МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
Лекция 9

Повтор

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 $p_A = \mathrm{P}\{\mathrm{A}\}$, $p_B = \mathrm{P}\{\mathrm{B}\}$, то гипотеза о независимости сводится к следующему: *H*: $\mathrm{P}\{\mathrm{AB}\} = p_A p_B$, (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:

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

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

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

задано α. Из условия $\alpha = P\{\tilde{X}_1^{\,2} > h\} \approx P\{\chi_1^{\,2} > h\}$ 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).

Конец повтора

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, $\sum_{i=1}^{m}\sum_{j=1}^{k} \nu_{ij} = n$.

	B_1	…	B_j	…	B_k	Σ
A_1	ν_{11}	⋯	ν_{1j}	⋯	ν_{1k}	$\nu_{1\bullet}$
⋮	⋮		⋮		⋮	⋮
A_i	ν_{i1}	⋯	ν_{ij}	⋯	ν_{ik}	$\nu_{i\bullet}$
⋮	⋮		⋮		⋮	⋮
A_{mi}	ν_{m1}	⋯	ν_{mj}	⋯	ν_{mk}	$\nu_{m\bullet}$
Σ	$\nu_{\bullet 1}$	⋯	$\nu_{\bullet j}$	⋯	$\nu_{\bullet k}$	n

По совокупности наблюдений $\{\nu_{ij}\}$ (таблица $m \times k$) требуется проверить гипотезу H о независимости признаков A и B. Независимость означает, что вероятность встретить сочетание A_i и B_j равна произведению:

$$P(A_i B_j) = P(A_i)\, P(B_j).$$

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 $(m - 1) + (k - 1)$. В выражение для статистики Пирсона входят неизвестные вероятности :

$$X^2 = \sum_{i=1}^{m}\sum_{j=1}^{k} \frac{\nu_{ij}^2}{nP(A_i)P(B_j)} - n \qquad (8)$$

Оценки этих вероятностей по минимуму X^2 приводят к вполне понятным значениям:

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

(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

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

Мы можем определить вероятность получить наше значение $\tilde{X}^2$ (-меры различия), которое мы получили, если H верна:

$$P\{\text{получить} \geq \tilde{X}^2 | H\} \approx P\{\chi_f^2 \geq \tilde{X}^2\} \leq \alpha; \qquad (12)$$

если она мала, 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.

8.7. Проверка гипотезы об однородности выборок

Пусть имеется k выборок объемами n_i, $i = 1, 2\ldots m$, извлеченных из различных совокупностей (например, имеем наблюдения доходов Z служащих в городе А и в городе В; можно ли считать, что распределения доходов одинаковы?).

Измеряемая величина в каждой из выборок может иметь k уровней B_j, $j = 1, 2\ldots k$. Пусть $p_{i\,j}$ — истинная (неизвестная) вероятность получить Z_j в i-й совокупности. Требуется проверить гипотезу H о том, что исходные совокупности распределены одинаково, т.е.

$$p_{ij} = q_j .$$

Обозначим ν_{ij} — число наблюдений в *i*-й выборке, имеющих уровень Z_j, причем $\sum_j \nu_{ij} \equiv \nu_{i\bullet} = n_i$, $\sum_i n_i = n$. Имеем таблицу наблюдений размером $m \times k$, аналогично предыдущему разделу 8.5.

