

tróci wynika, że $H_0: p = \frac{1}{3}$ przeciwko hipotezie $H_1: p < \frac{1}{3}$. Aby zweryfikować tę hipotezę zastosujemy test dla frekwencji. Wtedy

$$z_{obs} = \frac{\frac{45}{150} - \frac{1}{3}}{\sqrt{\frac{\frac{1}{3} \cdot \frac{2}{3}}{150}}} \cong -0,346.$$

Tymczasem lewostronny obszar krytyczny $Q^- = (-\infty, -n_\alpha)$, gdzie n_α jest rozwiązaniem równania

$$P(\{\omega \in \Omega: |\mathbb{X}|(\omega \geq n_\alpha)\}) = \alpha$$

równoważnego równaniu

$$2(1 - \Phi(n_\alpha)) = \alpha$$

dla $\alpha = 0,05$ ma postać

$$Q^- = (-\infty, -2,23).$$

Ponieważ $z_{obs} \notin Q^-$, więc wnosimy, że nie ma powodów do odrzucenia hipotezy zerowej. Należy więc przyjąć, że co najmniej trzecia część studentów badanej uczelni zdaje wszystkie egzaminy w pierwszym terminie.

Zaprezentujemy teraz podstawowe metody *nieparametrycznych testów istotności*. Na ogół testy te dotyczą trzech kwestii:

1. rozstrzygnięcia hipotezy czy rozkład cechy $\mathbb{X}$ populacji generalnej jest określonego typu F . Mówimy wtedy o tzw. *testach zgodności*,
2. zbadania, czy pobrany materiał statystyczny spełnia wymogi określone przez definicję próby prostej. Takie testy nazywamy *testami losowości*,
3. zbadanie współzależności dwóch i więcej cech tej samej populacji. Mówimy wtedy o *testach niezależności*.

Test zgodności chi-kwadrat (χ^2) Pearsona.

Niech $(x_1, \dots, x_n) = (\mathbb{X}_1, \dots, \mathbb{X}_n)(\omega_o)$ będzie próbą prostą cechy $\mathbb{X}$ populacji generalnej. Stawiamy hipotezę na temat nieznanego rozkładu F cechy $\mathbb{X}$ w postaci $H_0: F = F_o$, dla pewnego rozkładu F_o o k parametrach. Wtedy hipoteza alternatywna H_1 oznacza, że dystrybuanta F jest inna aniżeli F_o , czyli $H_1: F \neq F_o$. Dla celów weryfikacji hipotezy H_0 ustalamy poziom istotności α (na ogół $\alpha = 0,05$) oraz za statystykę testową bierzemy *statystykę χ^2 Pearsona*, o której więcej powiemy za chwilę.

Najpierw musimy zwrócić uwagę na kilka faktów związanych ze strukturą próby prostej. Przede wszystkim należy podkreślić, że test zgodności χ^2 Pearsona ma zastosowanie zarówno dla rozkładów ciągłych jak i dyskretnych. Poniżej

pokażemy na przykładach każdą z sytuacji. W przypadku rozkładów ciągłych należy elementy próby prostej podzielić na r rozłącznych *klas*. W tym celu zaleca się (mówi o tym Polska norma PN-85, N-01052.07), aby $n \geq 100$, $r \in \{10, \dots, 25\}$ oraz n_i – ilość elementów w i -tej klasie, była równa co najmniej 5. Wróćmy teraz do statystyki χ^2 . Wtedy jej wartość w punkcie ω_o , który wyznacza próbę prostą ma postać (patrz [4])

$$\chi(\omega_o)^2 = \sum_{i=1}^r \frac{(n_i - np_i)^2}{np_i},$$

gdzie r, n, n_i mają znaczenie jak wyżej. Przy założeniu hipotezy H_o , statystyka χ^2 ma rozkład *chi-kwadrat* o $r - k - 1$ stopniach swobody, gdzie k oznacza liczbę parametrów dystrybuanty F_o . Wtedy dla $2 \leq i \leq r$

$$p_i = \Phi(y_i) - \Phi(y_{i-1}), \text{ gdzie } y_i = \frac{x_i - \bar{x}}{s},$$

