Andrzej Dąbrowski

Analiza danych jakościowych

0.1. Wstęp

Table 1. Typy analizy wielowymiarowej statystycznej

typ	argumentu	kategoryczny	ilościowy	mieszany
odpowiedzi	*kategoryczny*	*tablice kontyngencji, logliniowy*	*logit*	*logit*
	ilościowy	ANOVA	regresja	ANCOVA

Kategoryczne:
— nominalne
— porządkowe
Ilościowe:
— przedziałowe
— ilorazowe
— dyskretne
The Framingham Heart Study is a long-term, ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The study began in 1948 with 5,209 adult subjects from Framingham, and is now on its third generation of participants.[1] Prior to it almost nothing was known about the "epidemiology of hypertensive or arteriosclerotic cardiovascular disease".[2] Much of the now-common knowledge concerning heart disease, such as the effects of diet, exercise, and common medications such as aspirin, is based on this longitudinal study. It is a project of the National Heart, Lung, and Blood Institute, in collaboration with (since 1971) Boston University.[3] Various health professionals from the hospitals and universities of Greater Boston staff the project.(patrz tab.2)

Table 2. Framingham Longitudinal Study of Coronary Heart Disease

Choroba wieńcowa	Cholesterol mg/cm^3	Ciśnienie skurczowe $mm\,Hg$			
		<127	127-146	147-166	167+
obecna	< 200	2	3	3	4
	200 - 219	3	2	0	3
	220 - 259	8	1	6	6
	≥ 260	7	12	11	11
brak	< 200	117	121	47	22
	200 - 219	85	98	43	20
	220 - 259	119	209	68	43
	≥ 260	67	99	46	33

0.2. Rozkłady zmiennych kategorycznych

Rozkład Bernouilliego

$$P(Y=1)=\pi,\ P(Y=0)=1-\pi$$

$$E(Y)=\pi,\ V(Y)=\pi(1-\pi)$$

Rozkład dwumianowy
$Y_1, Y_2, \ldots, Y_n$ są niezależne i o identycznym rozkładzie Bernouilliego. $Y = \sum_{i=1}^{n} Y_i$ jest liczbą sukcesów.

$$P(Y=y)=\left(\begin{array}{c} n \\ y \end{array}\right)\pi^y(1-\pi)^{n-y}$$

$$E(Y)=n\pi,\ V(Y)=n\pi(1-\pi)$$

Rozkład wielomianowy

Uogólnienie rozkładu dwumianowego

Każde z n niezależnych, identycznych doświadczeń kończy się jednym z c możliwych wyników.

Niech A_{ij} zdarzenie, że i-te doświadczenie zakończy sie j-tym typem wyniku.

$$Y_{ij} = \begin{cases} 1 & A_{ij} \\ 0 & A'_{ij} \end{cases}$$

$P(Y_{ij} = 1) = \pi_j$.

Zmienne $N_j = \sum_{i=1}^{n} Y_{ij}$ dla $ij = 1, 2, \ldots, c$ mają rozkład wielomianowy.

$$P(N_1 = n_1, N_2 = n_2, \ldots, N_c = n_c) = \frac{n!}{n_1! n_2! \ldots n_c!} \pi_1^{n_1} \pi_2^{n_2} \ldots \pi_c^{n_c}$$

N_j mają rozkład dwumianowy, stąd

$$E(N_j) = n\pi_j,\ V(N_j) = n\pi_j(1 - \pi_j),\ Cov(N_j, N_k) = -n\pi_j\pi_k$$

Rozkład Poissona

$$P(Y = y) = e^{-\mu} \frac{\mu^y}{y!}$$

$$E(Y) = V(Y) = \mu$$

Aproksymacja rozkładu dwumianowego za pomocą rozkładu Poissona

Związek między rozkładem wielomianowym a rozkładem Poissona

Niech $Y_1, Y_2, \ldots$ będą niezależnymi zmiennymi o rozkładzie Poissona, $E(Y_i) = \mu_i$.

$$P\left[(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c) \middle| \sum Y_j = n\right]$$

$$= \frac{P(Y_1 = n_1, Y_2 = n_2, \ldots, Y_c = n_c)}{P(\sum Y_j = n)}$$

$$= \frac{\prod_i \left[\exp(-\mu_i)\, \mu_i^{n_i}/n_i!\right]}{\exp(-\sum \mu_j)(\sum \mu_j)^n/n!} = \frac{n!}{\prod_i n_i!} \prod_i \pi_i^{n_i},$$

gdzie

$$\pi_i = \frac{\mu_i}{\sum \mu_i}$$

Jest to więc rozkład wielomianowy.

