

1 Wykład 8

Test wartości wariancji dla znanej średniej

Niech $X^T = (x_1, x_2, \dots, x_n)$ będzie próbą prostą z rozkładu $\mathcal{N}(\mu, \sigma)$, średnia μ jest znana. Testujemy hipotezy:

$$\begin{aligned} H_0 &: \sigma = \sigma_0, \\ H_1 &: \sigma \neq \sigma_0 \end{aligned}$$

Iloraz wariancji w tym przypadku ma postać:

$$L_x(H_0, H_1) = \frac{\sup \{l(x, \mu, \sigma) : \sigma\}}{l(x, \mu, \sigma_0)} = \frac{\left(2\pi \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2\right)^{-n/2} \exp(-n/2)}{(2\pi\sigma_0^2)^{-n/2} \exp\left(-\sum_{i=1}^n (x_i - \mu_0)^2 / 2\sigma_0^2\right)} =$$

Oznaczmy $u \stackrel{def}{=} \sum_{i=1}^n (x_i - \mu_0)^2 / n\sigma_0^2$. Wtedy

$$2 \log L_x(H_0, H_1) = n(u - 1 - \log u)$$

jest rosnącą funkcją u . Zbiór krytyczny będzie miał więc postać:

$$C = \left\{ X : \frac{\sum_{i=1}^n (x_i - \mu_0)^2}{n\sigma_0^2} > k \right\}$$

gdzie k jest tak wybrane, aby $P\left(\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{n\sigma_0^2} > k\right) = \alpha$. Jak nietrudno zauważyć, zmienna losowa $T = \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{n\sigma_0^2}$ ma rozkład χ_n^2 (z n stopniami swobody), gdy hipoteza H_0 jest prawdziwa, więc k należy wyznaczyć z dystrybucji rozkładu χ_n^2 .

Test równości dwóch średnich w rozkładzie normalnym ze znaną wariancją

Tym razem mamy dwie niezależne próby proste: $X^T = (x_1, x_2, \dots, x_m)$ i $Y^T = (y_1, y_2, \dots, y_n)$. Pierwsza pochodzi z rozkładu normalnego $\mathcal{N}(\mu_1, \sigma)$, druga z rozkładu $\mathcal{N}(\mu_2, \sigma)$. Testujemy hipotezy:

$$\begin{aligned} H_0 &: \mu_1 = \mu_2, \\ H_1 &: \mu_1 \neq \mu_2. \end{aligned}$$

Wyliczymy iloraz wiarygodności dla tych hipotez

$$\begin{aligned} L_x(H_0, H_1) &= \frac{\sup \{l(x, \mu_1, \sigma) l(y, \mu_2, \sigma) : \mu_1, \mu_2\}}{\sup \{l(x, \mu, \sigma) l(y, \mu, \sigma) : \mu\}} = \\ &= \frac{(2\pi\sigma^2)^{-(m+n)/2} \exp\left(-\sum_{i=1}^m (x_i - \bar{x})^2 / 2\sigma^2\right)}{(2\pi\sigma^2)^{-(m+n)/2} \exp\left(-\sum_{i=1}^m \left(x_i - \frac{m\bar{x} + n\bar{y}}{m+n}\right)^2 / 2\sigma^2\right)} \cdot \\ &\cdot \frac{\exp\left(-\sum_{i=1}^n (y_i - \bar{y})^2 / 2\sigma^2\right)}{\exp\left(-\sum_{i=1}^n \left(y_i - \frac{m\bar{x} + n\bar{y}}{m+n}\right)^2 / 2\sigma^2\right)} = \exp\left(\frac{1}{2\sigma^2} \frac{mn}{m+n} (\bar{x} - \bar{y})^2\right) \end{aligned}$$

jest więc rosnącą funkcją $|\bar{x} - \bar{y}|$. Zbiór krytyczny ma postać:

$$C = \{X, Y : |\bar{x} - \bar{y}| > k\}$$

Zmienna losowa $\bar{X} - \bar{Y}$, gdy prawdziwa jest hipoteza zerowa, ma rozkład $\mathcal{N}\left(0, \sigma\sqrt{\frac{1}{m} + \frac{1}{n}}\right)$. W związku z tym dla testowania na poziomie α zachodzić musi ciąg równości:

$$\begin{aligned} \alpha &= P(|\bar{x} - \bar{y}| > k) = P\left(\frac{|\bar{x} - \bar{y}|}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}} > \frac{k}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}\right) = \\ &= 2\left(1 - \Phi\left(\frac{k}{\sigma\sqrt{\frac{1}{m} + \frac{1}{n}}}\right)\right) \end{aligned}$$

Tak więc $k = z_\alpha \sigma\sqrt{\frac{1}{m} + \frac{1}{n}}$ gdzie $\Phi(z_\alpha) = 1 - \alpha/2$ i Φ jest dystrybucją rozkładu $\mathcal{N}(0, 1)$.

