Goodness-of-fit test dla dyskretnych modeli

Jan Molczyk

12 maja 2026

Table of Contents

1. Cel pracy i jej kształt
2. Postać statystyki testowej
3. Symulacje

Cel pracy i jej kształt

Głównym celem pracy jest przedstawienie statystyki testowej dla testu, który będzie badał czy dana próba pochodzi z konkretnego rozkładu dyskretnego. Sama w sobie praca będzie miała pięć głównych działów

- Wstęp;
- Postać statystyki testowej;
- Wyniki teoretyczne;
- Symulacje komputerowe;
- Krótkie podsumowanie.

Wszystkie teoretyczne rozważania oraz praktyczne symulacje są oparte na podstawie artykułu "Goodness-of-fit tests for discrete models based on the integrated distribution function" autorstwa Bernharda Klara.

Założenia problemu

Niech $X_1, X_2, X_3, \ldots, X_n$ będzie próbą na wspólnej przestrzeni $(\Omega, \mathcal{A}, P)$. Próba ta zawiera niezależne zmienne losowe o jednakowym rozkładzie, których wartości są w N_0. Niech F będzie ich wspólną dystrybuantą. Problem którego rozwiązania będziemy szukać ma postać

$$H_0 : F \in \mathcal{F}_\Theta \text{ vs } H_1 : F \notin \mathcal{F}_\Theta,$$

gdzie $\mathcal{F}_\Theta = \{F(\cdot, \theta) : \theta \in \Theta\}$. Dodatkowym założeniem jest skończoność $\mathbb{E}_\theta(X_i)$ dla każdego $\theta \in \Theta$, które jest potrzebne w dalszej części.

Skumulowana dystrybuanta

Aby rozwiązać powyższy problem potrzebna jest wiedza na temat skumulowanej dystrybuanty. Dla zmiennej losowej X z skończoną wartością oczekiwaną jest ona zdefiniowana jako

$$\begin{aligned}\Psi_n(t) &= \int_t^\infty \bar{F}_n(x)\,dx = \int_t^\infty (1 - F_n(x))\,dx = \\ &= \int_t^\infty \left(1 - \frac{1}{n}\sum_{i=1}^n 1_{\{X_i \leqslant x\}}\right)\,dx = \frac{1}{n}\int_t^\infty \sum_{i=1}^n 1_{\{X_i > x\}}\,dx = \\ &= \frac{1}{n}\sum_{i=1}^n (X_i - t)1_{\{X_i > t\}},\end{aligned}$$

gdzie $\bar{F}(x)$ to funkcja przeżycia, która jest postaci $\bar{F}(x) = 1 - F$, 1 jest funkcją charakterystyczną zbioru oraz $F_n(x) = \frac{1}{n}\sum_{i=1}^n 1_{\{X_i \leqslant t\}}$ to empiryczna dystrybuanta $X_1, \ldots, X_n$.

Własności skumulowanej dystrybuanty

Skumulowana dystrybuanta posiada następujące własności:

- jest funkcją wypukłą;
- malejąca na przedziale $[0, \infty)$;
- jej wartość w zerze jest równa wartości oczekiwanej zmiennej losowej X ($\Psi_X(0) = \mathbb{E}X$).

Przypadek dyskretny

W tym miejscu przechodzimy do momentu, który nas obchodzi, czyli kiedy X jest zmienną dyskretną z gęstością prawdopodobieństwa f, a co za tym idzie skumulowana dystrybuanta przyjmie postać

$$\Psi_X(t) = \sum_{k=[t]+1}^{\infty} (k-t)\, f(k) = ([t]+1-t)\, \bar{F}(\lceil t \rceil) + \sum_{k=[t]+1}^{\infty} F(k),$$

gdzie $[t]$ oznacza część całkowitą liczby t.

Statystyka testowa

Statystyka testowa będzie miała formę

$$T_n = \sup_{t \geqslant 0} \sqrt{n} \left| \Psi_n(t) - \hat{\Psi}(t) \right|,$$

która jest podobna do statystyki Kolmogorova-Smirnova dla teoretycznych dytrybuant. Z faktu iż obie funkcje $\Psi_n(t)$ i $\hat{\Psi}(t)$ są kawałkami liniowe, to ich różnica między punktami X_i również zachowuje się liniowo. Wynika z tego, że supremum jest osiągane w punktach $t = X_i$, a co za tym idzie mamy możliwość zamienienia supremum w maximum na przedziale od 0 do M, gdzie M jest zdefiniowane jako największa wartość z wszystkich X_i ($M = \max_{1 \leqslant i \leqslant n} X_i$). Niech $k \in N_0$. Otrzymujemy wtedy

$$T_n = \sup_{k \in N_0} \sqrt{n} \left| \Psi_n(k) - \hat{\Psi}(k) \right| = \max_{0 \leqslant k \leqslant M} \sqrt{n} \left| \Psi_n(k) - \hat{\Psi}(k) \right|.$$

