Test zgodności χ^2

- Dotychczas zajmowaliśmy się testami parametrycznymi dotyczącymi wartości nieznanych parametrów $\mu, \sigma^2, p, \rho, \ldots$
- zakładaliśmy wtedy, że próbka pochodzi z konkretnego rozkładu pr. (np. normalny, dwumianowy, ...)
- czasami takie założenie jest dokładnie tym co chcemy przetestować, np. czy próbka pochodzi z rozkładu normalnego albo Poissona, ...
- takie testy nazywamy testami zgodności
- testy zgodności należą do grupy testów nieparametrycznych
- ważnym przykładem testu zgodności jest test χ^2

Przykład 18.8 (Test zgodności z rozkładem dwumianowym)**.**
Pewien eksperyment zakończył się $n_1 = 470$ razy wynikiem W_1 oraz $n_2 = 1030$ razy wynikiem W_2

$$H_0: \text{ jest to rozkładem } \left(\frac{1}{3}, \frac{2}{3}\right) \text{ na } \{W_1, W_2\}$$

Szukamy statystyki która jest miarą dopasowania wartości n_1, n_2 do spodziewanych 500 i 1000

$$(n_1 - 500)^2 + (n_2 - 1000)^2$$

$$Z^2 = \frac{(n_1 - 500)^2}{500} + \frac{(n_2 - 1000)^2}{1000}$$

$$n_1 + n_2 = 1500 \to n_2 = 1500 - n_1$$

$$Z^2 = \frac{(n_1 - 500)^2}{500} + \frac{(1500 - n_1 - 1000)^2}{1000} = \frac{(n_1 - 500)^2}{500} + \frac{(n_1 - 500)^2}{1000} = \frac{(n_1 - 500)^2}{1000/3}$$

$$n_1 \sim Bin(n = 1500, p = 1/3) \to N(\mu = np = 500, \sigma^2 = np(1-p) = 1000/3)$$

$$Z^2 = Y^2, \qquad Y \sim N(0,1)$$

$$Z^2 \sim \chi^2_1$$

Przykład 18.9 (Test zgodności z rozkładem dwumianowym).
Pewien eksperyment zakończył się $n_1 = 470$ razy wynikiem W_1 oraz $n_2 = 1030$ razy wynikiem W_2

$$H_0: \text{ jest to rozkładem } \left(\frac{1}{3}, \frac{2}{3}\right) \text{ na } \{W_1, W_2\}$$

$$Z^2 = \frac{(470-500)^2}{500} + \frac{(1030-1000)^2}{1000} = 2,7$$
$$C = (u(\alpha, 1), \infty)$$

Dla $\alpha = 0,05$ mamy $C = (3,81, \infty)$ zatem nie mamy podstaw do odrzucenia H_0

Twierdzenie 18.3 (Test zgodności χ^2).
Załóżmy, ze wynik eksperymentu możemy przypisać do r różnych kategorii, z prawdopodobieństwem odpowiednio $p_1, p_2, ..., p_r$ jeśli eksperyment powtórzymy n razy a przezn_k oznaczymy ilość powtórzeń eksperymentu którego wynik możemy przypisać do kategorii k, gdzie $k = 1, 2, ..., r$ to:

$$\sum_{k=1}^{r} \frac{(n_k - np_k)^2}{np_k} \quad \sim \quad \chi^2_{r-1}$$

Wynik ten często przedstawiamy dla dowolnych rozkładów w postaci:

$$\chi^2 = \sum_{k=1}^{r} \frac{(O_k - E_k)^2}{E_k}$$

gdzie O_k jest częstością obserwowaną a E_k jest częstością spodziewaną. Uwaga: przybliżenie jest słuszne, jeśli $E_k \geqslant 5$ dla wszystkich k.

