POROČILO SEMINARSKE NALOGE

STATISTIKA

Sara Bizjak

Univerza v Ljubljani
Fakulteta za matematiko in fiziko
Oddelek za matematiko

JULIJ 2020

1. NALOGA

Podatki so vzeti iz datoteke `Kibergard`, kjer se nahajajo informacije o 43.886 družinah, ki stanujejo v mestu Kibergard. Za vsako družino so zabeleženi naslednji podatki:

- Tip družine (od 1 do 3)
- Število članov družine
- Število otrok v družini
- Skupni dohodek družine
- Mestna četrt, v kateri stanuje družina (od 1 do 4)
- Stopnja izobrazbe in vodje gospodinjstva (od 31 do 46: opisi v datoteki z navodili)

Nalogo sem reševala s pomočjo programa R. Koda, uporabljena za generiranje enostavnih slučajnih vzorcev in izračune, je priložena na koncu poročila pod naslovom *Dodatek A*, dostopna pa je tudi v priloženi datoteki *naloga1.R*.

Primer a
Vzamemo enostavni slučajni vzorec 200 družin in na njegovi podlagi ocenimo delež družin v Kibergardu, v katerih vodja gospodinjstva nima srednješolske izobrazbe (niti poklicne niti splošne mature). Opisan delež znaša $p = 0.195$.

Primer b
Ocenimo standardno napako in postavimo 95% interval zaupanja.
Standardno napako za delež izračunamo po formuli (na strani 210 v knjigi *John A. Rice: Mathematical Statistics and Data Analysis*):

$$\hat{se}(p) = \sqrt{\frac{p \cdot (1-p)}{n} \cdot \left(1 - \frac{n-1}{N-1}\right)},$$

kjer so $p = 0.195$, $n = 200$, $N = 43.886$.
Dobimo rezultat $\hat{se}(p) = 0.02795203$.
Interval zaupanja je enak: $[0.140215, 0.249785]$.

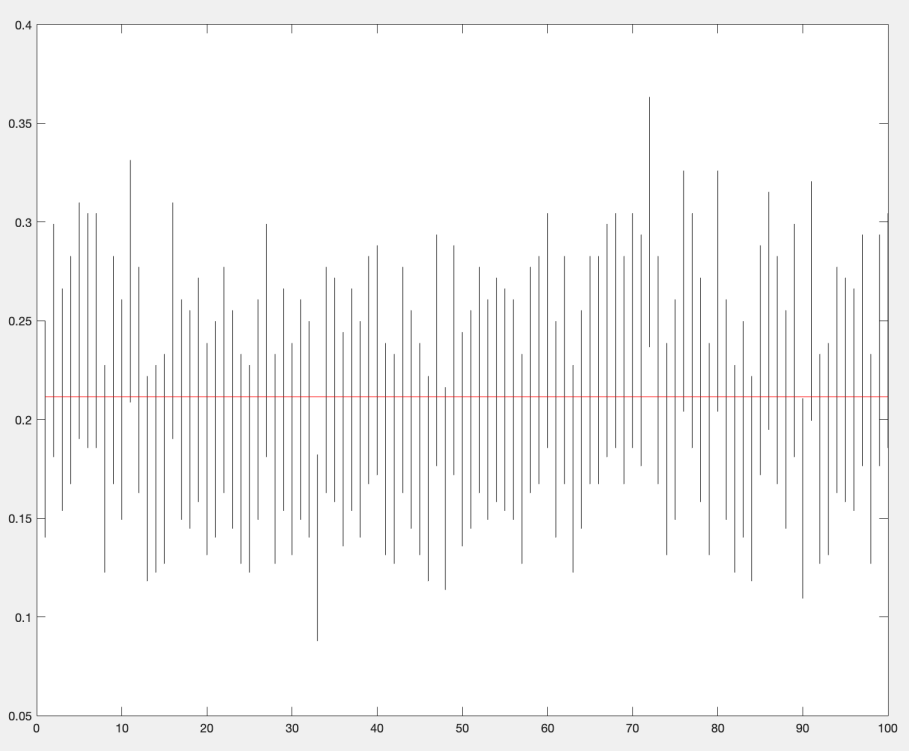
Primer c
Vzorčni delež in ocenjeno standardno napako primerjamo s populacijskim deležem in pravo standardno napako.

