000} + \frac{(16 - 0{,}019 \cdot 1000)^2}{0{,}019 \cdot 1000} =$$
$$= 1{,}6984 + 0{,}043 + 1{,}5706 + 0{,}8033 + 0{,}4737 = 4{,}58.$$

По таблице квантелей χ^2-распределения (см. табл. П.3) находим $\chi^2_{0{,}95}(4) \approx 9{,}49$. Так как $4{,}58 < 9{,}49$, то гипотеза H_0 принимается.