МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
Лекция 8

Идея.

Проверить гипотезу Н означает ответить на вопрос, обладает ли неизвестное распределение некоторым свойством.

Главное при ответе на вопрос состоит в том, чтобы НАЙТИ РЕШАЮЩУЮ СТАТИСТИКУ, которая отражала бы в концентрированном виде это свойство, и определить ее распределение, когда гипотеза верна.

Затем определить практически достоверный диапазон ее возможных значений; и если окажется, что наблюдаемое значение не попадает в этот диапазон, то отклонить гипотезу.

Дополнительно хотелось бы, чтобы, когда гипотеза неверна, распределение изменялось бы существенным образом так, чтобы вероятность отклонить Н была бы как можно больше.

Критерий Пирсона-это общий метод получить решающую статистику и определить ее распределение.

§ 8. Критерий хи-квадрат Пирсона проверки гипотез

8.1. Простая гипотеза о вероятностях (повтор)

Пусть результатом одного наблюдения могут быть

$A_1, A_2 ... A_m$ — m возможных исходов. Обозначим:

$p_1, p_2 ... p_m$ — соответствующие истинные (неизвестные) вероятности, $\sum_{i=1}^{m} p_i = 1$,

n — число независимых наблюдений при повторении опыта,

$v_1, v_2 \ldots v_m$ — число появлений соответствующих исходов в n опытах, $\sum_{i=1}^{m} \nu_i = n$;

$p_1^0, p_2^0 ... p_m^0$ — *теоретические* (*гипотетические*) значения вероятностей, $p_i^0 > 0$, $\sum_{i=1}^{m} p_i^0 = 1$.

Требуется по наблюдениям $v_1, v_2 \ldots v_m$ проверить гипотезу H о том, что истинные вероятности $p_1, \ldots, p_m$ имеют значения $p_1^0, p_2^0 ... p_m^0$, т.е.

$$H: \; p_i = p_i^0 \;, \; i=1, 2...m.$$

Оценим по наблюдениям $v_1, v_2 \ldots v_m$ вероятности $p_1, p_2 ... p_m$.

Пусть $\hat{p}_1 = \nu_1 / n \ldots \hat{p}_m = \nu_m / n$ — оценки вероятностей. Мерой расхождения между теоретическими (гипотетическими) $p_1^0, p_2^0 ... p_m^0$ и эмпирическими $\hat{p}_1, \hat{p}_2 \ldots \hat{p}_m$ вероятностями принимается величина

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

которая с точностью до множителя n есть усредненное (при истинности H) значение квадрата относительного отклонения оценок $\hat{p}_i$ от теоретических значений p_i^0. **Статистика X^2 называется статистикой хи-квадрат Пирсона**. Для ее вычисления используются две эквивалентные формулы:

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

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

если верна H ($p_i = p_i^0$), то $X^2/n = \sum_{i=1}^{m} p_i^0\left(\frac{\hat{p}_i - p_i^0}{p_i^0}\right)^2 \to 0$;

если же H неверна, то $X^2/n \to \varepsilon > 0$, и $X^2 = n\varepsilon \to \infty$ при увеличении n.

И потому процедура проверки гипотезы состоит в том, что если величина X^2 принимает «слишком большое, критическое)» значение h, т.е.

если $X^2 \geq h$, то гипотеза H отклоняется. (2)

Если это не так, будем говорить, что «*наблюдения не противоречат гипотезе*». «наблюдения не противоречат гипотезе»-это не значит, что гипотеза верна.

На вопрос, что означает «слишком большое» значение, отвечает теорема Пирсона.

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

$$P\{X^2 < x / H\} \to F_{m-1}(x) \equiv P\{\chi^2_{m-1} < x\}.$$

Покажем, как возникает распределение хи-квадрат. Рассмотрим частный случай $m = 2$. Действительно, так как $\nu_2 = n - \nu_1$, $p_2^0 = 1 - p_1^0$ имеем

$$\frac{(\nu_2 - np_2^0)^2}{np_2^0} = \frac{(\nu_1 - np_1^0)^2}{n(1 - p_1^0)},$$

и статистика X^2 принимает вид

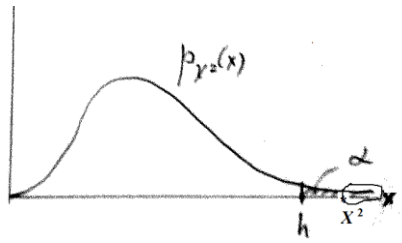
$$X^2 = \frac{(\nu_1 - np_1^0)^2}{np_1^0} + \frac{(\nu_2 - np_2^0)^2}{np_2^0} = \frac{(\nu_1 - np_1^0)^2}{n}\left(\frac{1}{p_1^0} + \frac{1}{1 - p_1^0}\right) = \frac{(\nu_1 - np_1^0)^2}{np_1^0(1 - p_1^0)}. \qquad (3)$$