- Vzorčni delež: 0.195
- Populacijski delež: 0.2115025
- Razlika obeh deležev: 0.01650253
- Ocenjena standardna napaka (iz vzorca): $0,02802185$
- Prava standardna napaka (iz celotne populacije): 0.02881085
- Razlika med ocenjeno in pravo standardno napako: 0.0008588181

Ker velja $0.2115025 \in [0.140215, 0.249785]$, interval zaupanja pokrije populacijski delež.

Primer d

Poleg vzorca iz točke a) vzamemo še 99 enostavnih slučajnih vzorcev po 200 družin in prav tako za vsakega določimo 95% interval zaupanja. Intervale zaupanja prikažemo na grafu in ugotovimo, koliko jih pokrije populacijski delež.

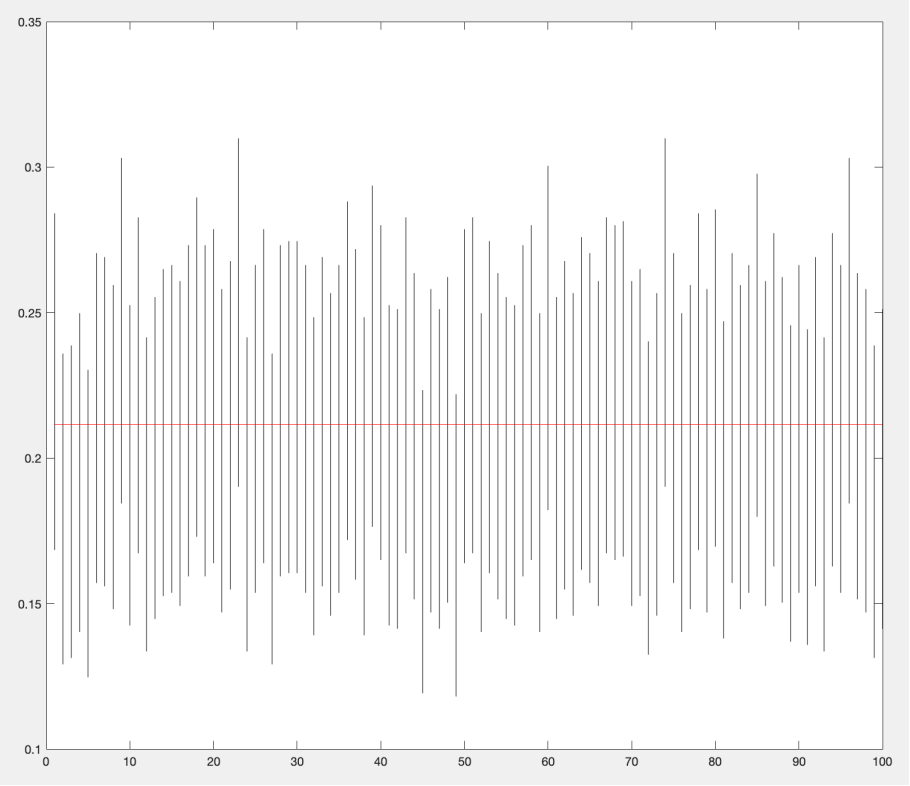


Slika 1: Intervali zaupanja za 100 slučajnih vzorcev velikosti 200.

Izmed 100 intervalov zaupanja jih 97 pokrije populacijski delež. To vidimo tudi na grafu, kjer rdeča črta označuje populacijski delež.

Primer e
Standardni odklon vzorčnih deležev za 100 prej dobljenih vzorcev je enak 0.00140308. Prava standardna napaka za vzorec velikosti 200 pa je 0.02881085. Razlikujeta se za 0.02740777.

Primer f
Vzamemo 100 slučajnih vzorcev po 800 družin. Za vsakega določimo 95% interval zaupanja in intervale prikažemo na grafu in ugotovimo, koliko jih pokrije populacijski delež.



Slika 2: Intervali zaupanja za 100 slučajnih vzorcev velikosti 800.

Izmed 100 intervalov zaupanja vzorcev po 800 jih vseh 100 pokrije poulacijski delež. To vidimo tudi na grafu, kjer rdeča črta označuje populacijski delež. Standardni odklon vzorčnih deležev za 100 dobljenih vzorcev velikosti 800 je enak 0.0008074025. Prava standardna napaka za vzorec velikosti 800 pa je 0.01430616. Razlikujeta se za 0.01349875.

