SdI30 W10: *TESTY NIEPARAMETRYCZNE*

1. Testy nieparametryczne

Właściwością parametrycznych testów istotności jest to, że postulowana hipoteza dotyczy parametrów rozkładu badanej cechy. Testy te często wymagają założenia o typie rozkładu cechy.

Nieparametryczne testy istotności rozszerzają stawiane hipotezy – dotyczą one zarówno postaci rozkładu, jak i ich parametrów. Nadal jednak pozostają testami istotności, a to oznacza ograniczenie wnioskowania do alternatywy: odrzucić stawianą hipotezę zerową – nie odrzucać, gdyż brak jest do tego wystarczających podstaw.

Siła testów nieparametrycznych polega na tym, że wymagają bardzo ogólnych założeń dotyczących liczebności oraz typu danych.

Testy nieparametryczne dotyczą różnorodnych właściwości populacji oraz prób losowych.

Ze względu na zakres zastosowań testy te dzielimy na grupy:

1) testy zgodności rozkładu empirycznego z teoretycznym,
2) testy zgodności dwóch lub kilku rozkładów empirycznych,
3) testy niezależności,
4) testy losowości próby,
5) testy stosowane w analizie regresji.

2. Dystrybuanta empiryczna i jej własności

Zgodność rozkładów jest badana za pomocą dystrybuant. Estymatorem dystrybuanty F_X cechy X w populacji jest dystrybuanta empiryczna $F_n(x|\boldsymbol{X})$ wyznaczona na podstawie próby losowej.

Niech $\boldsymbol{X} = (X_1, \ldots, X_n)$ będzie n-elementową prostą próbą losową z populacji, w której badana cecha X ma nieznany rozkład.

Dystrybuantą empiryczną nazywamy funkcję

$$F_n : \mathbb{R} \times \Theta \to [0, 1]$$

która dla $x \in \mathbb{R}$ oraz $\boldsymbol{X} \in \Theta$ (Θ *oznacza przestrzeń prób*) jest określona wzorem:

$$F_n(x|\boldsymbol{X}) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{I}_{(-\infty, x]}(X_k)$$

gdzie $\mathbb{I}_A(X)$ jest funkcją wskaźnikową.

Dla ustalonego $x \in \mathbb{R}$ funkcja $F_n(x|\cdot)$ jest częstością zdarzenia $\{X_k \leq x\}$, czyli zmienną losową.

Dla ustalonej realizacji próby $\boldsymbol{x} = (x_1, \ldots, x_n)$ funkcja $F_n(\cdot\,|\boldsymbol{x})$ jest dystrybuantą schodkową, mającą skoki wielkości $1/n$ w punktach $x_1, \ldots, x_n$.

Twierdzenie (o lokalnych własnościach dystrybuanty F_n).

Jeżeli F_n jest dystrybuantą empiryczną z n-elementowej próby $\boldsymbol{X}$ pobranej z rozkładu o dystrybuancie F, to dla $x \in \mathbb{R}$:

a) $nF_n(x|\boldsymbol{X})$ jest zm. l. o rozkładzie $\mathrm{bin}\big(n, F_n(x)\big)$;

b) $\mathrm{P}\left(\lim_{n\to\infty} F_n(x|\boldsymbol{X}) = F(x)\right) = 1$;

c) Ciąg zm. l. $\{F_n(x|\boldsymbol{X})\}$ jest asymptotycznie normalny

$$\mathcal{N}\left(F(x), \sqrt{\frac{F(x)\left(1-F(x)\right)}{n}}\right).$$

Dowód. Teza a) jest oczywista, teza b) jest bezpośrednim wnioskiem z prawa wielkich liczb, teza c) jest wnioskiem z CTG.

Bezpośrednim zastosowaniem dystrybuanty empirycznej jest estymacja punktowa dystrybuanty badanej cechy X w populacji. Interesująca może być również charakteryzacja globalnej estymacji dystrybuanty F estymatorem $F_n(x|\boldsymbol{X})$.

Dla takiej charakteryzacji bardzo pożytecznymi miarami zgodności dystrybuant są statystyki χ^2 *Pearsona* oraz *KS*1 i *KS*2 *Kołmogorowa-Smirnowa*.

3. Badanie normalności rozkładu

W parametrycznych testach statystycznych t, z, F, χ^2 istotnym założeniem jest to, że dane wejściowe w próbie pochodzą z populacji o rozkładzie normalny. W praktyce musimy tę hipotezę sprawdzić. Do jej weryfikacji służą narzędzia graficzne:

- histogram z naniesionym fitem rozkładu normalnego histfit,
- wykres wartości w próbie, wzg. prawd. uzyskania takiej wartości w rozkładzie normalnym normplot,

oraz testy nieparametryczne:

- test Lillieforsa (lillietest) – test oparty na badaniu maksymalnej różnicy pomiędzy dystrybuantą empiryczną a dystrybuantą rozkładu normalnego o takiej samej średniej i wariancji jak oszacowana z próby;
- test Kolmogorova-Smirnova (kstest) – parametry zadajemy.

4. Testy zgodności

Testy zgodności (*goodness-of-fit tests*) służą do weryfikacji hipotez orzekających o postaci rozkładu badanej cechy (lub kilku cech) w populacji.

Rozkład hipotetyczny zwykle jest określony za pomocą dystrybuanty $F_0(x)$ lub symbolicznego oznaczenia rozkładu.

Weryfikowana jest hipoteza zerowa

$$H_0: F_X(x|\theta) = F_0(x|\theta);$$

przeciw hipotezie alternatywnej

$$H_1: F_X(x|\theta) \neq F_0(x|\theta)$$

θ jest parametrem, którego wartość też może być weryfikowana.

Stosując symboliczne oznaczenia rozkładów „SYMBOL(θ)” hipotezę zerową zapisujemy $\mathrm{H}_0: X \sim \mathrm{SYMBOL}(\theta)$.

Hipoteza alternatywna jest zaprzeczeniem hipotezy zerowej.