gdzie Φ oznacza dystrybuantę standardowego rozkładu normalnego, x_i jest górną wartością i -tej klasy, $\bar{x}$ jest średnią z próby prostej, natomiast $s = \sqrt{s^2}$ odchyleniem standardowym. Wtedy $p_1 = 1 - \sum_{i=2}^r p_i$. Dla dostatecznie dużych n ($n \geq 100$) obszar krytyczny Q , który jest zawsze obszarem prawostronnym, wyznacza równość

$$P(\{\omega \in \Omega: \chi^2(\omega) \geq \chi_\alpha^2\}) = \alpha.$$

Wtedy

$$Q = Q^+ = (\chi_\alpha^2, +\infty).$$

Jeśli teraz $\chi_{obs}^2 = \chi^2(\omega_o) \in Q$, to odrzucamy hipotezę H_o przyjmując hipotezę H_1 . W przeciwnym razie wnioskujemy, że nie ma podstaw aby H_o odrzucić, a więc nie ma powodów do tego aby nie przyjąć, że $F = F_o$.

Przykład 6.4.11 W pewnej miejscowości wylosowano niezależnie 500 gospodarstw i zbadano miesięczne zużycie energii elektrycznej w każdym z nich. Zebrane dane przedstawia poniższa tabela

Zużycie energii w KWh	35 – 45	45 – 55	55 – 65	65 – 75	75 – 85
Liczba gospodarstw	70	100	140	110	80

Na poziomie istotności 0,05 należy zweryfikować hipotezę, że rozkład zużytej energii przez gospodarstwa jest normalny.

Sprawdzimy najpierw, czy spełnione są warunki umożliwiające stosowanie testu zgodności. Mamy odpowiednio:

$$n = 500, \quad r = 5, \quad n_1 = 70, \quad n_2 = 100, \quad n_3 = 140, \quad n_4 = 110, \quad n_5 = 80.$$

Jak widać wartość liczby klas jest trochę mała. Może to rzutować na wiarygodność całej procedury.

Niech x_i dla $i = 1, 2, \dots, 5$ oznacza górne wartości kolejnych klas. Wtedy

$$x_1 = 45, x_2 = 55, x_3 = 65, x_4 = 75, x_5 = 85.$$

Wyznamy wartości p_i . W tym celu należy standaryzować rozkład F_o . Wtedy

$$F_o(x_i) = \Phi\left(\frac{x_i - \bar{x}}{s}\right).$$

Ponieważ

$$\bar{x} = \frac{1}{n} \sum_{i=1}^r n_i x_i^o,$$

gdzie x_i^o oznacza środek i -tej klasy, więc po podstawieniu danych liczbowych otrzymamy

$$\bar{x} = \frac{1}{500} \left(70 \cdot \frac{35 + 45}{2} + 100 \cdot \frac{45 + 55}{2} + 140 \cdot \frac{55 + 65}{2} + 110 \cdot \frac{65 + 75}{2} + 80 \cdot \frac{75 + 85}{2} \right) = 60,6.$$

Podobnie

$$s^2 = \frac{1}{n} \sum_{i=1}^r (x_i^o - \bar{x})^2 n_i$$

skąd

$$s^2 = \frac{1}{500} \left((40 - 60,6)^2 \cdot 70 + (50 - 60,6)^2 \cdot 100 + (60 - 60,6)^2 \cdot 140 + (70 - 60,6)^2 \cdot 110 + (80 - 60,6)^2 \cdot 80 \right) = 161,65,$$

i dlatego $s = 12,71$. W takim razie

$$y_1 = \frac{x_1 - \bar{x}}{s} = \frac{45 - 60,6}{12,71} = -1,227, y_2 = \frac{x_2 - \bar{x}}{s} = \frac{55 - 60,6}{12,71} = -0,440,$$

$$y_3 = \frac{x_3 - \bar{x}}{s} = \frac{65 - 60,6}{12,71} = 0,346, y_4 = \frac{x_4 - \bar{x}}{s} = \frac{75 - 60,6}{12,71} = 1,132,$$

$$y_5 = \frac{x_5 - \bar{x}}{s} = \frac{85 - 60,6}{12,71} = 1,919.$$

Możemy teraz obliczyć wartości p_i :

$$p_2 = \Phi(y_2) - \Phi(y_1) = \Phi(-0,440) - \Phi(-1,227) =$$

$$\Phi(1,227) - \Phi(0,440) = 0,888767 - 0,670031 \cong 0,219,$$

$$p_3 = \Phi(y_3) - \Phi(y_2) = \Phi(0,346) - \Phi(-0,440) =$$

$$\Phi(0,346) - (1 - \Phi(0,440)) = 0,633072 - (1 - 0,670031) \cong 0,304,$$

$$p_4 = \Phi(y_4) - \Phi(y_3) = \Phi(1,132) - \Phi(0,346) = 0,870762 - 0,633072 \cong 0,238,$$