Przykład 18.10 (Test zgodności z rozkładem wielomianowym).
Test hipotezy:

$$H_0: \text{ rozkładem jest } (p_1, p_2, ..., p_r)$$
$$Z^2 = \sum_{k=1}^{r} \frac{(n_k - np_k)^2}{np_k} \sim \chi^2_{r-1}$$
$$C = (u(\alpha, r-1), \infty)$$

Przykład 18.11 (Test zgodności z rozkładem wielomianowym).
W wyniku $n = 1000$ rzutów kostką do gry uzyskano następujące liczby wyrzuconych oczek: 1-173, 2-149, 3-179, 4-181, 5-155, 6-163. Czy kostka jest rzetelna?

$$H_0: \text{ rozkładem jest } (1/6, 1/6, ..., 1/6)$$
$$\chi^2 = \frac{(173-166.7)^2}{166.7} + \frac{(149-166.7)^2}{166.7} + \cdots = 5,15$$
$$C = (u(0.05, 5), \infty) = (11,07, \infty)$$

Nie mamy podstaw do kwestionownia rzetelności kostki.

Przykład 18.12 (Test zgodności z rozkładem Poissona).
Poniższa tabela przedstawia rozkład ilości silnych trzęsień ziemi w ciągu roku: 0-15, 1-13, 2-4, 3-1, >3 - 0.

$$H_0: \text{ rozkładem jest } Poi(\mu = 1.5)$$

$$E_0 = 7.4; E_1 = 11; E_2 = 8.3; E_3 = 4.1; E_4 = 1.6; ...$$

Kategoria	0	1	2	$\geqslant 3$
Obserwowne	15	13	4	1
Spodziewane	7,4	11	8,3	6,3

$$\chi^2 = 14,9$$

$$C = (u(0.05, 3), \infty) = (7,82, \infty)$$

Możemy odrzucić hipotezę zerową.

Twierdzenie 18.4 (Test zgodności χ^2)**.**
Załóżmy, ze wynik eksperymentu możemy przypisać do r różnych kategorii, z prawdopodobieństwem odpowiednio $p_1(\theta), p_2(\theta), ..., p_r(\theta)$ to jeśli $\hat{\theta}$ jest estymatorem największej wiarygodności parametru θ to:

$$\sum_{k=1}^{r} \frac{(n_k - np_k(\hat{\theta}))^2}{np_k(\hat{\theta})} \quad \sim \quad \chi^2_{r-2}$$

Jeśli ilość parametrów wynosi p i każdy zastępujemy jego estymatorem to

$$\sum_{k=1}^{r} \frac{(O_k - E_k)^2}{E_k} \sim \chi^2_{r-1-p}$$

Przykład 18.13 (Test zgodności z rozkładem Poissona)**.**
Poniższa tabela przedstawia rozkład ilości silnych trzęsień ziemi w ciągu roku: 0-15, 1-13, 2-4, 3-1, >3 - 0.

$$H_0: \text{ rozkładem jest } Poi(\hat{\mu} = \frac{24}{33} = 0,73)$$

Kategoria	0	1	$\geqslant 2$
Obserwowne	15	13	5
Spodziewane	15,9	11,6	5,5

$$\chi^2 = 0,27$$

$$C = (u(0.05, 1), \infty) = (3,84, \infty)$$

Rozkład jest zgodny z rozkładem Poissona.
Test zgodności χ^2 możemy również stosować do testowania rozkładów ciągłych, dzieląc próbę losową na kategorie (np. różne przedziały ciągłej zmiennej los.)

Przykład 18.14 (Test zgodności z rozkładem ciągłym)**.**
Poniższa tabela przedstawia rozkład ilości studentów osiągających dane wyniki z egzaminu:

Kategoria	0-59	60-69	70-79	80-89	90 -100
Obserwowne	12	36	90	44	18

$$H_0: \text{ rozkładem jest } N(\mu = 75, \sigma = 8)$$

Kategoria	0-59	60-69	70-79	80-89	90 -100
Obserwowne	12	36	90	44	18
Spodziewane	5,7	43,8	94,4	49,5	7,0

$$\chi^2 = 26,22$$

$$C = (u(0.05, 4), \infty) = (7,11, \infty)$$

Możemy odrzucić hipoteze H_0.