Przykłady hipotez zerowych:

$X \sim \mathcal{N}(m, \sigma)$, gdzie m i σ są nieustalonymi parametrami,

$X \sim \mathcal{N}(m_0, \sigma)$, gdzie m_0 jest ustaloną wartością oczekiwaną, a σ nie interesującym nas parametrem,

Do podstawowych testów zgodności należą:

a) test zgodności chi-kwadrat Pearsona,
b) test zgodności λ Kołmogorowa,
c) test zgodności Kołmogorowa-Smirnowa,
d) test znaków,
e) test Smirnowa zgodności trzech rozkładów empirycznych.

Uwaga. W statystyce do sprawdzenia, czy dana zmienna ma pewien rozkład (np. rozkład normalny), używa się tzw. wykresów kwantylowych (Quantile to Quantile (Q-Q) plot), w którym na jednej osi (poziomej) umieszczone są kwantyle teoretyczne (parametry estymowane z próby) rozkładu badanej zmiennej (najczęściej $F^{-1}((i-0{,}5)/n)$, gdzie F oznacza dystrybuantę badanego rozkładu, a $i = 1, 2, \ldots, n$ kolejne numery obserwacji, a na drugiej osi kwantyle porównywanego rozkładu (czyli w zasadzie uszeregowane wartości z próby) – nanosimy więc punkty

$$\left(F^{-1}\left(\frac{i-0{,}5}{n}\right), x_{(i)}\right)$$

Jeśli zmienna ma idealnie zadany rozkład, wykres ten przedstawia dokładnie prostą.

Odchyłki od prostej wskazują na określone typy odchylenia, np. skośny, spłaszczony. Wykres ten pomaga nam również wykryć obserwacje odstające.

Najczęściej w praktyce stosuje się wykresy kwantylowe do porównania z rozkładem normalnym.

-------$\mathcal{R}$----------

dane = **rnorm**(10)
qqnorm(dane) # porównanie z rozkładem normalnym
qqline(dane) # dadaje linię przechodzącą przez Q_1 i Q_3

Można również porównać z dowolnym rozkładem (lub też dwa dowolne zbiory). Służy do tego procedura **qqplot**(x, y). jeszcze lepsze rozwiązanie możemy znaleźć w pakiecie **car**, który zawiera funkcję **qq.plot** umożliwiającą zadanie rozkładu, z którym chcemy porównywać (poprzez parametr **distribution**).

shapiro.test(dane) # Test Shapiro-Wilka

-------$\mathcal{R}$----------

5. Test zgodności chi-kwadrat Pearsona

Test zgodności *chi-kwadrat* oparty jest na szeregu rozdzielczym i przeprowadza się go tylko dla dużej próby prostej.

Próba prosta $\boldsymbol{X}$ pochodzi z populacji, w której badana cecha X ma nieznany rozkład ciągły lub skokowy.

Zasady podziału próby na k klas są typowe.

A. W przypadku cechy ciągłej oś liczbową dzielimy punktami $g_i, i = 1, 2, \ldots, k-1$ na rozłączne przedziały Δ_i.

Otrzymujemy w ten sposób k przedziałów,

$$\Delta_1 = (-\infty, g_1], \Delta_2 = (g_1, g_2], \ldots, \Delta_k = (g_{k-1}, \infty).$$

Przez N_i oznaczamy liczbę obserwacji w przedziale Δ_i (liczebność empiryczna). Liczebność próby spełnia warunek

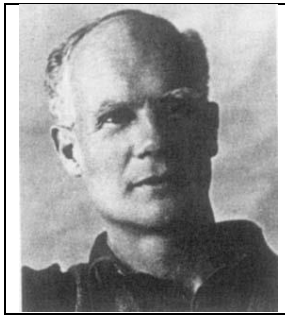
$$n = N_1 + N_2 + \ldots + N_k$$

Utworzonym przedziałom klasowym odpowiadają liczebności teoretyczne np_i, gdzie $p_i = \mathrm{P}(X \in \Delta_i) = F(g_i) - F(g_{i-1})$, $i = 1, 2, \ldots, k$ oraz $F(g_0) = 0$, $F(g_k) = 1$.

Statystykę

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

nazywamy statystyką χ^2 $Pearsona$[1]. Dla efektywnego obliczenia prawd. p_i należy znać wszystkie parametry dystrybuanty



[1] Karl Pearson (1857 – 1936) – matematyk, filozof i biolog angielski. Jeden z twórców współczesnej statystyki.

teoretycznej. Jeżeli parametry nie są znane, to estymujemy je metodą największej wiarygodności.

Twierdzenie o rozkładzie statystyki χ^2 *Pearsona*.
Jeżeli nieznane parametry dystrybuanty F są oszacowane metodą największej wiarygodności, to dystrybuanta statystyki χ^2 *Pearsona* jest zbieżna, dla $n \to \infty$, do dystrybuanty rozkładu chis$(k - r - 1)$, gdzie r jest liczbą estymowanych parametrów rozkładu teoretycznego.

Gdy liczebności teoretyczne dla pewnych klas są mniejsze od liczby 5, to należy je połączyć z sąsiednimi klasami, tak aby sumaryczne liczebności nowych klas wynosiły co najmniej 5. Dla klas ze środka szeregu rozdzielczego łączenie następuje z klasą, która jest bliższa skrajnej.

W wyniku łączenia klas nastąpi zmniejszenie ich liczby, co spowoduje zmniejszenie liczby stopni swobody $v = k - r - 1$, więc może się zdarzyć, że $v \leq 1$ i braknie stopni swobody. W takiej sytuacji zwiększamy liczebność próby lub zwiększamy liczbę klas przed ich łączeniem.

Obszar krytyczny: $R_\alpha = \left(\chi^2_{1-\alpha,k-r-1}, \infty\right)$, gdzie $\chi^2_{1-\alpha,k-r-1}$ jest kwantylem rzędu $1 - \alpha$ rozkładu chis$(k - r - 1)$.

Przykład 1. W celu zbadania rozkładu długości X produkowanego parkietu, została pobrana próba losowa o liczebności $n = 150$ klepek. Dokonano pomiaru ich długości.