$$p_5 = \Phi(y_5) - \Phi(y_4) = \Phi(1,919) - \Phi(1,132) = 0,971933 - 0,870762 \cong 0,101,$$

gdzie

$$p_1 = 1 - (p_2 + \dots + p_5) = 1 - 0,862 = 0,138.$$

uzyskane dane i pozostałe, niezbędne dla wyliczenia wartości zaobserwowanej statystyki χ^2 zebraliśmy w poniższej tabeli

x_i	n_i	y_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
45	70	-1,227	0,138	69	0,014
55	100	-0,440	0,219	109,5	0,824
65	140	0,346	0,304	152	0,947
75	110	1,132	0,238	119	0,680
85	80	1,919	0,101	50,5	17,232
Σ	500		1,000	500	$\chi_{obs}^2 = 19,694$

Ponieważ

$$P(\{\omega \in \Omega: \chi^2(\omega) \geq \chi_{\alpha, r-k-1}^2\}) = \alpha \Leftrightarrow P(\{\omega \in \Omega: \chi^2(\omega) \geq \chi_{0,05, 2}^2\}) = 0,05,$$

więc $P(\{\chi^2(\omega) \geq \chi_{0,05,2}^2\}) = 5,991 < \chi_{obs}^2$, co oznacza, że hipotezę H_0 należy odrzucić.

Przy okazji tego przykładu warto zauważyć, że np_i (patrz powyższa tabela) oznaczają liczebność teoretyczną kolejnych klas, w odróżnieniu od wartości n_i , które oznaczają liczebność zmierzoną.

Przejdziemy teraz do omówienia przykładu, w którym pokażemy sposób posługiwania się testem zgodności w przypadku rozkładu dyskretnego.

Przykład 6.4.12 Kandydatów na kierowcę poddano badaniom sprawdzającym refleks i uwagę. Każdy z nich miał do wykonania zadania na czterech stanowiskach. W wyniku przebadania 100 losowo wybranych osób otrzymano następujące dane:

Liczba wykonanych zadań	0	1	2	3	4
Liczba kandydatów	5	12	23	40	20

Na poziomie istotności 0,01 zweryfikować hipotezę, że rozkład ten jest rozkładem dwumianowym.

Rozkład dwumianowy zależy od dwóch parametrów, z których jeden (ilość powtórzeń) znamy. Aby ustalić wartość statystyki χ^2 musimy wyznaczyć prawdopodobieństwo sukcesu w tym rozkładzie. Sukcesem jest tutaj zdarzenie polegające na tym, że osoba przechodząca badania zaliczyła pomyślnie testy na wszystkich czterech stanowiskach. Z powyższych danych wynika zatem, że $p = \frac{20}{100} = \frac{1}{5}$. Odpowiednie dane potrzebne dla celów weryfikacji hipotezy $H_0: F \in B(n, p)$ przy hipotezie alternatywnej $H_1: F \notin B(n, p)$ zebraliśmy w poniższej tabeli

x_i	n_i	np_i	$(n_i - np_i)^2$	$\frac{(n_i - np_i)^2}{np_i}$
0	5	20	225	11,25
1	12	20	64	3,2
2	23	20	9	0,45
3	40	20	400	20
4	20	20	0	0
Σ	100	100		$\chi_{obs}^2 = 34,9$

Określimy obszar krytyczny dla tego testu. W tym celu należy rozwiązać równanie

$$P(\{\omega \in \Omega: \chi^2(\omega) \geq \chi_{\alpha, r-k-1}^2\}) = \alpha.$$

Ponieważ $r - k - 1 = 5 - 1 - 1 = 3$ oraz przyjęliśmy, że $\alpha = 0,01$, z tablic rozkładu χ^2 wynika, że $\chi_{0,01, 3}^2 = 5,841$. Stąd obszar krytyczny jest przedziałem $Q = (5,841, +\infty)$. Ponieważ $\chi_{obs}^2 = 34,9 \in Q$, więc należy odrzucić hipotezę zerową na korzyść hipotezy alternatywnej.