Korona zapiski

STATISTIKA

M. Perman

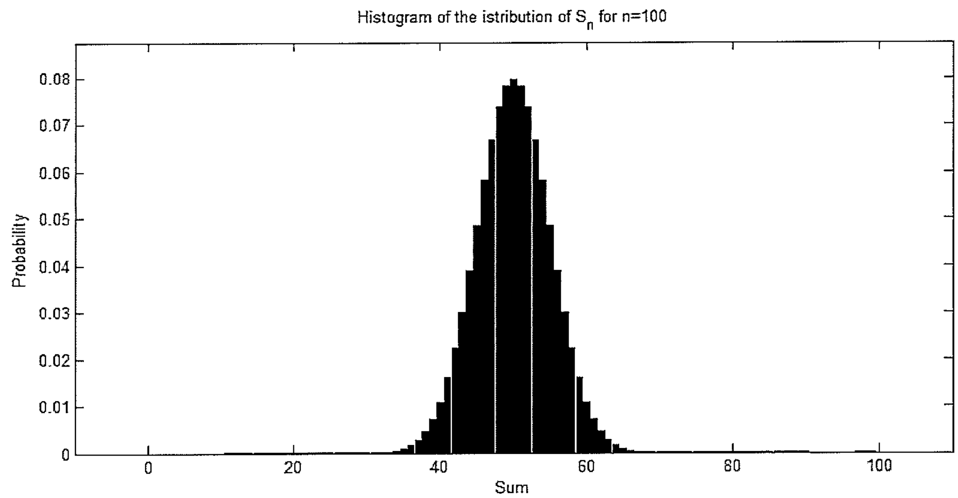
Poletni semester 2020

1.3. Centralni limitni izrek

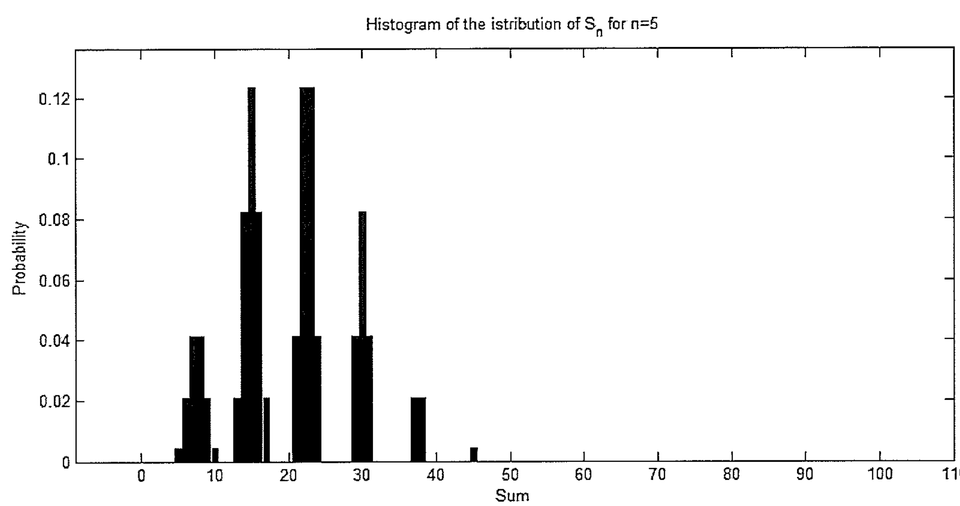
Pogosto nas zanimajo porazdelitve vsot slučajnih spremenljivk. Analitično se da porazdelitve izračunati le v nekaj primerih. Omejili se bomo na primere, ko bodo $X_1, X_2, \ldots$ neodvisne, enako porazdeljene, vsota pa bo $S_n = X_1 + X_2 + \cdots + X_n$. Oglejmo si nekaj primerov histogramov vsot S_n za ~~cele~~ celoštevilske $X_1, X_2, \ldots$