Если H верна, то ν_1 подчиняется биномиальному распределению $Bi(n, p_1^0)$ с параметрами n и p_1^0, а отношение $\dfrac{\nu_1 - np_1^0}{\sqrt{np_1^0(1-p_1^0)}}$, в силу теоремы Муавра-Лапласа, асимптотически нормально $N(0,1)$. В правой части (3) имеем квадрат этого отношения, что означает сходимость соответствующего распределения к распределению хи-квадрат с одной степенью свободы.

Для произвольного m теорема доказывается методом полной математической индукции.

Теорема означает, что если гипотеза H верна, то при достаточно большом n можно считать, что распределение статистики X^2 подчиняется хи-квадрат распределению.

Порог (критическое значение) h в (2) выберем из условия, что *вероятность ошибки первого рода*, т.е. ***вероятность отклонения гипотезы, когда она верна, должна быть достаточно малой,*** т.е. равной выбираемому значению α — уровню значимости (рис.10):



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

откуда $h = Q(1 - \alpha, m - 1)$ (3а)

- квантиль уровня $(1 - \alpha)$ распределения хи-квадрат с $(m - 1)$ степенями свободы.

Рис. 10. Выбор критического значения

Не обязательно определять квантиль h по α. Можно немного проще.

Процедуры (2) и (3а) проверки H может быть записана иначе в эквивалентном виде: гипотеза H отклоняется, если мала вероятность

$$P\{\chi^2_{m-1} \geq X^2\} \leq \alpha, \qquad (4)$$

т.е. вероятность получить значение статистики не меньшее, чем в опыте X^2, мала, меньше или равна α. Вероятность слева называется минимальным уровнем значимости, потому что при любом значении α, большем $P\{\chi^2_{m-1} \geq X^2\}$, гипотеза будет отклоняться.

Пояснение к (4):

$$X^2 \geq h \quad \Leftrightarrow \int_{X^2}^{\infty} p(x)dx \leq \alpha = \int_{h}^{\infty} p(x)dx$$

Правее h находится вероятность α. Если наблюдаемое значение X^2 правее h, $X^2 \geq h$, то правее X^2 находится вероятность меньше α (см. рис)

Замечание. Теорему Пирсона можно применять, если $n > 50$, все наблюдаемые частоты

$$\nu_i \geq 5, \quad i=1, 2 \ldots m \tag{5а}$$

и теоретические частоты

$$np_i^0 \geq 10, \quad i=1, 2 \ldots m. \tag{5б}$$

Если (5) не выполняется, необходимо объединять некоторые исходы из множества $A_1, A_2 \ldots A_m$.

Пример. Имеется механизм, который предназначен генерировать случайную величину, принимающую с равными вероятностями $p = 0{,}1$ значения 0, 1...9. В табл. 3 приведены количества цифр, появившихся в результате $n = 200$ независимых наблюдений.

Табл. 3. Результаты наблюдений

цифры	0	1	2	3	4	5	6	7	8	9
ν_i	35	16	15	17	17	19	11	16	30	24
$\nu_i - np_i^0$	- 15	- 4	- 5	- 3	- 3	- 1	- 9	- 4	10	4

Каждое значение должно быть в окрестности 20. Есть выделенные наибольшие отклонения: 15 , -9, 10. Не слишком ли они велики для H? Необходимо проверить гипотезу о том, что каждая цифра появляется с равной вероятностью $p = 0{,}1$. В этом примере $np_i^0 = 20$, значение $X^2 = [(-15)^2 + (-4)^2 + \ldots + (4)^2]/20 = 24{,}9$. Для уровня значимости $\alpha = 0{,}05$ порог h = 16,9, т.е.ь $P\{\chi_9^2 > 16{,}9\}$ = 0,05. Если H верна, то получить значение, большее 16.9 весьма моловероятно. Поскольку $X^2 > h$, гипотезу о равных вероятностях следует отклонить. Судя по табличным данным вероятности цифр 0 и 8 превосходят 1/10, а вероятность цифры 6 меньше $1/10$.