	B_1	…	B_j	…	B_k	Σ
1-я совокупность	ν_{11}	⋯	ν_{1j}	⋯	ν_{1k}	$\nu_{1\bullet} = n_1$
⋮	⋮		⋮		⋮	⋮
i-я совокупность	ν_{i1}	⋯	ν_{ij}	⋯	ν_{ik}	$\nu_{i\bullet} = n_i$
⋮	⋮		⋮		⋮	⋮
m-я совокупность	ν_{m1}	⋯	ν_{mj}	⋯	ν_{mk}	$\nu_{m\bullet} = n_m$
Σ	$\nu_{\bullet 1}$	⋯	$\nu_{\bullet j}$	⋯	$\nu_{\bullet k}$	n

Если *H* верна, то имеем неизвестные параметры q_1, $q_2 \ldots q_k$. Если бы они были известны, можно было бы определить теоретические частоты и составить статистику

$$X^2(q_1,\ldots, q_m) = \sum_{i=1}^{m}\left(\sum_{j=1}^{k}\frac{(\nu_{ij} - n_i q_j)^2}{n_i q_j}\right) = \sum_{i=1}^{m}\left(\sum_{j=1}^{k}\frac{\nu_{ij}^2}{n_i q_j} - n_i\right), \qquad (13)$$

причем $\sum_{j=1}^{k}\frac{(\nu_{ij} - n_i q_j)^2}{n_i q_j}$ при истинности *H* приближено подчинялась бы распределению хи-квадрат с $(k-1)$ степенями свободы, а вся сумма в (13) — распределению хи-квадрат с $m(k-1)$ степенями свободы в силу независимости выборок.

Но вероятности q_1, $q_2 \ldots q_m$ неизвестны. Определяем оценки для них, полагая, что H верна:

$$\hat{q}_j = \frac{1}{n}\sum_{i=1}^{m}\nu_{ij}$$

и после подстановки оценок количество степеней свободы уменьшится на число$(k-1)$ оцененных параметров и станет равным

$$f = m(k-1) - (k-1) = (k-1)(m-1).$$

Статистика $\tilde{X}^2$ и вся процедура проверки гипотезы примет вид, аналогичный (11), (12) при проверке гипотезы о независимости. Решающая статистика: $\tilde{X}^2 = n\left(\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{\nu_{ij}^2}{\nu_{i\bullet}\nu_{\bullet j}} - 1\right)$.

Если $P\{\chi_f^2 \geq \tilde{X}^2\} \leq \alpha$, то гипотезу об однородности следует отклонить.

§ 9. Критерий согласия Колмогорова

Проверяется гипотеза *H* о том, что последовательность независимых наблюдений $\xi_1, \xi_2 \ldots \xi_n$ извлечена из совокупности с непрерывной функцией распределения $F(x)$.

Эту задачу мы уже решали с помощью критерия хи-квадрат Там была одна мера различия между гипотезой и наблюдениями, здесь – другая. Там был общий метод (для произвольных гипотез) построения решающей статистики, здесь – частный метод (и более простой) для проверки Н о заданной непрерывной ф.р.

Построим вариационный ряд $\xi_{(1)} \le \xi_{(2)} \le \ldots \le \xi_{(n)}$ и функцию эмпирического распределения $F_n^*(x)$. Мерой различия (можно говорить качества согласования) эмпирического и гипотетического распределения примем максимальное отклонение функции эмпирического распределения $F_n^*(x)$ от гипотетического $F(x)$, то есть верхняя грань модуля разности

$$D_n = \sup_x \left| F_n^*(x) - F(x) \right|.$$

Замечателен тот факт, что распределение статистики D_n при истинности *H* не зависит от $F(x)$ [4, таблицы Бльшева и Смирнова], и потому критическое значение определяется.

Процедура проверки гипотезы *H*:

— вычисляется D_n;

— выбирается порог λ_1 (о выборе порога, см. ниже);

— если $D_n \ge \lambda_1$, то *H* отклоняется.

Критическое значение λ_1 выбирается так, чтобы обеспечить заданную вероятность ошибки при истинности *H:*

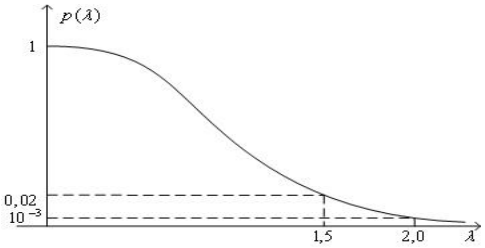
$$P\{\text{отклонить } H \mid H\} = P\{D_n \ge \lambda_1 \mid H\} = \alpha.$$

α - задаваемый уровень значимости. В таблицах [4] имеются значения $\lambda_1(n, \alpha)$ для значений α = 1, 2, 5, 10, 20 %, $n = 1(1)100$. Там же имеются асимптотические формулы.