Test zgodności

Doświadczenie ma k możliwych wyników o prawdopodobieństwach $p_1, p_2, \dots, p_k$. Niezależnie, n -krotnie powtarzamy to doświadczenie. Otrzymamy wektor wyników tej obserwacji postaci $X^T = (x_1, x_2, \dots, x_n)$ gdzie x_i oznacza liczbę powtórzeń, w których pojawił się i -ty wynik doświadczenia. Oczywiście $\sum_{i=1}^n x_i = n$. Chcemy na podstawie tych obserwacji testować hipotezę, że p_i przyjmują konkretne wartości. Dopuszczamy możliwość, że p_i mogą być funkcjami nieznanymi parametrów θ ($p_i = p_i(\theta)$). Chcemy więc testować hipotezę:

$$H_0 : p_i = p_i(\theta) \text{ dla } \theta \in \Theta_0,$$

$$H_1 : \text{są dowolne}$$

Rozkład wektora X jest wielomianowy:

$$P(X^T = (x_1, x_2, \dots, x_n)) = n! \prod_{i=1}^k \frac{p_i^{x_i}}{x_i!}$$

Aby obliczyć logarytm ilorazu wiarygodności musimy obliczyć

$$\log L_x(H_1) = \sup \left\{ \log \left(n! \prod_{i=1}^k \frac{p_i^{x_i}}{x_i!} \right) : 0 \leq p_i, \sum_{i=1}^k p_i = 1 \right\}$$

Zgodnie z zasadą mnożników Lagrange'a wystarczy szukać maksimum wyrażenia

$$\sum_{i=1}^k x_i \log p_i + \lambda \sum_{i=1}^k p_i$$

względem zmiennych p_i . Maksimum jest osiągnięte w punkcie $\hat{p}_i = x_i/n$ i wynosi

$$\log L_x(H_1) = \log \left(\frac{n!}{\prod_{i=1}^k x_i!} \right) + \sum_{i=1}^k x_i \log \left(\frac{x_i}{n} \right)$$

Podobnie,

$$\log L_x(H_0) = \log \left(\frac{n!}{\prod_{i=1}^k x_i!} \right) + \sum_{i=1}^k x_i \log(p_i(\hat{\theta}))$$

gdzie $\hat{\theta}$ jest estymatorem największej wiarygodności parametru θ w obszarze Θ_0 . w takim razie

$$\begin{aligned} 2 \log L_x(H_0, H_1) &= 2(\log L_x(H_1) - \log L_x(H_0)) = \\ &= 2 \left(\sum_{i=1}^k x_i \log \left(\frac{x_i}{n} \right) - \sum_{i=1}^k x_i \log(p_i(\hat{\theta})) \right) = \\ &= 2 \left(\sum_{i=1}^k x_i \log \left(\frac{x_i}{np_i(\hat{\theta})} \right) \right) \end{aligned}$$

Wielkości $e_i \stackrel{def}{=} np_i(\hat{\theta})$ są *oczekiwanymi*, zaś x_i są obserwowanymi liczbami za-
jść i -tego wyniku podczas n powtórzeń doświadczenia. Wyrażenie $D(x, e) \stackrel{def}{=} 2 \left(\sum_{i=1}^k x_i \log \left(\frac{x_i}{e_i} \right) \right)$ nazywamy *odchyleniem rozkładu teoretycznego i obser-
wowanego*. Zgodnie z twierdzeniem o ogólnym, asymptotycznym teście ilorazu
wariancji $D(x, e)$ ma rozkład χ^2 z $k-1-p$ stopniami swobody, gdzie p jest liczbą
stopni swobody dla estymacji $\theta \in \Theta_0$. Taka liczba stopni swobody wynika z
faktu, że w obszarze hipotezy H_1 występuje k nieznanymi parametrów z jednym
ograniczeniem $\sum_{i=1}^k p_i = 1$. Tak więc zbiór krytyczny będzie miał postać:

$$C = \{X : D(x, e) > k\}$$

gdzie k musi spełniać warunek

$$P(\chi_{k-1-p}^2 > k) = \alpha$$

W praktyce stosuje się przybliżoną postać odchylenia rozkładu¹, zwaną *statystyką*
 χ^2 *Pearsona*.

Z rozwinięcia Taylora funkcji $\log(1+x)$ otrzymamy przybliżoną równość dla
małych x

$$\log(1+x) \approx x - \frac{x^2}{2}$$

Tak więc zauważywszy, że

$$\frac{x_i}{e_i} = 1 + \frac{x_i - e_i}{e_i} \stackrel{def}{=} 1 + \frac{\delta_i}{e_i}$$