Wyniki w mm są zebrane w postaci szeregu rozdzielczego:

Przedział i	Liczebność N_i
$(-\infty, 249{,}20]$	11
$(249{,}200; 249{,}600]$	22
$(249{,}600; 250{,}000)$	44
$(250{,}000; 250{,}400]$	38
$(250{,}400; 250{,}800]$	26
$(250{,}800; \infty)$	9
	$\sum = 150$

Tabela 1. Dane pomiarowe.

Cel poznawczy: Sprawdzenie hipotez dotyczących zgodności rozkładu długości parkietu z rozkładami:

a) $\mathcal{N}(249{,}9;\ 0{,}5)[mm]$,

b) $\mathcal{N}(\mu, \sigma)[mm]$.

Rozwiązanie. a) Sprawdzana jest hipoteza o normalności rozkładu długości klepek ze znanymi parametrami, tj.

$$\mathrm{H}_0: X \sim \mathcal{N}(249{,}9;\, 0{,}5)$$

Przyjmujemy, że prawdziwa jest hipoteza H_0 i wyznaczamy liczebności teoretyczne:

$$p_1 = \mathrm{P}(X \leq 249{,}2) = \Phi(-1{,}4) = 0{,}08076,$$

$$p_2 = \mathrm{P}(249{,}2 < X \leq 249{,}6) = \Phi(-0{,}6) - \Phi(-1{,}4) = 0{,}2743 - 0{,}08076 = 0{,}19354, \text{ itd.}$$

i	Przedział w $[mm]$	N_i	np_i	$Chi - kwadrat$
1	$(-\infty,$ 249,20]	11	12,1135	0,10
2	(249,200; 249,600]	22	29,0245	1,70
3	(249,600; 250,000)	44	45,7510	0,07
4	(250,000; 250,400]	38	39,3128	0,04

5	(250,400; 250,800]	26	18,4087	3,13
6	(250,800; ∞)	9	5,3895	2,42
	$\sum$	150	150,0000	7,46

Rozważany szereg rozdzielczy ma sześć klas $(k = 6)$ spełnia warunki poprawnego stosowania testu $Chi - kwadrat$.

Parametry rozkładu są znane, więc statystyka testowa ma rozkład chi-kwadrat z 5. stopniami swobody.

Wartość statystyki testowej wynosi $\chi_0^2 = 7{,}46$.

Obszar krytyczny $R_{0,05} = (11{,}1; \infty)$.

Decyzja: Ponieważ statystyka testowa nie należy do obszaru krytycznego, więc nie mamy podstaw do odrzucenia hipotezy, że rozkład długości klepek parkietowych jest normalny z podanymi parametrami, tj. $X \sim \mathcal{N}(249{,}9; 0{,}5)\ [mm]$.

b) Weryfikowana jest hipoteza $\mathrm{H}_0: X \sim \mathcal{N}(m, \sigma)$ z nieznanymi parametrami m i σ. Obydwa parametry estymujemy z próby. Oceny tych parametrów wynoszą:

$$\hat{m} = \bar{\mathbf{x}}_n = 249{,}984,\ \hat{\sigma} = s_n = 0{,}55287\ .$$

Statystyka testowa $\chi^2_{150} \sim \mathrm{chi2}(k-3)$, gdzie k jest liczbą klas po połączeniu wszystkich klas, których liczebności teoretyczne są mniejsze od liczby 5.

Statystyka ta jest obliczona podobnie jak w podpunkcie a).

$$p_1 = \mathrm{P}(X \leq 249{,}2) \overset{STD}{\approx} \Phi(-1{,}418) = 0{,}078\ ,$$

$$p_2 = \mathrm{P}(249{,}2 < X \leq 249{,}6) \overset{STD}{\approx} \Phi(-0{,}695) - \Phi(-1{,}418) = 0{,}1653, \text{ itd.}$$

i	Przedział	n_i	np_i	*Chi-kwadrat*
1	(–∞, 249,20]	11	11,71	0,0434
2	(249,200; 249,600]	22	24,84	0,3340
3	(249,600; 250,000)	44	40,18	0,3629
4	(250,000; 250,400]	38	39,38	0,0497
5	(250,400; 250,800]	26	23,39	0,2919
6	(250,800; ∞)	9	10,50	0,2135
	Σ	150	150,00	1,2844

Wniosek. Ponieważ $\chi_0^2 = 1{,}28 \notin R_{0{,}05} = (7{,}81, \infty)$, więc również nie mamy podstaw do odrzucenia hipotezy zerowej.

B. Test zgodności dla cechy skategoryzowanej

Szczególnym przypadkiem testu zgodności jest test dla **rozkładu wielomianowego** (*multinomial distribution*). Rozkład wielomianowy jest uogólnieniem rozkładu dwumianowego.

W przypadku wielomianowym mamy $k > 2$ kategorii danych $E_1, \ldots, E_k$. Punkt danych należy tylko do jednej z k kategorii i prawd., że punkt ten należy do i-tej kategorii (gdzie $i = 1, \ldots, k$) jest stałe i równe p_i. Oczywiście, $p_1 + \ldots + p_k = 1$.

Bezpośrednie stosowanie rozkładu wielomianowego jest trudne i rozkład $chi - kwadrat$ jest bardzo dobrą alternatywą, gdy liczebność próby jest dostatecznie duża.

Hipoteza zerowa ma postać: $\mathrm{H}_0: \mathrm{P}(E_i) = p_i$, dla $i = 1, \ldots, k$. Hipoteza alternatywna jest zaprzeczeniem hipotezy zerowej.

Do sprawdzenia hipotezy zerowej korzystamy ze statystyki $chi - kwadrat$ z $k - 1$ stopniami swobody.

Rozkład $chi - kwadrat$ stosujemy wówczas, gdy liczebność teoretyczna w każdej klasie wynosi co najmniej 5.

Przykład 2. Na drugim roku studiów kierunku informatyka są cztery grupy ćwiczeniowe. Stawiamy pytanie: Czy zainteresowanie wykładem z metod probabilistycznych w dniu 28 marca było takie samo dla wszystkich grup?

Rozwiązanie. Postawione pytanie jest równoważne pytaniu:

Czy rozkład liczby studentów obecnych na wykładzie jest proporcjonalny do liczebności poszczególnych grup?

