# Prawdopodobieństwo & Statystyka.

Ekonometria WNE UW

Sebastian Zalas

s.zalas@uw.edu.pl

Proszę nie rozpowszechniać - Wersja niekompletna.

Zadania przygotowane do powtórzenia.

## Spis treści

# I. Rachunek prawdopodobieństwa.

## I.A. *Wektor losowy*

- **Wektor losowy** to wektor którego elementy są zmiennymi losowymi.

## I.B. *Wartość oczekiwana*

- Dla dyskretnej zmiennej losowej $X$ z funkcją masy prawdopodobieństwa $f$ (ang. PMF), jeżeli $\sum_x |x| f(x) < \infty$, wtedy **wartość oczekiwana** zmiennej losowej równa się:

  $$\mathbb{E}[X] = \sum_x x f(x)$$

  Dla ciągłej zmiennej losowej $X$ z funkcją gęstości prawdopodobieństwa (ang. PDF) $f$, jeżeli $\int_x |x| f(x) dx < \infty$, wtedy **wartość oczekiwana** zmiennej losowej równa się:

  $$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x) dx$$

- O **wartości oczekiwanej** zmiennej losowej można myśleć jako o wartości jaką byśmy otrzymali jeżeli wzięlibyśmy średnią z wielu realizacji tej zmiennej losowej. Jest to najbardziej znana miara "środka"rozkładu prawdopodobieństwa. Wartość oczekiwana przyjmuje zmienną losową, a zwraca skalar (liczbę).

- Własności wartości oczekiwanej:

  - $\forall c \in \mathbb{R},\ \mathbb{E}[c] = c$
  - $\forall a \in \mathbb{R},\ \mathbb{E}[aX] = a\,\mathbb{E}[X]$

- Możemy uogólnić pojęcie wartości oczekiwanej do dwuwymiarowego przypadku (także analogicznie do trójwymiarowego i więcej). Ponieważ każdy element wektora losowego jest zmienną losową, wartość oczekiwana wektora losowego jest zdefiniowana jako wektor wartości oczekiwanych.

  - Dla wektora losowego $[X, Y]$ wartość oczekiwana równa się:

    $$\mathbb{E}[\begin{bmatrix} X & Y \end{bmatrix}] = \begin{bmatrix} \mathbb{E}[X] & \mathbb{E}[Y] \end{bmatrix}$$

- Liniowość wartości oczekiwanej. Niech $X$ oraz $Y$ będą zmiennymi losowymi. Wtedy $\forall a,\ b,\ c \in \mathbf{R}$,

  $$\mathbb{E}[aX + bY + c] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y] + c$$

## I.C. Warunkowa wartość oczekiwana

- Dla dyskretnych zmiennych losowych $X$ i $Y$ z łącznym rozkładem masy prawdopodobieństwa $f$, warunkowa wartość oczekiwana $Y$ pod warunkiem, że $X = x$ to:

$$\mathbb{E}[Y|X=x] = \sum_y y f_{Y|X}(y|x), \quad \forall x \in Supp[X]$$

  Dla ciągłych zmiennych losowych $X$ i $Y$ z łącznym rozkładem gęstości prawdopodobieństwa $f$, warunkowa wartość oczekiwana $Y$ pod warunkiem, że $X = x$ to:

$$\mathbb{E}[Y|X=x] = \int_y y f_{Y|X}(y|x)dy, \quad \forall x \in Supp[X]$$

  Innymi słowy, warunkowa wartość oczekiwana to wartość oczekiwana zmiennej losowej pod warunkiem, że inna zmienna losowa przyjmuje pewną wartość. Dzięki warunkowej własności oczekiwanej możemy opisać związek dwóch rozkładów.

- Wzór na całkowitą wartość oczekiwaną. Dla dwóch zmiennych losowych $X$ i $Y$:

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$$

- Wartość oczekiwana - przypadek wielowymiarowy. Dla dyskretnych zmiennych losowych $X_1 \ldots, X_k$ oraz $Y$ z łącznym rozkładem masy prawdopodobieństwa $f$, warunkowa wartość oczekiwana $Y$ pod warunkiem, że $\boldsymbol{X} = \boldsymbol{x}$ to:

$$\mathbb{E}[Y|\boldsymbol{X}=\boldsymbol{x}] = \sum_y y f_{Y|X}(y|\boldsymbol{x}), \quad \forall \boldsymbol{x} \in Supp[\boldsymbol{X}]$$

  Dla ciągłych zmiennych losowych $X_1 \ldots, X_k$ oraz $Y$ z łącznym rozkładem gęstości prawdopodobieństwa $f$, warunkowa wartość oczekiwana $Y$ pod warunkiem, że $\boldsymbol{X} = \boldsymbol{x}$ to:

$$\mathbb{E}[Y|X=x] = \int_y y f_{Y|X}(y|\boldsymbol{x})dy, \quad \forall \boldsymbol{x} \in Supp[\boldsymbol{X}]$$

## I.D. Wariancja. Macierz wariancji kowariancji.

- Wartość oczekiwana opisuje "środek"rozkładu, natomiast **wariancja** opisuje zmienność rozkładu lub jego rozpiętość. Formalnie, wariancja mierzy wartość oczekiwaną kwadratu różnicy między obserwowaną wartością zm. losowej $X$ oraz średnią.