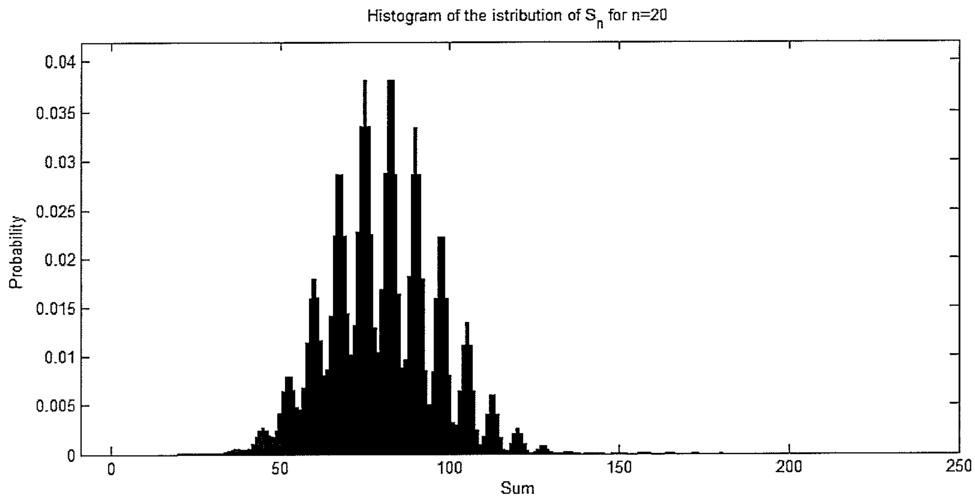
We will look at a few examples of distributions of S_n for different distributions of X_1 and different n .

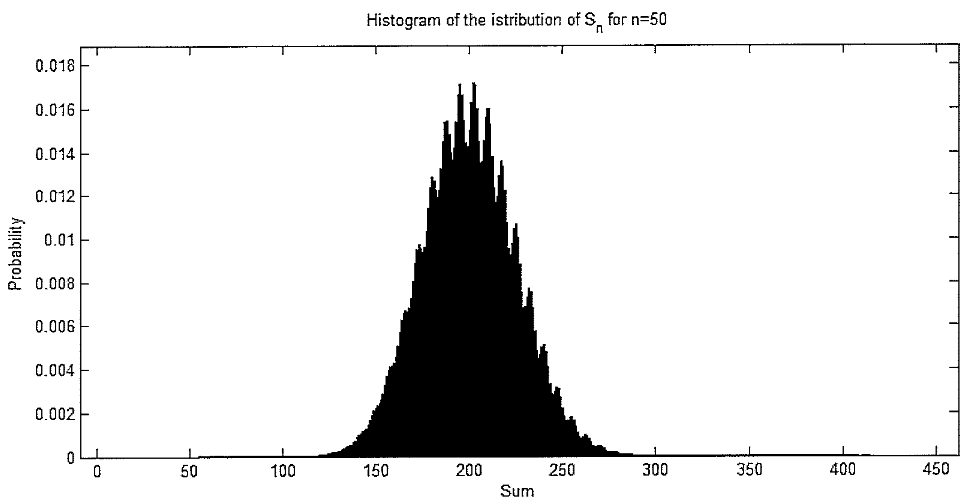
1. Let $P(X_1 = 0) = P(X_1 = 1) = \frac{1}{2}$. Take $n = 100$. Let $S_n = X_1 + \cdots + X_n$. The histogram of the distribution of S_n is:

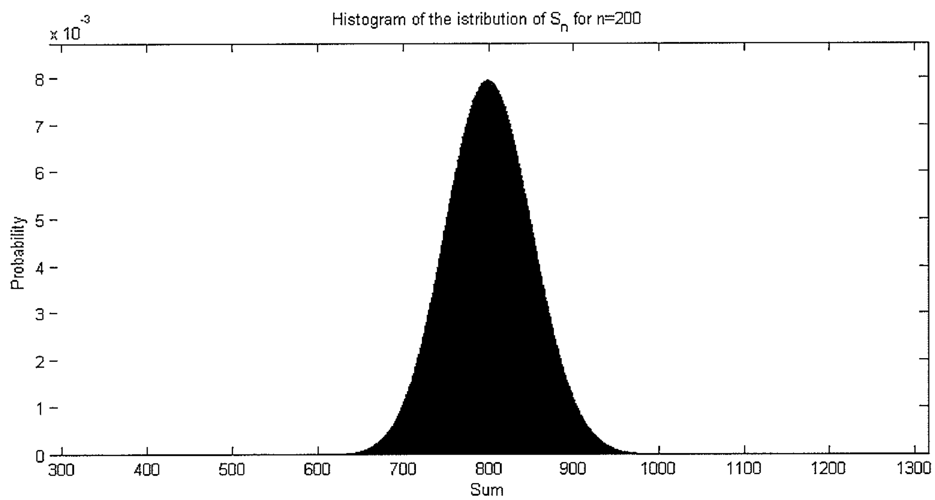


2. Let $P(X_1 = 1) = P(X_1 = 2) = P(X_1 = 9) = 1/3$. Let $n = 5{,}20{,}50{,}200$.

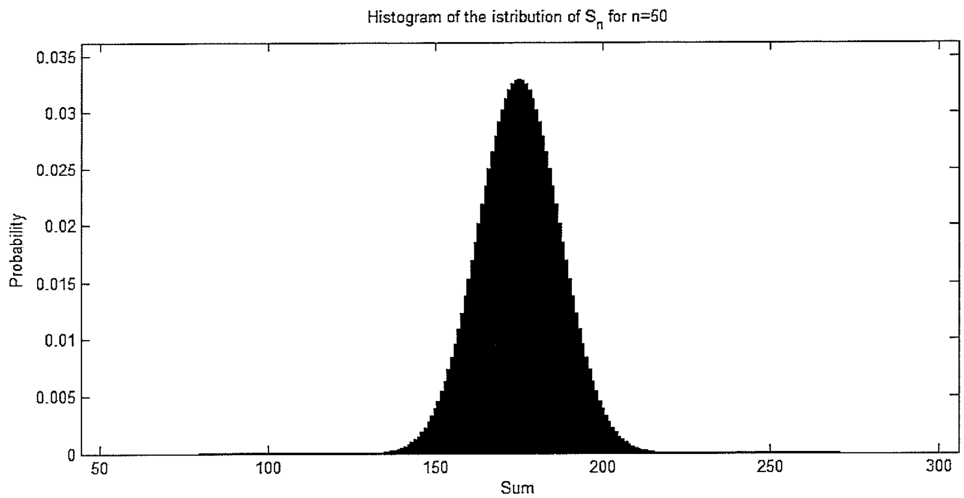



Histogram of the istribution of S_n for n=20
Probability
Sum
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0
50
100
150
200
250

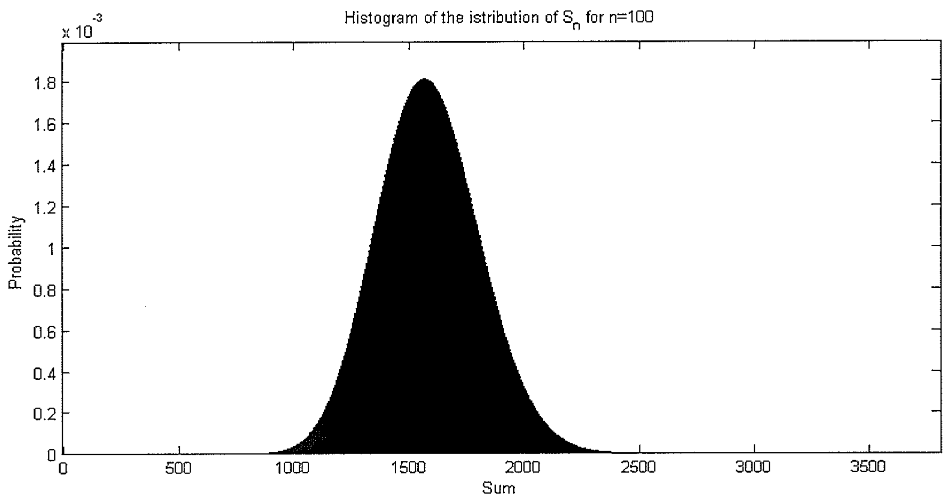

Histogram of the istribution of S_n for n=50
Probability
Sum
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014
0.016
0.018
0
50
100
150
200
250
300
350
400
450