Tabl. Informacje o licznościach grup oraz o obecnościach

$grupa$	$liczebność$	$prawd.$	$obecnych$
1.	22	$22/104$	18
2.	31	$31/104$	17
3.	26	$26/104$	22
4.	25	$25/104$	16
Σ	**104**	**1**	**73**

Prawd. p_i dla $i = 1, \ldots, 4$ oznacza prawd. zdarzenia, że wylosowany student ω spośród obecnych jest z i–tej grupy.

Hipoteza zerowa o identycznym zainteresowaniu wykładem:

$$\mathrm{H}_0: p_1 = \frac{22}{104}, p_2 = \frac{31}{104}, p_3 = \frac{26}{104}, p_4 = \frac{25}{104}$$

Hipoteza alt. orzeka zróżnicowane zainteresowanie wykładem.

Hipotezę H_0 sprawdzamy za pomocą statystyki chi-kwadrat

$$\chi_0^2 = \sum_{i=1}^{4} \frac{(n_i - np_i)^2}{np_i}, n = n_1 + n_2 + n_3 + n_4$$

Ustalamy obecności teoretyczne np_i dla wszystkich grup i obliczamy statystykę chi-kwadrat.

Tabl. Wyniki obliczeń

$numer\ spec.$	$prawd.\ p_i$	$obecności\ empiryczne\ n_i$	$obecności\ teoretyczne\ np_i$	$chi - kwadrat$
1	$22/104$	18	15,44	0,424
2	$31/104$	17	21,76	1,041
3	$26/104$	22	18,25	0,771
4	$25/104$	16	17,55	0,137
Σ	1	$n = 73$	73	$\chi_0^2 = 2{,}373$

Wyznaczamy obszar krytyczny CR: $R_{0,05} = \left(\chi^2_{0,95;\,3};\ \infty\right)$.

Z tablic odczytujemy $\chi^2_{0,95;\,3} = 7{,}815$.

Decyzja: Ponieważ $\chi^2_0 \notin R_{0,05}$, więc nie było istotnych różnic między grupami w zainteresowaniu wykładem.

6. Test zgodności Kołmogorowa

Test zgodności λ *Kołmogorowa* można stosować dla dużych prób $\boldsymbol{X}$, ale tylko dla populacji X o rozkładzie absolutnie ciągłym. Istotę testu nie stanowi porównanie liczebności empirycznych z teoretycznymi, jak to było w teście $Pearsona$, lecz wprost porównanie wartości dystrybuanty empirycznej i teoretycznej. Statystyka testowa λ Kołmogorowa mierzy odległość dystrybuanty emp. $F_n(x)$ od dystrybuanty teoretycznej $F(x)$, tj.

$$\lambda = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|$$

Aby wyznaczyć statystykę λ Kołmogorowa trzeba znać dane szczegółowe. Rozkład graniczny statystyki λ nazywamy rozkładem Kołmogorowa. Jest on stablicowany.

7. Test niezależności chi-kwadrat

Populację generalną Ω badamy ze względu na dwie cechy X i Y. Chcemy sprawdzić, czy cechy te są niezależne, tj. hipotezę:

$$\mathrm{H}_0\colon X \text{ i } Y \text{ są niezależne,}$$

tj., że dla każdego x i $y \in \mathbb{R}$, $F_{XY}(x,y) = F_X(x)F_Y(y)$, przeciw hipotezie $\mathrm{H}_1\colon F_{XY}(x,y) \neq F_X(x)F_Y(y)$

Dużą próbę o liczebności n (zwykle $n \geq 100$) dzielimy ze względu na cechę X na r klas, a ze względu na cechę Y na s klas. W ten sposób wszystkie elementy z próby dzielimy na rs klas, otrzymując tak zwaną tablicę wielodzielczą.

Hipotezę H_0 łatwiej jest sprawdzać w postaci: $p_{ij} = p_{i\bullet}p_{\bullet j}$, gdzie $i \in \{1, \dots, r\}$, $j \in \{1, \dots, s\}$.

Niech N_{ij} oznacza liczbę elementów, które ze względu na cechę X znajdą się w i-tej klasie, a ze względu na cechę Y w j-tej klasie.

Określamy liczebności brzegowe

$$N_{i\bullet} = N_{i1} + N_{i2} + \ldots + N_{is},$$

$$N_{\bullet j} = N_{1j} + N_{2j} + \ldots + N_{rj}$$

Z liczebności brzegowych szacujemy prawd. brzegowe

$$p_{i\bullet} = \frac{N_{i\bullet}}{n},\ p_{\bullet j} = \frac{N_{\bullet j}}{n}$$

Przy założeniu niezależności cech X i Y, obliczamy prawdopodobieństwa hipotetyczne

$$p_{ij} = p_{i\bullet} p_{\bullet j}$$

Do sprawdzenia hipotezy o niezależności cech stosujemy statystykę

$$(*) \quad \chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - np_{ij}\right)^2}{np_{ij}}$$

Statystyka $(*)$ ma rozkład podany w tw. Pearsona (1916).

Twierdzenie o rozkładzie statystyki χ^2.

Rozkład statystyki χ^2 określonej wzorem $(*)$, jest zbieżny do rozkładu chi-kwadrat o $(r-1)(s-1)$ stopniach swobody, przy $n \to \infty$.

Obszar krytyczny: $R_\alpha = \left(\chi^2_{1-\alpha;(r-1)(s-1)}, \infty\right)$.

Założenia testu: $n \geq 100$, wszystkie $np_{ij} \geq 8$.

Jeżeli liczebności $n_{ij} < 8$, to łączymy dwie klasy w jedną.

W szczególności, jeżeli $r = s = 2$, to otrzymujemy czteropolową tablicę danych

$$\begin{pmatrix} n_{11} & n_{12} \\ n_{21} & n_{22} \end{pmatrix}$$

Statystyka $(*)$ χ_n^2 ma wówczas tylko jeden stopień swobody.

W takich przypadkach, zalecane jest „skorygowanie" wartości statystyki tak, aby jej rozkład dyskretny był lepiej przybliżony przez ciągły rozkład $chi - kwadrat$.