Оказывается, что для любой непрерывной $F(\cdot)$ при $n \to \infty$ для распределения статистики D_n при истинности гипотезы H справедливо соотношение:

$$P\{D_n \sqrt{n} \ge \lambda \mid H\} \xrightarrow[n \to \infty]{} P(\lambda) \equiv 1 - K(\lambda),$$

где $K(\lambda) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2\lambda^2}$ — функция распределения Колмогорова.



Этот факт при $n \geq 20$ позволяет простым способом приближенно обеспечить заданную вероятность ошибки первого рода (уровень значимости)

$$\alpha = P\{D_n\sqrt{n} \geq \lambda | H\} \approx P(\lambda).$$

Рис. 12. График функции *P*(λ)

График функции $P(\lambda)$ условно показан на рис. 12. Итак, λ выбирается по таблицам как функция α : $\lambda(\alpha)$, т.е.

$$P\{D_n\sqrt{n} \geq \lambda(\alpha)|H\} \approx \alpha \Rightarrow \lambda_1 \approx \frac{\lambda(\alpha)}{\sqrt{n}},$$

и если $D_n \geq \lambda_1 = \frac{\lambda(\alpha)}{\sqrt{n}}$, то H отклоняется.

На рис.12 показаны две характерные точки функции P(λ):

при α = 0,02 $\lambda(\alpha) \approx 1{,}5$,

при $\alpha = 10^{-3}$ $\lambda(\alpha) \approx 2{,}0$.

Для $\lambda > 2{,}5$ хорошим приближением является $P(\lambda) \approx 2e^{-2\lambda^2}$, причем с ростом n точность улучшается.

§ 10. Различение двух простых гипотез

10.1. Фиксированный объем наблюдений

Пусть имеется совокупность наблюдений x = (x_1, x_2...x_n), являющихся реализацией случайных величин $\xi \equiv$ (ξ_1, ξ_2…ξ_n), относительно которой имеется два предположения (гипотезы):

1. H_0: ξ распределена по закону $p_0(x)$;
2. H_1: ξ распределена по закону $p_1(x)$.

(если ξ — непрерывна, то $p_0(x)$, $p_1(x)$ — плотности, если дискретна — вероятности). По наблюдениям x требуется принять одно из двух решений:

или «*верна* H_0» (это решение обозначим 0),

или «*верна* H_1» (решение 1).

Ясно, что дело сводится к определению решающей функции $\delta(x)$, имеющей два значения 0 и 1, т.е. к определению разбиения $\Gamma = (\Gamma_0, \Gamma_1)$ пространства X всех возможных значений x:

$$\delta(x) = \begin{cases} 0, \text{ если } x \in \Gamma_0, \\ 1, \text{ если } x \in \Gamma_1. \end{cases}$$

При использовании любой решающей функции $\delta(x)$ возможны ошибки двух типов:

1) ошибка первого рода — принятие H_1 при истинности H_0;
2) ошибка второго рода — принятие H_0 при истинности H_1.

Для любой решающей функции имеем две условные вероятности:

$$\alpha = P(\text{принять } H_1 \mid H_0) = \int_{\Gamma_1} p_0(x)dx, \qquad (1)$$

$$\beta = P(\text{принять } H_0 \mid H_1) = \int_{\Gamma_0} p_1(x)dx,$$

которые называются вероятностями ошибок первого и второго рода соответственно (рис. 13). Хотелось бы иметь α и β близкими к нулю, но из (1) ясно, что если одна из них уменьшается, например, α (за счет уменьшения Γ_1), то другая, β, увеличивается (за счет увеличения Γ_0; $\Gamma_0 \cup \Gamma_1 = X$, $\Gamma_0 \setminus \Gamma_1 = \varnothing$. Существуют различные подходы к определению оптимального правила.

