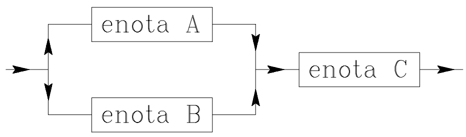
Statistika z elementi informatike
Osnove verjetnostnega računa in statistike

11.6.1999

1. Naloga: Zanesljivost sistema črpalk

Sistem črpalk je shematično prikazan na sliki. Sistem deluje, če deluje vsaj ena izmed enot A in B ter istočasno tudi enota C.



Vzemimo, da dogodki A, B in C predstavljajo delovanje enot A, B in C, dogodki $\bar{A}$, $\bar{B}$ in $\bar{C}$ pa predstavljajo njihovo nedelovanje oziroma okvaro. Zapišite dogodek X, ki predstavlja delovanje sistema, v odvisnosti od dogodkov A, B in C. Vzemimo, da so dogodki A, B in C medsebojno neodvisni. Izračunajte verjetnost, da sistem deluje, če so verjetnosti, da so posamezne enote okvarjene enake:
$P[\bar{A}] = 0.10$,
$P[\bar{B}] = 0.10$,
$P[\bar{C}] = 0.02$.

Rešitev: Zapišimo najprej dogodek X v odvisnosti od dogodkov A, B in C. Iz slike in besedila sledi, da je

$$X = (A \cup B) \cap C.$$

Ker so dogodki A, B in C medsebojno neodvisni, lahko izračunamo verjetnost dogodka X po naslednji enačbi:

$$P[X] = (P[A] + P[B] - P[A \cap B])\,P[C] = (P[A] + P[B] - P[A]\,P[B])\,P[C]$$

Verjetnosti, da posamezne črpalke delujejo, izračunamo s preprostimi enačbami

$$P[A] = 1 - P[\bar{A}] = 0.90,$$
$$P[B] = 1 - P[\bar{B}] = 0.90,$$
$$P[C] = 1 - P[\bar{C}] = 0.98.$$

Sedaj lahko izračunamo tudi verjetnost dogodka X, da sistem deluje

$$P[X] = (0.90 + 0.90 - 0.90 \cdot 0.90) \cdot 0.98 = 0.9702.$$

To verjetnost lahko določimo tudi tako, da dogodek X izrazimo z dogodki $\bar{A}$, $\bar{B}$ in $\bar{C}$. Upoštevamo zvezi

$$(\overline{Y \cup Z}) = \bar{Y} \cap \bar{Z} \qquad \longrightarrow \qquad Y \cup Z = (\overline{\bar{Y} \cap \bar{Z}}),$$
$$(\overline{Y \cap Z}) = \bar{Y} \cup \bar{Z} \qquad \longrightarrow \qquad Y \cap Z = (\overline{\bar{Y} \cup \bar{Z}})$$

in zapišemo

$$X = (\overline{\bar{A} \cap \bar{B}}) \cap C = \overline{(\bar{A} \cap \bar{B}) \cup \bar{C}}.$$

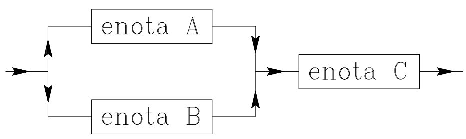
Verjetnost, da sistem deluje, izračunamo na osnovi zadnjega izraza

$$\begin{aligned} P[X] &= 1 - P[(\bar{A} \cap \bar{B}) \cup \bar{C}] = \\ &= 1 - (P[\bar{A} \cap \bar{B}] + P[\bar{C}] - P[\bar{A} \cap \bar{B} \cap \bar{C}]) = \\ &= 1 - P[\bar{A}]\, P[\bar{B}] - P[\bar{C}] + P[\bar{A}]\, P[\bar{B}]\, P[\bar{C}] = \\ &= 1 - 0.1 \cdot 0.1 - 0.02 + 0.1 \cdot 0.1 \cdot 0.02 = 0.9702. \end{aligned}$$

Vidimo, da smo na oba načina dobili enako verjetnost delovanja sistema črpalk – zanesljivost sistema.

2. Naloga: srednja vrednost in varianca

Na sliki je shematično prikazan sistem cevi, kjer so enote A, B in C merilci pretoka.