V primeru, ko smo vzeli 100 vzorcev velikosti 800, smo dobili boljše rezultate kot pri izbiri 100 vzorcev velikosti 200. Praviloma lahko z izbiro večjega vzorca dobimo boljše napovedi, vendar moramo pri tem paziti še na dober vzorčni načrt. Vzorčni načrt je v naprej predpisan postopek izbiranja vzorca iz vnaprej določene in natančno opredeljene populacija. Način vzorčenja bistveno vpliva na zanesljivost ocen. Če je izbira enot iz populacije naključna, potem takemu vzorčenju pravimo verjetnostno vzorčenje. Z vpeljavo naključnosti se najboljse izognemo pristranskosti. Primer slabega vzorčnega načrta in posledično napačnih napovedi je napoved izida volitev v ZDA leta 1936.

2. NALOGA

Sklepi in izpeljave, ki jih bom uporabila pri vseh podnalogah, sledijo dokazu izreka v knjigi *John A. Rice: Mathematical Statistics and Data Analysis* (stran 232, 233).
Naredimo raziskavo na populaciji, ki ima K stratumov z velikostmi $N_1, N_2, \ldots, N_K$. Denimo, da lahko izberemo vzorec velikosti n.

Primer a
Denimo, da so stroški raziskave enaki $C = C_0 + nC_1$, kjer je n število enot v vzorcu (C_0 je začetni stršek, C_1 pa je nadaljnji strošek na enoto). Pri danih sredstvih za raziskavo v višini C poiščemo velikosti podvzorcev $n_1, n_2, \ldots, n_K$, pri katerih je varianca standardne cenilke populacijskega povprečja minimalna.
Če imamo stratume velikosti $N_1, N_2, \ldots, N_K$, se moramo odločiti, kako velike vzorce bomo izbrali iz posameznih stratumov. Izbiramo tako, da bo varianca cenilke $\overline{X}$ čim manjša. Poiščemo torej take $n_1, n_2, \ldots, n_K$, kjer $n = n_1 + n_2 + \ldots + n_K$, da bo

$$\mathrm{var}(\overline{X}) = \sum_{k=1}^{K} W_k^2 \left(\frac{\sigma_k^2}{n_k}\right)\left(\frac{N_k - n_k}{N_k - 1}\right)$$

čim manjša. Ker so v večini praktičnih situacij korekturni faktorji $\frac{N_k - n_k}{N_k - 1} \approx 1$, jih lahko zanemarimo. Rešujemo torej problem vezanega ekstrema:

$$f(n_1, n_2, \ldots, n_K) = \sum_{k=1}^{K} \frac{\sigma_k^2}{n_k} W_k^2$$

$$\text{z vezjo } C = C_0 + nC_1.$$

Sestavimo Lagrangeovo funkcijo:

$$F(n_1, n_2, \ldots, n_K, \lambda) = f(n_1, n_2, \ldots, n_K) + \lambda(C_0 + \sum_{k=1}^{K} n_k C_1 - C).$$

Zapišemo parcialne odvode funkcije F in jih enačimo z 0.

$$\frac{\partial F}{\partial n_i} = -\frac{\sigma_i^2}{n_i^2} W_i^2 + \lambda C_1 = 0 \quad \text{za } i = 1, 2, \ldots, K.$$

Izrazimo n_i in dobimo sistem enačb

$$n_i = \frac{W_i \sigma_i}{\sqrt{\lambda C_1}} \tag{1}$$

Da določimo λ, naredimo vsoto enačbe (1) po i, $i = 1, 2, \ldots, K$.

$$n = \frac{1}{\sqrt{\lambda C_1}} \sum_{k=1}^{K} W_k \sigma_k$$

$$\Rightarrow \sqrt{\lambda} = \frac{\sum_{k=1}^{K} W_k \sigma_k}{n\sqrt{C_1}} \tag{2}$$

Če združimo (1) in (2), dobimo:

$$n_i = n \frac{W_i \sigma_i}{\sum_{k=1}^{K} W_k \sigma_k}.$$

Rezultat je enak, kot če bi iskali minimalno varianco brez danega začetnega pogoja – ni pomembno, kakšne $n_1, n_2, \ldots, n_K$ izberemo. Stroški raziskave C bodo vedno enaki, ker je cena C_1 vedno enaka.

