1. STATYSTYKA MATEMATYCZNA – TESTY ZGODNOŚCI

Testy zgodności należą do grupy testów w których weryfikowane są hipotezy nieparametryczne dotyczące zgodności próby z pewnym rozkładem teoretycznym (np. normalnym), najczęściej stosowanymi testami zgodności są: test χ^2 oraz test Kołmogorowa – Smirnowa.

1.1. Test zgodności χ^2

W teście χ^2 dla zweryfikowania hipotezy o zgodności rozkładu próby z pewnym rozkładem teoretycznym porównuje się liczebności szeregu empirycznego (próby) z liczebnościami szeregu teoretycznego. Do wyznaczenia wartości statystyki testowej niezbędne jest pogrupowanie wyników próby w *szereg rozdzielczy*. Następnie wyznaczana jest wartość statystyki:

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

gdzie: r – liczba klas szeregu rozdzielczego; n_i – liczebność i–tej klasy; n – liczebność próby; p_i – prawdopodobieństwo teoretyczne, że zmienna losowa o weryfikowanym typie rozkładu przyjmie wartość należącą do i–tej klasy.

Przy założeniu prawdziwości hipotezy H_0 (rozkład z próby jest zgodny z rozkładem teoretycznym) statystyka χ^2 ma rozkład χ^2 o $\nu = r - k - 1$ stopniach swobody (k – liczba szacowanych parametrów rozkładu). Obszar krytyczny w teście budowany jest jako prawostronny.

Uwagi: Liczebność przedziałów szeregu nie powinna być mniejsza od 5. Końce przedziałów pierwszego i ostatniego przyjmuje się w nieskończoności.

Przykład 1.

Wykonano 100 pomiarów długości detalu. Średnia długość wyniosła $\bar{x} \approx 20.96$ a odchylenia standardowe $s \approx 0.69$. Dane zebrano w postaci szeregu rozdzielczego. Zweryfikować na poziomie istotności $\alpha = 0.01$ hipotezę, że rozkład długości jest rozkładem normalnym.

Długość	Liczność
[19, 19.5]	1
[19.5 20]	6
[20 20.5]	18
[20.5 21]	29
[21 21.5]	26
[21.5 22]	12
[22 22.5]	6
[22.5 23]	2

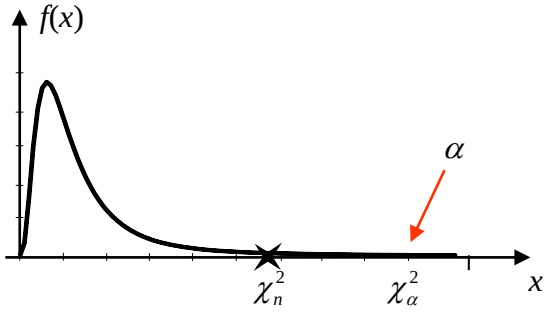
Obliczenia wartości statystyki χ^2 wygodnie jest przedstawiać w formie tabelarycznej. Ze względu na to, że liczebności w dwóch przedziałach były mniejsze od 5, przedziały te zostały połączone z sąsiednimi.

Lp	**Odchylenie od nominalnej średnicy**		n_i	p_i	n^*p_i	$n_i-n^*p_i$	$(n_i-n^*p_i)^2$	$\frac{(n_i-n^*p_i)^2}{n^*p_i}$
	od	**do**						
1		20	7	0.0828	8.28	-1.28	1.639	0.1962
2	20	20,5	18	0.1699	16.99	1.01	1.017	0.0599
3	20,5	21	29	0.2690	26.90	2.10	4.399	0.1635
4	21	21,5	26	0.2591	25.91	0.09	0.008	0.0003
5	21,5	22	12	0.1518	15.18	-3.18	10.095	0.6651
6	22		8	0.0674	6.74	1.26	1.590	0.2360
Σ	100			1			χ^2	1.3228

Obliczona na podstawie wyników z próby wartość statystyki testowej wyniosła więc:

$$\chi_n^2 \approx 1.3228.$$

Obszar krytyczny wyznacza się wykorzystując rozkład χ^2 o 3 stopniach swobody (statystyka χ^2 została zbudowana na podstawie $r = 6$ klas szeregu rozdzielczego danych, 2 parametry rozkładu były szacowane)



$$\chi_\alpha^2 = F^{-1}_{\chi^2(6-2-1)}(1-\alpha) = F^{-1}_{\chi^2(3)}(0.99) \approx 11.35$$

Wartość statystyki testowej leży poza obszarem krytycznym, nie można odrzucić hipotezy H_0.

