Statistika z elementi informatike
Osnove verjetnostnega računa in statistike

21.1.1999

1. Naloga: momenti porazdelitve

Obravnavamo porazdelitev, katere gostota verjetnosti je linearna funkcija:

$$f_X(x) = \begin{cases} a\,x & \ldots \quad 0 \le x \le 2 \\ 0 & \ldots \quad \text{drugje} \end{cases}$$

a) Določite porazdelitveno funkcijo $F_X(x)$.
b) Določite pričakovano vrednost $\mathrm{E}[X]$ in varianco $\mathrm{VAR}[X]$.
c) Ali je centralni moment tretjega reda enak nič, ali različen od nič?
d) Določite mediano $\tilde{m}_X$.

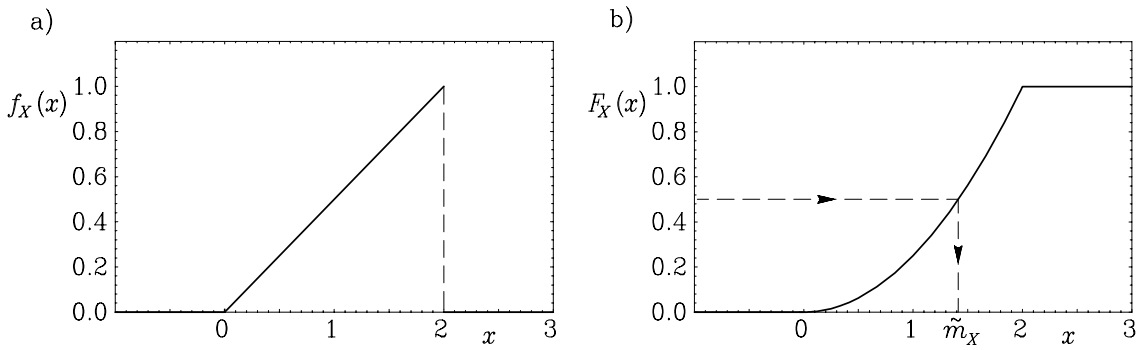
Rešitev: Najprej določimo paramter a iz pogoja, da je integral gostote verjetnosti po celotnem območju realnih števil enak ena:

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_0^2 a\,x\,dx = a\frac{x^2}{2}\Big|_0^2 = 2\,a = 1 \quad \longrightarrow \quad a = \frac{1}{2}.$$

Porazdelitveno funkcijo določimo z integriranjem gostote verjetnosti

$$F_X(x) = \int_{-\infty}^{x} f_X(\bar{x})\,d\bar{x} = \begin{cases} 0 \\ \int_0^x \frac{\bar{x}}{2}\,d\bar{x} \\ 1 \end{cases} = \begin{cases} 0 \\ \frac{\bar{x}^2}{4}\Big|_0^x \\ 1 \end{cases} = \begin{cases} 0 & \ldots \quad x < 0 \\ \frac{x^2}{4} & \ldots \quad 0 \le x \le 2 \\ 1 & \ldots \quad 2 < x \end{cases}$$

Na sliki 1 prikazujemo gostoto verjetnosti in porazdelitveno funkcijo slučajne spremenljivke X.



Slika 1: Gostota verjetnosti $f_X(x)$ in porazdelitvena funkcija $F_X(x)$ slučajne spremenljivke X

Tudi pričakovano vrednost in varianco izračunamo z integriranjem gostote verjetnosti

$$\mathrm{E}[X] = m_X = \int_{-\infty}^{\infty} x\, f_X(x)\, dx = \int_0^2 x\, \frac{x}{2}\, dx = \int_0^2 \frac{x^2}{2}\, dx = \left.\frac{x^3}{6}\right|_0^2 = \frac{8}{6} = \frac{4}{3},$$

$$\mathrm{E}[X^2] = \int_{-\infty}^{\infty} x^2\, f_X(x)\, dx = \int_0^2 x^2\, \frac{x}{2}\, dx = \int_0^2 \frac{x^3}{2}\, dx = \left.\frac{x^4}{8}\right|_0^2 = \frac{16}{8} = 2,$$

$$\mathrm{E}[X^3] = \int_{-\infty}^{\infty} x^3\, f_X(x)\, dx = \int_0^2 x^3\, \frac{x}{2}\, dx = \int_0^2 \frac{x^4}{2}\, dx = \left.\frac{x^5}{10}\right|_0^2 = \frac{32}{10} = \frac{16}{5},$$

$$\mathrm{VAR}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2 = 2 - \left(\frac{4}{3}\right)^2 = \frac{2}{9}.$$