Primer b

Sedaj se lahko stroški opažanja spreminjajo od stratuma do strauma. Če je n_k število enot iz $k-$tega stratuma, ki so zajete v vzorec, naj bodo stroški raziskave enaki:

$$C = C_0 + \sum_{k=1}^{K} n_k C_k.$$

Pri danih sredstvih za raziskavo v višini C poiščemo tiste velikosti podvzorcev, pri katerih je varianca cenilke populacijskega povprečja minimalna.
Rešujemo podoben primer kot prej, le da sedaj nimamo več fiksne C_1, ampak

se spreminja z vsakim stratumom. Z enakim razmislekom kot prej sestavimo Lagrangeovo funkcijo:

$$F(n_1, n_2, \ldots, n_K, \lambda) = f(n_1, n_2, \ldots, n_K) + \lambda(C_0 + \sum_{k=1}^{K} n_k C_k - C).$$

Zapišemo parcialne odvode funkcije F in jih enačimo z 0.

$$\frac{\partial F}{\partial n_i} = -\frac{\sigma_i^2}{n_i^2} W_i^2 + \lambda C_i = 0 \quad \text{za } i = 1, 2, \ldots, K.$$

Izrazimo n_i in dobimo sistem enačb

$$n_i = \frac{W_i \sigma_i}{\sqrt{\lambda C_i}} \tag{3}$$

Da določimo λ, naredimo vsoto enačbe (3) po i, $i = 1, 2, \ldots, K$.

$$n = \frac{1}{\sqrt{\lambda}} \sum_{k=1}^{K} \frac{W_k \sigma_k}{\sqrt{C_k}}$$

$$\Rightarrow \sqrt{\lambda} = \frac{1}{n} \sum_{k=1}^{K} \frac{W_k \sigma_k}{\sqrt{C_k}} \tag{4}$$

Če združimo (3) in (4), dobimo:

$$n_i = \frac{n}{\sqrt{C_i}} \frac{W_i \sigma_i}{\sum_{k=1}^{K} \frac{W_k \sigma_k}{\sqrt{C_k}}}.$$

Primer c

Naj se stroški raziskave izražajo na enak način kot v prejšnji točki, predpisano pa imamo natančnost raziskave, torej varianco cenilke. Poiščemo tiste vrednosti podvzorcev, pri katerih bodo stroški najmanjši.

Želimo torej minimizirati stroške pri pogoju $\text{var}(\overline{X}) = \sum_{k=1}^{K} \frac{W_k^2 \sigma_k^2}{n_k}$. Označimo $\text{var}(\overline{X}) = V$, kjer je V predpisana vrednost.

Rešujemo torej vezani ekstrem za funkcijo:

$$f(n_1, n_2, \ldots, n_K) = C_0 + \sum_{k=1}^{K} n_k C_k$$

$$\text{z vezjo } \sum_{k=1}^{K} \frac{W_k^2 \sigma_k^2}{n_k} - V.$$

Sestavimo Lagrangeovo funkcijo:

$$F(n_1, n_2, \ldots, n_K, \lambda) = C_0 + \sum_{k=1}^{K} n_k C_k + \lambda \left(\sum_{k=1}^{K} \frac{W_k^2 \sigma_k^2}{n_k} - V \right)$$

Zapišemo parcialne odvode funkcije F in jih enačimo z 0.

$$\frac{\partial F}{\partial n_i} = C_i - \lambda \, \frac{W_i^2 \sigma_i^2}{n_i^2} = 0 \quad \text{za } i = 1, 2, \ldots, K.$$

Izrazimo n_i in dobimo sistem enačb

$$n_i = \sqrt{\frac{\lambda}{C_i}} \, W_i \sigma_i \tag{5}$$

Da določimo λ, naredimo vsoto enačbe (5) po i, $i = 1, 2, \ldots, K$.

$$n = \sqrt{\lambda} \, \sum_{k=1}^{K} \frac{W_k \sigma_k}{\sqrt{C_k}}$$

$$\Rightarrow \sqrt{\lambda} = \frac{n}{\sum_{k=1}^{K} \frac{W_k \sigma_k}{\sqrt{C_k}}} \tag{6}$$

Če združimo (5) in (6), dobimo:

$$n_i = \frac{n}{\sqrt{C_i}} \, \frac{W_i \sigma_i}{\sum_{k=1}^{K} \frac{W_k \sigma_k}{\sqrt{C_k}}}.$$

