Spis treści

1 Rozkłady

1.1 Rozkład dwumianowy

Rozkład dwumianowy to rozkład sumy n zmiennych losowych o rozkładzie Bernoulliego. Zmienna losowa X ma rozkład dwumianowy z parametrami n i p, zatem:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

gdzie $\binom{n}{k}$ to liczba kombinacji k sukcesów w n próbach.

1.2 Rozkład normalny

Rozkład normalny (Gaussa) jest jednym z najważniejszych rozkładów statystycznych. Jest on określony przez dwa parametry: wartość oczekiwaną μ i wariancję σ^2. Gęstość rozkładu normalnego jest dana wzorem:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

1.3 Rozkład chi-kwadrat

Niech $X_1, X_2, \ldots, X_n$ będą niezależnymi zmiennymi losowymi o rozkładzie normalnym $N(0,1)$. Wtedy zmienna losowa $X = \sum_{i=1}^{n} X_i^2$ ma rozkład chi-kwadrat ($X \sim \chi(n)$) z n stopniami swobody.

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x/2}$$

1.4 Rozkład wykładniczy

Jest to rozkład zmiennej, która opisuje czas między zdarzeniami w procesie Poissona. Zmienna losowa X ma rozkład wykładniczy z parametrem λ, zatem:

$$f(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{dla } x \geq 0 \\ 0 & \text{dla } x < 0 \end{cases}$$

1.5 Rozkład t-Studenta

Niech $X \sim N(0,1) oraz Y \sim \chi^2(n)$ będą niezależnymi zmiennymi losowymi. Wtedy zmienna losowa:

$$\frac{X}{\sqrt{Y/n}}$$

ma rozkład t-Studenta z n stopniami swobody. Funkcja gęstości rozkładu t-Studenta jest dana wzorem:

$$f(x, n) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\Gamma(\frac{n}{2})} (1 + \frac{x^2}{n})^{-\frac{n+1}{2}}$$

1.6 F-Snedecora

Niech $X \sim \chi^2(n)$, $Y \sim \chi^2(m)$ będą niezależnymi zmiennymi losowymi. Wtedy zmienna losowa:

$$\frac{X/n}{Y/m}$$

ma rozkład F-Snedecora z n i m stopniami swobody. Funkcja gęstości rozkładu F-Snedecora jest oznacza przez $F(n, m)$.

2 Statystyka Opisowa

Niech $X' = \{x_1, x_2, \ldots, x_n\}$ będzie zbiorem n obserwacji zmiennej losowej X. Zadaniem statystyki opisowej jest prezentacja rozkładu zmiennej losowej X w próbce X'.

2.1 Rodzaje statystyk opisowe

- Klasyczne - uśredniające wartość próbki. Na przykład momenty zwykłe r-tego rzędu:

$$m_r = \frac{1}{n} \sum_{i=1}^{n} x_i^r$$

- Pozycyjne - oparte na pozycjach obserwacji w próbce. Na przykład mediana, kwartyle, percentyle.

2.2 Tendencja centralnej rozkładu empirycznego

- Średnia arytmetyczna:

$$\overline{X'} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Mediana

2.3 Charakterystyki rozrzutu rozkładu empirycznego

- Odchylenie standardowe:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X'})^2}$$

- Współczynnik zmienności:

$$v = \frac{s}{\overline{X'}} \cdot 100\%$$

3 Model statystyczny

Jeżeli próba X' jest reprezentatywna, to można na jej podstawie wnioskować na temat populacji z której pochodzi. Aby określić zachowanie zmiennej losowej X w populacji, stosuje się model statystyczny. Zatem traktujemy wektor X' jako realizację zmiennej losowej X.

4 Estymacja Punktowa

Niech X' będzie próba populacji o rozkładzie P_θ gdzie $\theta \in \Theta$ jest parametrem. Estymatorem parametru θ nazywamy statystykę $\hat{\theta}$: $X' \to \Theta$ która pozwala na oszacowanie wartości parametru θ.

4.1 Metoda momentów

Metoda momentów polega na przyrównaniu kolejnych d momentów $m_1, \ldots, m_d$ do odpowiednich momentów rozkładu populacji $E(X^i) : i \in [1, d]$

4.2 Metoda największej wiarygodności

Funkcję $L(\theta, x) = p_\theta(x)$ nazywamy funkcją wiarygodności. Estymatorem największej wiarygodności parametru θ nazywamy nazywamy statystykę $\hat{\theta}$ która maksymalizuje funkcję wiarygodności.

$$\forall_{x \in X} L(\hat{\theta}, x) = \sup_{\theta \in \Theta} L(\theta, x)$$

4.3 Przykład

Estymatorem największej wiarygodności oraz metody momentów dla rozkładu wykładniczego z parametrem λ jest:

$$\hat{\lambda} = \frac{1}{\overline{X}}$$

4.4 Estymatory nieobciążone

Estymator $\hat{\theta}$ nazywamy nieobciążonym, jeżeli $E(\hat{\theta}) = \theta$