19 Testowanie hipotez dla dwóch próbek

- Konstrukcja testów hipotez dotyczących dwóch niezależnych próbek normalnych przepiega analogicznie do konstrukcji przedziałów ufności.
- Niech $\vec{x}_1$ będzie realizacją n_1-wymiarowego wektora losowego $\vec{X}_1$ przy czym $X_{1,i} \sim N(\mu_1, \sigma_1^2)$;
- Niech $\vec{x}_2$ będzie realizacją n_2-wymiarowego wektora losowego $\vec{X}_2$ przy czym $X_{2,i} \sim N(\mu_2, \sigma_2^2)$;

Jeśli hipoteza dotyczy wartości oczekiwanych to odpowiednią statystyką jest:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu = \mu_1 - \mu_2, \sigma^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Rozkłady normalne o znanych odchyleniach: σ_1**,** σ_2

Jeśli znane są: σ_1, σ_2 to znamy $\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = (\sigma^\star)^2$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma^\star} \sim N(0, 1)$$

$$H_0 : \mu_1 - \mu_2 = D_0$$

Zbiorami krytycznymi są:

- dla $H_1 : \mu_1 - \mu_2 \neq D_0$ $C = (-\infty, -z(\alpha/2)) \cup (z(\alpha/2), \infty)$
- dla $H_1 : \mu_1 - \mu_2 < D_0$ $C = (-\infty, -z(\alpha))$
- dla $H_1 : \mu_1 - \mu_2 > D_0$ $C = (z(\alpha), \infty)$

Próbki o dużej liczebności i dowolnych rozkładach

Jeśli znane są: σ_1, σ_2 to dla $n_1 > 30$ i $n_2 > 30$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sigma^\star} \sim N(0, 1) \quad \sigma^\star = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

Jeśli nie znamy σ_1, σ_2 to dla $n_1 > 100$ i $n_2 > 100$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s^\dagger} \sim N(0, 1); \quad s^\dagger = \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

$$H_0 : \mu_1 - \mu_2 = D_0$$

Zbiorami krytycznymi są:

- dla $H_1 : \mu_1 - \mu_2 \neq D_0$ $C = (-\infty, -z(\alpha/2)) \cup (z(\alpha/2), \infty)$
- dla $H_1 : \mu_1 - \mu_2 < D_0$ $C = (-\infty, -z(\alpha))$
- dla $H_1 : \mu_1 - \mu_2 > D_0$ $C = (z(\alpha), \infty)$

Próbki o dużej liczebności z rozkładów dwumianowych

Jeśli dwie próbki z rozkładów dwumianowych są duże ($n_1p_1 > 5,\ \ n_1(1-p_1) > 5,\ \ n_2p_2 > 5,\ \ n_2(1-p_2) > 5$) to

$$H_0 : p_1 - p_2 = p_0 = 0$$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2)}{\breve{S}} \sim N(0,1) \quad \breve{S} = \sqrt{\bar{p}(1-\bar{p})(1/n_1 + 1/n_2)} \quad \bar{p} = (n_1\hat{p}_1 + n_2\hat{p}_2)/(n_1 + n_2)$$

$$H_0 : p_1 - p_2 = p_0 \neq 0$$

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - p_0}{\tilde{S}} \sim N(0,1) \quad \tilde{S} = \sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$$

Zbiorami krytycznymi są:

- dla $H_1 : p_1 - p_2 \neq p_0 \ \ C = (-\infty, -z(\alpha/2)) \cup (z(\alpha/2), \infty)$
- dla $H_1 : p_1 - p_2 < p_0 \ \ C = (-\infty, -z(\alpha))$
- dla $H_1 : p_1 - p_2 > p_0 \ \ C = (z(\alpha), \infty)$

Rozkłady normalne o nieznanych odchyleniach: σ_1, σ_2 ale $\sigma_1^2 = \eta\sigma_2^2$