Graniczny poziom istotności p–value dla testu prawostronnego wynosi:

$$p\text{–}value = 1 - F_{\chi^2(3)}(\chi_n^2) = 1 - F_{\chi^2(5)}(1.3228) \approx 0.72$$

Założony poziom istotności α jest niższy od poziomu granicznego nie można więc odrzucić hipotezy H_0.

1.2. Test zgodności λ Kołmogorowa (Kołmogorowa – Smirnowa)

W teście λ dla zweryfikowania hipotezy o zgodności rozkładu próby z pewnym rozkładem teoretycznym porównuje się dystrybuanty empiryczną i teoretyczną i jeśli obydwie dystrybuanty mają we wszystkich badanych punktach zbliżone wartości to uznaje się, że hipoteza o zgodności rozkładu próby z badanym rozkładem teoretycznym nie może być odrzucona. W przypadku testu Kołmogorowa zakłada się, że dystrybuanta teoretyczna jest ciągła – test ten nie może być więc stosowany do zbadania zgodności rozkładu z próby z rozkładem skokowym. Ograniczenia tego nie ma omówiony powyżej test χ^2.

Do budowy statystyki testowej λ Kołmogorowa wykorzystywana jest największa różnica pomiędzy dystrybuantami:

$$\lambda = D\sqrt{n},$$

gdzie: D – maksymalna różnica pomiędzy dystrybuantami empiryczną i teoretyczną, definiowana jako: $D = \sup_x |F_n(x) - F(x)|$, sup – supremum, kres górny zbioru, $F_n(x)$ – dystrybuanta empiryczna (wyznaczana na podstawie rozkładu empirycznego przypisującego każdej wartości z próby prawdopodobieństwo $1/n$), $F(x)$ – dystrybuanta teoretyczna.

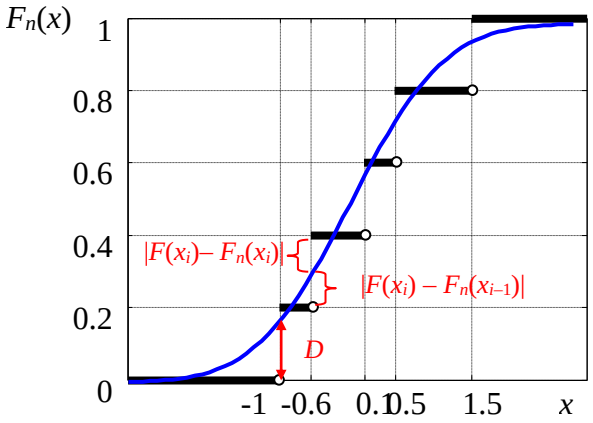
Statystyka λ przy prawdziwości hipotezy o zgodności rozkładu z wybranym rozkładem teoretycznym ma rozkład λ Kołmogorowa, obszar krytyczny w teście budowany jest jako prawostronny.

Przykład 2

Wykonano 5 pomiarów długości detalu: -1.0, 0.1, -0.6, 0.5, 1.5. Zweryfikować na poziomie istotności $\alpha = 0.01$ hipotezę, że rozkład długości jest rozkładem normalnym standaryzowanym.

Dystrybuantę empiryczną wyznacza się przypisując każdej wartości z próby prawdopodobieństwo $p = 1/n$, czyli w tym przypadku $p = 1/5 = 0.2$. Porządkując wyniki pomiarów dystrybuantę empiryczną można przedstawić w postaci tabelarycznej oraz w postaci wykresu:

x_i	-1.0	-0.6	0.1	0.5	1.5
p_i	0.2	0.2	0.2	0.2	0.2
F_n	0.2	0.4	0.6	0.8	1.0

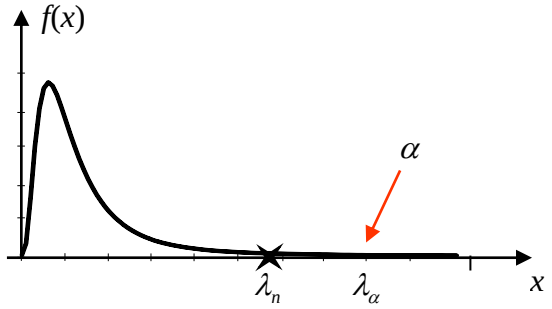


Na powyższym wykresie, dodatkowo linią w kolorze niebieskim, zaznaczono dystrybuantę standaryzowanego rozkładu normalnego. W teście Kołmogorowa wyznaczana jest największa różnica pomiędzy dystrybuantami empiryczną a teoretyczną. Odległości pomiędzy dystrybuantami wyznaczane są dla wszystkich danych z próby. Dla każdego punktu liczone są właściwie dwie odległości: odległość pomiędzy dystrybuantą teoretyczną $F(x_i)$ a dystrybuantą empiryczną $F_n(x_i)$ oraz odległość pomiędzy $F(x_i)$ a $F_n(x_{i-1})$. Obliczenia wartości statystyki λ wygodnie jest przedstawiać w formie tabelarycznej.

