# Statistika z elementi informatike
# Osnove verjetnostnega računa in statistike

## 5.2.1999

### 1. Naloga: disktretna slučajna spremenljivka

Janko in Tomaž streljata na isto tarčo. Verjetnost, da Janko zadene v posameznem poskusu je 0.7. Verjetnost, da Tomaž zadene v posameznem poskusu je 0.8. Vzemimo, da je imel vsak na voljo dva poskusa. Določite zalogo vrednosti, verjetnostno in porazdelitveno funkcijo slučajne spremenljivke $X$, ki predstavlja skupno število zadetkov. Narišite grafa verjetnostne in porazdelitvene funkcije. (Dodatno vprašanje: Vzemimo, da zmaga tisti, ki zadene več kot drugi. Določite verjetnost, da je zmagal Janko.)

**Rešitev:** Označimo verjetnosti, da v posameznem poskusu zadene Janko oziroma Tomaž:
$P[\text{Janko zadene}] = p_J = 0.7$, $P[\text{Janko zgreši}] = q_J = 1 - p_J = 0.3$,
$P[\text{Tomaž zadene}] = p_T = 0.8$, $P[\text{Tomaž zgreši}] = q_T = 1 - p_T = 0.2$.

Zaloga vrednosti slučajne spremenljivke $X$, ki predstavlja skupno število zadetkov, če oba strelca poskusita dvakrat, je 0, 1, 2, 3 in 4.

Predpostavimo, da so posamezni poskusi medsebojno neodvisni. Tako lahko izračunamo verjetnosti, da sta skupaj zadela 0, 1, 2, 3 ali 4 krat:

$$\begin{aligned}
P[X=0] &= q_J^2 \cdot q_T^2 = 0.3^2 \cdot 0.2^2 = 0.0036,\\
P[X=1] &= 2 \cdot q_J \cdot p_J \cdot q_T^2 + 2 \cdot q_J^2 \cdot q_T \cdot p_T = \\
&= 2 \cdot 0.3 \cdot 0.7 \cdot 0.2^2 + 2 \cdot 0.3^2 \cdot 0.2 \cdot 0.8 = 0.0456,\\
P[X=2] &= 4 \cdot q_J \cdot p_J \cdot q_T \cdot p_T + p_J^2 \cdot q_T^2 + q_J^2 \cdot p_T^2 = \\
&= 4 \cdot 0.3 \cdot 0.7 \cdot 0.2 \cdot 0.8 + 0.7^2 \cdot 0.2^2 + 0.3^2 \cdot 0.8^2 = 0.2116,\\
P[X=3] &= 2 \cdot q_J \cdot p_J \cdot p_T^2 + 2 \cdot p_J^2 \cdot q_T \cdot p_T = \\
&= 2 \cdot 0.3 \cdot 0.7 \cdot 0.8^2 + 2 \cdot 0.7^2 \cdot 0.2 \cdot 0.8 = 0.4256,\\
P[X=4] &= p_J^2 \cdot p_T^2 = 0.7^2 \cdot 0.8^2 = 0.3136.
\end{aligned}$$

Preverimo lahko, ali je vsota verjetnosti, da sta skupaj zadela 0, 1, 2, 3 ali 4 krat, enaka ena