4.5 Estymator modelu wykładniczego

Dla modelu wykładniczego, parametryzowanego przez λ, estymatorem nieobciążonym jest:

$$\hat{\lambda} = \frac{n-1}{n}\frac{1}{\overline{X}}$$

4.6 Estymator modelu normalnego

Dla modelu normalnego, parametryzowanego przez μ i σ^2, estymatorem nieobciążonym jest:

$$\hat{\mu} = \overline{X}$$

$$\hat{\sigma^2} = S^2$$

4.7 Metoda monte carlo

Niech X' będzie próbą populacji o rozkładzie P_θ, oraz niech $\hat{\theta}$ będzie estymatorem parametru θ. Załóżmy też, że mamy k niezależnych realizacji próby $x_1, \dots x_k$. Wtedy histogram wartości $\hat{x_n}$: $n \in [1, k]$ jest przybliżeniem rozkładu $\hat{\theta}$.

4.8 Metoda bootstrap

Dystrybuanta empiryczna to statystyka o następującej postaci:

$$\hat{F}(x) = \frac{\#\{k : X_k \leq x\}}{n}$$

Dla takiej dystrybuanty i próby X' zachodzi:

$$\sup_{x \in \mathbb{R}} |\hat{F}(x) - F(x)| \xrightarrow{1} 0$$

Próba bootstrapową X^* to próba losowa z rozkładu empirycznego. Ta próba musi powstać w wyniku n-krotnego losowania z zwracaniem. Rozkład statystyki $T(X^*) - \hat{\theta}$ jest bliski rozkładowi statystyki $T(X) - \theta$.

Mając k realizacji prób bootstrapowych $X_1^*, \dots, X_k^*$, możemy przybliżyć rozkład statystyki $\hat{\theta} - \theta$, poprzez stworzenie histogramu $\hat{\theta} *_n : n \in [0, k]$

5 Przedziały ufności

Przedział ufności to przedział $[L, R]$ określony parą statystyk, takich, że:

$$P_\theta(L < \theta < R) = 1 - \alpha$$

gdzie α to poziom ufności, a θ to parametr modelu.

Funkcję $Q(X, \theta)$ nazywamy funkcją centralną dla parametru θ, jeżeli rozkład prawdopodobieństwa zmiennej Q jest absolutnie ciągły i nie zależy od parametru θ, oraz funkcja Q jest ciągła i ściśle monotoniczna względem θ.

Obieranie przedziału ufności następuje poprzez rozwiązanie nierówności:

$$a < Q(X, \theta) < b$$

gdzie a i b się z reguły dobiera tak aby:

$$P(Q \leq a) = P(Q \geq b) = \frac{\alpha}{2}$$

6 Testy statystyczne

6.1 Hipotezy statystyczne

Mając do czynienia z daną hipotezą (H), np: parametr θ jest równy θ_0, chcemy sprawdzić, czy hipoteza jest prawdziwa. W tym celu stosujemy test statystyczny, który jest funkcją $T(X')$ z próby X'. Pierwsza utworzona hipoteza to hipoteza zerowa(H_0), podczas kolejnych testów tworzymy hipotezy alternatywne, lub przyjmujemy, że hipoteza zerowa jest prawdziwa.

6.2 Obszar krytyczny

Podczas wyznaczania procedury testowej danej hipotezy, określamy jej obszar krytyczny, czyli zbiór wartości statystyki testowej, dla których odrzucamy hipotezę. Najbardziej typowy jest prawostronny obszar krytyczny, zdefiniowany jako:

$$\mathcal{R} = \{x : T(x) \geq k\}$$

6.3 Błędy

Rozróżniamy dwa rodzaje błędów:

- **Błąd I rodzaju**: Odrzucenie hipotezy zerowej, gdy jest ona prawdziwa. Prawdopodobieństwo popełnienia błędu I rodzaju to α.
- **Błąd II rodzaju**: Nieodrzucenie hipotezy zerowej, gdy jest ona fałszywa. Prawdopodobieństwo popełnienia błędu II rodzaju to β.

6.4 p-wartość

p-wartość to najmniejsza wartość poziomu istotności α, dla której odrzucamy hipotezę zerową. Jest to prawdopodobieństwo uzyskania wartości statystyki testowej, która jest równa lub bardziej ekstremalna niż obserwowana wartość statystyki testowej, zakładając, że hipoteza zerowa jest prawdziwa.

7 Test t-Studenta

Test t-Studenta jest testem statystycznym, który służy do porównywania średnich wartości dwóch prób lub jednej próby z wartością oczekiwaną.

7.1 Dla jednej próby

- **Hipoteza zerowa**: $H_0 : \mu = \mu_0$
- Hipotezy alternatywne: $H_1 : \mu \neq \mu_0$, $H_1 : \mu > \mu_0$, $H_1 : \mu < \mu_0$
- Statystyka testowa:

$$t = \frac{\overline{X} - \mu_0}{S}\sqrt{n}$$

- Rozkład statystyki testowej: $t|_{H_0} \sim t(n-1)$

7.2 Dla dwóch prób

Posiadamy obserwacje jednej zmiennej (cechy) na jednostkach eksperymentalnych pochodzących z dwóch populacji (grup) lub posiadamy dwukrotne obserwacje tej samej zmiennej na tych samych jednostkach eksperymentalnych jednej populacji. Wyróżniamy dwa rodzaje prób: niezależne oraz zależne.