Histogram of the istribution of S_n for n=200
x 10^{-3}
Probability
Sum
0
1
2
3
4
5
6
7
8
300
400
500
600
700
800
900
1000
1100
1200
1300

3. Take $P(X_1 = k) = {}^1\!/_6$ for $k = 1,2, \dots, 6$. Take $n = 50$.



4. Take $P(X_1 = 2^k) = {}^1\!/_7$ for $k = 0,1,23,4,5,6$. Take $n = 100$.



~~1.3 Centralni limitni izrek~~

Centralni limitni izrek je eden od ključnih izrekov za statistiko. Kot smo večkrat opazili, se vsote neodvisnih, enako porazdeljenih slučajnih spremenljivk naravno pojavijo v številnih situacijah. Formulacija centralnega limitnega izreka je naslednja:

Izrek 1.10: Naj bodo $X_1, X_2, \dots$ med sabo neodvisne, enako porazdeljene slučajne spremenljivke. Predpostavimo, da $E(|X_1|) < \infty$ in $\mathrm{var}(X_1) < \infty$. Naj bo $S_n = X_1 + X_2 + \cdots + X_n$. Za poljuben $x \in \mathbb{R}$ velja

$$\lim_{n\to\infty} P\left(\frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \le x\right) = \Phi(x)\,,$$

kjer je Φ porazdelitvena funkcija standardizirano normalne porazdelitve.

Preden se lotimo formalnega dokaza, dodajmo še nekaj komentarjev.

1. Odštevanje $E(S_n)$ in deljenje z $\sqrt{\mathrm{var}(S_n)}$ ima za posledico, da ima kvocient, ki nastopa v formulaciji izreka, pričakovano vrednost 0 in varianco 1. Temu se pogosto reče, da smo vsoto S_n standardizirali.

2. Iz Izreka 1.10 sledi, da za $\alpha < \beta$ velja

$$\lim_{n\to\infty} P\left(\alpha \le \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \le \beta\right) = \Phi(\beta) - \Phi(\alpha)\,.$$

3. Izrek 1.10 v vsej splošnosti dokažemo s pomočjo karakterističnih funkcij, ki so za potrebe verjetnosti prilagojene verzije Fourierove transformacije. Elementarni dokaz, ki ga bomo navedli, zahteva dodatno predpostavko, da je $E(|X_1 - E(X_1)|^3) < \infty$. Ta predpostavka bo v vseh primerih praktične uporabe izpolnjena.

Dokaz centralnega limitnega izreka temelji na Lindeberg-Bergströmovi neenačbi.

Izrek 1.11: Naj bodo $X_1, \dots X_n$ med sabo neodvisne slučajne spremenljivke z $E(X_i) = 0$ in $\mathrm{var}(X_1 + \cdots + X_n) = 1$. Označimo $X := X_1 + \dots X_n$ in naj bo $Z \sim \mathrm{N}(0,1)$. Naj bo f trikrat zvezno odvedljiva funkcija na $\mathbb{R}$ z $|f|, |f'|, |f''|, |f'''| \leq M < \infty$. Velja

$$\left|E(f(X)) - E(f(Z))\right| \leq \frac{1}{6}\left(1 + \sqrt{\frac{8}{\pi}}\right) M\big(E(|X_1|^3) + ... + E(|X_n|^3)\big) .$$

DOKAZ: Naj bodo $Z_1, \ldots Z_n$ porazdeljene normalno z istimi pričakovanimi vrednostmi in variancami kot X_i ter neodvisne med seboj in od $X_1, \ldots, X_n$. Taka izbira slučajnih spremenljivk je vedno možna. Privzamemo lahko, da je $Z = Z_1 + \ldots + Z_n$. Zapišimo

$$E(f(X)) - E(f(Z)) = a_1 + ... + a_n ,$$

kjer je

$$\begin{aligned} a_i = E\left[f(X_1 + \cdots + X_i + Z_{i+1} + ... + Z_n)\right] \\ -E\left[f(X_1 + \cdots + X_{i-1} + Z_i + ... + Z_n)\right] . \end{aligned}$$