Chapter 1

Wnioskowanie statystyczne dla danych kategorycznych

1.1. Funkcje wiarygodności i estymacja największej wiarygodności

Metoda estymacji - *największej wiarygodności.* Przy bardzo słabych założeniach (???) np takich że zbiór parametrów jest skończenie wymiarowy, prawdziwa wartość parametru leży wewnątrz obszaru dopuszczalnego, estymatory mają własności
— są asymptotycznie normalne
— są asymptotycznie zgodne
— są asymptotycznie efektywne (minimalny błąd standardowy)
Funkcja wiarygodności ....

Oznaczenia:
— β - parametr
— $\hat{\beta}$ - estymator ML
— $l(\beta)$- funkcja wiarygodności
— $L(\beta) = \lg(l(\beta))$ - logarytm funkcji wiarygodności
Dla wielu modeli $L(\beta)$ jest wklęsła i minimum jest rozwiązaniem równania

$$\frac{\partial L(\beta)}{\partial \beta} = 0$$

Przy założeniach

ASSUMPTION 1. The derivatives $d \log p/d\theta$, $d^2 \log p/d\theta^2$, and $d^3 \log p/d\theta^3$ esist, for almost all x in an interval A of θ including the true value.

ASSUMPTION 2. At the true value of θ,

$$E\left[\frac{p'(x,\theta)}{p(x,\theta)}\middle|\theta\right] = 0, \qquad E\left[\frac{p''(x,\theta)}{p(x,\theta)}\middle|\theta\right] = 0$$

$$E\left[\frac{p'(x,\theta)^2}{p(x,\theta)}\middle|\theta\right] > 0,$$

where the primes denote differentiation with respect to θ.

ASSUMPTION 3. For every θ in A,

$$\left|\frac{d^3 \log p}{d\theta^3}\right| < M(x), \qquad E[M(x)|\theta] < K,$$

where K is independent of θ.

Let $x_1, \ldots, x_n$ be n independent observations (r.v.'s) and put

$$l(\theta|x) = \sum \log p(x_i, \theta)$$

$$\frac{dl}{d\theta} = \sum \frac{p'(x_i,\theta)}{p(x_i,\theta)}.$$

$cov\left(\hat{\beta}\right)$ jest asymptytycznie odwrotnością macierzy informacyjnej, postaci

$$-E\left(\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \beta_j \partial \beta_k}\right).$$

1.1.1. Estymacja dla rozkładu dwumianowego

$$L(\pi) = \log\left[\pi^{y}(1-\pi)^{n-y}\right] = y\log(\pi) + (n-y)\log(1-\pi).$$

$$\partial L(\pi)/\partial\pi = y/\pi - (n-y)/(1-\pi) = (y-n\pi)/\pi(1-\pi).$$

$$\hat{\pi} = \frac{y}{n}$$

$$-E\left[\partial^2 L(\pi)/\partial\pi^2\right] = E\left[y/\pi^2 + (n-y)/(1-\pi)^2\right] = n/[\pi(1-\pi)].$$

Stąd

$$SE(\pi) = \sqrt{\frac{\pi(1-\pi)}{n}}$$

co daje się wyliczyć bezpośrednio

$$E(\hat{\pi}) = \pi, \qquad \sigma(\hat{\pi}) = \sqrt{\frac{\pi(1-\pi)}{n}}\,.$$

1.2. Trójka: testy Walda - ilorazu wariancji - punktowy (*score test*)

Testowana jest hipoteza $H_1 : \beta \neq \beta_0$, $H_0 : \beta = \beta_0$
Statystyka testowa
$z = \left(\hat{\beta} - \beta_0\right)/\mathrm{SE}$

gdy $\beta = \beta_0$ ma asymptotycznie standardowy rozkład normalny
z^2 ma rozkład χ^2 z jednym stopniem swobody (jest to *statystyka Walda*)
Wielowymiarowy wariant statystyki Walda:
$W = \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right)'\left[\mathrm{cov}(\hat{\boldsymbol{\beta}})\right]^{-1}\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\right).$

W ma wielowymiarowy rozkład χ^2 z liczbą stopni swobody, równą rzędowi macierzy $cov\left(\hat{\beta}\right)$.

