Statistika z elementi informatike
Osnove verjetnostnega računa in statistike

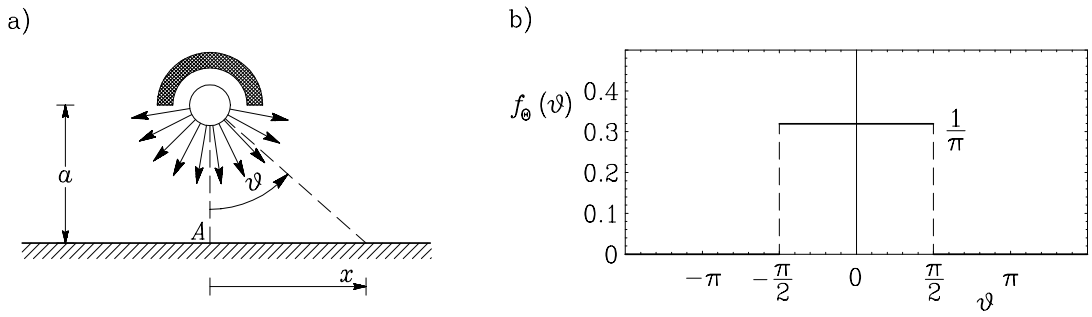
9.4.1999

1. Naloga: Izpeljana porazdelitev

Obravnavamo ravninski primer sevanja radioaktivnega telesa, ki seva na vse strani enako. To pomeni, da je slučajna spremenljivka Θ, ki predstavlja kot posameznega žarka, enakomerno porazdeljena. Zaradi svinčene zaščite seva le navzdol (glej sliko). Tako lahko gostoto verjetnosti zapišemo takole:

$$f_\Theta(\vartheta) = 1/\pi \qquad \ldots \qquad -\pi/2 \leq \vartheta \leq \pi/2.$$

Določi gostoto verjetnosti slučajne spremenljivke X, da posamezen žarek pade na raven ekran neskončnih razsežnosti v oddaljenosti x od točke A. Skiciraj gostoto verjetnosti te porazdelitve, ki je poznana kot *Cauchyjeva porazdelitev*. Razdalja radioaktivnega telesa do ekrana a = dan vašega rojstva [cm].



SLIKA 1: Radioaktivno sevanje v ravnini

Rešitev: Iz slike 1a lahko določimo zvezo med kotom ϑ in vodoravno razdaljo x:

$$\frac{x}{a} = \mathrm{tg}\vartheta \qquad \longrightarrow \qquad x = g(\vartheta) = a\,\mathrm{tg}\vartheta.$$

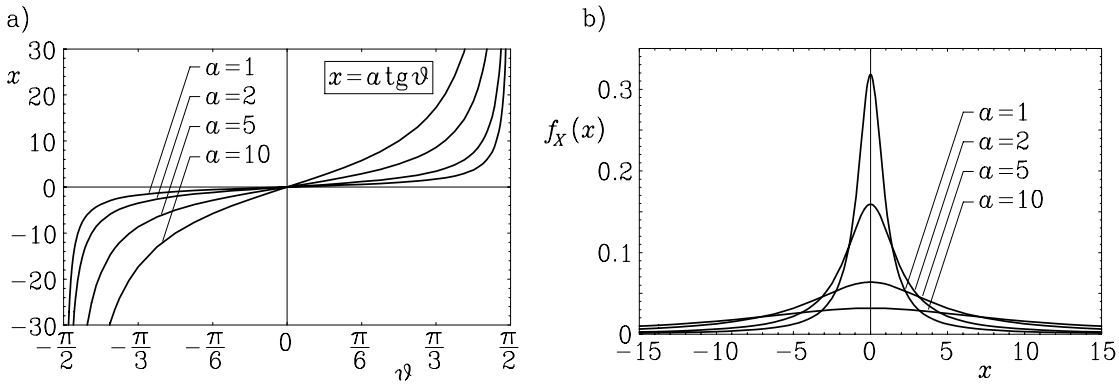
Iz slik 1a in 2a lahko sklepamo, da območju $-\pi/2 \leq \vartheta \leq \pi/2$ ustreza neskončno območje $-\infty \leq x \leq \infty$. Ker je na območju od $-\pi/2 \leq \vartheta \leq \pi/2$ funkcija monotona (glej sliko 2a), lahko zapišemo obratno zvezo

$$\vartheta = \mathrm{arctg}\frac{x}{a} = g^{-1}(x)$$

in izpeljemo gostoto verjetnosti po enačbi:

$$f_X(x) = f_\Theta\left(g^{-1}(x)\right) \cdot \frac{dg^{-1}(x)}{dx} = \frac{1}{\pi} \cdot \frac{a}{a^2 + x^2} \qquad \ldots \qquad -\infty \leq x \leq \infty.$$

Na sliki 2b prikazujemo graf gostote verjetnosti slučajne spremenljivke X za nekaj vrednosti parametra a.