Centralni moment tretjega reda je merilo za simetričnost. Ker je obravnavana porazdelitev nesimetrična, je centralni moment tretjega reda $\mu_X^{(3)}$ različen od nič. Izračunamo ga z enačbo

$$\begin{aligned}\mu_X^{(3)} &= \mathrm{E}[(X - m_X)^3] = \mathrm{E}[X^3 - 3\,X^2\, m_X + 3\,X\, m_X^2 - m_X^3] = \\ &= \mathrm{E}[X^3] - 3\,\mathrm{E}[X^2]\,\mathrm{E}[X] + 2\,\mathrm{E}[X]^3 = \frac{16}{5} - 3\cdot 2\cdot \frac{4}{3} + 2\cdot\left(\frac{4}{3}\right)^3 = -\frac{8}{135}.\end{aligned}$$

Mediana $\tilde{m}_X$ je vrednost, za katero velja, da je verjetnost, da je X manjša od nje, enaka 0.5. Mediano določamo iz izraza za porazdelitveno funkcijo

$$0.5 = P[X < \tilde{m}_X] = F_X(\tilde{m}_X) = \frac{\tilde{m}_X^2}{4} \quad \longrightarrow \quad \tilde{m}_X = \sqrt{2}$$

Na sliki 1 lahko vidimo, kako mediano določimo grafično iz grafa porazdelitvene funkcije.

2. Naloga: normalna porazdelitev

Predpostavimo, da se teža slovenskih prvošolčkov X porazdeljuje normalno. Raziskave kažejo, da je 15% prvošolčkov lažjih od 25 kg, 10% pa težjih od 33 kg.

a) Določite srednjo vrednost m_X in standardno deviacijo σ_X.
b) Tina tehta 22 kg. Ali je med 5% najlažjih?

Rešitev: Zapišimo podatke z enačbami:

$$P[X < 25] = 0.15, \qquad P[X > 33] = 0.10 \;\rightarrow\; P[X < 33] = 0.90.$$

Neenačbe v izrazih za verjetnosti transformiramo tako, da dobimo standardno normalno porazdelitev

$$P\left[\frac{X - m_X}{\sigma_X} < \frac{25 - m_X}{\sigma_X}\right] = P\left[U < \frac{25 - m_X}{\sigma_X}\right] = F_U\left(\frac{25 - m_X}{\sigma_X}\right) = 0.15,$$

$$P\left[\frac{X - m_X}{\sigma_X} < \frac{33 - m_X}{\sigma_X}\right] = P\left[U < \frac{33 - m_X}{\sigma_X}\right] = F_U\left(\frac{33 - m_X}{\sigma_X}\right) = 0.90,$$

kjer je $F_U(u)$ porazdelitvena funkcija standardne normalne porazdelitve. Vrednosti inverzne funkcije $F_U^{-1}(0.15)$ in $F_U^{-1}(0.90)$ lahko odčitamo iz preglednic ali pa uporabimo računalniški program (na primer EXCEL: ukaza `NORMINV(0.15;0;1)` in `NORMINV(0.90;0;1)`)

ali `NORMSINV(0.15)` in `NORMSINV(0.90)`. Ti vrednosti sta $F_U^{-1}(0.15) = -1.0364$ in $F_U^{-1}(0.90) = 1.2816$. Sedaj lahko zapišemo

$$\frac{25 - m_X}{\sigma_X} = F_U^{-1}(0.15) = -1.0364, \qquad \frac{33 - m_X}{\sigma_X} = F_U^{-1}(0.90) = 1.2816.$$

Če zgornji enačbi pomnožimo s σ_X, dobimo sistem dveh linearnih enačb z dvema neznankama (m_X in σ_X). Rešitev tega sistema je

$$m_X = 28.5770\,\text{kg}, \qquad \sigma_X = 3.4513\,\text{kg}.$$

Določiti moramo še verjetnost $P[X < 22]$, kar lahko določimo tako, da neenačbo v izrazu za verjetnost transformiramo

$$P\left[\frac{X - 28.5770}{3.4513} < \frac{22 - 28.5770}{3.4513}\right] = P\left[U < -1.9057\right] = 0.0283,$$

kar lahko odčitamo iz preglednic za standardno normalno porazdelitev ali pa z uporabo računalniškega programa (na primer EXCEL: če računamo za standardno normalno porazdelitev, uporabimo ukaz `NORMSDIST(-1.9057)` ali `NORMDIST(-1.9057;0;1;TRUE)`, če pa računamo za poljubno normalno porazdelitev, uporabimo `NORMDIST(22;28.5770;3.4513;TRUE)`). Vidimo torej, da je Tina med 5% najlažjih prvošolčkov, saj je le 2.83% njenih sošolcev lažjih.