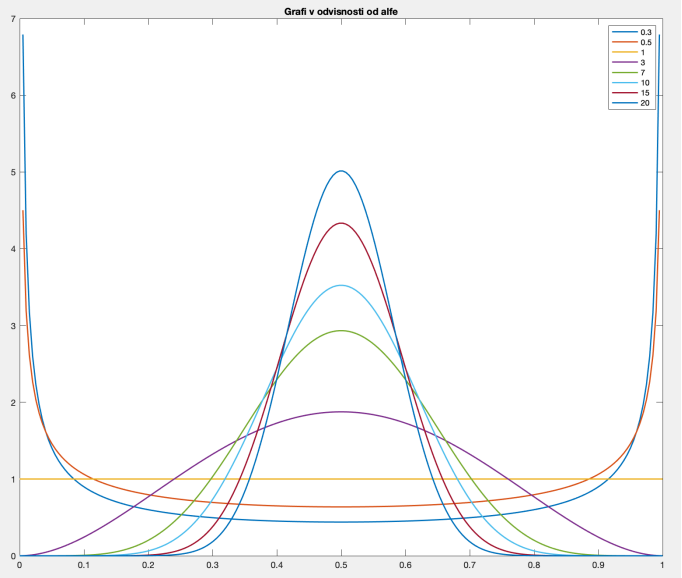
3. NALOGA

Opazimo n neodvisnih realizacij zvezne porazdelitve z gostoto:

$$f(x \mid \alpha) = \begin{cases} \frac{\Gamma(2\alpha)}{(\Gamma(\alpha))^2} \, [x(1-x)]^{\alpha - 1} & ; \ 0 < x < 1 \\ 0 & ; \ \text{sicer,} \end{cases}$$

kjer je $\alpha > 0$ neznam parameter. Če je X slučajna spremenljivka s to gostoto, se da izračunati:

$$\mathrm{E}(X) = \frac{1}{2} \, , \qquad \mathrm{var}(X) = \frac{1}{4(2\alpha + 1)}.$$



Slika 3: Grafi podane funkcije za različne α.

PRIMER A

Določimo obliko porazdelitve v odvisnosti od α. Vidimo, da je funkcija $f(x \mid \alpha)$ simetrična glede na $x = \frac{1}{2}$ in ima v tej točki ekstrem ne glede na vrednost α. Opazimo tudi, da je funkcija za $\alpha = 1$ kar vodoravna premica. Za $\alpha < 1$ je funkcija konveksna, za $\alpha > 1$ pa konkavna. Z večanjem α postaja funkcija vse bolj strma in *"ožja"*.

PRIMER B

Ocenimo α po metodi momentov.
Iz podane pričakovane vrednosti in variance lahko izračunamo vrednosti prvega in drugega momenta.

$$\mathrm{E}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{2},$$

$$\mathrm{E}(X^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mathrm{var}(X) + \mathrm{E}(X)^2 = \frac{1}{4(2\alpha+1)} + \frac{1}{4}.$$

Iz enačbe drugega momenta izrazimo α.

$$\begin{aligned}
\frac{1}{4(2\alpha+1)} + \frac{1}{4} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 \\
\frac{1}{4(2\alpha+1)} &= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{4} \\
\frac{1}{2\alpha+1} &= \frac{4}{n}\sum_{i=1}^{n} X_i^2 - 1 \\
2\alpha+1 &= \frac{1}{\frac{4}{n}\sum_{i=1}^{n} X_i^2 - 1} \\
\Rightarrow \hat{\alpha} &= \frac{1}{2}\left(\frac{1}{\frac{4}{n}\sum_{i=1}^{n} X_i^2 - 1} - 1\right).
\end{aligned}$$

Primer c

Poiščemo enačbo, ki določa cenilko po metodi največjega verjetja. Pogledamo, kdaj ta cenilka obstaja.

$$L_1(\alpha \mid x_1) = \frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}\, x_1^{\alpha-1}(1-x_1)^{\alpha-1}$$

$$l_1(\alpha \mid x_1) = \ln(\mathrm{L}_1(\alpha|x_1))$$

$$\Rightarrow l_1(\alpha \mid x_1) = \ln(\Gamma(2\alpha)) - 2\ln(\Gamma(\alpha)) + (\alpha-1)\ln(x_1) + (\alpha-1)\ln(x_1-1)$$

$$l(\alpha \mid x) = \sum_{i=1}^{n} \mathrm{l}_1(\alpha|x_1)$$

$$\Rightarrow l(\alpha \mid x) = n\cdot\ln(\Gamma(2\alpha)) - 2n\cdot\ln(\Gamma(\alpha)) + (\alpha-1)\sum_{i=1}^{n}\ln(x_i) + (\alpha-1)\sum_{i=1}^{n}\ln(x_i-1)$$

Funkcijo $l(\alpha \mid x)$ odvajamo po α.

$$\frac{\partial l}{\partial \alpha} = n\cdot\frac{1}{\Gamma(2\alpha)}\cdot 2\Gamma'(2\alpha) - 2n\cdot\frac{1}{\Gamma(\alpha)}\cdot\Gamma'(\alpha) + \sum_{i=1}^{n}\ln(x_i) + \sum_{i=1}^{n}\ln(x_i-1)$$

Cenilka bo obstaja natanko tedaj, ko bo imela enačba

$$\frac{1}{\Gamma(2\alpha)}\cdot\Gamma'(2\alpha) - \frac{1}{\Gamma(\alpha)}\cdot\Gamma'(\alpha) = \frac{1}{2n}\sum_{i=1}^{n}\ln\left(\frac{1}{x_i(x_i-1)}\right)$$

rešitev.