Statystyka testu ilorazu wiarygodności:
$-2\log\Lambda = -2\log\left(\ell_0/\ell_1\right) = -2(L_0 - L_1),$

gdzie

$$l_0 = \max\left(l(\beta) : \beta \in H_0\right),$$

$$l_1 = \max\left(l(\beta) : \beta \in H_0 \cup H_1\right),$$

Wilks pokazał, że statystyka testu ilorazu wiarygodności ma asymptotycznie rozkład χ^2 z liczbą stopni swobody, równą różnicy wymiaru przestrzeni $H_0 \cup H_1$ i wymiaru przestrzeni H_0.

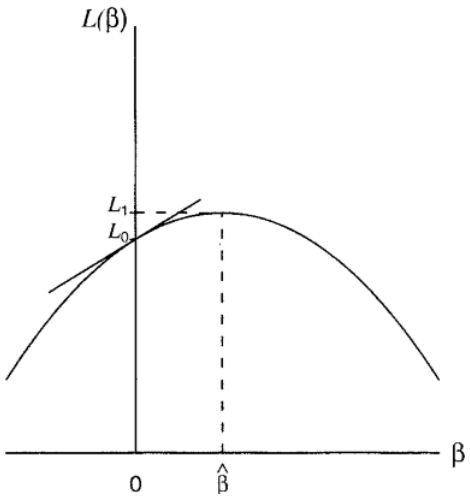
Statystyka punktowa (Fisher,Rao)
Funkcja punktowa
$u(\beta) = \partial L(\beta)/\partial\beta.$

Wtedy

$$\frac{\left[u(\beta_0)\right]^2}{\iota(\beta_0)} = \frac{\left[\partial L(\beta)/\partial\beta_0\right]^2}{-E\left[\partial^2 L(\beta)/\partial\beta_0^2\right]}$$

ma rozkład χ^2 z jednym stopniem swobody. Postasć normalna jest pierwiastkiem z tej statystyki.

Jest wielowymiarowa wersja tego testu.



Wykres funkcji L w przypadku jednowymiarowym.

Test Walda bada iloraz estymatora $\hat{\beta}$ i zakrzywienia funkcji L.

Test punktowy bada nachylenie i krzywiznę $L(\beta)$ w punkcie 0.

Test ilorazu wariancji jest różnicą $-2(L_0 - L_1)$. Jest on najbardziej elastyczny, gdyż uwzględnia zachowanie dwóch wymienionych wcześniej statystyk. Dla umiarkowanych wartości n statystyka ilorazu wiarygodności jest bardziej stabilna niz statystyka Walda.

1.3. Przedziały ufności

Przedziały ufności są zazwyczaj bardziej czytelne dla odbiorcy niż testowanie hipotez.

Można je otrzymać przez zastowanie odpowiedniości: test na poziomie α odpowiada przedziałowi ufności na poziomie $1-\alpha$, otrzymanemu przez rozwiązanie nierówności ze statystyką testową i wartością krytyczną testu.

Przedział ufności Walda powstaje z rozwiązania nierówności $\left|\hat{\beta} - \beta_0\right| / SE < z_{\alpha/2}$ czyli ma postać $\hat{\beta} \pm z_{\alpha/2} SE$.

Przedział ufności ilorazu wiarygodności spełnia nierówność $-2\left[L(\beta_0) - L\left(\hat{\beta}\right) < \chi_1^2(\alpha)\right]$. Ten przedział jest zalecany dla małych i umiarkowanych wartości n.