3. Naloga: Linearna regresija

V naslednji preglednici so podatki o povprečni koncentraciji SO_2 v Ljubljani za obdobje 1988–1994, ki jih je izdalo Ministrstvo za okolje in prostor. Predpostavimo, da je koncentracija SO_2 linerano odvisna od leta.

Leto	1988	1989	1990	1991	1992	1993	1994
Koncentracija SO_2	67	72	78	52	41	35	24

a) Izračunajte ocene parametrov regresijske premice B_0, B_1 in σ.
b) Določite, ali je parameter B_1 statistično značilno različen od nič. Stopnja tveganja je $\alpha = 5\%$.
c) Narišite graf linearne funkcije, katere parametre ste določili v točki a). Narišite tudi točke, ki predstavljajo meritve.

Rešitev: Zaradi lažjega računanja za spremenljivko X vzamemo zaporedna števila let v obravnavanem obdobju, kar pomeni, da zalogo vrednosti spremenljivke X tvorijo naravna števila od 1 do 7. Prepišimo zgornjo preglednico, v kateri dodamo stolpec za spremenljivko X ter stolpce za vrednosti X^2, Y^2 in XY.

Leto	X	Y	X^2	Y^2	XY
1988	1	67	1	4489	67
1989	2	72	4	5184	144
1990	3	78	9	6084	234
1991	4	52	16	2704	208
1992	5	41	25	1681	205
1993	6	35	36	1225	210
1994	7	24	49	576	168
Vsota	28	369	140	21943	1236
Povprečje	4.000	52.714	20.000	3134.714	176.571

Momente S_X^2, S_Y^2 in S_{XY} izračunamo iz zadnje preglednice.

$$\begin{aligned} S_X^2 &= 20 - 4^2 = 4, \\ S_Y^2 &= 3134.714 - 52.714^2 = 355.918, \\ S_{XY} &= 176.571 - 4 \cdot 52.714 = -34.286. \end{aligned}$$

Ocene parametrov B_0, B_1 in σ izračunamo po naslednjih enačbah:

$$\hat{B}_1 = \frac{S_{XY}}{S_X^2} = \frac{-34.286}{4} = -8.571,$$

$$\hat{B}_0 = \bar{Y} - \bar{X}\,\hat{B}_1 = 52.714 + 8.571 \cdot 4 = 87.000,$$

$$\hat{\sigma} = \sqrt{\frac{n}{n-2}\left(S_Y^2 - \frac{S_{XY}^2}{S_X^2}\right)} = \sqrt{\frac{7}{5}\left(355.918 - \frac{(-34.286)^2}{4}\right)} = 9.320.$$

Ugotoviti moramo, ali je parameter B_1 statistično značilno večji od nič. Postavimo ničelno in alternativno hipotezo:
H_0: parameter $B_1 = 0$,
H_1: parameter $B_1 \neq 0$.

Testna statistika

$$T = \frac{B_1}{\hat{\sigma}/\sqrt{S_X^2 n}}$$

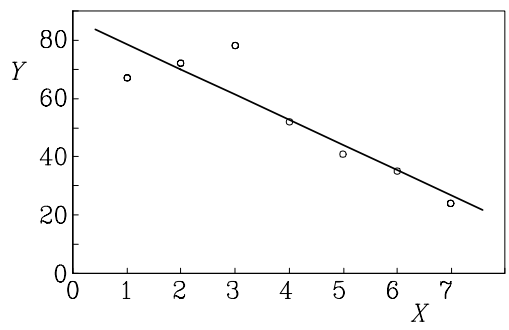
se porazdeljuje po porazdelitvi t z $\nu = n-2$ prostostninimi stopnjami. Pogoj za zavrnitev ničelne hipoteze je:

$$\hat{B}_1 > t_{\alpha/2,n-2}\,\frac{\hat{\sigma}}{\sqrt{S_X^2\, n}} \quad \text{ali} \quad \hat{B}_1 < -t_{\alpha/2,n-2}\,\frac{\hat{\sigma}}{\sqrt{S_X^2\, n}}.$$

