Badanie zgodności z rozkładem Poissona

Michał Porczyński

2026

Plan pracy

Najważniejsze informacje o rozkładzie Poissona

Definicja

Mówimy, że zmienna losowa X ma rozkład Poissona z parametrem $\theta > 0$, gdy jej rozkład masy prawdopodbieństwa jest postaci

$$P_\theta(X = x) = e^{-\theta}\frac{\theta^x}{x!},$$

dla $x = 0, 1, ...$

Ponadto dla rozkładu Poissona zachodzi

$$\mathbb{E}[X] = Var(X) = \theta.$$

Estymatorem największej wiarogodności parametru θ jest

$$\hat{\theta} = \bar{X}$$

Funkcja tworząca prawdopodobieństwa

Definicja

Funkcja tworzącą prawdopodobieństwa (funkcję tworzącą) dyskretnej zmiennej losowej X definiujemy jako

$$G_X(t) = \mathbb{E}[t^X] = \sum_{k=0}^{\infty} t^k \mathbb{P}(X = k),$$

dla $t \in [0, 1]$

Dla funkcji tworzącej dowolnej zmiennej losowej X zachodzi

$$G_X(0) = \mathbb{P}(X = 0), \qquad G_X(1) = 1, \qquad G'_X(1) = \mathbb{E}[X]$$

Funkcja tworząca dla zmiennej losowej X pochodzącej z rozkładu Poissona z parametrem θ wyraża się wzorem

$$G_X(t) = \phi(t, \theta) = e^{\theta(t-1)}$$

Postać hipotezy zerowej

Rozważać będziemy następujący problem testowania

$$H_0 : F \in \{P(\theta)\}_{\theta>0},$$

$$H_1 : F \notin \{P(\theta)\}_{\theta>0}.$$

W następnej części przedstawiać będziemy kolejne statystyki testowe oparte na funkcji tworzącej. W tym celu zdefiniujmy jej empiryczną wersję. Mianowicie

$$\phi_n(t) = \frac{1}{n}\sum_{i=1}^{n} t^{X_i}.$$

Statystyka testowa Kocherlakoty i Kocherlakoty

Test opiera się na badaniu różnicy między empiryczną a teoretyczną funkcją tworzącą. Niech

$$\xi(t) = \phi_n(t) - \phi(t, \hat{\theta}).$$

Wtedy przy H_0

$$\text{Var}(\xi(t)) = \frac{1}{n} e^{\bar{X}(t^2-1)} - \frac{1}{n}(1 + \bar{X}(t-1)^2) e^{2\bar{X}(t-1)}.$$

Statystykę testową Kocherlakoty i Kocherlakoty definiujemy jako

$$K_n = \frac{\phi_n(t) - \phi(t, \bar{X})}{\sqrt{\text{Var}(\xi(t))}} = \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} t^{X_i} - \mathbb{E}\left[t^X\right]\right)}{\sqrt{e^{\bar{X}(t^2-1)} - (1 + \bar{X}(t-1)^2) e^{2\bar{X}(t-1)}}}.$$

Modyfikacja testu Kocherlakoty i Kocherlakoty

Poprzedni test można zmodyfikować tak, aby badał wartość błędu nie dla jednej konkretnej wartości t, lecz dla q wartości t. Niech

$$\xi = \xi(t_1, ..., t_q) = \begin{pmatrix} \phi_n(t_1) - e^{\hat{\theta}(t_1 - 1)} \\ \vdots \\ \phi_n(t_q) - e^{\hat{\theta}(t_q - 1)} \end{pmatrix}$$

Oznaczając przez Y macierz kowariancji tego wektora losowego dostajemy, że pojedynczy wyraz macierz Y ma wartość

$$Y_{ij} = \frac{1}{n}\left(e^{\hat{\theta}(t_i t_j - 1)} - e^{\hat{\theta}(t_i - 1)} e^{\hat{\theta}(t_j - 1)} - \hat{\theta}(t_i - 1)(t_j - 1) e^{\hat{\theta}(t_i - 1)} e^{\hat{\theta}(t_j - 1)}\right),$$

dla $i, j = 1, ..., q$. Ostatecznie możemy skonstruować statystykę testową postaci

$$K_n^* = \xi' Y^{-1} \xi,$$

Statystyka testowa Ruedy

Statystyka testowa zaproponowana przez Ruedę ma postać

$$R_n = n \int_0^1 (\phi_n(t) - \phi(t, \theta))^2 \, dt.$$

Do symulacji wykorzystujemy rozwiniętą postać tej statystyki testowej. Mamy

$$R_n = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{1}{X_i + X_j + 1} - 2e^{-\hat{\theta}} \sum_{i=1}^{n} \int_0^1 e^{\hat{\theta} t} t^{X_i} \, dt + \frac{n}{2\hat{\theta}} \left(1 - e^{-2\hat{\theta}}\right).$$