Inna postać statystyki testowej

Inną formą w jakiej możemy zapisać statystykę T_n jest za pomocą funkcjonału estymowanego dyskretnego procesu empirycznego zdefiniowanego jako $Z_{n,k} = \sqrt{n}\left(F_n(k) - F(k;\hat{\theta}_n)\right)$. Intuicyjnie proces ten mierzy błąd dopasowania w punkcie k, przeskalowany przez $\sqrt{n}$. Przekształcając wcześniej otrzymaną postać statystyki testowej dostajemy

$$T_n = g((Z_{n,k})_{k\geqslant 0}) = \sup_{k\geqslant 0}\left|\sum_{j\geqslant k} Z_{n,k}\right|,$$

gdzie dla wektora $\mathbf{x}$ $g(\mathbf{x}) = \sup_{k\geqslant 0}\left|\sum_{j\geqslant k} x_j\right|$.

Bootstrap parametryczny

Jest to metoda, dzięki której możemy szacować wyniki na podstawie podanej próby, poprzez wielokrotne losowanie nowych wartości z parametrem opartym na posiadanych danych. W taki sposób otrzymujemy nowy rozkład wyników, który w dalszej części możemy użyć do testów.

Bootstrap parametryczny

Kroki algorytmu

1. Wylosuj dane z zadanego rozkładu;
2. Na podstawie posiadanych danych oblicz wartość statystyki testowej oraz wyestymuj wartość parametru np. za pomocą metody wiarygodności;
3. Wygeneruj B nowych bootstrapowych prób z rozkładu, który posiada parametr wyestymowany we wcześniejszym kroku;
4. Dla każdej nowej próby policz wartość statystyki;
5. Na podstawie otrzymanych statystyk utwórz przybliżony rozkład i wyznaccz interesującą cię wartość, którą w przypadku tej pracy będzie kwantyl rzędu $1 - \alpha$;
6. Sprawdź, czy wartość statystyki testowej z kroku 2 jest większa od kwantyla z kroku 5. Jeśli tak to odrzuć hipotezę zerową;
7. Powtórz kroki 1-6 R razy, a następnie oblicz średnią liczbę odrzuceń hipotezy zerowej.

Wstępne informacje

W pracy zostaną przyjęte wartości $R, B = 1000$. Do policzenia statystyki testowej użyjemy następujących praktycznych wzorów wyprowadzonych z teorii:

$$T_n = \sqrt{n} \sup_{1 \leqslant k \leqslant M} \left| \bar{X}_n - E_{\hat{\theta}_n}(X) + \sum_{j=0}^{k-1} (F_n(j) - F(j, \hat{\theta}_n)) \right|,$$

$$\tilde{T}_n = \sum_{k=0}^{M} |Z_{n,k}| + \sqrt{n} E_{\hat{\theta}}(X) - \sqrt{n} \sum_{k=0}^{M} (1 - F(k, \hat{\theta})),$$

$$W^2_{mod} = \sum_{k=0}^{M} Z^2_{n,k}.$$

Przykład 1

Będziemy badać, czy dana próba pochodzi z rozkładu Poissona. Każdy test będzie na poziomie istotności $\alpha = 10\%$. W testach próby początkowe będą miały rozmiar $n = 50$ i $n = 200$, a same testy będą przeprowadzone na wszystkich trzech wcześniej wymienionych statystykach testowych.

Wyniki

	T		$\tilde{T}$		W_{mod}	
Rozkład	$n = 50$	$n = 200$	$n = 50$	$n = 200$	$n = 50$	$n = 200$
$P(3)$	9.1	10.1	8.4	9.8	8.9	9.7
$P(7)$	9.8	11.0	9.9	11.3	9.5	11.3
$Bin(10, 0.2)$	26.8	65.6	21.0	55.9	19.2	52.5
$Bin(10, 0.5)$	95.5	100.0	90.6	100.0	85.8	100.0
$NB(5, 0.71)$	46.4	92.0	44.6	89.2	41.4	87.4
$NB(10, 0.5)$	91.9	100.0	91.1	100.0	88.7	100.0
$NB(1, 0.3)$	100.0	100.0	100.0	100.0	99.7	100.0
$NB(1, 0.9)$	10.6	27.5	12.2	27.8	14.6	31.9
$NA(5, 0.2)$	22.2	52.3	21.7	48.3	20.4	45.7
$NA(2, 0.5)$	57.3	99.1	56.6	98.7	56.2	98.4
$P \circ L(0.5, 0.5)$	24.4	63.4	23.4	60.5	22.0	53.8
$P \circ L(0.8, 0.2)$	14.4	21.7	14.1	20.9	13.2	18.2
$P0(0.1, 3)$	42.2	91.5	44.9	91.9	48.7	94.4
$P0(0.3, 1)$	40.9	89.8	41.9	89.3	44.3	89.6
$PB(0.9, 2, 0.9)$	20.1	41.4	18.7	43.1	18.8	44.0
$PB(0.9, 20, 0.9)$	19.8	37.1	17.6	38.1	18.4	40.1
$PNB(0.1, 2, 0.5)$	84.7	100.0	82.5	100.0	80.4	100.0

Przykład 2

Podobnie jak w poprzednim przykładzie będziemy testować, czy dana próba pochodzi z rozkładu Poissona, jednak w tym przypadku rozmiar początkowych prób ulegnie zmniejszeniu. Próby te będą miały rozmiary odpowiednio $n = 20$ i $n = 31$. Poziom istotności ulegnie lekkiej zmianie, ponieważ będziemy testować zarówno dla wartości $\alpha = 10\%$, jak również $\alpha = 5\%$.