Свойство состоятельности критерия. Решающее правило, описанное формулами (2) и (3а), обладает важным свойством *состоятельности*: *если гипотеза H неверна, то с ростом числа наблюдений оно отклоняет гипотезу с вероятностью, стремящейся к* 1. Это свойство выражается следующим соотношением для мощности $W(p)$ критерия (p - истинные вероятности) :

$$W(p) \equiv P\{\text{отклонить } H \,|\, \bar{H} : p, p \neq p^0\} \xrightarrow[n\to\infty]{} 1.$$

Важную характеристику, мощность $W(p)$, можно определить приближенно, опираясь на следующую теорему.

Теорема (б.д.). Если гипотеза неверна, то при $n \to \infty$ распределение статистики X^2 сходится к распределению $\chi^2_{m-1}(a)$ — нецентраль-

ному хи-квадрат с числом степеней свободы (m – 1) и параметром нецентральности a, причем

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

(б/д).

Для запоминания: эта формула получается изосновной (1) заменой наблюдаемых частот ν_i на истинные $np_i = M\nu_i$.

И потому

$$W(p) = P\{X^2 \ge h \,|\, \bar{H} : p, p \ne p^0\} \approx P\{\chi^2_{m-1}(a) \ge h\}$$

Справка о нецентральном распределении хи-квадрат $\chi^2_k(a)$.

Пусть α_1, $\alpha_2 \ldots \alpha_k$ —k штук независимых и нормально распределенных по $N(0,1)$. Составим случайную величину

$$(\alpha_1 + a_1)^2 + (\alpha_2 + a_2)^2 + \ldots + (\alpha_k + a_k)^2 \ ,$$

где a_1, $a_2 \ldots a_k$ — произвольные константы. Нетрудно увидеть, что распределение этой случайной величины зависит не от k параметров a_1, $a_2 \ldots a_k$, а только от одного - суммы их квадратов:

$$a = \sum_{i=1}^{k} a_i^2 \ .$$

Это означает, что составленную случайную величину можно представить в виде:

$$\chi^2_k(a) = (\alpha_1 + \sqrt{a})^2 + \alpha_2^{\,2} + \ldots + \alpha_k^2 \ .$$

Первые два момента получим, вычислив М.О. и дисперсию:

$$M\chi^2_k(a) = k + a, \qquad D\chi^2_k(a) = 2k + 4a.$$

При увеличении k, согласно центральной предельной теореме, распределение асимптотически нормально $N(k + a,\ 2k + 4a)$. Если воспользоваться этим приближением, то

для функции мощности приближенно будем иметь

$$W(p) = P\{X^2 \ge h \,|\, \bar{H} : p, p \ne p^0\} \approx P\{\chi^2_{m-1}(a) \ge h\} \approx 1 - \Phi\left(\frac{h - (m - 1 + a)}{\sqrt{2(m-1) + 4a}}\right).$$

Т.е. мы можем посчитать, с какой вероятностью процедура отклоняет Н, если истинные вероятности равны $p \ne p^0$.

Существует более точное приближение для нецентрального $\chi^2_k(a)$ — **приближение Патнайка**. Оно основано на приближении

$\chi^2_k(a)$ величиной $c\chi^2_m$,

где c и m подбираются из условий равенства первых двух моментов:

$$k + a = cm, \qquad 2k + 4a = c^2 2m;$$

результат решения уравнений:

$$c = (k+2a)/(k+a),\ m = (k+a)^2/(k+2a).$$

Тогда функция распределения приближенно:

$$P\{\chi_k^2(a) < x\} \approx P\{c\chi_m^2 < x\} = P\{\chi_m^2 < x/c\}.$$

8.2. Сложная гипотеза о вероятностях.

Пусть $A_1, A_2 \ldots A_m$ — m исходов некоторого опыта, n — число независимых повторений опыта, $\nu_1, \nu_2 \ldots \nu_m$ — числа появлений исходов. Проверяемая гипотеза H предполагает, что вероятности исходов $P(A_i)$ являются известными функциями $p_i^0(a)$ параметра a (произвольной размерности k: $a = (a_1, a_2 \ldots a_k)$), т.е.

$$H:\ P(A_i) = p_i^0(a),\quad i = 1, 2 \ldots m,$$

но значение a неизвестно.