Statystyka testowa Nakamury i Pereza Abreu

Test ten wykorzystuje własność funkcji tworzącej dla rozkładu Poissona. Mianowicie

$$\frac{\partial^2 \log(\phi(t,\theta)}{\partial t^2} = \frac{\partial^2 \log(e^{\theta(t-1)})}{\partial t^2} = \frac{\partial^2 \theta(t-1)}{\partial t^2} = 0$$

Niech $Y_n(t) = \log(\phi_n(t))$. Wtedy

$$\frac{\partial}{\partial t} Y_n(t) = \frac{1}{\phi_n(t)} \frac{\partial}{\partial t} \phi_n(t),$$

$$\frac{\partial^2}{\partial t^2} Y_n(t) = \frac{\phi_n(t) \frac{\partial^2}{\partial t^2} \phi_n(t) - (\frac{\partial}{\partial t} \phi_n(t))^2}{(\phi_n(t))^2}.$$

$$\frac{\partial}{\partial t} \phi_n(t) = \frac{1}{n} \sum_{i=1}^{n} X_i t^{X_i - 1} \qquad \frac{\partial^2}{\partial t^2} \phi_n(t) = \frac{1}{n} \sum_{i=1}^{n} X_i (X_i - 1) t^{X_i - 2}$$

Test Nakamury i Pereza Abreu c.d.

Chcielibyśmy aby również $\frac{\partial^2}{\partial t^2} Y_n(t) = 0$. Niech

$$N_n(t) = \phi_n(t)\frac{\partial^2}{\partial t^2}\phi_n(t) - (\frac{\partial}{\partial t}\phi_n(t))^2$$

Po podstawieniach otrzymujemy

$$N_n(t) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i(X_i - X_j - 1)t^{X_i+X_j-2}.$$

Oznaczmy $X_{(n)} = \max\{X_1, ..., X_n\}$. $N_n(t)$ jest wielomianem stopnia $d(n) = 2X_{(n)} - 2$, to znaczy $N_n(t) = \sum_{k=0}^{d(n)} a_k t^k$, gdzie

$$a_k = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} (X_i(X_i - X_j - 1))\, \mathbb{1}_{\{X_i+X_j-2=k\}}.$$

Test Nakamury i Pereza Abreu c.d.

Badając czy $N_n(t)$ jest bliskie zeru można badać poszczególne współczynniki a_k. W związku z tym rozważmy sumę kwadratów współczynników wielomianu. Mamy

$$T_n = \sum_{k=0}^{d(n)} a_k^2.$$

Korzystając z reprezentacji a_k uzyskujemy

$$T_n = \sum_{k=0}^{d(n)} \frac{1}{n^4} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} X_i(X_i - X_j - 1) \mathbb{1}_{\{X_i + X_j - 2 = k\}} \right.$$
$$\left. \times \sum_{l=1}^{n} \sum_{m=1}^{n} X_l(X_l - X_m - 1) \mathbb{1}_{\{X_l + X_m - 2 = k\}} \right).$$

Test Nakamury i Pereza Abreu c.d.

Zauważmy, że poszczególne składniki będą niezerowe wtedy i tylko wtedy, gdy $X_i + X_j - 2 = k$ oraz $X_l + X_m - 2 = k$, czyli gdy $X_i + X_j = X_l + X_m$. Korzystając z tego warunku otrzymujemy

$$T_n = \frac{1}{n^4} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{l=1}^{n} \sum_{m=1}^{n} X_i(X_i - X_j - 1)X_l(X_l - X_m - 1)\mathbb{1}_{\{X_i+X_j=X_l+X_m\}}.$$

Ostatecznie otrzymujemy statystykę testową postaci

$$T_n^* = \frac{nT_n}{\bar{X}_n^{\,1.45}}.$$

Statystyka testowa Meintanisa i Nikitina

Ostatni z testów wykorzystuje następującą własność funkcji tworzącej dla rozkładu Poissona. Mianowicie

$$\frac{\partial \phi}{\partial t}(t,\theta) = \theta \phi(t,\theta)$$

Dla statystyki testowej Meintanisa i Nikitina mamy

$$MN = \sqrt{n} \int_0^1 \left(\frac{d\phi_n(t)}{dt} - \bar{X}\phi_n(t) \right) t^a dt,$$

Oznaczmy przez

$$\epsilon_a = \mathbb{E}\left[\frac{X}{X+a}\right] - \theta\, \mathbb{E}\left[\frac{1}{X+a+1}\right]$$

Statystyka testowa Meintanisa i Nikitina c.d.