Vrednost $t_{0.025,5}$ preberemo iz preglednic za porazdelitev t, ali pa z uporabo računalniškega programa (na primer EXCEL: ukaz `TINV(0.05,5)` in je enaka $t_{0.025,5} = 2.5706$. Kritična vrednost ocene parametra B_1 je

$$t_{\alpha/2,n-2}\,\frac{\hat{\sigma}}{\sqrt{S_X^2\, n}} = 2.5706 \cdot \frac{9.320}{\sqrt{4 \cdot 7}} = 4.527.$$

Ker je $\hat{B}_1 = -8.571 < -4.527$, je pogoj za zavrnitev ničelne hipiteze izpolnjen in lahko zaključimo: *S stopnjo tveganja 5% lahko trdimo, da je paramter B_1 statistično zančilno različen od nič. Torej lahko sklepamo, da leto linerano vpliva na koncentracijo SO_2.* Vidimo tudi, da koncentracija SO_2 z leti pada. Na naslednji sliki prikazujemo graf linearne funkcije, s katero aproksimiramo rezultate meritev.



Slika 2: Linearna regresija

4. Naloga: Analiza variance

Na preglednici podajamo podatke o povprečni koncentraciji SO_2 za Ljubljano, Maribor in Koper, ki jih je izdalo Ministrstvo za okolje in prostor. Z analizo variance ugotovite, ali je kraj statistično značilno pomemben faktor (tveganje je 5%).

Kraj	**1988**	**1989**	**1990**	**1991**	**1992**	**1993**	**1994**
Ljubljana	67	72	78	52	41	35	24
Maribor	67	71	66	76	28	34	28
Koper	17	19	17	12	14	17	11

Rešitev: Najprej izračunamo aritmetične sredine koncentracij za posamezne kraje in skupno aritmetično sredino:

$$\bar{X}_{LJ} = \bar{X}_1 = \frac{67+72+78+52+41+35+24}{7} = 52.714,$$

$$\bar{X}_{MB} = \bar{X}_2 = \frac{67+71+66+76+28+34+28}{7} = 52.857,$$

$$\bar{X}_{KP} = \bar{X}_3 = \frac{17+19+17+12+14+17+11}{7} = 15.286,$$

$$\bar{X}_{skupno} = \bar{X} = \frac{52.714+52.857+15.286}{3} = 40.286.$$

Upoštevamo, da je število različnih vrednosti obravnavanega faktorja $a = 3$, število ponovitev pa je enako $n = 7$. Sedaj izračunamo še vsote kvadratov SS_F, SS_E in SS_T

$$SS_F = n\sum_{i=1}^{a}(\bar{X}_i - \bar{X})^2 = 7 \cdot \left[(52.714 - 40.286)^2 + \cdots\right] = 6562.57,$$

$$SS_E = \sum_{i=1}^{a}\sum_{i=1}^{n}(X_{ij} - \bar{X}_i)^2 = (67 - 52.714)^2 + \cdots = 5373.71,$$

$$SS_T = \sum_{i=1}^{a}\sum_{i=1}^{n}(X_{ij} - \bar{X})^2 = (67 - 40.286)^2 + \cdots = 11936.29.$$

Preverimo lahko, da velja zveza $SS_T = SS_E + SS_F$. Sedaj lahko opravimo testiranje hipoteze. Ničelna in alternativna hipoteza sta:

H_0: Kraj ne vpliva na koncentracijo SO_2;
H_1: Kraj vpliva na koncentracijo SO_2.

Kritično vrednost statistike $F_{1-\alpha,\nu_1,\nu_2}$ odčitamo iz preglednic za porazdelitev F ali pa uporabimo računalniški program (na primer EXCEL: ukaz `=FINV(0.05;2;18)`. Števili prostostnih stopenj sta $\nu_1 = a - 1 = 2$ in $\nu_2 = a\,(n-1) = 18$. Če je statistika $F = MS_F/MS_E$ večja od kritične vrednosti $F_{0.95,2,18} = 3.555$, moramo ničelno hipotezo zavrniti. Zapišimo preglednico analize variance (ANOVA).

Vir	SS	n_{ps}	MS	F	F_{krit}
Faktor	6562.571	2	3281.286	10.991	3.555
Napaka	5373.714	18	298.540		
Skupaj	11936.286	20			

Ker je testna statistika $F = 10.991 > 3.555 = F_{krit}$, moramo ničelno hipotezo zavrniti. Zaključimo lahko: *S tveganjem 5% lahko trdimo, da kraj vpliva na koncentracijo SO_2 v zraku* ali z drugimi besedami: *vpliv kraja na koncentracijo SO_2 je statistično značilen s stopnjo tveganja 5%*