Пусть $\hat{a}$ — оценка для a, минимизирующая значение статистики X^2. Определим статистику

$$\tilde{X}^2 = \min_a \sum_{i=1}^m \frac{(\nu_i - np_i^0(a))^2}{np_i^0(a)} = \sum_{i=1}^m \frac{(\nu_i - np_i^0(\hat{a}))^2}{np_i^0(\hat{a})}. \qquad (6)$$

Оказывается справедливой теорема Фишера.

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

$$\tilde{X}^2 \geq h, \qquad (7)$$

где $h = Q((1-\alpha), f)$ — квантиль уровня $(1-\alpha)$ распределения хи-квадрат с числом степеней свободы f. Такой порог обеспечивает выбранный уровень α вероятности P (отклонить H / H) ошибки первого рода. Если (7) не выполняется, делаем вывод, что ***наблюдения не противоречат гипотезе***.

Важное замечание. Распределению хи-квадрат с $f = m - 1 - k$ степенями свободы асимптотически подчиняется также статистика (6), если вместо оценок $\hat{a}$ по минимуму X^2 используется a^* — оценка максимального правдоподобия для a.

Процедура (7) может быть записана иначе. Определяем по таблицам $P\{\chi_f^2 \geq \tilde{X}^2\}$, и если

$$P\{\chi_f^2 \geq \tilde{X}^2\} \leq \alpha, \qquad (8)$$

то гипотеза H отклоняется. Иначе «не отклоняется»

8.3. Критерий хи-квадрат при полностью определенном гипотетическом распределении.

Пусть по выборке $\xi_1, \xi_2 \ldots \xi_n$ требуется проверить гипотезу H о том, что функция распределения случайной величины ξ равна $F(x)$. Разобьем диапазон значений случайной величины ξ на m промежутков E_1, $E_2 \ldots E_m$ без общих точек (рис. 11). Такое разбиение осуществляется числами b_1, $b_2 \ldots$ b_{m-1}. При этом правый конец каждого промежутка исключается из соответствующего промежутка, а левый — включается.
Обозначим через A_i событие «попадание наблюдения в E_i»:

$$A_i = \{\xi \in E_i\},\ i = 1, 2 \ldots m.$$

Обозначим через $p_i^0 = P\{A_i\} = P\{\xi \in E_i\}$, $i = 1, 2 \ldots m$. Эти вероятности мы можем определить, т.к. $F(x)$ известна. Пусть ν_i — число элементов выборки, попавших в E_i. Если верна H, то $P\{A_i\} = p_i^0$ - гипотетические вер-ти

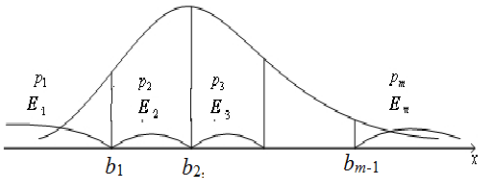


Рис. 11. Разбиение множества значений СВ

Таким образом, требуется проверить гипотезу о вероятностях p_1^0, $p_2^0 \ldots$ p_m^0 по наблюдениям $\nu_1, \nu_2 \ldots \nu_m$. Задача сведена к задаче раздела 8.1.

Как выбрать разбиение на промеюутки, их количество и конкретные точки ?

Обычно промежутки выбирают так, чтобы вероятности были одинаковы и равны

$$p_i^0 = 1/m,$$

причем, чтобы

$$np_i^0 = n/m \geq 10, \text{ то есть } m \leq n/10,$$

однако, слишком малые значения m нежелательны, так как статистика не будет реагировать на изменение распределения внутри интервала.
При равных вероятностях $p_i^0 = 1/m$ точки b_1, $b_2 \ldots$ b_{m-1} деления, есть квантили

$$b_i = Q(i/m) \text{ уровня } i/m,\ i = 1, 2 \ldots m$$

8.4. Гипотеза о типе распределения

Пусть требуется проверить гипотезу о том, что выборка $\xi_1, \xi_2 \ldots \xi_n$ извлечена из совокупности, распределенной по некоторому закону, известному с точностью до параметров $a = (a_1, a_2 \ldots a_k)$, значения которых неизвестны, с функцией распределения $F(x; a)$ (например, требуется проверить гипотезу о нормальности или о пуассоновости).