Dejanske pretoke označimo z m_A, m_B in m_C, izmerjeni pretoki pa so slučajne spremenljivke X_A, X_B in X_C s srednjimi vrednostmi m_A, m_B in m_C in enakimi standardnimi deviacijami $\sigma_A = \sigma_B = \sigma_C = \sigma$. Predpostavimo, da sistem cevi na obravnavanem delu ne pušča. To pomeni, da je $m_A + m_B = m_C$. Po metodi najmanjših kvadratov lahko ugotovimo, da je najboljša ocena za pretok skozi enoto A enaka

$$\hat{m}_A = \frac{2\,X_A - X_B + X_C}{3}.$$

Pokaži, da je prikazana ocena nepristranska ($\mathrm{E}[\hat{m}_A] = m_A$), in določi njeno varianco $\mathrm{VAR}[\hat{m}_A]$. Dodatna (neobvezna) naloga: po metodi najmanjših kvadratov določi oceni $\hat{m}_B$ in $\hat{m}_C$ za pretoke skozi enoti B in C.

Rešitev: Izračunajmo srednjo vrednost in varianco ocene pretoka $\hat{m}_A$ skozi enoto A. Pri tem upoštevamo naslednji zvezi za srednjo vrednost in varianco funkcij slučajnih spremenljivk:

$$\mathrm{E}[a\,X - b\,Y] = a\,\mathrm{E}[X] - b\,\mathrm{E}[Y], \qquad \mathrm{VAR}[a\,X - b\,Y] = a^2\,\mathrm{VAR}[X] + b^2\,\mathrm{VAR}[Y].$$

Če zadnji enačbi upoštevamo v izrazu za oceno pretoka $\hat{m}_A$ skozi enoto A in upoštevamo tudi, da v obravnavanem delu sistema ne pušča ($m_A + m_B = m_C$), dobimo

$$\begin{aligned} \mathrm{E}[\hat{m}_A] &= \mathrm{E}\left[\frac{2\,X_A - X_B + X_C}{3}\right] = \frac{2\,\mathrm{E}[X_A] - \mathrm{E}[X_B] + \mathrm{E}[X_C])}{3} = \\ &= \frac{2\,m_A - m_B + m_C}{3} = \frac{2\,m_A - m_B + m_A + m_B}{3} = \frac{2\,m_A + m_A}{3} = m_A. \end{aligned}$$

Ker je pričakovana oziroma srednja vrednost ocene $\hat{m}_A$ enaka dejanskemu pretoku, lahko zaključimo, da je ocena nepristranska. Na podoben način izračunamo tudi varianco:

$$\begin{aligned} \mathrm{VAR}[\hat{m}_A] &= \mathrm{VAR}\left[\frac{2\,X_A - X_B + X_C}{3}\right] = \\ &= \frac{4\,\mathrm{VAR}[X_A] + \mathrm{VAR}[X_B] + \mathrm{VAR}[X_C])}{9} = \frac{4\,\sigma^2 + \sigma^2 + \sigma^2}{9} = \frac{2\,\sigma^2}{3}. \end{aligned}$$

Vidimo, da ima ocena $\hat{m}_A$ manjšo varianco od meritve na enoti A, kar pomeni, da je boljša ocena pretoka skozi enoto A od same meritve X_A.

3. Naloga: Območje zaupanja

Ne prav spreten košarkar je v pretekli sezoni v 150 poskusih zadel 90 prostih metov. Določite območje zaupanja za vrednost p, to je verjetnost, da košarkar v posameznem poskusu zadene prosti met. Stopnja tveganja je $\alpha = 2\%$.

Rešitev: Ocena verjetnosti, da košarkar zadene koš, je

$$\hat{p} = \frac{90}{150} = 0.60.$$

Ker je število poskusov veliko, lahko predpostavimo, da je porazdelitev ocene $\hat{p}$ normalna s srednjo vrednostjo $m_P = p$ in varianco $\sigma_P = p(1-p)/n$, kjer je n število poskusov. Ocena standardne deviacije je

$$\hat{\sigma}_P = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.60 \cdot (1-0.60)}{150}} = 0.04.$$

Območje zaupanja določimo z naslednjim izrazom

$$P[\hat{p} - k_{\alpha/2}\,\sigma_P \leq p \leq \hat{p} + k_{\alpha/2}\,\sigma_P] = 1 - \alpha.$$