PRIMER D

Poiščemo asimptotično varianco cenilke po metodi največjega verjetja.
Za asimptotično varianco bomo potrebovali Fischerjevo informacijo za cenilko, saj velja

$$\mathrm{var}(\hat{\alpha}) \approx \frac{1}{n \cdot I_1(\hat{\alpha})},$$

kjer (iz predavanj)

$$I_1(\hat{\alpha}) = -\mathrm{E}\left[\frac{\partial^2\, l_1(\alpha \mid x_1)}{\partial \alpha^2}\right] \tag{7}$$

V $\frac{\partial l}{\partial \alpha}$ vstavimo $n = 1$ in izračunamo še drugi odvod. Dobimo:

$$\frac{\partial^2 l_1}{\partial \alpha^2} = \frac{4\ \Gamma''(2\alpha)\ \Gamma(2\alpha) - 4\ \Gamma'(2\alpha)^2}{\Gamma(2\alpha)^2} - \frac{2\ \Gamma''(\alpha)\ \Gamma(\alpha) - 2\Gamma'(\alpha)^2}{\Gamma(\alpha)^2}$$

Pričakovana vrednost drugega odvoda je kar enaka drugemu odvodu, ker v njem ne nastopa slučajna spremenljivka X in je torej konstanta. Zato lahko takoj zapišemo varianco, ki je enaka

$$\mathrm{var}(\hat{\alpha}) = \frac{1}{n} \cdot \frac{1}{\frac{2\ \Gamma''(\alpha)\ \Gamma(\alpha) - 2\Gamma'(\alpha)^2}{\Gamma(\alpha)^2} - \frac{4\ \Gamma''(2\alpha)\ \Gamma(2\alpha) - 4\ \Gamma'(2\alpha)^2}{\Gamma(2\alpha)^2}}\ .$$

4. NALOGA V tabeli imamo podatke Ameriškega nacionalnega centra za statistiko zdravja o številu samomorov v ZDA v letu 1970 po mesecih.

Mesec	Število samomorov	Število dni
Januar	1867	31
Februar	1789	28
Marec	1944	31
April	2094	30
Maj	2097	31
Junij	1981	30
Julij	1887	31
Avgust	2024	31
September	1928	30
Oktober	2032	31
November	1978	30
December	1859	31

Slika 4: Podatki o samomorih

PRIMER A
Narišemo histogram, pri katerem so širine stolpcev sorazmerne dolžinam mesecev.
PRIMER B
V knjigi *John A. Rice: Mathematical Statistics and Data Analysis* (stran 343, 344) preberemo o Pearsonovem χ^2 testu, ki ga bomo uporabili v našem primeru.
Označimo z 1 mesec januar, z 2 febaruar in po vrsti naprej do 12 za december.
Preveriti želimo predpostavko, da je samomorilnost skozi celo leto konstanta, kar pomeni, da je konstantna za vseh 12 mesecev (razmerje samomorov v posameznem mesecu je torej $\frac{1}{12}$). Naš test bo v tem primeru enak:

$$H_0 : p_1 = p_2 = \ldots = p_{12} = \frac{1}{12}$$

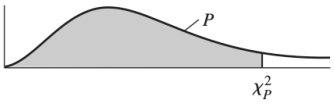
$$H_1 : \text{vsaj ena izmed } p_1, \ldots, p_{12} \text{ ni enaka } \frac{1}{12}$$

Uporabili bomo Parsonovo χ^2 statistiko:

$$\chi^2 = \sum_{i=1}^{12} \frac{(O_i - E_i)^2}{E_i} \xrightarrow{n\to\infty} \chi^2_{m-k-1}, \tag{8}$$

kjer je $m = 12$, k število ocenjenih parametrov, kar je v napem primeru 0. Imamo torej 11 prostostnih stopenj. O_i je opazovan rezultat, torej podatek, koliko samomorov se v mesecu zgodi.
Iz spodnje tabele (slika iz *John A. Rice: Mathematical Statistics and Data Analysis*) vidimo, da je točka, ki določa zgornjih 5% χ^2 testa z 11 prostostnimi stopnjami enaka 26.76. To pomeni, da test ovržemo, če $\chi^2 > 26.76$.