SLIKA 2: Zveza med ϑ in x ter gostota verjetnosti $f_X(x)$

2. Naloga: diskretna slučajna spremenljivka

Nek stroj lahko dnevno proizvede 1, 2, 3 ali 4 izdelke. Verjetnostna funkcija števila proizvedenih izdelkov N je:

$$p_N(n) = \begin{cases} 0.1 & \dots \quad n = 1 \\ 0.2 & \dots \quad n = 2 \\ 0.5 & \dots \quad n = 3 \\ 0.2 & \dots \quad n = 4 \end{cases}$$

Določite verjetnostno funkcijo skupnega števila M proizvedenih izdelkov, ki jih proizvedeta dva neodvisno delujoča stroja z enakimi lastnostmi. Narišite verjetnostno funkcijo in določite srednjo vrednost $\mathrm{E}[M]$ in varianco $\mathrm{VAR}[M]$ slučajne spremenljivke M.

Rešitev: Ker dva stroja, ki delujeta neodvisno, lahko izdelata po 1 do 4 izdelkov, je skupno število izdelkov lahko 2, 3, 4, 5, 6, 7 ali 8 izdelkov. Verjetnostno funkcijo $p_M(m)$ določimo z naslednjo preglednico:

1. stroj N_1	2. stroj N_2	Skupaj M	Verjetnost
1	1	2	$0.1 \cdot 0.1 = 0.01$
1	2	3	$0.1 \cdot 0.2 = 0.02$
1	3	4	$0.1 \cdot 0.5 = 0.05$
1	4	5	$0.1 \cdot 0.2 = 0.02$
2	1	3	$0.2 \cdot 0.1 = 0.02$
2	2	4	$0.2 \cdot 0.2 = 0.04$
2	3	5	$0.2 \cdot 0.5 = 0.10$
2	4	6	$0.2 \cdot 0.2 = 0.04$
3	1	4	$0.5 \cdot 0.1 = 0.05$
3	2	5	$0.5 \cdot 0.2 = 0.10$
3	3	6	$0.5 \cdot 0.5 = 0.25$
3	4	7	$0.5 \cdot 0.2 = 0.10$
4	1	5	$0.2 \cdot 0.1 = 0.02$
4	2	6	$0.2 \cdot 0.2 = 0.04$
4	3	7	$0.2 \cdot 0.5 = 0.10$
4	4	8	$0.2 \cdot 0.2 = 0.04$

Iz te preglednice lahko povzamemo verjetnostno funkcijo $p_M(m)$

$$p_M(m) = \begin{cases} 0.01 & \dots & m = 2 \\ 0.04 & \dots & m = 3 \\ 0.14 & \dots & m = 4 \\ 0.24 & \dots & m = 5 \\ 0.33 & \dots & m = 6 \\ 0.20 & \dots & m = 7 \\ 0.04 & \dots & m = 8 \end{cases},$$

ki jo narišemo na sliki 3. Srednjo vrednost in varianco slučajne spremenljivke M lahko izračunamo po enačbah

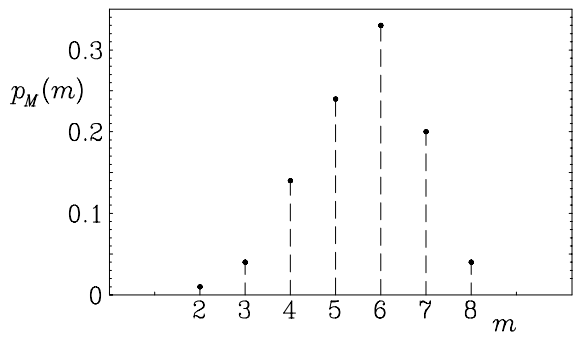
$$\begin{aligned} \mathrm{E}[M] &= \sum_{m=2}^{8} m \cdot p_M(m) = 5.60, \\ \mathrm{E}[M^2] &= \sum_{m=2}^{8} m^2 \cdot p_M(m) = 32.88, \\ \mathrm{VAR}[M] &= \mathrm{E}[M^2] - \mathrm{E}[M]^2 = 1.52. \end{aligned}$$