$$\sum_{i=0}^{4} P[X=i] = 0.0036 + 0.0456 + 0.2116 + 0.4256 + 0.3136 = 1.0000.$$

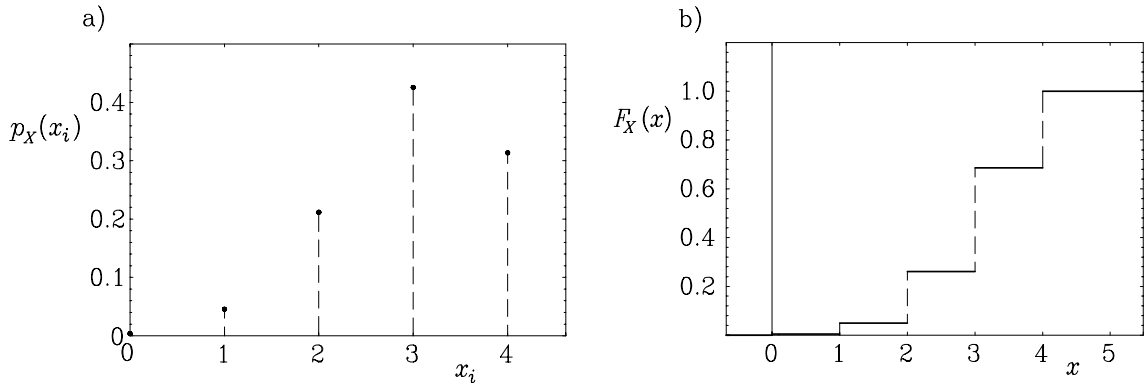
Verjetnostna funkcija je torej

$$p_X(x_i) = P[X=x_i] = \begin{cases} 0.0036 & \dots\ X=0,\\ 0.0456 & \dots\ X=1,\\ 0.2116 & \dots\ X=2,\\ 0.4256 & \dots\ X=3,\\ 0.3136 & \dots\ X=4. \end{cases}$$

Porazdelitvena funkcija je

$$F_X(x) = P[X < x] = \begin{cases} 0 & \dots \quad x \leq 0, \\ 0.0036 & \dots \quad 0 < x \leq 1, \\ 0.0492 & \dots \quad 1 < x \leq 2, \\ 0.2608 & \dots \quad 2 < x \leq 3, \\ 0.6864 & \dots \quad 3 < x \leq 4, \\ 1 & \dots \quad 4 < x. \end{cases}$$

Na naslednji sliki prikazujemo verjetnostno in porazdelitveno funkcijo.




SLIKA 1: Verjetnostna funkcija $p_X(x_i)$ in porazdelitvena funkcija $F_X(x)$

Janko zmaga v primeru, da zadene več kot Tomaž. To se zgodi, če Janko zadene enkrat ali dvakrat, Tomaž obakrat zgreši ali če Janko zadene dvakrat, Tomaž pa enkrat. Verjetnost, da zmaga Janko, je enaka

$$\begin{aligned} P[\text{Janko}] &= 2 \cdot q_T^2 \cdot p_J \cdot q_J + q_T^2 \cdot p_J^2 + 2 \cdot p_T \cdot q_T \cdot p_J^2 = \\ &= 2 \cdot 0.2^2 \cdot 0.7 \cdot 0.3 + 0.2^2 \cdot 0.7^2 + 2 \cdot 0.8 \cdot 0.2 \cdot 0.7^2 = 0.1932. \end{aligned}$$

Podobno izračunamo tudi verjetnost, da zmaga Tomaž, in verjetnost, da zadeneta enako tarč:

$$\begin{aligned} P[\text{Tomaž}] &= 2 \cdot q_J^2 \cdot p_T \cdot q_T + q_J^2 \cdot p_T^2 + 2 \cdot p_J \cdot q_J \cdot p_T^2 = \\ &= 2 \cdot 0.3^2 \cdot 0.8 \cdot 0.2 + 0.3^2 \cdot 0.8^2 + 2 \cdot 0.7 \cdot 0.3 \cdot 0.8^2 = 0.3552, \\ P[\text{enaka}] &= q_J^2 \cdot q_T^2 + 4 \cdot q_J \cdot p_J \cdot q_T \cdot p_T + p_J^2 \cdot p_T^2 = \\ &= 0.3^2 \cdot 0.2^2 + 4 \cdot 0.3 \cdot 0.7 \cdot 0.2 \cdot 0.8 + 0.7^2 \cdot 0.8^2 = 0.4516. \end{aligned}$$

Vsota verjetnosti, da zmaga Janko, Tomaž ali sta si enaka, mora biti enaka ena:

$$P[\text{Janko}] + P[\text{Tomaž}] + P[\text{enaka}] = 0.1932 + 0.3552 + 0.4516 = 1.0000.$$

## 2. Naloga: izpeljana porazdelitev

Izpeljite gostoto verjetnosti logaritemske porazdelitve $f_Y(y)$, ki je definirana z naslednjo definicijo:

*Če je $X$ enakomerna porazdelitev od 0 do 1 in velja* $\ln Y = X$, *potem se $Y$ porazdeljuje po logaritemski porazdelitvi.* Izpeljava je zelo podobna izpeljavi lognormalne porazdelitve. Ne pozabite določiti zaloge vrednosti slučajne spremenljivke $Y$.