Jeśli nie znamy: σ_1, σ_2 ale wiemy, że $\sigma_1^2 = \eta\sigma_2^2$ to

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S^\star} \sim t(n^\star); \quad n^\star = n_1 + n_2 - 2$$

$$(S^\star)^2 = \frac{(n_1 - 1)S_1^2/\eta + (n_2 - 1)S_2^2}{(n_1 - 1) + (n_2 - 1)} \left(\frac{\eta}{n_1} + \frac{1}{n_2} \right)$$

$$H_0 : \mu_1 - \mu_2 = D_0$$

Zbiorami krytycznymi są:

- dla $H_1 : \mu_1 - \mu_2 \neq D_0 \ \ C = (-\infty, -t(\alpha/2, n^\star)) \cup (t(\alpha/2, n^\star), \infty)$
- dla $H_1 : \mu_1 - \mu_2 < D_0 \ \ C = (-\infty, -t(\alpha, n^\star))$
- dla $H_1 : \mu_1 - \mu_2 > D_0 \ \ C = (t(\alpha, n^\star), \infty)$

Rozkłady normalne o nieznanych odchyleniach: σ_1, σ_2

Jeśli nie znamy: σ_1, σ_2 to

$$C = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S^\dagger} \sim C(n_1, n_2)$$

$$(S^\dagger)^2 = S_1^2/n_1 + S_2^2/n_2$$

$$H_0 : \mu_1 - \mu_2 = D_0$$

Zbiorami krytycznymi są:

- dla $H_1 : \mu_1 - \mu_2 \neq D_0 \ \ C = (-\infty, -c(\alpha/2, n_1, n_2)) \cup (c(\alpha/2, n_1, n_2), \infty)$

- dla $H_1 : \mu_1 - \mu_2 < D_0 \;\; C = (-\infty, -c(\alpha, n_1, n_2))$
- dla $H_1 : \mu_1 - \mu_2 > D_0 \;\; C = (c(\alpha, n_1, n_2), \infty)$

W praktyce wartości krytyczne rozkładu Cohrana-Coxa przybliżamy średnią ważoną wartości krytycznych rozkładu Studenta:

$$c\left(\alpha, n_1, n_2\right) \approx \frac{\frac{S_1^2}{n_1} t\left(\alpha, n_1 - 1\right) + \frac{S_2^2}{n_2} t\left(\alpha, n_2 - 1\right)}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Test na relacje między wariancjami dwóch próbek normalnych

$$H_0 : \sigma_1^2 = \eta \sigma_2^2$$

$$F = \frac{S_1^2/\eta}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Zbiorami krytycznymi są:

- dla $H_1 : \sigma_1^2 \neq \eta\sigma_2^2 \;\; C = (0, 1/f(\alpha/2, n_1 - 1, n_2 - 1)) \cup (f(\alpha/2, n_1 - 1, n_2 - 1), \infty)$
- dla $H_1 : \sigma_1^2 < \eta\sigma_2^2 \;\; C = (0, 1/f(\alpha, n_1 - 1, n_2 - 1))$
- dla $H_1 : \sigma_1^2 > \eta\sigma_2^2 \;\; C = (f(\alpha, n_1 - 1, n_2 - 1), \infty)$

Test na relacje między wariancjami dwóch próbek normalnych

Przykład 19.1. Rozpatrzmy dwie niezależne próby losowe pochodzące z rozkładów normalnych.

$$n_1 = 25, \quad \bar{x}_1 = 410, \quad s_1^2 = 95, \quad n_2 = 16, \quad \bar{x}_1 = 310, \quad s_2^2 = 300$$

Przeprowadź na poziomie istotności $\alpha = 0.2$ test hipotezy $H_0 : \sigma_1^2 = \sigma_2^2$ wobec hipotezy alternatywnej $H_1 : \sigma_1^2 \neq \sigma_2^2$

$$F = \frac{s_1^2}{s_2^2} = \frac{95}{300} = 0.32$$

$$C = (0, 1/f(0.1, 24, 15)) \cup (f(0.1, 24, 15), \infty) = (0, 0.56) \cup (1.9, \infty)$$

$$F \in C$$

Mamy podstawy do odrzucenia hipotezy o równości wariancji. Nie częściej niż w 20 % przypadków popełnimy błąd polegający na odrzuceniu hipotezy prawdziwej.