Enako izračunamo tudi, če po podobnih enačbah izračunamo najprej srednjo vrednost in varianco slučajne spremenljivke N:

$$\begin{aligned} \mathrm{E}[N] &= \sum_{n=1}^{4} n \cdot p_N(n) = 2.8, \\ \mathrm{E}[N^2] &== \sum_{n=1}^{4} n^2 \cdot p_N(n) = 8.6 \\ \mathrm{VAR}[N] &= \mathrm{E}[N^2] - \mathrm{E}[N]^2 = 0.76. \end{aligned} \tag{1}$$

ter upoštevamo zveze za momente vsote več neodvisnih slučajnih spremenljivk:

$$\begin{aligned} \mathrm{E}[M] &= \mathrm{E}[N + N] = \mathrm{E}[N] + \mathrm{E}[N] = 2.8 + 2.8 = 5.60, \\ \mathrm{VAR}[M] &= \mathrm{VAR}[N + N] = \mathrm{VAR}[N] + \mathrm{VAR}[N] = 0.76 + 0.76 = 1.52. \end{aligned}$$



SLIKA 3: Verjetnostna funkcija $p_M(m)$

3. Naloga: Analiza variance

Na odlagališču odpadkov so merili koncentracije kroma v ppm (part per million). Odlagališče so razdelili na tri območja in v vsakem območju izmerili koncentracijo kroma na šestih slučajno izbranih merilnih mestih. Ugotovite ali je območje odlagališča statistično značilen faktor. Stopnja tveganja naj bo 1%. Podatke podajamo v naslednji preglednici:

območje	**Količina Cr [ppm]**					
prvo območje	35	36	30	29	32	32
drugo območje	14	21	19	24	26	37
tretje območje	13	6	21	23	21	33

Rešitev: Najprej izračunajmo povprečno vrednost vseh meritev $\bar{x}$ ter povprečne vrednosti meritev po območjih $\bar{x}_1$, $\bar{x}_2$ in $\bar{x}_3$:

$$\bar{x} = \frac{1}{3 \cdot 6} \sum_{i=1}^{3} \sum_{j=1}^{6} x_{ij} = \frac{35 + 36 + \cdots + 21 + 33}{18} = 25.11,$$

$$\bar{x}_1 = \frac{1}{6} \sum_{j=1}^{6} x_{1j} = \frac{35 + 36 + 30 + 29 + 32 + 32}{6} = 32.33,$$

$$\bar{x}_2 = \frac{1}{6} \sum_{j=1}^{6} x_{2j} = \frac{14 + 21 + 19 + 24 + 26 + 37}{6} = 23.50,$$

$$\bar{x}_3 = \frac{1}{6} \sum_{j=1}^{6} x_{3j} = \frac{13 + 6 + 21 + 23 + 21 + 33}{6} = 19.50.$$

Izračunajmo še vsoto kvadratov odstopanj od povprečnih vrednosti:

$$\begin{aligned} SS_A &= 6 \cdot \sum_{i=1}^{3} (\bar{x}_i - \bar{x})^2 = \\ &= (32.33 - 25.11)^2 + (23.5 - 25.11)^2 + (19.5 - 25.11)^2 = 517.44, \end{aligned}$$

$$\begin{aligned} SS_E &= \sum_{i=1}^{3} \sum_{j=1}^{6} (x_{ij} - \bar{x}_i)^2 = \\ &= (35 - 32.33)^2 + (36 - 32.33)^2 + \cdots + \\ &+ (14 - 23.50)^2 + (21 - 23.50)^2 + \cdots + \\ &+ (13 - 19.50)^2 + (6 - 19.5)^2 + \cdots = 766.33, \end{aligned}$$

$$\begin{aligned} SS_T &= \sum_{i=1}^{3} \sum_{j=1}^{6} (x_{ij} - \bar{x})^2 = \\ &= (35 - 25.11)^2 + (36 - 25.11)^2 + \cdots + (33 - 25.11)^2 = 1283.78. \end{aligned}$$

Ta izračun preverimo tudi z izrazom

$$SS_T = SS_A + SS_E \qquad \longrightarrow \qquad 1283.78 = 517.44 + 766.33.$$

Število prostostnih stopenj za faktor je $3 - 1 = 2$, za napako oziroma nepojasnjena odstopanja $6 \cdot 3 - 3 = 15$ in za skupna odstopanja $6 \cdot 3 - 1 = 17$.

Sedaj lahko pripravimo preglednico analize variance:

Vzrok	SS	n_{ps}	MS
Faktor	517.44	2	258.72
Napaka	766.33	15	51.09
Skupaj	1283.78	17	

Postavimo ničelno in alternativno hipotezo:
H_0: Območje ne vpliva na koncentracijo kroma.
H_1: Območje vpliva na koncentracijo kroma.