V definiciji a_1 je v odštevancu vsota vseh Z_i, za $i = n$ pa je v prvem členu v a_n vsota vseh X_i. Funkcijo f lahko razvijemo po Taylorju. Za vsak $x \in \mathbb{R}$ velja

$$f(x + h) = f(x) + f'(x)h + \tfrac{1}{2}f''(x)h^2 + r$$

kjer je $|r| \leq \frac{1}{6}M|h|^3$. Če zdaj oba člena v a_i razvijemo okoli:

$$Y_i := X_1 + ... + X_{i-1} + Z_{i+1} + ... + Z_n$$

dobimo

$$a_i = E\left[f'(Y_i)(X_i - Z_i)\right] + \tfrac{1}{2}E\left[f''(Y_i)(X_i^2 - Z_i^2)\right] + r_i .$$

kjer je

$$|r_i| \leq \tfrac{1}{6}\, M\big(E(|X_i|^3) + E(|Z_i|^3)\big)$$

Zaradi neodvisnosti X_i, Y_i, Z_i in predpostavk o enakosti pričakovanih vrednosti in varianc, sta pričakovani vrednosti v razvoju za a_i enaki 0, zato je $a_i = r_i$. Z uporabo Jensenove neenakosti za konveksno funkcijo $\phi(x) = x^{3/2}$ na $(0, \infty)$ in elementarnim dejstvom, da je

$$E(|Z_i|^3) = \sqrt{\frac{8}{\pi}}\,\mathrm{var}(Z_i)^{3/2} ,$$

ki ga preverimo z integriranjem, lahko ocenimo

$$E(|Z_i|^3) = \sqrt{\frac{8}{\pi}}\,\mathrm{var}(Z_i)^{3/2} = \sqrt{\frac{8}{\pi}}\,E(|X_i|^2)^{3/2} \leq \sqrt{\frac{8}{\pi}}E(|X_i|^3)\,.$$

Končno lahko ocenimo

$$\big|E(f(X)) - E(f(Z))\big| \leq \frac{1}{6}\left(1 + \sqrt{\frac{8}{\pi}}\right) M\big(E(|X_1|^3) + \ldots + E(|X_n|^3)\big)$$

KOMENTAR: V Izreku 1.11 je dovolj privzeti samo $|f'''| \leq M < \infty$, vendar smo privzeli več, da ni treba utemeljevati obstoja pričakovanih vrednosti, ki nastopajo v dokazu izreka.

V posebnem primeru, ko so vse slučajne spremenljivke enako porazdeljene, nadomestimo X_i z

$$\tilde{X}_i = \frac{X_i - E(X_i)}{\sqrt{\mathrm{var}(S_n)}}\,.$$

Te nove slučajne spremenjivke ustrezajo pogojem Izreka 1.11 in velja

$$\tilde{X}_1 + \cdots + \tilde{X}_n = \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}}\,.$$

Vstavimo in zaradi $\mathrm{var}(S_n) = n\mathrm{var}(X_1)$ sledi

$$\left|E\left(f\left(\frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}}\right)\right) - E(f(Z))\right| \leq \frac{1}{6}\left(1 + \sqrt{\frac{8}{\pi}}\right)\frac{ME\left(|X_1 - E(X_1)|^3\right)}{\sqrt{n}\mathrm{var}(X_1)^{3/2}}\,. \tag{1}$$

V zanimivem primeru, ko je $\mathrm{var}(X_1) > 0$, gre desna stran zgornje neenačbe proti 0, ko $n \to \infty$.

Za dokaz centralnega limitnega izreka potrebujemo še nekaj Analize 1. Označimo z $\chi_{(-\infty,x]}$ karakteristično funkcijo intervala $(-\infty, x]$. Za poljubna $x \in \mathbb{R}$ in $\delta > 0$ obstajata trikrat zvezno odvedljivi funkciji $f_{-\delta}$ in f_δ z vrednostmi na $[0, 1]$, taki da veljajo neenakosti

$$\chi_{(-\infty,x-\delta]} \leq f_{-\delta} \leq \chi_{(-\infty,x]} \leq f_\delta \leq \chi_{(-\infty,x+\delta]}\,.$$

Ker sta $f_{\pm\delta}$ nekonstantni samo na $[x - \delta, x + \delta]$, imata zaradi zvezne odvedljivosti omejene prve tri odvode.