Wyniki

	$\alpha = 0.1$		$\alpha = 0.05$	
Rozkład	$n = 20$	$n = 31$	$n = 20$	$n = 31$
$P(3)$	9.1	10.2	5.2	3.7
$P(7)$	9.8	10.2	3.7	4.7
$Bin(6, 0.5)$	62.0	79.0	44.3	65.9
$Bin(3, 0.4)$	40.1	58.1	25.6	38.2
$NB(10, 0.7)$	28.2	40.8	19.9	28.3
$NB(1, 0.4)$	74.3	89.7	69.4	83.8
$U(0..4)$	15.4	19.0	5.0	9.5
$U(0..8)$	67.0	85.0	59.0	77.9
$L^-(0.6)$	57.9	72.9	48.8	64.8
$L^-(0.8)$	90.5	98.2	87.1	96.8
$Z^-(0.8)$	100.0	100.0	100.0	100.0
$\frac{1}{2}(P(2) + P(6))$	76.6	89.7	66.6	86.3

Przykład 3

Przykład ten ma na celu pokazanie jak test zachowuje się w przypadku zbadania zgodności z innym rozkładem jak Poissona. Badać będziemy zgodność dla rozkładu geometrycznego, logarytmicznego oraz positive Poissona. Przyjęte zostaną początkowe próby rozmiaru $n = 50$ oraz $n = 200$. Poziom istotności, jaki zostanie przyjęty, będzie wynosił $\alpha = 5\%$.

Wyniki

H_0: Geometric		
Distribution	$n = 50$	$n = 200$
$G(0.5)$	4.0	4.6
$NA(2, 0.5)$	14.9	56.2
$NA(4, 0.5)$	72.9	100.0
$NA(5, 0.2)$	44.5	98.6
$NA(10, 0.2)$	95.4	100.0
$P(0.3)$	11.8	53.3
$P(0.5)$	29.2	89.8
$P(0.7)$	50.3	99.0
$P \circ L(0.5, 0.7)$	13.6	35.0
$P \circ L(0.5, 0.5)$	19.8	71.6
$P \circ L(0.5, 0.3)$	28.6	84.7
$P0(0.2, 1)$	25.3	87.5
$P0(0.2, 2)$	64.2	100.0
$P0(0.5, 1)$	4.1	9.8

H_0: Logseries		
Distribution	$n = 50$	$n = 200$
$L(0.5)$	4.0	4.8
$P^+(1.05)$	51.7	99.5
$P^+(1.2)$	59.6	100.0
$P^+(1.3)$	67.6	100.0
$G^+(0.4)$	30.5	91.1
$G^+(0.33)$	37.0	97.4
$G^+(0.25)$	51.6	99.9
$Z(0.7)$	96.5	100.0
$Z(1.0)$	77.0	100.0
$Z(1.3)$	40.5	100.0
$Z(2.0)$	99.0	100.0

H_0: Positive Poisson		
Distribution	$n = 50$	$n = 200$
$P^+(3)$	4.3	4.7
$Bin^+(10, 0.2)$	58.8	99.7
$L(0.3)$	20.3	48.0
$L(0.5)$	44.9	92.3
$L(0.7)$	87.9	100.0
$G^+(0.4)$	83.8	100.0
$G^+(0.33)$	94.4	100.0
$NA^+(0.5, 2)$	99.0	100.0
$Z(1.0)$	100.0	100.0
$Z(2.0)$	97.0	100.0

Przykład 4

Ostatni przykład będzie pokazaniem, jak można w praktyce wykorzystać test. W tym celu wykorzystane zostaną prawdziwe dane na temat liczby transakcji na akcjach amerykańskiej firmy American Home Products Corporation. Obserwacje były przeprowadzone dla 243 dni w godzinach od 13:00 do 13:30. Zaobserwowane zostało 0,1,. . . ,8,12 transakcji w odpowiednio 33, 55, 68, 38, 20, 11, 8, 7, 2, i 1 dniu. Na podstawie tych danych będziemy chcieli przetestować, czy pochodzą one z rozkładu Poissona. Rezultatem obliczeń jest aproksymowana p-wartość, która powstaje za pomocą bootstrapu parametrycznego. Po wykonaniu 1000 bootstrapowych powtórzeń zwracana zostaje p-wartość $= 0$, co oznacza, że można odrzucić hipotezę zerową mówiącą o tym, że dane pochodzą z rozkładu Poissona.