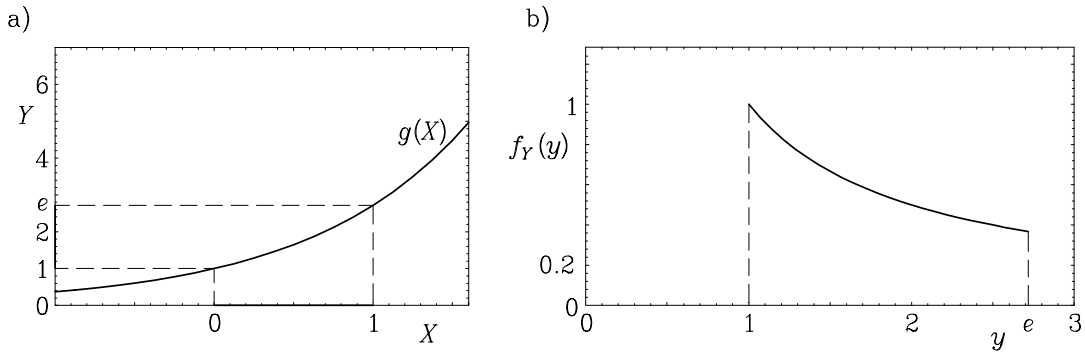
**Rešitev:** Zapišimo najprej gostoto verjetnosti slučajne spremneljivke $X$

$$f_X(x) = \begin{cases} 1 & \ldots \ 0 \leq x \leq 1 \\ 0 & \ldots \ \text{drugje} \end{cases}$$

in funkcijo $Y = g(X)$

$$X = g^{-1}(Y) = \ln Y \quad \longrightarrow \quad Y = g(X) = e^X.$$

Inverzno funkcijo $g^{-1}(Y)$ lahko zapišemo zato, ker je funkcija $g(X)$ monotona funkcija, kar lahko vidimo tudi iz slike 2a.




SLIKA 2: Funkcijska zveza med $X$ in $Y$ ter gostota verjetnosti $f_Y(y)$

Iz slike 2a določimo tudi zalogo vrednosti slučajne spremenljivke $Y$. Območje, kjer je gostota verjetnosti slučajne spremenljivke $Y$ različna od nič, je med $e^0 = 1$ in $e^1 = e = 2.71828$. Zalogo vrednosti po enačbah določimo tako, da neenačbo, ki opisuje zalogo vrednosti slučajne spremenljivke $X$, transformiramo s funkcijo $g(X)$:

$$0 \leq x \leq 1 \quad \longrightarrow \quad g(0) \leq g(x) \leq g(1) \quad \longrightarrow \quad e^0 \leq e^x \leq e^1 \quad \longrightarrow \quad 1 \leq y \leq e.$$

Ker je funkcija $g(X)$ monotona, lahko gostoto verjetnosti slučajne spremenljivke $Y$ določimo po naslednji enačbi:

$$f_Y(y) = f_X\left(g^{-1}(y)\right) \frac{dg^{-1}(y)}{dy} = 1 \cdot \frac{1}{y} = \frac{1}{y} \quad \ldots \quad 1 \leq y \leq e$$

Gostoto verjetnosti slučajne spremenljivke $Y$ prikazujemo na sliki 2b. Preverimo lahko, ali funkcija $f_Y(y)$ zadošča pogojem za gostoto verjetnosti. Vidimo, da je funkcija povsod nenegativna. Drugi pogoj pa zapišemo z enačbo

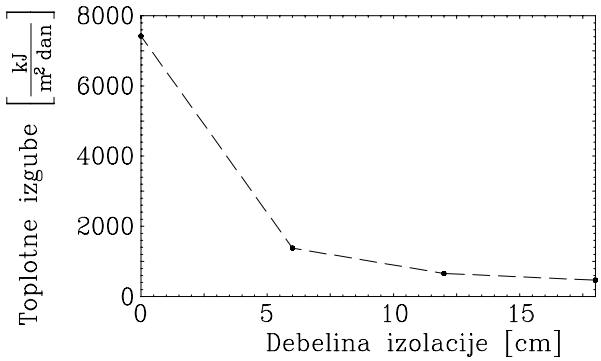
$$\int_{-\infty}^{\infty} f_Y(y)dy = \int_1^e \frac{1}{y}\,dy = \ln y\Big|_1^e = 1 - 0 = 1.$$

Vidimo, da je izpolnjen tudi pogoj, da je integral gostote verjetnosti po celotnem območju realnih števil enak ena.