Korekta ta, zwana poprawką *Yatesa*, polega na odjęciu liczby ½ od modułu różnicy liczebności zaobserwowanych i hipotetycznych przed podniesieniem tej różnicy do kwadratu.

Przykład 3. Wybrano losowo próbę 100 firm i dla każdej z nich zanotowano, czy miała zysk, czy straty oraz czy należy do sektora usług, czy nie. Dane podsumowane w postaci tablicy 2×2 są przedstawione w tablicy 1.

Tablica 1.

Zysk/strata	Rodzaj działalności		Suma
	usługi	inne	
Zysk	42	18	60
Strata	6	34	40
Suma	48	52	100

Zbadać, czy obydwa zdarzenia „firma przyniosła zysk” i „firma działa w sektorze usług” są niezależne.

Rozwiązanie. Wyznaczamy liczebności hipotetyczne

$$np_{11} = np_{1\bullet}p_{\bullet 1} = n\left(\frac{n_{1\bullet}}{n}\right)\left(\frac{n_{\bullet 1}}{n}\right) = \frac{n_{1\bullet}n_{\bullet 1}}{n}$$
$$= 60 \cdot 48/100 = 28{,}8;$$
$$np_{12} = n_{1\bullet}n_{\bullet 2}/n = 60 \cdot 52/100 = 31{,}2;$$
$$np_{21} = n_{2\bullet}n_{\bullet 1}/n = 40 \cdot 48/100 = 19{,}2;$$
$$np_{22} = n_{2\bullet}n_{\bullet 2}/n = 40 \cdot 52/100 = 20{,}8;$$

i zestawiamy je w tablicę

	Usługi	Inne
Zysk	28,8 (42)	31,2 (18)
Strata	19,2 (6)	20,8 (34)

Obliczamy statystykę chi-kwadrat z poprawką Yatesa.

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(\left|n_{ij} - np_{ij}\right| - 0{,}5\right)^2}{np_{ij}}$$

$$\chi_0^2 = \frac{(13{,}2-0{,}5)^2}{28{,}8} + \frac{(13{,}2-0{,}5)^2}{31{,}2} + \frac{(13{,}2-0{,}5)^2}{19{,}2} + \frac{(13{,}2-0{,}5)^2}{20{,}8} = 26{,}92$$

Obszar krytyczny dla $\alpha = 0{,}01$, $R_{0{,}01} = \left(\chi^2_{0{,}99;\,1}, \infty\right)$.

Kwantyl rozkładu $chi - kwadrat$ $\chi^2_{0{,}99;\,1} = 6{,}63$.

Wniosek: Obliczona wartość statystyki 26,92 jest o wiele większa od wartości krytycznej 6,63, więc odrzucamy hipotezę zerową i wnioskujemy, że badane cechy, zysk/strata oraz typ działalności firmy, nie są niezależne.

8. Test losowości próby

Losowość próby odgrywa kluczową rolę we wnioskowaniu statystycznym. Najczęściej stosowanymi testami losowości próby są testy oparte na liczbie serii. ***Serią*** nazywamy każdy podciąg ciągu elementów a i b o takiej własności, że wszystkie kolejne elementy podciągu są tego samego typu.

Test serii.

Założenia: 1) badana cecha ma dowolny rozkład,
2) pobrano n-elementową próbę $x_1, \ldots, x_n$.

Hipotezy: H_0: próba ma charakter losowy,
H_1: kolejne wartości próby są zależne.

Statystyka: $k =$ liczba serii w ciągu elementów a, b,
gdzie a – zdarzenie $x_i < me$, b – zdarzenie $x_i > me$,
$me =$ mediana empiryczna.

W teście serii zachowując kolejność pobierania elementów próby $X_1, \dots, X_n$ obliczamy znaki odchyleń od mediany X_M wyznaczonej z próby:

$$\operatorname{sgn}(x_k - x_M), k = 1, \dots, n$$

i otrzymujemy ciąg utworzony z liczb $-1, 0, 1$.
Statystyką jest losowa liczba K serii w próbie.

Jeżeli hipoteza zerowa jest prawdziwa, to statystyka K ma rozkład zależny tylko od liczby n_1 jedynek w ciągu.

Obszar krytyczny $R_\alpha = (0, k_\alpha)$. Wartości kwantyli k_α w tablicy 52a w [Ryszard Zieliński -Tablice statystyczne].

Przykład (Bobrowski). Badając jakość wyrobu pobrano próbę, po jednej wykonanej sztuce w odstępach półgodzinnych, w ciągu jednej zmiany. Badana cecha miała wartości kolejno równe:

1,04; 1,07; 1,08; 0,96; 1,02; 1,03; 1,03; 0,92; 1,00; 0,97; 0,95; 0,99; 0,96; 1,04; 0,98.

Problem: Czy ze względu na badaną cechę produkcja była stabilna?

Rozwiązanie. Stabilność produkcji charakteryzuje losowość sekwencyjnej próby i ta losowość jest sprawdzana testem serii.

Ponieważ mediana $x_M = 1{,}00$, więc ciąg liczb przyporządkowany próbie ma postać:

$$1, 1, 1, -1, 1, 1, 1, -1, 0, -1, -1, -1, -1, 1, -1$$

Liczba serii w próbie $k = 7$, przy czym $n_1 = 7$.

Dla $\alpha = 0{,}05$ obszar krytyczny $R_{0,05} = (0, 4)$.

Decyzja: Ponieważ $k = 7 \notin R_{0,05}$, więc nie ma podstaw do odrzucenia hipotezy o losowości próby (na przyjętym poziomie istotności), a tym samym o rozregulowaniu procesu produkcyjnego ze względu na badaną cechę.