x_i	$F(x_i) = \Phi(x_i)$	$F_n(x_i)$	$\lvert F(x_i) - F_n(x_i)\rvert$	$\lvert F(x_i) - F_n(x_{i-1})\rvert$
-1.0	0.159	0.2	0.041	**0.159**
-0.6	0.274	0.4	0.125	0.074
0.1	0.540	0.6	0.060	0.140
0.5	0.692	0.8	0.109	0.092
1.5	0.933	1.0	0.067	0.133

Największa odległość pomiędzy dystrybuantami wystąpiła dla pierwszego punktu próby i wyniosła $D = 0.159$. Wartość statystyki testowej wynosi więc $\lambda_n = D\sqrt{n} = 0.159\sqrt{5} \approx 0.355$.

Obszar krytyczny wyznacza się wykorzystując odwrotność dystrybuanty rozkładu testowej λ Kołmogorowa:



$$\lambda_\alpha = F_\lambda^{-1}(1-\alpha) = F_\lambda^{-1}(0.99) \approx 1.628$$

Wartość statystyki testowej leży poza obszarem krytycznym, nie można odrzucić hipotezy H_0.

Graniczny poziom istotności p–value dla testu prawostronnego wynosi:

$$p\text{–}value = 1 - F_\lambda(\lambda_n) = 1 - F_\lambda(0.355) \approx 0.999$$

Założony poziom istotności α jest niższy od poziomu granicznego nie można więc odrzucić hipotezy H_0.

2. STATYSTYKA MATEMATYCZNA – DOPASOWYWANIE I TRANSFORMACJA ROZKŁADÓW

Dopasowywanie rozkładu jest procedurą poszukiwania rozkładu teoretycznego, który najlepiej odpowiada danym empirycznym. Dobrze zidentyfikowany rozkład teoretyczny pozwala na prawidłowe szacowanie prawdopodobieństw wystąpienia określonych zdarzeń, pozwala więc na podejmowanie właściwych decyzji.

Identyfikacja rozkładu teoretycznego, który najlepiej odpowiada zaobserwowanym danym to:

- wybór typu rozkładu,
- ustalenie wartości parametrów tego rozkładu.

Parametry rozkładu teoretycznego są na ogół szacowane przy pomocy:

- *metody największej wiarygodności*
  Idea metody sprowadza się do spostrzeżenia, że dane z próby są najbardziej prawdopodobne dla parametrów rozkładu odpowiadających parametrom rozkładu populacji generalnej. W celu znalezienia parametrów rozkładu budowana jest *funkcja wiarygodności*, która odpowiada prawdopodobieństwu uzyskania wartości otrzymanych z próby. Ostatecznie, poszukiwane są takie wartości parametrów dla których funkcja wiarygodności osiąga wartość maksymalną.
- *metody momentów*
  Idea metody sprowadza się do założenia, że momenty teoretyczne rozkładu odpowiadają momentom z próby. W celu znalezienia parametrów rozkładu należy znaleźć związki opisujące zależność momentów teoretycznych od parametrów rozkładu. Ostatecznie, parametry rozkładu są

znajdowane w wyniku rozwiązania układu równań, w którym momenty teoretyczne zastępowane są momentami z próby.

Zgodność rozkładu empirycznego z wybranym rozkładem teoretycznym może być oceniana poprzez analizę wykresów: histogramów z nałożoną funkcją gęstości rozkładu teoretycznego, wykresów *Q–Q* czy wykresów *P–P* lub poprzez testowanie hipotez o zgodności rozkładów.

W wielu przypadkach znajomość rozkładu danych jest niewystarczająca. Większość analiz statystycznych dotyczy zmiennych o rozkładzie normalnym. W praktyce rozkład zmiennych często odbiega od rozkładu normalnego – w takim przypadku przed wykonaniem odpowiedniej analizy konieczne jest takie przekształcenie zmiennych aby po wykonaniu przekształcenia ich rozkład był bliski normalnemu.