7.2.1 Błąd

Błąd to różnica między wartością zmiennej a wartością przewidywaną przez model. Zakładamy, że błąd:

- ma rozkład normalny
- jest niezależny od zmiennej
- ma wartość oczekiwaną równą 0
- ma stałą wariancję

7.2.2 Próby niezależne z jednorodnymi wariancjami

$$\hat{\mu_1} = \overline{X}_1$$
$$\hat{\mu_2} = \overline{X}_2$$
$$\hat{\sigma^2} = S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$$
$$S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(x_{ij}-\overline{X}_i)^2$$

- **Hipoteza zerowa**: $H_0 : \mu_1 = \mu_2$
- Hipotezy alternatywne: $H_1 : \mu_1 \neq \mu_2$, $H_1 : \mu_1 > \mu_2$, $H_1 : \mu_1 < \mu_2$
- Statystyka testowa:
$$t = \frac{\overline{X}_1 - \overline{X}_2}{S}\sqrt{\frac{n_1 n_2}{n_1+n_2}}$$
- Rozkład statystyki testowej: $t|_{H_0} \sim t(n_1+n_2-2)$

7.2.3 Próby niezależne z różnymi wariancjami

$$\hat{\mu_1} = \overline{X}_1$$
$$\hat{\mu_2} = \overline{X}_2$$
$$\hat{\sigma_1^2} = S_1^2 = \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_{1i}-\overline{X}_1)^2$$
$$\hat{\sigma_2^2} = S_2^2$$

- **Hipoteza zerowa**: $H_0 : \mu_1 = \mu_2$
- Hipotezy alternatywne: $H_1 : \mu_1 \neq \mu_2$, $H_1 : \mu_1 > \mu_2$, $H_1 : \mu_1 < \mu_2$
- Statystyka testowa:
$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$$
- Rozkład statystyki testowej: $t|_{H_0} \sim t(m)$ (test Welch)

8 Test F

Test F jest testem statystycznym, który służy do porównywania wariancji dwóch prób. Jest to test parametryczny, który zakłada, że dane mają rozkład normalny. Test F jest często stosowany w analizie wariancji (ANOVA). Będziemy zakładać dwie różne wariancje dla dwóch prób.

- **Hipoteza zerowa**: $H_0 : \sigma_1^2 = \sigma_2^2$
- Hipotezy alternatywne: $H_1 : \sigma_1^2 \neq \sigma_2^2$, $H_1 : \sigma_1^2 > \sigma_2^2$, $H_1 : \sigma_1^2 < \sigma_2^2$
- Statystyka testowa:
$$F = \frac{S_1^2}{S_2^2}$$
- Rozkład statystyki testowej: $F|_{H_0} \sim F(n_1-1, n_2-1)$

9 Analiza wariancji (ANOVA)

Analiza wariancji (ANOVA) jest techniką statystyczną służącą do porównywania średnich wartości więcej niż dwóch grup. ANOVA w istocie jest generalizacją testu t-Studenta dla więcej niż dwóch grup. ANOVA pozwala na sprawdzenie, czy istnieją istotne różnice między średnimi wartościami grup, a także określenie, które grupy różnią się między sobą.

9.1 Jednoczynnikowa ANOVA

- **Hipoteza zerowa**: $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$
- Hipotezy alternatywne: $H_1 : \mu_i \neq \mu_j$ dla co najmniej jednej pary (i, j)
- Statystyka testowa:

$$F = \frac{n-k}{k-1} \frac{\sum_{i=k}^{k} n_i (\overline{X}_i - \overline{X})^2}{\sum_{i=k}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{X}_i)^2}$$

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{k} n_i \overline{X}_i$$

$$\overline{X}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

- Założenia:
  - dane są niezależne
  - dane mają rozkład normalny
  - wariancje są jednorodne (homogeniczne)

10 Test Barletta

Test Barletta jest testem statystycznym służącym do porównywania średnich wartości dwóch prób, gdy nie możemy założyć, że wariancje są jednorodne. Jest to test nieparametryczny, który nie zakłada rozkładu normalnego danych.

- **Hipoteza zerowa**: $H_0 : \mu_1^2 = \mu_2^2 = \ldots = \mu_k^2$
- Hipotezy alternatywne: $H_1 : \mu_i^2 \neq \mu_j^2$ dla co najmniej jednej pary (i, j)
- Statystyka testowa:

$$B = \frac{1}{C}(n-k) \ln MSE - \sum_{i=1}^{k} (n_i - 1) \ln S_i^2$$

$$C = 1 + \frac{1}{3(k-1)} (\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{n-k})$$