TABLE 3 Percentiles of the χ^2 Distribution—Values of χ^2_P Corresponding to P



df	$\chi^2_{.005}$	$\chi^2_{.01}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.10}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$
1	.000039	.00016	.00098	.0039	.0158	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.1026	.2107	4.61	5.99	7.38	9.21	10.60
3	.0717	.115	.216	.352	.584	6.25	7.81	9.35	11.34	12.84
4	.207	.297	.484	.711	1.064	7.78	9.49	11.14	13.28	14.86
5	.412	.554	.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	.676	.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
120	83.85	86.92	91.58	95.70	100.62	140.23	146.57	152.21	158.95	163.64

Slika 5: Hi kvadrat porazdelitev glede na prostostne stopnje.

Izračunajmo χ^2. Seštejmo vse samomore v letu in seštevek označimo z N. Velja $N = 1867 + 1789 + \ldots + 1859 = 23480$. Po ničelni hipotezi velja, da je $p_i = \frac{1}{12},\ \forall i = 1, 2, \ldots, 12$. Velja torej

$$\mathrm{E}_i = N \cdot p_i = 1956.67,\ \forall i = 1, 2, \ldots, 12.$$

Za lažji pregled zapišimo podatke za rabo izračuna (8) v tabelo:

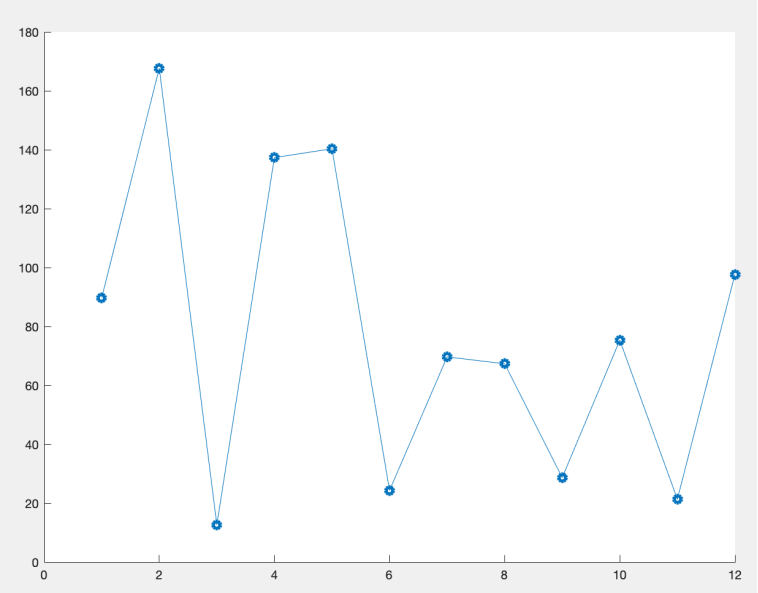
i	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
O_i	1867	1789	1944	2094	2097	1981	1887	2024	1928	2032	1978	1859
$\frac{(O_i-E_i)^2}{E_i}$	4.11	14.37	0.08	9.64	10.06	0.30	2.48	2.32	0.42	2.90	0.23	4.88

Vrednost χ^2 statistike je vsota števil v zadnji vrstici tabele. Velja torej

$$\chi^2 = 4.11 + 14.37 + \ldots + 4.88 = 51.79$$

Iz zgornjih sklepov sledi, da H_0 zavržemo, saj je $51.79 > 26.76$. Zavrnili smo torej hipotezo, da je razmerje samomorov v posameznih mesecih konstantno.

Če še malo analiziramo podatke v tabeli, vidimo, da k testni statistiki največ prispevajo meseci februar, april in maj, saj imajo glede na druge mesece zelo visoko vrednost. Pri teh treh mesecih se pričakovano število samomorov in dejansko število samomorov za največ razlikujeta.



Slika 6: Absolutne razlike med pričakovanim in dejanskim številom samomorv za posamezni mesec

Iz grafa ne vidimo nobenega vzorca, ki bi nam lahko kaj povedal o številu samomorov glede na posamezni mesec.

5. NALOGA

X in Y sta slučajni spremenljivki, za kateri velja:

- $\mathrm{E}(X) = \mu_x$,
- $\mathrm{E}(Y) = \mu_y$,
- $\mathrm{var}(X) = \sigma_x^2$,
- $\mathrm{var}(Y) = \sigma_y^2$,
- $\mathrm{cov}(X, Y) = \sigma_{x,y}$.