### 3. Naloga: Nelinearna regresija

V naslednji preglednici so podatki o odvisnosti med dnevno toplotno izgubo $Y$ betonskega zida in debelino izolacije $X$.

| $X_i$ debelina izolacije [cm] | $Y_i$ toplotne izgube [kJ/m$^2$dan] |
|---|---|
| 0 | 7421.86 |
| 6 | 1379.76 |
| 12 | 655.82 |
| 18 | 468.78 |




SLIKA 3: Odvisnost dnevne toplotne izgube $Y$ od debeline izolacije $X$

Iz slike 3 je razvidno, da zveza ni linearna. Predpostavimo, da je ustrezna naslednja zveza med $X$ in $Y$

$$Y = a\, e^{bX}.$$

Določite oceni parametrov $a$ in $b$. Navodilo: Enačbo logaritmirajte in jo preoblikujte tako, da bo zveza med $\ln Y$ in $X$ linearna. Nato določite oceni parametrov $B_0 = \ln a$ in $B_1 = b$. Določite približno vrednost dnevnih toplotnih izgub za 10 cm sloj izolacije.

**Rešitev:** Eksponentno zvezo logaritmiramo in dobimo naslednjo enačbo

$$Z = \ln Y = \ln a + b\, X \quad \longrightarrow \quad Z = B_0 + B_1\, X,$$

kjer sta $B_0 = \ln a$ in $B_1 = b$. S tem smo problem nelinearne regresije spremenili v problem linearne regresije med spremenljivkama $X$ in $Z$. Prepišimo torej preglednico podatkov in dopišimo še stolpec $Z_i = \ln Y_i$

| $X_i$ | $Y_i$ | $Z_i$ |
|---|---|---|
| 0 | 7421.86 | 8.912 |
| 6 | 1379.76 | 7.230 |
| 12 | 655.82 | 6.486 |
| 18 | 468.78 | 6.150 |

Oceni parametrov $B_0$ in $B_1$ izračunamo po enačbah

$$\hat{B}_1 = \frac{S_{XZ}}{S_{X^2}}, \qquad \hat{B}_0 = \bar{Z} - \bar{X}\, \hat{B}_1.$$

Momenti $\bar{X}$, $\bar{Z}$, $S_{X^2}$ in $S_{XZ}$ so

$$\begin{aligned}
\bar{X} &= \frac{\sum_{i=1}^{4} X_i}{4} = \frac{0+6+12+18}{4} = 9,\\
\bar{Z} &= \frac{\sum_{i=1}^{4} Z_i}{4} = \frac{8.912+7.230+6.486+6.150}{4} = 7.194,\\
S_{X^2} &= \frac{\sum_{i=1}^{4}(X_i-\bar{X})^2}{4} = \frac{(0-9)^2+\cdots+(18-9)^2}{4} = 45,\\
S_{XZ} &= \frac{\sum_{i=1}^{4}(X_i-\bar{X})(Z_i-\bar{Z})}{4} =\\
&= \frac{(0-9)(8.912-7.194)+\cdots+(18-9)(6.150-7.194)}{4} = -6.772
\end{aligned}$$

oceni parametrov $B_0$ in $B_1$ pa sta

$$\hat{B}_1 = \frac{-6.772}{45} = -0.1505, \qquad \hat{B}_0 = 7.194 - 9\cdot 0.1505 = 8.549.$$

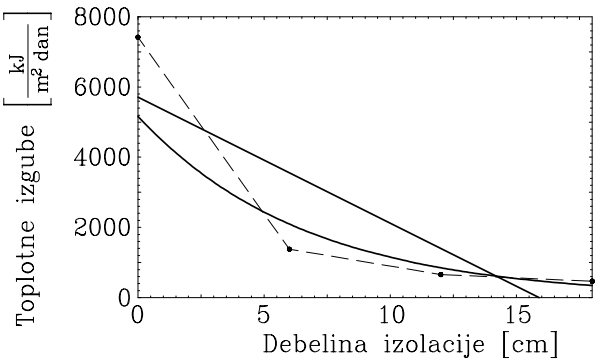
Sedaj lahko izračunamo tudi oceni parametrov $a$ in $b$