Po kolejnych obliczeniach oznaczmy przez

$$\bar{T}_{MN} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i}{X_i + a} - \frac{\theta}{X_i + a + 1} - \epsilon_a - (X_i - \theta)\, \mathbb{E}\left[\frac{1}{X + a + 1} \right] \right).$$

$$V(\theta) = \mathbb{E}\left[\left(\frac{X}{X + a} \right)^2 \right] + \theta^2 \mathbb{E}\left[\frac{1}{(X + a + 1)^2} \right] + \theta \left(\mathbb{E}\left[\frac{1}{X + a + 1} \right] \right)^2 -$$
$$2\theta \mathbb{E}\left[\frac{X}{(X + a)(X + a + 1)} \right] - 2\mathbb{E}\left[\frac{1}{X + a + 1} \right] \mathbb{E}\left[\frac{X(X - \theta)}{X + a} \right] +$$
$$2\theta \mathbb{E}\left[\frac{1}{X + a + 1} \right] \mathbb{E}\left[\frac{X - \theta}{X + a + 1} \right].$$

Ostateczna statystyka testowa ma wtedy postać

$$MN_a = \frac{\bar{T}_{MN}}{\sqrt{V(\bar{X})}}.$$

Inne wykorzystywane statystyki testowe

Test Fishera

$$F_n = \frac{(n-1)S_n^2}{\bar{X}_n}.$$

Test Kołmogorowa-Smirnowa

$$KS = \max_{X_i} \left\{ \left| F_n(X_i) - F(X_i) \right| \right\},$$

Test χ^2

$$\chi^2 = \frac{\sum_{j=1}^{k} (O_j - E_j)^2}{E_j},$$

Inne wykorzystywane statystyki testowe c.d.

Test Gupty

$$G = \frac{1}{2}\sqrt{\frac{n}{1+24\bar{X}+6\bar{X}^2}}\,\frac{m_2(m_4-3m_2^2)-m_3^2}{\bar{X}^2},$$

gdzie $m_r = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^r$.

Bootstrap parametryczny

Algorytm symulacji

1. *Gemerujemy* $\underline{X} = (X_1, ..., X_n)$, *czyli próbę z zadanego rozkładu*
2. *Obliczamy wartość danej statystyki testowej -* $T(\underline{X})$ *- na podstawie wygenerowanej próby.*
3. *Wyznaczamy* $\hat{\theta} = \bar{X}$ *na podstawie próby* $\underline{X}$.
4. *Generujemy próbę* $\underline{Y} = (Y_1, ..., Y_n)$ *i.i.d, gdzie* $Y_i \sim Pois(\hat{\theta})$.
5. *Obliczamy wartość danej statystyki testowej na podstawie próby* $Y_1, ..., Y_n$ *uzyskując* $T(\underline{Y}) = T_j$.
6. *Powtarzamy korki 3-5 B razy otrzymując* $T_1, ..., T_B$.

Bootstrap parametryczny c.d.

7. Wyznaczamy prawostronny obszar krytyczny przez wyznaczenie kwantylu rzędu $1-\alpha$*, tzn.*

$$\hat{q}_{1-\alpha} = \alpha T_{((1-\alpha)B)} + (1-\alpha) T_{((1-\alpha)B+1)},$$

gdzie $T_{(1)} \leq T_{(2)} \leq ... \leq T_{(B)}$*. W niektórych sytuacjach należy wyznaczyć dwustronny obszar krytyczny. Mamy*

$$\hat{q}_{\alpha/2} = \alpha T_{(B\alpha/2-1)} + (1-\alpha) T_{(B\alpha/2)},$$

$$\hat{q}_{1-\alpha/2} = \alpha T_{((1-\alpha/2)B)} + (1-\alpha) T_{((1-\alpha/2)B+1)},$$