1.4. Wnioskowanie statystyczne dla parametru w rozkładzie dwumianowym

Statystyka Walda

$$z_W = \frac{\hat{\pi} - \pi_0}{\mathrm{SE}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1-\hat{\pi})/n}}$$

Statystyka punktowa

$$u(\pi_0) = \frac{y}{\pi_0} - \frac{n-y}{1-\pi_0}, \qquad \iota(\pi_0) = \frac{n}{\pi_0(1-\pi_0)}$$

postać normalna

$$z_S = \frac{u(\pi_0)}{[\iota(\pi_0)]^{1/2}} = \frac{y - n\pi_0}{\sqrt{n\pi_0(1-\pi_0)}} = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

Statystyka ilorazu wiarygodności

$$L_0 = y \log \pi_0 + (n-y) \log (1-\pi_0)$$

$$L_1 = y \log \hat{\pi} + (n-y) \log (1-\hat{\pi})$$

$$-2(L_0 - L_1) = 2\left(y \log \frac{\hat{\pi}}{\pi_0} + (n-y) \log \frac{1-\hat{\pi}}{1-\pi_0} \right)$$

Równoważnie

$$-2(L_0 - L_1) = 2\left(y \log \frac{y}{n\pi_0} + (n-y) \log \frac{n-y}{n-n\pi_0} \right)$$

Można to wyrazić formułą

$$2\left(sukcesy \log \frac{sukcesy}{E\,(sukcesy)} + porazki \log \frac{porazki}{E\,(porazki)} \right)$$

1.4.1. Przedziały ufności dla rozkładu dwumianowego

Odwrócenie testu Walda

$$\hat{\pi} \pm z_{\alpha/2} \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}$$

Przedział ten był jednym z pierwszych w historii (Laplace 1812) przedziałów ufności.

Ma on jednak złe własności i może być poprawiony (Wilson, Agresti i Coull - patrz zadania na ćwiczenia)

Odwrócenie testu ilorazu wiarygodności

Przykład

Wśród losowo wybranych 25 osób nikt nie okazał się wegetarianinem. Jakie jest prawdopodobiństwo spotkania wegetarianina w populacji?

Stosując przedział Walda mamy $\hat{\pi} = 0$ a więc przedział będzie miał postać $(0,0)$!!!

95% przedział ufności powstały z odwrócenia testu ilorazu wiarygodności:

$$\begin{aligned} -2(L_0 - L_1) &= -2\left[L(\pi_0) - L(\hat{\pi}) \right] \\ &= -50 \log(1-\pi_0) \leq \chi^2_1(0.05) = 3.84. \end{aligned}$$

,

gdzie $L\,(\pi) = 25 \ln (1-\pi)$

Przedział ufności ma więc postać

$$(0, 1 - \exp(-3.84/50)) = (0, 0.074)$$

■

1.5. Wnioskowanie statystyczne dla parametrów rozkładu wielomianowego

Logarytm funkcji wiarygodności dla jądra rozkładu:

$$L\,(\boldsymbol{\pi}) = \sum_{j=1}^{c} n_j \ln (\pi_j)\,,$$

$$\sum_{j=1}^{c} \pi_j = 1$$

Rozwiązanie jest postaci

$$\hat{\pi_j} = \frac{n_j}{n},\ n = \sum_{j=1}^{c} n_j$$

1.5.1. Statystyka χ^2 Pearsona (Pearson, 1900)

Rozważa się problem

$$H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$$
$$H_1 : \boldsymbol{\pi} \neq \boldsymbol{\pi}_0$$

Statystyka testowa

$$\chi^2 = \frac{\sum_{j=1}^{c} (n_j - \mu_j)^2}{\mu_j}$$

gdzie $\mu_j = n\pi_{j0}$.

Niech χ_0^2 będzie zaobserwowaną wartoscią próbki. p-wartość testu jest równe prawdopodobieństwu $P\left(\chi^2 \geq \chi_0^2\right)$, które można obliczyć jako sumę prawdopodobieństw w rozkładzie wielomianowym, odpowiadającym wszystkim układom $n_1, n_2, \ldots, n_c$ że $\sum n_j = n$oraz spełniony jest warunek $\chi^2 \geq \chi_0^2$.