$$\hat{a} = e^{\hat{B}_0} = 5161.36, \qquad \hat{b} = \hat{B}_1 = -0.1505$$

in zapišemo izraz, s katerim lahko izračunamo približno vrednost toplotnih izgub za poljubno debelino toplotne izolacije

$$Y = 5161.36\, e^{-0.1505\, X}.$$

Na primer: dnevne toplotne izgube pri 10 cm topotne izolacije so približno 1145 kJ/m$^2$dan. Na sliki 4 primerjamo nelinearno aproksimacijo z pravimi vrednostmi toplotnih izgub. Vidimo, da so odstopanja precej velika in torej lahko zaključimo, da eksponentna funkcija, ki smo jo tu uporabili, ni prava. Na sliki 4 prikazujemo tudi linearno aproksimacijo in lahko opazimo, da je linearna aproksimacija še slabša.




SLIKA 4: Linearna in nelinearna aproksimacija zveze med $Y$ in $X$

## 4. Naloga: Testiranje hipotez

Znani so podatki o uspešnosti pridelave naftnih derivatov iz surove nafte. Uspešnost pridelave merimo v relativni količini naftnih derivatov glede na surovo nafto. Na primer: 50% uspešnost bi pomenila, da iz ene tone surove nafte pridelamo pol tone naftnih derivatov. Obravnavana sta dva načina pridelave. Predpostavite, da standardni deviaciji poznamo in da sta $\sigma_{X1} = 0.9$ za prvi način in $\sigma_{X2} = 0.6$ za drugi način pridelave. Ugotovite, ali lahko zavrnemo ničelno hipotezo, ki pravi, da sta uspešnosti obeh načinov enaki. Tveganje naj bo 1%. Podajte zaključek. Podatki iz predhodnih poskusov so prikazani v naslednji preglednici.

| Prvi način [%] | 24.2 | 26.6 | 25.7 | 24.8 | 25.9 | 26.5 |
|---|---|---|---|---|---|---|
| Drugi način [%] | 21.0 | 22.1 | 21.8 | 20.9 | 22.4 | 22.0 |

**Rešitev:** Postavimo ničelno in alternativno hipotezo:
$H_0$: Uspešnosti obeh načinov sta enaki: $m_{X1} = m_{X2}$,
$H_1$: Uspešnosti obeh načinov se razlikujeta: $m_{X1} \neq m_{X2}$.

Ker sta standardni deviaciji znani, je statistika

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_{X1}^2}{6} + \frac{\sigma_{X2}^2}{6}}}$$

porazdeljena standardno normalno. Za stopnjo tveganja $\alpha = 1\%$ določimo območje zavrnitve ničelne hipoteze:

$$k_{\alpha/2} = 2.5758,$$

kar lahko odčitamo iz preglednice za normalno porazdelitev ali pa z računalniškim programom (na primer EXCEL: ukaz `NORMINV(0.995;0;1)`. Če je statistika $Z$ manjša od $-k_{\alpha/2} = -2.5758$ ali večja od $k_{\alpha/2} = 2.5758$, moramo ničelno hipotezo zavrniti.

Povprečni vrednosti za oba načina sta

$$\bar{X}_1 = \frac{\sum_{i=1}^{6} X_{1i}}{6} = \frac{24.2 + 26.6 + 25.7 + 24.8 + 25.9 + 26.5}{6} = 25.62,$$

$$\bar{X}_2 = \frac{\sum_{i=1}^{6} X_{2i}}{6} = \frac{21.0 + 22.1 + 21.8 + 20.9 + 22.4 + 22.0}{6} = 21.70.$$

Statistika $Z$ pa je enaka

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_{X1}^2}{6} + \frac{\sigma_{X2}^2}{6}}} = \frac{25.62 - 21.70}{\sqrt{\frac{0.9^2}{6} + \frac{0.6^2}{6}}} = 8.8695 > k_{\alpha/2} = 2.5758,$$

zato moramo ničelno hipotezo zavrniti in lahko trdimo: *Za stopnjo tveganja 1% lahko trdimo, da se uspešnosti obeh načinov statistično značilno razlikujeta.*