Разобьем, аналогично разделу 8.3, диапазон значений с.в. ξ на m промежутков E_1, $E_2 \ldots E_m$ без общих точек. Определяем вероятности попадания наблюдения в E_i:

$$P\{\xi \in E_i\} = P\{A_i\} = p_i^0(a), \ , \ i = 1, 2 \ldots m.$$

они являются вполне конкретными функциями параметров, и возникает сложная гипотеза о вероятностях (см. раздел 8.2).

Итак, наши действия таковы:

1. Получим значение оценки a^* методом максимального правдоподобия (или по минимуму X^2).
2. Разобьем весь диапазон наблюдений на m интервалов.
3. Определяем вероятности $p_i^0(a^*)$ или $p_i^0(\hat{a})$ попадания в i-й интервал.
4. Определяем значения ν_i (число наблюдений в i-м интервале).
5. Вычисляем $\tilde{X}^2$ по (6) и принимаем решение по (7).

8.5. Проверка независимости бинарных признаков (таблица 2×2)

Пусть имеется n объектов, выбранных случайно из большой совокупности. Каждый объект характеризуется двумя признаками, которые могут или присутствовать, или отсутствовать в каждом из объектов. Возникает вопрос, имеется ли связь между этими признаками. Например, влияет ли форма собственности на рентабельность производства (A— частная собственность, $\bar{A}$— общественная собственность, B — рентабельно, $\bar{B}$ — нерентабельно), влияет ли курение на легочные заболевания и т.д.?

Проверяется гипотеза H о независимости признаков A и B. Если обозначить $p_A = \mathrm{P}\{A\}$, $p_B = \mathrm{P}\{B\}$, то гипотеза о независимости сводится к следующему: H: $\mathrm{P}\{AB\} = p_A p_B$, (9)

т. е. вероятность встретить сочетание признаков A и B равна произведению вероятностей встретить A и встретить B. В результате анализа на присутствие признаков могут появиться события

$$AB, \ A\bar{B}, \ \bar{A}B, \ \bar{A}\bar{B}.$$

Пусть эти комбинации появились соответственно:

$$\nu_1, \quad \nu_2, \quad \nu_3, \quad \nu_4$$

число раз.

Гипотеза о независимости A и B сводится к гипотезе о том, что вероятности этих событий

$$p_1^0 = p_A p_B, \quad p_2^0 = p_A(1 - p_B), \quad p_3^0 = (1 - p_A)p_B, \quad p_4^0 = (1 - p_A)(1 - p_B)$$

соответственно. Вероятности p_A и p_B — два неизвестных параметра.

Образуем случайную величину X^2, зависящую от неизвестных параметров:

$$X^2(\nu_1, \nu_2, \nu_3, \nu_4, p_A, p_B) = \sum_{i=1}^{4} \frac{(\nu_i - np_i^0(p_A, p_B))^2}{np_i^0(p_A, p_B)}$$

Решив задачу на минимум X^2 по p_A и p_B, найдем оценки, которые получатся такими, как мы ожидаем:

$$\hat{p}_A = \frac{\nu_1 + \nu_2}{n},\ \hat{p}_B = \frac{\nu_1 + \nu_3}{n}$$

Подставив $\hat{p}_A$ и $\hat{p}_B$ в X^2 получим:

$$\tilde{X}^2 = X^2(\nu_1, \nu_2, \nu_3, \nu_4, \hat{p}_A, \hat{p}_B).$$

Примечание [MA1]: Поправить в скобках ν_1, ν_2, ν_3, ν_4

При истинности *H* статистика $\tilde{X}^2$ асимптотически распределена по $\chi^2_{4-2-1} = \chi^2_1$ — хи-квадрат с одной (4-1-2=1) степенью свободы . Пусть задано α. Из условия $\alpha = P\{\tilde{X}^2 > h\} \approx P\{\chi^2 > h\}$ находим порог *h*. Итак, если $\tilde{X}^2 \geq h$, то гипотезу *H* отклоняем.