Za izbrano vrednost $\alpha = 2\%$ lahko vrednost $k_{\alpha/2}$ določimo iz preglednice za standardno normalno porazdelitev, ali pa z uporabo računalniškega programa (na primer EXCEL: ukaz `NORMSINV(0.99)` ali `NORMINV(0.99;0;1)`): $k_{\alpha/2} = 2,3263$. Sedaj lahko izračunamo območje zaupanja za verjetnost p, da košarkar zadene prosti met:

$$P[0.507 \leq p \leq 0.693] = 0.98.$$

4. Naloga: Linearna regresija

Pričakujemo lahko, da je količina letno uskladiščenih radioaktivnih odpadkov odvisna od količine električne energije, ki jo lahko proizvedejo jedrske centrale. Na naslednji preglednici so podatki za šest držav. Predpostavite, da je povezava med količino električne energije in letno uskladiščenih radioaktivnih odpadkov linearna, in izračunajte ocene za parametre linearne regresije B_0 in B_1. Ugotovite tudi, ali je parameter B_1 pri stopnji tveganja 5% statistično značilno **večji** od nič.

Država	**Proizvodnja električne energije** [MW]	**Uskladiščeni radioaktvini odpadki** [m^3/leto]	**Uskladiščeni radioaktvini odpadki** [m^3]	**Obdobje skladiščenja radioaktvinih odpadkov**
Finska	2400	167	1000	1992-1997
Francija	60000	20690	600000	1969-1997
Japonska	42400	1167	7000	1992-1997
Nemčija	22300	1500	30000	1978-1997
Španija	7200	1000	6000	1992-1997
Švedska	10000	2400	24000	1988-1997

Rešitev: Za neodvisno spremenljivko X izberemo proizvodnjo električne energije, odvisna spremenljivka Y pa je količina letno uskladiščenih radioaktivnih podatkov. Prepišimo sedaj preglednico s podatki in dodajmo še račun X^2, Y^2 in XY ter povprečne vrednosti vseh teh količin.

	X	Y	X^2	Y^2	XY
1	2400	166.67	5760000	27777.78	400000.00
2	60000	20689.66	3600000000	428061831.15	1241379310.34
3	42400	1166.67	1797760000	1361111.11	49466666.67
4	22300	1500.00	497290000	2250000.00	33450000.00
5	7200	1000.00	51840000	1000000.00	7200000.00
6	10000	2400.00	100000000	5760000.00	24000000.00
Povprečje	24050	4487.16	1008775000	73076786.67	225982662.84

Z rezultati zgornje preglednice lahko izračunamo S_X, S_Y, S_{XY}

$$\begin{aligned} S_X^2 &= 1008775000 - 24050^2 = 430372500, \\ S_Y^2 &= 73076786.67 - 4487.16^2 = 52942139.17, \\ S_{xy} &= 225982662.84 - 24050 \cdot 4487.16 = 118066350.57 \end{aligned}$$

in nato tudi oceni parametrov linearne regresije

$$\hat{B}_1 = \frac{S_{XY}}{S_X^2} = \frac{118066350.57}{430372500} = 0.2743,$$

$$\hat{B}_0 = \bar{Y} - \bar{X}\,\hat{B}_1 = 4487.16 - 0.2743 \cdot 24050 = -2110.60,$$

$$\hat{\sigma} = \sqrt{\frac{n}{n-2}\left(S_Y^2 - \frac{S_{XY}^2}{S_X^2}\right)} = \sqrt{\frac{6}{4}\left(52942139.17 - \frac{118066350.57^2}{430372500}\right)} = 5552.35.$$

Nazadnje moramo ugotoviti, ali je parameter B_1 statistično značilno večji od nič. Postavimo ničelno in alternativno hipotezo:

H_0: parameter $B_1 = 0$,
H_1: parameter $B_1 > 0$.

Vidimo, da bomo morali opraviti enostranski test. 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:

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

Vrednost $t_{0.05,4}$ preberemo iz tabel za porazdelitev t, ali pa z uporabo računalniškega programa (na primer EXCEL: ukaz `TINV(0.10,4)` in je enaka $t_{0.05,4} = 2.1318$. Kritična vrednost je

$$t_{\alpha,n-2}\,\frac{\hat{\sigma}}{\sqrt{S_X^2 n}} = 2.1318 \cdot \frac{5552.35}{\sqrt{430372500 \cdot 6}} = 0.2329.$$

Vidimo, da je pogoj za zavrnitev ničelne hipiteze izpolnjen in lahko zaključimo: *S stopnjo tveganja 5% lahko trdimo, da je paramter* B_1 *statistično zančilno večji od nič. Torej lahko sklepamo, da faktor* X *(proizvodnja električne energije) vpliva na spremenljivko* Y *(količino letno uskladiščenih radioaktivnih odpadkov).*