Opazimo X in želimo napovedati Y.

Primer a
Poiščemo napoved, ki je oblike $\hat{Y} = \alpha + \beta \cdot X$, kjer α in β izberemo tako, da je srednja kvadratična napaka $\mathrm{E}\left[\left(Y - \hat{Y}\right)^2\right]$ minimalna.
Uporabimo namig

$$\mathrm{E}\left[\left(Y - \hat{Y}\right)^2\right] = \left[\mathrm{E}(Y) - \mathrm{E}(\hat{Y})\right]^2 + \mathrm{var}(Y - \hat{Y}). \tag{9}$$

Ker sta oba člena desne strani enačbe večja od 0, je dovolj, da poiščemo vrednosti α in β, ki minimizirata ta dva člena – potem bo najmanjša možna tudi njuna vsota.
Poglejmo si najprej prvi člen enačbe. Vrednost enačbe

$$\left[\mathrm{E}(Y) - \mathrm{E}(\hat{Y})\right]^2 = [\mu_y - \alpha - \beta\mu_x]^2$$

bo najmanjša, ko bo $\mu_y - \alpha - \beta \cdot \mu_x = 0$. To pa bo res natanko takrat, ko bo $\alpha = \mu_y - \beta\mu_x$.
Poglejmo si še drugi člen enačbe. Vidimo, da je

$$\begin{aligned}\mathrm{var}(Y - \hat{Y}) &= \mathrm{var}(Y - \alpha - \beta X) = \mathrm{var}(Y - \beta X) = \mathrm{var}(Y) - 2\beta\mathrm{cov}(X, Y) + \beta^2\mathrm{var}(X) \\ &= \sigma_y^2 - 2\beta\sigma_{x,y} + \beta^2\sigma_x^2\end{aligned}$$

funkcija spremenljivke β. Za izračun minimuma enačbo najprej odvajamo po β in enačimo z 0.

$$\frac{\partial}{\partial\beta}(\mathrm{var}(Y-\hat{Y})) = -2\sigma_{x,y} + 2\beta\sigma_x^2 = 0$$

To bo res, ko bo $\beta = \frac{\sigma_{x,y}}{\sigma_x^2}$.

Vrednosti α in β, pri katerih bo izraz $\mathrm{E}\left[\left(Y-\hat{Y}\right)\right)^2\Big]$ minimalen, sta

$$\alpha = \mu_y - \mu_x\frac{\sigma_{x,y}}{\sigma_x^2} \qquad \text{in} \qquad \beta = \frac{\sigma_{x,y}}{\sigma_x^2}$$

PRIMER B

Pokažemo, da se pri tako izbranih koeficientih *determinacijski koeficient* (kvadrat korelacijskega koeficienta) izraža v obliki

$$r_{x,y}^2 = 1 - \frac{\mathrm{var}(Y-\hat{Y})}{\mathrm{var}(Y)}.$$

Spomnimo se najprej formule za *korelacijski koeficient*:

$$\mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}} = r_{x,y}.$$

Ker je determinacijski koeficient enak kvadratu korelacije, je enak

$$r_{x,y}^2 = \frac{\mathrm{cov}(X,Y)^2}{\mathrm{var}(X)\mathrm{var}(Y)}.$$

Želimo torej pokazati

$$1 - \frac{\mathrm{var}(Y-\hat{Y})}{\mathrm{var}(Y)} = \frac{\mathrm{cov}(X,Y)^2}{\mathrm{var}(X)\mathrm{var}(Y)}.$$

Če upoštevamo vrednosti α in β iz prvega dela naloge, lahko zapišemo

$$\begin{aligned} 1 - \frac{\mathrm{var}(Y-\hat{Y})}{\mathrm{var}(Y)} = \frac{\mathrm{var}(Y) - \mathrm{var}(Y-\hat{Y})}{\mathrm{var}(Y)} &= \frac{\sigma_y^2 - \sigma_y^2 + \frac{\sigma_{x,y}^2}{\sigma_x^2}}{\sigma_y^2} \\ = \frac{\sigma_{x,y}^2}{\sigma_x^2\sigma_y^2} = \frac{\mathrm{cov}(X,Y)^2}{\mathrm{var}(X)\mathrm{var}(Y)} &= r_{x,y}^2. \end{aligned}$$

Pokazali smo, da je determinacijski koeficient res enak $1 - \frac{\mathrm{var}(Y-\hat{Y})}{\mathrm{var}(Y)}$.