Statistika, na osnovi katere bomo ničelno hipotezo morda lahko zavrnili, je:

$$\frac{MS_A}{MS_E} = 5.064.$$

To vrednost primerjamo z vrednostjo porazdelitvene funkcije porazdelitve $F_{2,15}$ in sicer za stopnjo zaupanja 0.99. Iz preglednice za porazdelitev F lahko odčitamo približno vrednost $F_{2,15,0.99}$. Natančneje to vrednost določimo z računalniškim programom (na primer EXCEL: ukaz `FINV(0.01;2;15)`):

$$F_{2,15,0.99} = 6.359.$$

Ker je $MS_A/MS_E < F_{2,15,0.99}$, ničelne hipoteze ne moremo zavrniti in moramo zaključiti: *Za stopnjo statističnega tveganja 1% velja, da vpliv območij ni statistično značilen.*

Če bi izbrali večje tveganje (na primer 5%), bi bil zaključek v tem primeru drugačen. Vrednost porazdelitvene funkcije je namreč $F_{2,15,0.95} = 3.682 < MS_A/MS_E$. Tedaj bi zaključili: *Za stopnjo statističnega tveganja 5% velja, da je vpliv območij statistično značilen.*

4. Naloga: Določanje parametra porazdelitve gama

Gostota verjetnosti porazdelitve gama (Γ), ki opisuje čas, ko se zgodi k-ti uspeh, je:

$$f_X(x) = \frac{\lambda(\lambda x)^{k-1}e^{-\lambda x}}{(k-1)!} \qquad x \geq 0.$$

Srednja oziroma pričakovana vrednost te porazdelitve je:

$$\mathrm{E}[X] = m_X = \frac{k}{\lambda}.$$

Določite oceno parametra porazdelitve λ po metodi momentov in po metodi največje verjetnosti. Vzorec podatkov sestavlja pet vrednosti x_i. Vzemimo, da je drugi parameter porazdelitve $k = 10$.

0.7816	0.9398	0.6834	0.8141	1.5637

Rešitev: Iz podanega vzorca lahko izračunamo njegovo povprečno vrednost $\bar{X}$, ki predstavlja oceno srednje vrednosti $\hat{m}_X$ slučajne spremenljivke X:

$$\hat{m}_X = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{0.7816 + 0.9398 + 0.6834 + 0.8141 + 1.5637}{5} = 0.9565.$$

Ob upoštevanju podane enačbe o zvezi med m_X in parametrom λ lahko podamo oceno parametra $\hat{\lambda}$ po *metodi momentov*

$$\hat{\lambda} = \frac{k}{\hat{m}_X} = \frac{10}{0.9565} = 10.45.$$

Oceno po metodi največje verjetnosti moramo še izpeljati. Definirajmo funkcijo L

$$L(\lambda) = \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \frac{\lambda(\lambda x_i)^{k-1} e^{-\lambda x_i}}{(k-1)!}.$$

Ekstrem te funkcije določimo s pogojem, da je odvod $dL/d\lambda$ enak nič. Zaradi preprostejše izpeljave funkcijo L logaritmiramo, saj zaradi monotonosti logaritemske funkcije velja, da imata ekstremne vrednosti za iste vrednosti argumenta. Po logaritmiranju moramo upoštevati še lastnosti logaritmov ($\ln(ab) = \ln a + \ln b$ in $\ln a^c = c \ln a$)

$$\begin{aligned}
\ln L(\lambda) &= \ln \left[\prod_{i=1}^{n} \frac{\lambda(\lambda x_i)^{k-1} e^{-\lambda x_i}}{(k-1)!} \right] = \\
&= \sum_{i=1}^{n} \ln\lambda + \sum_{i=1}^{n} \ln(\lambda x_i)^{k-1} + \sum_{i=1}^{n} (-\lambda x_i) - \sum_{i=1}^{n} \ln(k-1)! = \\
&= n \ln\lambda + (k-1) \sum_{i=1}^{n} \ln(\lambda x_i) - \sum_{i=1}^{n} \lambda x_i - n \ln(k-1)!.
\end{aligned}$$

Če zadnjo enačbo odvajamo po λ, dobimo

$$\begin{aligned}
\frac{d \ln L(\lambda)}{d\lambda} &= n \frac{1}{\lambda} + (k-1) \sum_{i=1}^{n} \left(\frac{1}{\lambda x_i} x_i \right) - \sum_{i=1}^{n} x_i = n \frac{1}{\lambda} + (k-1)\, n\, \frac{1}{\lambda} - \sum_{i=1}^{n} x_i = \\
&= \frac{n\, k}{\lambda} - \sum_{i=1}^{n} x_i
\end{aligned}$$

Zahtevajmo sedaj, da je ta odvod enak nič, in dobimo izraz za oceno parametra $\hat{\lambda}$ po *metodi največje verjetnosti*:

$$\frac{d \ln L(\lambda)}{d\lambda} = 0 = \frac{n\, k}{\hat{\lambda}} - \sum_{i=1}^{n} x_i \qquad \longrightarrow \qquad \hat{\lambda} = \frac{n\, k}{\sum_{i=1}^{n} x_i} = \frac{k}{\bar{X}} = \frac{10}{0.9565} = 10.45$$

Dobili smo torej enak rezultat kot po metodi momentov.