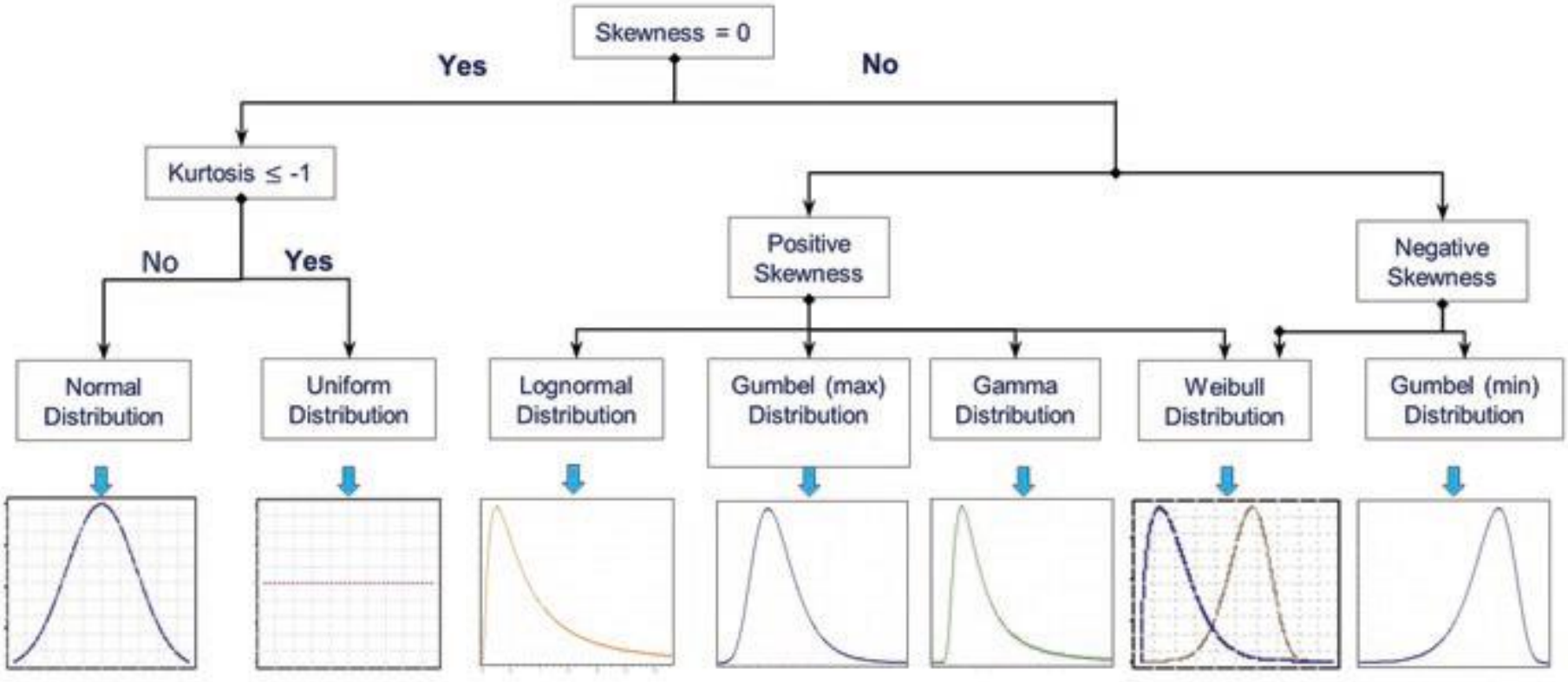
Identyfikacja rozkładu

Wybór rozkładów statystycznych wydaje się nieograniczony, ale identyfikację można zawęzić biorąc pod uwagę następujące własności:

- Czy dane są typu dyskretnego czy ciągłego?
- Jaka jest skośność i kurtoza zbioru danych?
- Jakie jest prawdopodobieństwo zaobserwowania skrajnych wartości w rozkładzie?

Identyfikując powyższe właściwości, łatwiej jest wybrać rozkład statystyczny dla danego zbioru danych.

Uproszczoną procedurę sprawdzania odpowiednich rodzajów rozkładów z uwzględnieniem tych kryteriów przedstawiono na rysunku.


Skewness = 0
Yes
No
Kurtosis ≤ -1
No
Yes
Positive Skewness
Negative Skewness
Normal Distribution
Uniform Distribution
Lognormal Distribution
Gumbel (max) Distribution
Gamma Distribution
Weibull Distribution
Gumbel (min) Distribution

9. Testy zgodności dwóch rozkładów

Do nieparametrycznych testów istotności, za pomocą których można weryfikować hipotezy, że dystrybuanty dwóch (lub większej liczby) rozpatrywanych zmiennych losowych są tożsamościowo równe należą:

- Test znaków
- Test serii
- Test rangowanych znaków
- Test mediany

A. Test znaków. Niech X oraz Y będą cechami typu ciągłego o nieznanych dystrybuantach F_X i F_Y. Chcemy sprawdzić hipotezę, że rozkłady tych cech nie różnią się istotnie, czyli hipotezę zerową postaci:

$$\mathrm{H}_0\colon\ F_X = F_Y$$

przy alternatywie $\mathrm{H}_1\colon\ F_X \neq F_Y$.

Dane: ciąg $(X_i, Y_i), i = 1, \dots, n$ niezależnych par obserwacji.

Wprowadzamy nową zm. l. $D = X - Y$.

Dla par (X_i, Y_i) wyznaczamy ich różnice $D_i = X_i - Y_i$ i otrzymujemy ciąg różnic $D_1, \dots, D_n$.

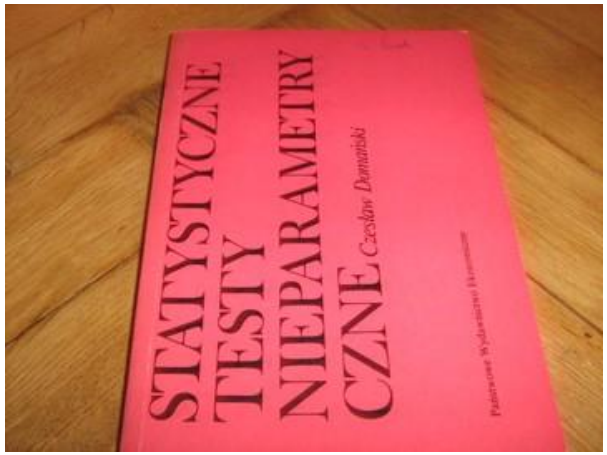
Sprawdzianem hipotezy zerowej jest liczba znaków dodatnich, które występują w ciągu realizacji statystyki D

$$L_n = |\{i\colon D_k > 0, i = 1, \dots, n\}|$$

Tw. Jeżeli hipoteza H_0 jest prawdziwa, to prawd. zaobserwowania k dodatnich różnic jest dane rozkładem $\mathrm{bin}(n;\ 1/2)$, tj.