Rzeczywisty rozkład zmiennej decyduje o transformacji, którą należy zastosować, np.:

a) przekształcenie logarytmiczne

wykorzystywane dla zmiennych o rozkładzie zbliżonym do rozkładu logarytmiczno-normalnego (wariancja jest zbliżona do pierwiastka ze średniej),

transformacja przeprowadzana zgodnie ze wzorem:

$$Y = \log(X)$$

lub gdy zmienna *X* przyjmuje wartości < 0

$$Y = \log(X + c), \quad (c - \text{przesunięcie zmiennej}),$$

b) przekształcenie pierwiastkowe

wykorzystywane dla zmiennych o rozkładzie prawostronnie skośnym, dla których wariancja jest zbliżona do średniej,

transformacja przeprowadzana zgodnie ze wzorem:

$$Y = \sqrt{X},$$

lub gdy zmienna *X* przyjmuje wartości < 0

$$Y = \sqrt{X + c}, \quad (c - \text{przesunięcie zmiennej}),$$

c) przekształcenie Blissa

wykorzystywane dla zmiennych reprezentujących dane dotyczące proporcji w sytuacji, gdy proporcje te są mniejsze od 0,2 lub większe od 0,8,

transformacja przeprowadzana zgodnie ze wzorem:

$$Y = \arcsin\left(\sqrt{X}\right),$$

lub gdy zmienna *X* przyjmuje wartości < 0

$$Y = \arcsin\left(\sqrt{X + c}\right), \quad (c - \text{przesunięcie zmiennej}).$$

Przekształcenia *a*) – *c*) są szczególnymi przypadkami przekształceń należącymi do rodziny *przekształceń Boxa – Coxa*.

Przekształcenie Boxa – Coxa

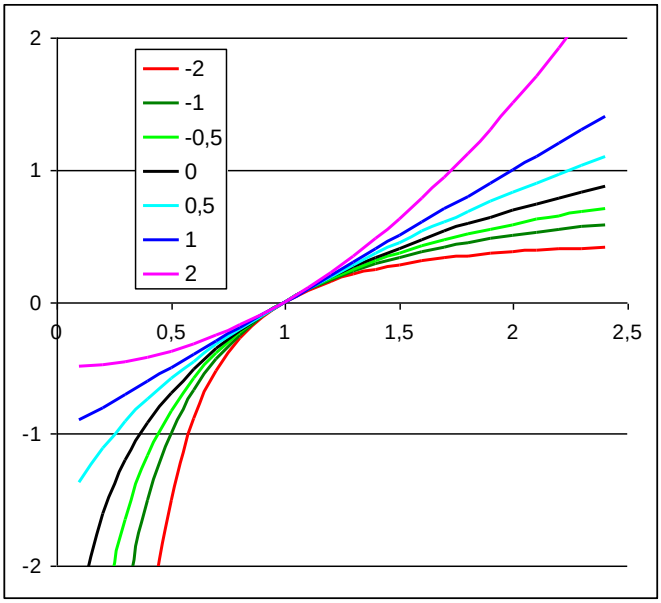
Transformacja przeprowadzana jest zgodnie ze wzorem:

$$Y = \frac{X^{\lambda} - 1}{\lambda},$$

lub gdy zmienna X przyjmuje wartości < 0:

$$Y = \frac{(X + c)^{\lambda} - 1}{\lambda},$$

(c – przesunięcie zmiennej).



Parametr λ decyduje o rodzaju przekształcenia, np.:

- dla $\lambda = 1$ *brak przekształcenia,*
- dla $\lambda = 0{,}5$ *przekształcenie pierwiastkowe,*
- dla $\lambda = 0$ *przekształcenie logarytmiczne,*
- dla $\lambda = -0{,}5$ *odwrotność pierwiastka,*
- dla $\lambda = -1$ *odwrotność.*

Przekształcenie dla $\lambda = 0$ można pokazać wykorzystując *regułę de l'Hospitala*:

$$\lim_{\lambda \to 0} \frac{X^{\lambda} - 1}{\lambda} \overset{H}{=} \lim_{\lambda \to 0} \frac{X^{\lambda} \log(X)}{1} = \log(X).$$

Nieznany parametr λ potrzebny do przekształcenia zmiennej której rozkład odbiega od normalnego może być wyznaczony na kilka sposobów:

- na podstawie wykresu normalności zmiennej Y (wykres ten przedstawia zależność kwantyli empirycznych od kwantyli rozkładu normalnego, którego parametry są uzależnione od wartości przekształconej zmiennej Y, maksymalizacja współczynnika korelacji liniowej zmiennych przedstawianych na wykresie prowadzi do wyznaczenia optymalnej wartości parametru λ),
- maksymalizując *funkcję największej wiarygodności* (stosując *metodę największej wiarygodności*) postaci:

$$L(\lambda) = -\frac{n}{2}\log(s^2) + (\lambda - 1)\sum_{i=1}^{n}\log(X_i),$$

gdzie: s^2 – odchylenie standardowe zmiennej Y, n – rozmiar próby.

Parametr λ nie może być wyznaczony analityczne – konieczne jest wykorzystanie numerycznej metody optymalizacyjnej (można zastosować np. *metodę złotego podziału*, która wykonuje minimalizację funkcji wyznaczając kolejne przybliżenia minimum w zadanym przedziale, na rys. $[a, c]$).