Приведем окончательные рабочие формулы. Для этого представим данные в следующей таблице 2×2:

	A	$\bar{A}$	
B	ν_{11}	ν_{12}	$\nu_{1\bullet} = \nu_{11} + \nu_{12}$
$\bar{B}$	ν_{21}	ν_{22}	$\nu_{2\bullet} = \nu_{21} + \nu_{22}$
	$\nu_{\bullet 1} = \nu_{11} + \nu_{21}$	$\nu_{\bullet 2} = \nu_{12} + \nu_{22}$	n

Если провести необходимые выкладки, получим следующую формулу:

$$\tilde{X}^2 \equiv n\frac{\left(\nu_{11}\nu_{22} - \nu_{12}\nu_{21}\right)^2}{\nu_{\bullet 1}\nu_{\bullet 2}\nu_{1\bullet}\nu_{2\bullet}}, \qquad (10)$$

где в круглых скобках в числителе стоит квадрат определителя матрицы, а в знаменателе — произведение частных сумм (точка в обозначениях — суммирование по соответствующему индексу).

Пример. Медики испытывали очередную противогриппозную сыворотку на некоторой группе из 25 человек, 5 из которых не приняли сыворотку, 20 приняли. Из 5 человек, не принявших сыворотку, заболели 4, т.е. почти все, из 20 принявших заболели только 6 человек (это реальные данные по одному из магазинов в Москве, из публикации 60-х годов). Медики сделали вывод о том, что получили подтверждение действенности сыворотки. Однако, проверим

гипотезу ***H о бесполезности сыворотки, т.е. о независимости признаков Б (заболеваемость) и С (принятие сыворотки)***.

Исходные данные записаны в таблице.

	Б	$\bar{Б}$	Σ
С	6	14	20
$\bar{С}$	4	1	5
Σ	10	15	25

Вычисляем значение статистики по формуле (10):

$$\tilde{X}^2 \equiv 25\frac{(6\cdot 1-14\cdot 4)^2}{20\cdot 5\cdot 10\cdot 15}=\frac{50^2}{600}=4\frac{1}{6}\ .$$

Много это или мало?

Какую вероятность α ошибки мы можем допустить, если:

α = P{откл. H | H верна} = P{сыв. хор. | сыв. бесполезна }?

Если такая ошибка произойдёт, то сотни тысяч людей останутся без медицинской помощи. В этой медицинской ситуации, связанной со здоровьем тысяч людей, значение α = 0,05 слишком велико. Возможно, подходит α = 0,003. Соответствующий порог $h = Q((1 - \alpha), 1) = Q(0{,}97, 1) = 3^2 = 9$, и мы имеем $\tilde{X}^2 < h$, так что приходится сделать вывод, что

***наблюдения не противоречат гипотезе о бесполезности сыворотки*.**

8.6. Обобщение. Проверка гипотезы о независимости признаков (таблица сопряженности признаков)

Предположим, имеется большая совокупность объектов, каждый из которых обладает двумя признаками A и B. Признак A имеет m уровней: $A_1, A_2 ... A_m$, а признак B — k уровней: $B_1, B_2 ... B_k$. Пусть уровень A_i встречается с вероятностью $P(A_i)$, а уровень B_j — с вероятностью $P(B_j)$. Независимость признаков A и B означает, что

$$P(A_i B_j) = P(A_i)\cdot P(B_j),\quad i = 1, 2...m,\ j = 1, 2...k,$$

т.е. вероятность встретить комбинацию A_iB_j равна произведению вероятностей. Пусть признаки определены на n объектах, случайно извлеченных из совокупности; ν_{ij} — число объектов, имеющих комбинацию A_iB_j, $\sum_{i=1}^{m}\sum_{j=1}^{k}\nu_{ij} = n$.

По совокупности наблюдений $\{\nu_{ij}\}$ (таблица $m \times k$) требуется проверить гипотезу H о независимости признаков A и B. Задача сводится к случаю с неизвестными параметрами, которыми являются вероятности

$$P(A_i),\ i = 1, 2...m \text{ и } P(B_j),\ j = 1, 2...k,$$

всего $(m - 1) + (k - 1)$. Оценки этих вероятностей

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

(точка в обозначениях — суммирование по соответствующему индексу) и статистика (8) принимает вид:

$$\tilde{X}^2=\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\nu_{ij}^2}{n\hat{P}(A_i)\hat{P}(B_j)}-n=n\left(\sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\nu_{ij}^2}{\nu_{i\bullet}\nu_{\bullet j}}-1\right). \qquad (11)$$

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

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

и потому, если

$$P\{\chi^2_f \geq \tilde{X}^2\} \leq \alpha, \qquad (12)$$

то гипотезу о независимости признаков следует отклонить.

Ясно, что по (11), (12) можно проверять независимость двух случайных величин, разбив диапазоны их значений на *m* и *k* частей.