$$P(L_n \leq k) = \frac{1}{2^n} \sum_{i=0}^{k} \binom{n}{i}$$

Więcej na temat testów nieparametrycznych można znaleźć w:



Jeżeli nie znamy rozkładu prawdopodobieństwa danych albo nie chcemy nic o nim zakładać i testujemy hipotezę zerową, że dwie próby, które ze sobą porównujemy pochodzą z populacji o takiej samej medianie, to korzystamy z klasycznego testu nieparametrycznego np.:

- **Wilcoxon rank sum test** – ranksum – próby nie są połączone w pary,

https://www.mathworks.com/help/stats/ranksum.html#bti4rqg-2

- **Wilcoxon signed rank test** – signrank – próby są połączone w pary (są sparowane),

https://www.mathworks.com/help/stats/signrank.html

- **Test znaków** signtest – próby są sparowane.

https://www.mathworks.com/help/stats/signtest.html

10. Zestaw zadań W10

1. Firma rozważa pięć projektów nazw swojego nowego produktu. Przed wybraniem jednej z nich firma postanowiła sprawdzić, czy wszystkie pięć nazw równie silnie przyciąga klientów. Wybrano próbę losową 100 osób i każdą z nich poproszono o wskazanie najlepszej spośród pięciu nazw. Liczby osób, które wybrały kolejne nazwy są podane poniżej:

Nazwa		A	B	C	D	E
Liczba osób	a)	8	16	30	34	12
	b)	4	12	34	40	10
	b)	12	30	15	15	28

Przeprowadzić test. **Odp.: b)** $\chi_0^2 = 50{,}8$, odrzucamy hipotezę zerową.

2. Istnieje przekonanie, że zwroty z pewnej inwestycji mają rozkład normalny z wartością oczekiwaną 11% (roczna stopa) i odchyleniem standardowym 2%. Firma brokerska chcąc przetestować hipotezę zerową, że przekonanie to jest prawdziwe, zebrała następujące dane o zwrotach w % (zakładamy, że jest to próba losowa):

i) 8,0; 9,0; 9,5; 9,5; 8,6; 13,0; 14,5; 12,0; 12,4; 19,0; 9,0; 10,0; 10,0; 11,7; 15,0; 10,1; 12,7; 17,0; 8,0; 9,9; 11,0; 12,5; 12,8; 10,6; 8,8; 9,4; 10,0; 12,3; 12,9; 7,0.

ii) Do wszystkich wartości dodajemy 0,2.

iii) Do nieparzystych wartości dodajemy 0,5, a od parzystych wartości odejmujemy 0,5.

a) Przeprowadź analizę statystyczną i sformułuj wniosek.

b) Przeprowadź test hipotezy zerowej, że zwroty z inwestycji mają rozkład normalny, ale z nieznaną wartością oczekiwaną.

c) Przeprowadź test hipotezy zerowej, że zwroty z inwestycji mają rozkład normalny, ale z nieznaną wartością oczekiwaną i nieznanym odchyleniem standardowym.

Odp.: i) c) $\bar{x} = 11{,}21$, $s = 2{,}71$, statystyka chi-kwadrat jest na tyle mała, że nie możemy odrzucić hipotezy zerowej.

3. Zaplanować doświadczenie do zbadania symetryczności monety. Powiedzmy, że w 100 rzutach monetą otrzymano

a) 55 orłów, b) 65 orłów.

Co możemy powiedzieć o symetryczności monety na poziomie istotności 0,05? **Odp.:** a) $\chi_0^2 = 1$; b) $\chi_0^2 = 9$; $\chi^2_{0{,}95;\,1} = 3{,}8415$.

4. Zaplanować doświadczenie do zbadania prawidłowości kostki do gry. Powiedzmy, że w 120 rzutach kostką otrzymano następujące wyniki:

Liczba oczek		1	2	3	4	5	6
Liczba rzutów	i)	14	18	23	22	28	15
	ii)	11	30	14	10	33	22

Co możemy powiedzieć o prawidłowości kostki na poziomie istotności 0,05? **Odp.: ii)** $\chi_0^2 = 24{,}5 > \chi^2_{0{,}95;5} = 11{,}0705.$

5. Do każdej z 20 tarcz oddano po 5 niezależnych strzałów i zanotowano liczbę trafień. Wyniki strzelania podane są w tabeli:

Liczba trafień	0	1	2	3	4	5
Liczba tarcz	1	2	3	10	3	1

Na poziomie istotności 0,1 zweryfikować hipotezę:

„liczba trafień do tarczy ma rozkład dwumianowy”.

Odp.: Łączymy klasy i szacujemy parametr p, $\bar{p}_n = 0{,}55$. Odrzucamy hipotezę zerową.

6. W celu zbadania rozkładu liczby awarii w sieci wodno-kanalizacyjnej pewnego miasta przeprowadzono obserwacje w ciągu 100 dni. Wyniki obserwacji podane są w tabeli.

Dzienna liczba awarii		0	1	2	3	4
Liczba dni	i)	22	30	22	16	10
	ii)	10	22	30	22	16

Na poziomie istotności 0,05 zbadać, czy rozkład liczby awarii jest rozkładem równomiernym.

7. Dane z próby zostały pogrupowane w tabeli:

Przedział	Liczba wyników
$(-\infty, -1)$	11
$[-1,1]$	76
$(1, \infty)$	13

Na poziomie istotności $\alpha = 0{,}05$ zweryfikować hipotezy:

a) dane pochodzą z rozkładu $\mathcal{N}(1;\ \sigma)$;
b) dane pochodzą z rozkładu $\mathcal{N}(m; \sigma)$,
c) dane pochodzą z rozkładu $\mathcal{N}(0;\ 1)$.