  - Wariancja zmiennej losowej $X$:

$$\mathbb{V}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

- Alternatywny wzór na wariancję zm. losowej $X$:

$$\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

- Własności wariancji zmiennej losowej $X$:
  - $\forall c \in \mathbb{R},\ \mathbb{V}[X + c] = \mathbb{V}[X]$
  - $\forall a \in \mathbb{R},\ \mathbb{V}[aX] = a^2\,\mathbb{V}[X]$
- **Odchylenie standardowe** zmiennej losowej $X$

$$\sigma[X] = \sqrt{\mathbb{V}[X]}$$

- Własności odchylenia standardowego zmiennej losowej $X$:
  - $\forall c \in \mathbb{R},\ \sigma[X + c] = \sigma[X]$
  - $\forall a \in \mathbb{R},\ \sigma[aX] = |a|\,\sigma[X]$
- Naturalnym uogólnieniem wariancji w przypadku dwuwymiarowym jest **kowariancja** dwóch zmiennych losowych:

$$\mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Kowariancja mierzy w jaki sposób dwie zmienne są "związane"ze sobą. Jeżeli $X$ i $Y$ mają dodatnią kowariancję to oznacza to że wartości $X$ mają tendencję do zwiększania się gdy wartości $Y$ rosną oraz maleją, gdy wartości $Y$ maleją. Jeśli kowariancja jest ujemna, wtedy przeciwieństwo jest prawdą, gdy wartości $X$ maleją, $Y$ ma tendencję wzrostową.

- Alternatywny wzór na wariancję:

$$\mathrm{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

- Własności kowariancji. Dla zmiennych losowych $X$, $Y$, $Z$ i $W$:

$$\forall c, d \in \mathbb{R}, \mathrm{Cov}[c, X] = \mathrm{Cov}[X, c] = \mathrm{Cov}[c, d] = 0$$
$$\mathrm{Cov}[X, Y] = \mathrm{Cov}[Y, X]$$
$$\mathrm{Cov}[X, X] = \mathbb{V}[X]$$
$$\forall a, b, c, d \in \mathbb{R}, \mathrm{Cov}[aX + c, bY + d] = ab\,\mathrm{Cov}[X, Y]$$
$$\mathrm{Cov}[X + W, Y + Z] = \mathrm{Cov}[X, Y] + \mathrm{Cov}[X, Z] + \mathrm{Cov}[W, Y] + \mathrm{Cov}[W, Z]$$

- **Korelacja** dwóch zmiennych losowych $X$ i $Y$ gdy $\sigma[X] > 0$ oraz $\sigma[Y] > 0$:

$$\rho[X,Y] = \frac{\mathrm{Cov}[X,Y]}{\sigma[X]\,\sigma[Y]}$$

- **Macierz wariancji-kowariancji**. Dla wektora losowego $\boldsymbol{X}$ o długości $k$, macierz wariancji-kowariancji $\mathbb{V}[\boldsymbol{X}]$ to macierz której poszczególne elementy $(i,i)$ są równe $\mathrm{Cov}[X_i, X_i]$:

$$\mathbb{V}[\boldsymbol{X}] = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])'] \begin{bmatrix} \mathbb{V}[X_1] & \mathrm{Cov}[X_1, X_2] & \cdots & \mathrm{Cov}[X_1, X_k] \\ \mathrm{Cov}[X_2, X_1] & \mathbb{V}[X_2] & \cdots & \mathrm{Cov}[X_2, X_k] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_k, X_1] & \mathrm{Cov}[X_k, X_2] & \cdots & \mathbb{V}[X_K] \end{bmatrix}$$

Macierz wariancji-kowariancji jest wielowymiarowym uogólnieniem wariancji. Jej cechą charakterystyczną jest to, że na przekątnej znajduje się wariancja poszczególnych zmiennych losowych.

- Wariancja sumy $k$ zmiennych losowych:

$$\mathbb{V}[X_1 + X_2 + \cdots + X_k] = \mathbb{V}[\sum_{i=1}^{k} X_i] = \sum_{i=1}^{k}\sum_{ij=1}^{k} \mathrm{Cov}[X_i, X_j]$$

## I.E. *Własności rozkładu normalnego, rozkładu $\chi^2$, rozkładu t i F.*

- Chyba najważniejszym ze znanych rozkładów jest tak zwany rozkład normalny, określany niekiedy jako rozkład Gaussa. Rozkład $P$ nazywamy rozkładem normalnym, jeżeli istnieją takie liczby rzeczywiste $\mu$ oraz $\sigma > 0$, że funkcja $f\colon \mathbb{R} \to \mathbb{R}$, określona wzorem:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\frac{1}{2}(\frac{x-m}{\sigma})^2} \quad \text{dla } x \in \mathbb{R},$$

Notacja jest następująca: $\mathcal{N}(\mu,\sigma)$ oznacza rozkład normalny o parametrach $\mu$ oraz $\sigma$ - jego dystrybuantę oznaczamy przez $\Phi_{\mu,\sigma}$. Wykres gęstości rozkładu normalnego nosi nazwę krzywej Gaussa.