Prawdopodobieństwo to można w przybliżeniu obliczyć, dla dużych n_j z rozkładu χ^2z $c-1$ stopniami swobody : $p = P\left(\chi_{c-1}^2 \geq \chi_0^2\right)$.

> Among its many applications, Pearson's test was used in genetics to test Mendel's theories of natural inheritance. Mendel crossed pea plants of pure yellow strain with plants of pure green strain. He predicted that second-generation hybrid seeds would be 75% yellow and 25% green, yellow being the dominant strain. One experiment produced $n = 8023$ seeds, of which $n_1 = 6022$ were yellow and $n_2 = 2001$ were green. The expected frequencies for H_0: $\pi_{10} = 0.75$, $\pi_{20} = 0.25$ are $\mu_1 = 8023(0.75) = 6017.25$ and $\mu_2 = 2005.75$. The Pearson statistic $X^2 = 0.015$ (df = 1) has a P-value of $P = 0.90$. This does not contradict Mendel's hypothesis.
>
> Mendel performed several experiments of this type. In 1936, R. A. Fisher summarized Mendel's results. He used the reproductive property of chi-squared: If $X_1^2, \ldots, X_k^2$ are independent chi-squared statistics with degrees of freedom $\nu_1, \ldots, \nu_k$, then $\sum_i X_i^2$ has a chi-squared distribution with df = $\sum_i \nu_i$. Fisher obtained a summary chi-squared statistic equal to 42, with df = 84. A chi-squared distribution with df = 84 has mean 84 and standard deviation $(2 \times 84)^{1/2} = 13.0$, and the right-tailed probability above 42 is $P = 0.99996$. In other words, the chi-squared statistic was so small that the fit seemed *too* good.

Fisher pokazał, że

$$\chi^2 = \sum_{j=1}^{c} r_j^2,$$

gdzie wielkości, zwane *resztami Pearsona*

$$r_j = \frac{n_j - \mu_j}{\sqrt{\mu_j}}$$

mają asymptotycznie standardowy rozkład normalny.

1.5.2. Test χ^2 ilorazu wiarygodności

Uwzględniając estymatory NW dla hipotez $H_0 : \beta = \beta_0$ przeciwko hipotezie $H_1 : \beta \neq \beta_0$ iloraz wiarygodności jest postaci:

$$\Lambda = \frac{\prod_j \pi_{j0}^{n_j}}{\prod_j (n_j/n)^{n_j}}$$

Statystyka ilorazu wiarygodności wyraża się wzorem:

$$G^2 = -2 \log \Lambda = 2 \sum n_j \log \left(n_j / n\pi_{j0} \right)$$

Przestrzeń parametryczna zawiera wektor$\pi = [\pi_1, \pi_2, \ldots, \pi_c]$ z warunkiem $\sum \pi_j = 1$ a więc jest wymiaru $c-1$. W przypadku hipotezy H_0 zawiera ona pojedynczy wektor π_0 a więc ma wymiar 0. Asymptotycznie,w przypadku hipotezy H_0,G^2 ma rozkład χ^2 z $c-1$ stopniami swobody.

Podobnie asymptotycznie,w przypadku hipotezy H_0, test Pearsona χ^2 ma rozkład χ^2 z $c-1$ stopniami swobody. Co więcej, mozna pokazać, że w tym przypadku, $\chi^2 - G^2$ zbiega według prawdopodobieństwa do 0.

Gdy zachodzi hipoteza H_1, obie statystyki rosną proporcjonalnie do n (!!!) i nawet dla dużych wartosci n nie osiągaja podobnych wartości.

Dla ustalonego c statystyka χ^2 Pearsona zbiega do rozkładu χ^2 szybciej, niż statystyka G^2.

Przybliżenie przez rozkład χ^2 rozkładu statystyki G^2 jest słabe, gdy zachodzi nierówność $n < 5c$.