¹chyba z powodu niedogodności obliczania logarytmów

oraz, że dla dużych n wyrażenie $\delta_i = x_i - e_i$ ma małe wartości² otrzymamy

$$\begin{aligned} D(x, e) &= 2 \left(\sum_{i=1}^k x_i \log \left(\frac{x_i}{e_i} \right) \right) \approx 2 \left(\sum_{i=1}^k x_i \log \left(1 + \frac{\delta_i}{e_i} \right) \right) \approx \\ &\approx 2 \left(\sum_{i=1}^k (\delta_i + e_i) \left(\frac{\delta_i}{e_i} - \frac{1}{2} \left(\frac{\delta_i}{e_i} \right)^2 \right) \right) = 2 \left(\sum_{i=1}^k \frac{1}{2} \frac{\delta_i^2}{e_i} - \frac{1}{2} \frac{\delta_i^3}{e_i^2} + \delta_i \right) = \\ &= \sum_{i=1}^k \frac{\delta_i^2}{e_i} - \sum_{i=1}^k \frac{\delta_i^3}{e_i^2} + 2 \sum_{i=1}^k \delta_i \approx \sum_{i=1}^k \frac{\delta_i^2}{e_i} + 2 \sum_{i=1}^k (x_i - e_i) = \sum_{i=1}^k \frac{\delta_i^2}{e_i} + 2(n - n) = \\ &= \sum_{i=1}^k \frac{(x_i - e_i)^2}{e_i} \end{aligned}$$

Statystyka χ^2 Pearsona wyraża się wzorem

$$\chi^2 = \sum_{i=1}^k \frac{(x_i - e_i)^2}{e_i}$$

i ma on w przybliżeniu, dla dużych n rozkład χ^2 z $k - 1 - p$ stopniami swobody³
Wzór Pearsona może mieć też inną interpretację

$$\chi^2 = \sum_{i=1}^k \frac{(x_i - e_i)^2}{e_i} = \sum_{i=1}^k e_i \left(\frac{x_i - e_i}{e_i} \right)^2 = n \sum_{i=1}^k p_i \left(\frac{x_i - e_i}{e_i} \right)^2$$

Tak więc χ^2/n jest oczekiwanym kwadratowym błędem względnym licznosci obserwowanych x_i względem oczekiwanych e_i .

Co więcej, asymptotycznie, gdy prawdziwa jest hipoteza zerowa, wielkości $\frac{x_i - e_i}{\sqrt{e_i}}$ są w przybliżeniu niezależne i mają rozkład normalny $\mathcal{N}(0, 1)$ ⁴. Jeżeli więc w sumie χ^2 wystąpią składniki większe od $1.96^2 = 3.85$ lub jeszcze bardziej - większe od $2.58^2 = 6.66$ ⁵ to te wyniki są najbardziej odpowiedzialne za niezgodność z hipotezą zerową. Pozwala to na bardziej szczegółową interpretację otrzymanych wyników.

Przykład 1.1 Weźmy dane Mendla, o których była mowa na początku rozdziału o testowaniu hipotez

typ groszku	liczba obserwacji	prawdopodobieństwo teoretyczne	oczekiwana liczba ziaren
gładki żółty	315	9/16	312.75
gładki zielony	108	3/16	104.25
pomarszczony żółty	102	3/16	104.25
pomarszczony zielony	31	1/16	34.75

W tej tablicy mamy licznosci obserwowane w drugiej kolumnie i oczekiwane, zgodne z modelem genetycznym Hardy'ego-Weinberga w czwartej. Zauważmy,

² z prawa wielkich liczb

³ w praktyce uważa się, że n jest dostatecznie duże, gdy $e_i \geq 5$ dla wszystkich i

⁴ to tłumaczy, że rozkład statystyki Pearsona jest χ^2

⁵ w rozkładzie normalnym standardowym wartości większe co do modułu od 1.96 występują z prawdopodobieństwem 0.05 a większe co do modułu od 2.58 z prawdopodobieństwem 0.01

że liczności oczekiwane są większe od 5 co pozwala na posłużenie się przybliżeniem Pearsona. Policzmy jednak dla porównania odchylenie i statystykę χ^2 Pearsona:

$$\begin{aligned} D(x, e) &= 2 \left(315 \log \left(\frac{315}{312.75} \right) + 108 \log \left(\frac{108}{104.25} \right) + 102 \log \left(\frac{102}{104.25} \right) + 31 \log \left(\frac{31}{34.75} \right) \right) = \\ &= 0.6184 \end{aligned}$$

Podobnie,

$$\begin{aligned} \chi^2 &= \frac{(315 - 312.75)^2}{312.75} + \frac{(108 - 104.25)^2}{104.25} + \frac{(102 - 104.25)^2}{104.25} + \frac{(31 - 34.75)^2}{34.75} = \\ &= 0.01619 + 0.1349 + 0.0486 + 0.4047 = 0.6043 \end{aligned}$$

Czy możemy odrzucić hipotezę o zgodności z modelem Hardy'ego-Weinberga? Obliczmy poziom krytyczny dla naszych danych:

$$\begin{aligned} P(\chi_{4-1-0}^2 > 0.6184) &< P(\chi_3^2 > 0.584) = 0.9, \\ P(\chi_3^2 > 0.6184) &> P(\chi_3^2 > 6.251) = 0.1 \end{aligned}$$

Jest on zawarty między 0.9 a 0.1⁶ więc zbyt ryzykowne byłoby odrzucenie hipotezy zerowej. Jak widać, żaden ze składników w statystyce Pearsona nie jest większy od 3.85 co wskazuje, że żaden z wyników nie groziłby zgodności obserwacji z hipotezą Hardy'ego-Weinberga. ¥

⁶dokładnie, 0.892