Odp.: c) $\chi_0^2 = 2{,}89 < \chi^2_{0{,}95;2} = 5{,}9915.$

8. Przeprowadzono badanie wytrzymałości betonu na ściskanie. Uzyskane wyniki pomiarów (w N/cm^2) są podane w tabeli:

Wytrzymałość	Liczba próbek	
	i)	ii)
(1900 – 2000]	14	10
(2000 – 2100]	26	26
(2100 – 2200]	52	56
(2200 – 2300]	58	64
(2300 – 2400]	33	30
(2400 – 2500]	17	14

Na poziomie istotności 0,05 sprawdzić, czy wytrzymałość betonu na ściskanie

a) ma rozkład normalny;

b) ma rozkład $\mathcal{N}(2200;\ \sigma)$;

c) ma rozkład $\mathcal{N}(2200;\ 100)$.

Odp.: ii) a) Szacujemy dwa parametry, stąd 3 stopnie swobody, $\chi_0^2 = 2{,}91 < \chi^2_{0{,}95;3}$=7,8147.

9. Dane z próby zostały pogrupowane w tabeli:

Przedział	(0, 1]	(1, 2]	(2, 3]	(3, 4]	(4, 5]	(5, 6]	(6, 7]	(7, 8]	(8, 9]	(9, 10]
liczba i)	52	38	20	12	7	6	5	5	4	1
wyników ii)	33	36	19	14	9	8	4	6	5	4

Na poziomie istotności 0,02 zweryfikować hipotezę, że dane te pochodzą z rozkładu o gęstości $f(x)$ określonej wzorem (wyznaczyć a):

$$f(x) = \begin{cases} a(10 - x) & \text{dla} \quad x \in [0, 10] \\ 0 & \text{w p. p.} \end{cases}$$

10. Badaniu poddano ciało radioaktywne ze względu na ilość emitowanych przez nie cząstek. Badanie polegało na obserwacji tego ciała w ciągu 2608 jednakowych okresów (po 7,5 sekundy każdy). Dla każdego okresu rejestrowano ilość cząstek wpadających do licznika. Wynik rejestracji jest zestawiony w postaci ciągu par i, n_i, gdzie $i = 0, 1, 2, \ldots, 10$ oznacza liczbę cząstek wpadających do licznika, natomiast n_i liczbę okresów, w których zaobserwowano i cząstek wpadających do licznika:

[0, 57], [1, 203], [2, 383], [3, 525], [4, 532], [5, 408], [6, 273], [7, 139], [8, 45], [9, 27], [10, 16].

Zbadać, czy ilość cząstek emitowanych przez badane ciało radioaktywne jest zgodna z rozkładem Poissona.

Odp.: Chisquare = 20.6105 with 10 d.f. Sig. level = 0.023979

11. (KA 6.18). Wynikami pięcioelementowej próby są: 1.37, 0.18, 0.56, 2.46, 0.87. Zapisać wyniki w zmiennej *wyn*. Na poziomie istotności $\alpha = 0{,}05$ zweryfikować hipotezę, że próba została pobrana z populacji $X \sim \exp(1)$.

12. (KA 6.19). Sprawdzić na poziomie istotności $\alpha = 0{,}05$ hipotezę, że próby $pba1$ i $pba2$, gdzie

$pba1 = \{0.46, 0.14, 2.45, -0.32, -0.07, 0.3\}$,

$pba2 =$

$\{0.06, -2.53, -0.53, -0.19, 0.54, -1.56, 0.19, -1.19, 0.02\}$

pochodzą z populacji X i Y o tym samym rozkładzie.

Odp.: zaobserwowana wartość statystyki $KS2(\boldsymbol{x}; \boldsymbol{y}) = \frac{8}{18}$

13. Wygenerować próby o liczebności 100 obserwacji według rozkładów:

i) $\mathcal{N}(900; 50)$,

ii) $TR(725; 1075)$, następnie

a) obliczyć podstawowe statystyki,
b) sporządzić wykresy histfit, normplot, Q-Q,
c) przeprowadzić testy losowości,
d) przeprowadzić testy normalności,
e) przeprowadzić testy zgodności z innymi rozkładami,
f) przeprowadzić test zgodności dla wygenerowanych prób.

14. Populacja pewnych elementów badana jest ze względu na dwie cechy, cechę Y przyjmującą wartości z przedziału $(0, 1)$ oraz cechę X przyjmującą tylko dwie wartości 0 i 1. Dane z próby losowej zebrane w czteropolowej tablicy są następujące:

$X \backslash Y$	$< 0{,}5$	$\geq 0{,}5$
0	72	29
1	53	26

Na poziomie istotności 0,05 zweryfikować hipotezę o niezależności cech X i Y.

Odp.: Nie ma podstaw do odrzucenia hipotezy o niezależności badanych cech.

15. Dla zbadania wpływu palenia tytoniu (zmienna X) na zachorowania na raka (zmienna Y) wylosowano 874 osoby z dorosłej populacji ludzi i uzyskano następujące dane:

$X \backslash Y$	ma raka	nie ma raka
pali	412	299
nie pali	32	131

Na poziomie istotności 0,02 zweryfikować hipotezę o niezależności palenia i zachorowalności na raka.

Odp.: Odrzucamy hipotezę o niezależności palenia i zachorowalności na raka.

16. Wyrób produkowany w dwóch zakładach A i B może być uznany jako wadliwy z dwóch przyczyn: J – niskiej jakości wykonania lub S – użycia gorszego surowca. Analizując losową próbę wyrobów uzyskano wyniki podane w tablicy:

$X\backslash Y$	A	B
J	33	46
S	16	5

Na poziomie istotności 0,01 zweryfikować hipotezę o niezależności między miejscem powstania wyrobu (zm. Y) a przyczyną uznania wyrobu za wadliwy (zm. X).

17. Zbadać, czy udziały w rynku firm A, B, C, D, E wynajmujących samochody zmieniły się, jeśli dane z dwóch lat dla prób losowych są następujące:

Rok \ Firma	A	B	C	D	E
I	39	26	18	14	3
II	29	25	16	19	11

18. Wygenerować dużą próbę według jednego z rozkładów: beta, gamma, Weibulla lub logarytmiczno-normalnego i przekazać uzyskane dane drugiej osobie do identyfikacji rozkładu – nie informując o mechanizmie generowania. Dokonać oceny jakości dokonanej identyfikacji.