8. Odrzucamy H_0*, gdy* $T(\underline{X}) > \hat{q}_{1-\alpha}$ *(lub gdy* $T(\underline{X}) > \hat{q}_{1-\alpha/2}$ *albo* $T(\underline{X}) < \hat{q}_{\alpha/2}$ *- w sytuacji, gdy analizujemu dwustronny obszar krytyczny).*
9. Powtarzamy kroki 1-8 N razzy notując 1 w przypadku odrzucenia H_0 *oraz 0 w przeciwnym przypadku.*
10. Otrzymując N-elementowy wektor stworzony z 0 oraz 1 wyznaczamy moc testu ze wzoru $\text{moc}(T) = \frac{\#1}{N}$*.*

Badanie mocy testów - dane z rozkładu jednostajnego

Parametry		Moce testów								
a	b	K_n	K_n^*	R_n	T_n^*	MN_a	F_n	χ^2	KS	G
0	4	0.139	0.179	0.125	0.248	0.050	0.025	0.155	0.186	0.071
0	8	0.628	0.686	0.772	0.685	0.666	0.581	0.575	0.498	0.624
0	10	0.735	0.831	0.882	0.820	0.852	0.802	0.723	0.629	0.797
0	20	0.954	0.990	0.997	0.992	0.996	0.996	0.977	0.926	0.986
1	7	0.000	0.020	0.001	0.210	0.011	0.013	0.125	0.138	0.055
5	15	0.000	0.000	0.001	0.172	0.005	0.007	0.120	0.114	0.050
1	10	0.011	0.234	0.317	0.577	0.398	0.410	0.418	0.387	0.496

Table 1. Moce testów dla $n = 20$

Badanie mocy testów - dane z rozkładu geometrycznego

Parametry	Moce testów								
p	K_n	K_n^*	R_n	T_n^*	MN_a	F_n	χ^2	KS	G
0.2	0.055	0.819	0.773	0.946	0.891	0.955	0.894	0.864	0.896
0.4	0.147	0.625	0.116	0.479	0.147	0.327	0.132	0.372	0.257
0.5	0.542	0.834	0.451	0.295	0.257	0.101	0.101	0.264	0.071
0.6	0.892	0.969	0.837	0.289	0.592	0.020	0.320	0.259	0.012
0.8	1.000	1.000	1.000	0.857	0.990	0.000	0.751	0.613	0.000

Table 2. Moce testów dla $n = 20$

Badanie mocy testów - dane z rozkładu ujemnego dwumianowego

Parametry		Moce testów								
r	p	K_n	K_n^*	R_n	T_n^*	MN_a	F_n	χ^2	KS	G
5	0.2	0.755	0.957	0.995	0.992	0.993	0.996	0.887	0.911	0.966
5	0.5	0.237	0.384	0.509	0.576	0.564	0.683	0.321	0.342	0.536
5	0.9	0.069	0.058	0.065	0.096	0.072	0.103	0.052	0.062	0.081
10	0.2	0.293	0.000	0.996	0.994	0.994	0.997	0.957	0.902	0.965
10	0.5	0.223	0.414	0.590	0.592	0.571	0.696	0.253	0.333	0.541
10	0.9	0.067	0.047	0.067	0.098	0.071	0.102	0.050	0.059	0.089
20	0.2	0.027	0.000	0.997	0.995	0.995	0.997	0.898	0.886	0.967
20	0.5	0.137	0.355	0.628	0.613	0.585	0.706	0.308	0.309	0.539
20	0.9	0.061	0.051	0.068	0.088	0.068	0.101	0.049	0.062	0.088

Table 3. Moce testów dla $n = 20$

Badanie mocy testów - dane z rozkładu ZIP

Parametry		Moce testów								
θ	π	K_n	K_n^*	R_n	T_n^*	MN_a	F_n	χ^2	KS	G
1	0.1	0.065	0.051	0.061	0.082	0.061	0.090	0.062	0.061	0.077
1	0.5	0.282	0.260	0.275	0.313	0.268	0.352	0.292	0.279	0.281
2	0.1	0.137	0.117	0.137	0.130	0.113	0.148	0.121	0.107	0.119
2	0.5	0.791	0.720	0.779	0.733	0.734	0.762	0.720	0.766	0.720
5	0.1	0.646	0.710	0.707	0.380	0.507	0.410	0.702	0.234	0.307
5	0.5	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	0.994
10	0.1	0.882	0.883	0.884	0.691	0.831	0.696	0.885	0.312	0.565
10	0.5	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000

Table 4. Moce testów dla $n = 20$