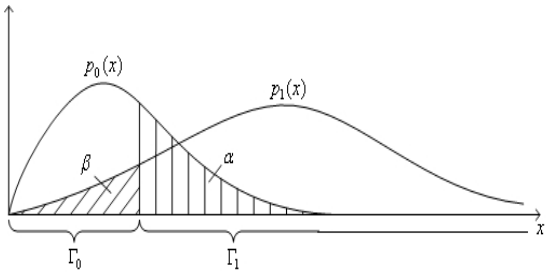


Рис. 13. Вероятности ошибок α и β

Байесовский подход

Будем считать, что многократно сталкиваемся с проблемой выбора между H_0 и H_1. В этом случае можно говорить о частоте, с которой истинна H_0 (или H_1), т.е. о том, что истинность H_0 (или H_1) — событие случайное, причем вероятность события, когда верна H_0, а когда верна H_1, известны:

$$P(H_0) = q_0, \quad P(H_1) = q_1, \quad q_0 + q_1 = 1.$$

Кроме того, будем считать, что
при каждой ошибке первого рода несем потери W_0 (вероятность этого события $P(H_0)$ $P(\text{принять } H_1 \mid H_0)$),
а при ошибке второго рода — потери W_1 (с вероятностью $P(H_1)$ $P(\text{принять } H_0 \mid H_1)$).

Если пользуемся правилом δ (с разбиением Γ), то средний штраф от однократного использования:

$$R(\Gamma) = q_0 \cdot \alpha(\Gamma) \cdot W_0 + q_1 \cdot \beta(\Gamma) \cdot W_1. \qquad (1a)$$

Назовем правило δ (соответственно, разбиение $\Gamma \equiv (\Gamma_0, \Gamma_1)$) в байесовском смысле оптимальным, если

$$R(\Gamma) = \min_{\Gamma'} R(\Gamma').$$

Оказывается справедливой следующая теорема.

Теорема. *Оптимальным является правило, для которого область* Γ_1 *принятия гипотезы* H_1 *определяется соотношением*:

$$\Gamma_1 = \left\{ x: \frac{p_1(x)}{p_0(x)} \geq h \equiv \frac{q_0 W_0}{q_1 W_1} \right\}. \qquad (2)$$

Действительно, пусть T = (T_0, T_1) — произвольное разбиение; для него средние потери

$$R(T) = q_0 \cdot \alpha(T) \cdot W_0 + q_1 \cdot \beta(T) \cdot W_1 = q_0 W_0 \int_{T_1} p_0(x)dx + q_1 W_1 \int_{T_0} p_1(x)dx -$$
$$- q_1 W_1 \int_{T_1} p_1(x)dx + q_1 W_1 \int_{T_1} p_1(x)dx .$$

Здесь во второй строке добавлено и вычтено одно и то же слагаемое. После объединения интегралов получаем

$$R(T) = q_0 W_0 \int_{T_1} [q_0 W_0 p_0(x) - q_1 W_1 p_1(x)]dx + q_1 W_1 .$$

Очевидно, для того чтобы получить минимальное значение средних потерь, в область интегрирования T_1 нужно включить те точки, в которых подынтегральная функция в квадратных скобках отрицательна, т.е.

$$q_0 W_0 p_0(x) - q_1 W_1 p_1(x) < 0,$$

что дает в эквивалентной записи оптимальную область Γ_1 в (2).

Замечание. В частном случае, если $W_0 = W_1 = 1$, то $R(\Gamma)$ в (1а) имеет смысл безусловной вероятности ошибки, а соответствующее оптимальное правило называется правилом *“идеального наблюдателя”* или правилом Зигерта- Котельникова.