1.5.3. Testy z prawdopodobieństwami, zależnymi od parametru

W estymacji prawdopodobieństwa π mogą być zależne od parametru θ, $\pi = \pi(\theta)$.W takim przypadku estymacja NW polega na znalezieniu estymatora $\hat{\theta}$ parametru θ, wyznaczającego estymator NW parametru π: $\hat{\pi} = \pi\left(\hat{\theta}\right)$. Wpływa to na liczbę stopni swobody testu χ^2 Pearsona[1]: jeżeli $\dim(\theta) = p$ to liczba stopni swobody t4estu jest równa $df = (c-1) - p$.

Przykład

Cielęta mleczne były obserwowane, czy w ciągu pierwszych 60 dni życia przeszły zapalenie płuc, a następnie czy wśród tych, które zachorowały, wystąpiło w ciągu kolejnych dwóch tygodni wtórne zapalenie płuc. Pytanie: czy prawdopodobieństwo wtórnego zachorowanie jest inne od prawdopodobieństwa zachorowania pierwotnego.

	Secondary Infection[a]	
Primary Infection	Yes	No
Yes	30 (38.1)	63 (39.0)
No	0 (—)	63 (78.9)

Source: Data courtesy of Thang Tran and G. A. Donovan, College of Veterinary Medicine, University of Florida.

[a] Values in parentheses are estimated expected frequencies.

Uwaga na zero strukturalne!!

Niech π_{ij} oznacza prawdopodobieństwo konfiguracji zdarzeń w tabeli. Wtedy

$$H_1 : \pi_{1+} \neq \frac{\pi_{11}}{\pi_{1+}}$$

przeciwko

$$H_0 : \pi_{1+} = \frac{\pi_{11}}{\pi_{1+}}$$

Oznaczmy przez $\theta = \pi_{1+}$. Wtedy, przy założeniu H_0 rozkład prawdopodobieństwa w tabeli zależy wyłącznie od parametru θ:

Infekcja pierwotna	Infekcja wtórna		
	Tak	Nie	Razem
Tak	θ^2	$\theta(1-\theta)$	θ
Nie	-	1-θ	1-θ

[1] Pearson nie zdawał sobie z tego sprawy! Poprawne rozwiązanie tego problemu należy do Fishera.

Jądro funkcji wiarygodności ma postać:

$$\left(\theta^2\right)^{n_{11}} \left(\theta\left(1-\theta\right)\right)^{n_{12}} \left(1-\theta\right)^{n_{22}}.$$

Logarytm tego jądra

$$L\left(\theta\right) = \left(2n_{11}+n_{12}\right)\log\left(\theta\right) + \left(n_{12}+n_{22}\right)\log\left(1-\theta\right) = \left(n_{11}+n_{1+}\right)\log\left(\theta\right) + n_{+2}\log\left(1-\theta\right).$$

Po przyrównaniu pochodnej do 0 otrzymamy

$$\hat{\theta} = \frac{n_{11}+n_{1+}}{n_{11}+n_{1+}+n_{+2}} = \frac{30+93}{30+93+126} = 0.49$$

Tabela wartości oczekiwanych

Infekcja pierwotna	Infekcja wtórna		
	Tak	Nie	Razem
Tak	$0.49^2 * 156$	$0.49 * 0.51 * 156$	$0.49 * 156$
Nie	-	0.51*156	0.51*156

Infekcja pierwotna	Infekcja wtórna		
	Tak	Nie	Razem
Tak	38.1	39.0	77.1
Nie	-	78.9	78.9

Reszty Pearsona

Infekcja pierwotna	Infekcja wtórna	
	Tak	Nie
Tak	−1.31	3.84
Nie	-	−1.79

$\chi^2 = 19.7$ z $df = 3-1-1 = 1$ stopni swobody. Wartość p tego testu jest równa 0.000009 co w zdecydowanym stopniu potwierdza podejrzenie, ze prawdopodobieństwa pierwotnego i wtórnego zakażenia są istotnie różne.

Jedyna istotna różnica Pearsona (3.84) wskazuje co jest tego przyczyną. Istotnie więcej jest cieląt, które przeszły pierwotne zakażenie i nie zakaziły się powtórnie, prawdopodobnie na skutek uodpornienia się.

■