- Własności. Niech $X_1$ oraz $X_2$ będą niezależnymi zmiennymi losowymi o rozkładach normalnych, odpowiednio $\mathcal{N}(m_1, \sigma_1)$ oraz $\mathcal{N}(m_2, \sigma_2)$. Wtedy:
  - $X_1 + X_2 \sim \mathcal{N}(m_1 + m_2, \sqrt{\sigma_1^2 + \sigma_2^2})$
  - $aX_1 + b \sim N(am_1 + b, |a|\sigma_1)$ dla wszystkich $a, b \in \mathbb{R}$.

- **Wielowymiarowy rozkład normalny**. Jednowymiarowy rozkład normalny ma dwa parametry - $\mu$ oraz $\sigma$. W wersji wielowymiarowej $\mu$ jest zastąpione przez wektor $\boldsymbol{\mu}$, oraz $\sigma$ jest zastąpiona przez macierz wariancji-kowariancji $\boldsymbol{\Sigma}$.

- **Rozkład** $\chi^2$ – rozkład zmiennej losowej, która jest sumą $k$ kwadratów niezależnych zmiennych losowych o standardowym rozkładzie normalnym. Liczbę $k$ nazywamy liczbą stopni swobody rozkładu zmiennej losowej. Jeżeli zmienne losowe $X_i \sim \mathcal{N}(0,\, 1)$ to: $Y \sim \chi^2(k)$.

# II. Statystyka.

Wnioskowanie statystyczne to proces w którym używamy danych do wnioskowania o rozkładzie który wygenerował te dane. We wnioskowaniu statystycznym zwykle zadajemy następujące pytanie: mając losową próbkę danych $(X_1, ...X_n)$ wylosowaną z rozkładu $F$ jaki jest rozkład $F$? Często chcemy dowiedzieć się czegoś o cechach rozkładu $F$, na przykład o jego wartości oczekiwanej.

## II.A. *Pojęcie estymatora*

- W procesie estymacji (punktowej) używamy danych do wyznaczenia pojedynczej wartości, znanej jako **estymator (punktowy)**, która ma być najlepszym przybliżeniem (*best guess*) nieznanego parametru w populacji (np. średniej w populacji).

- Zwykle oznaczamy estymator nieznanego parametru $\theta$ jako $\hat{\theta}$.

- Estymator punktowy $\hat{\theta}$ nieznanego parametru $\theta$ jest funkcją danych, a więc jest jest on zmienną losową.

$$\hat{\theta} \; = \; g(X_1, ...X_n)$$

  Wartość którą przyjmuje estymator nazywamy **oszacowaniem**.

## II.B. *Nieobciążoność estymatora, wariancja estymatora i efektywność*

- Estymator $\hat{\theta}$ jest zmienną losową (statystyką), czyli istnieje wartość oczekiwana $\mathbb{E}[\hat{\theta}]$, wariancja $\mathbb{V}[\hat{\theta}]$ (*sampling variance*). Istnieje także jego odchylenie standardowe, $\sigma[\hat{\theta}]$ które często jest nazywane **błędem standardowym**.

- Estymator $\hat{\theta}$ parametru $\boldsymbol{\theta}$ nazywamy **nieobciążonym** gdy jego wartość oczekiwana jest równa prawdziwemu parametrowi:

$$\mathbb{E}[\hat{\theta}] \; = \; \theta$$

- **Obciążenie** estymatora

$$bias(\hat{\theta}) = \mathbb{E}[\hat{\theta}] \; - \theta$$

- **Błędem średniokwadratowym** (*Mean Square Error, MSE*) estymatora $\hat{\theta}$ parametru $\theta$ nazywamy

$$\mathbb{E}[(\hat{\theta} - \theta)^2]$$

Alternatywny wzór:

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{V}[\hat{\theta}] - (\mathbb{E}[\hat{\theta}] - \theta)^2$$

- Estymator $\hat{\theta}$ parametru $\theta$ nazywamy **efektywnym** gdy jego wariancja, $\mathbb{V}[\boldsymbol{\theta}]$ jest najmniejsza w danej klasie estymatorów. Na przykład, niech $\hat{\theta}_A$ i $\hat{\theta}_B$ będą estymatorami parametru $\theta$. Bardziej efektywny jest ten estymator który ma niższą wariancję.

## *II.C. Przedziały ufności*

## *II.D. Testowanie hipotez statystycznych, wartości krytyczne i wartości p*