Dokaz izreka 1.10: Izrek bomo, kot rečeno, dokazali pod dodatno predpostavko, da je $E(|X_i - E(X_i)|^3) < \infty$. Označimo

$$\tilde{S}_n = \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}}\,.$$

Fiksirajmo $x \in \mathbb{R}$. Naj bo $\epsilon > 0$. Ker je Φ zvezna, obstaja $\delta > 0$, da iz $|x-y|$ sledi $|\Phi(x) - \Phi(y)| < \epsilon$. Ker je

$$E\left[\chi_{(-\infty,x]}(\tilde{S}_n)\right] = P(\tilde{S}_n \leq x)\,,$$

dobimo neenačbi

$$E\left[f_{-\delta}(\tilde{S}_n)\right] \leq P(\tilde{S}_n \leq x) \leq E\left[f_{\delta}(\tilde{S}_n)\right]\,.$$

Iz posledice (1) sledi, da je

$$\lim_{n\to\infty} E\left[f_{\pm\delta}(\tilde{S}_n)\right] = E\left[f_{\pm\delta}(Z)\right]\,,$$

kjer je $Z \sim \mathrm{N}(0,1)$. Sledi

$$E\left[f_{-\delta}(Z)\right] \leq \liminf_{n\to\infty} P\left(\tilde{S}_n \leq x\right) \leq \limsup_{n\to\infty} P\left(\tilde{S}_n \leq x\right) \leq E\left[f_{\delta}(Z)\right]\,.$$

Iz neenačb za funkciji $f_{\pm\delta}$ sledi še

$$E\left[f_{-\delta}(Z)\right] \geq P(Z \leq x - \delta) \quad \text{in} \quad E\left[f_{\delta}(Z)\right] \leq P(Z \leq x + \delta)\,.$$

Neenačbe združimo v

$$\Phi(x-\delta) \leq \liminf_{n\to\infty} P\left(\tilde{S}_n \leq x\right) \leq \limsup_{n\to\infty} P\left(\tilde{S}_n \leq x\right) \leq \Phi(x+\delta)\,.$$

Limsup in liminf se razlikujeta kvečjemu za 2ϵ. Ker je bil ϵ poljuben, limita obstaja in je enaka $\Phi(x)$.

Tipičen primer uporabe centralnega limitnega izreka je, da aproksimiramo verjetnosti $P(a \leq S_n \leq b)$, kjer je S_n vsota med sabo neodvisnih, enako porazdeljenih slučajnih spremenljivk. Aproksimacijo dobimo kot

$$\begin{aligned} P(a \leq S_n \leq b) &= P\left(\frac{a - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \leq \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \leq \frac{b - E(S_n)}{\sqrt{\mathrm{var}(S_n)}}\right) \\ &\approx \Phi\left(\frac{b - E(S_n)}{\sqrt{\mathrm{var}(S_n)}}\right) - \Phi\left(\frac{a - E(S_n)}{\sqrt{\mathrm{var}(S_n)}}\right)\,. \end{aligned}$$

Za velike n torej aproksimiramo člen zaporedja z limito zaporedja. Če so slučajne spremenljivke $X_1, X_2, \ldots$ celoštevilske, lahko aproksimacijo še nekoliko izboljšamo s *korekcijo za zveznost*, kar pomeni, da a nadomestimo z $a-\frac{1}{2}$ in b z $b+\frac{1}{2}$.

Kot pri vsaki aproksimaciji se postavi vprašanje napake. Na to vprašanje odgovarja izrek, ki ga tukaj navajamo brez dokaza. [1]

Izrek 1.12 (Berry-Esséen) Naj veljajo enake oznake in predpostavke kot v izreku 1.10. Označimo $\rho = E(|X_1 - E(X_1)|^3)$. Velja neenačba

$$\sup_{x\in\mathbb{R}} \left| P\left(\frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \leq x\right) - \Phi(x)\right| \leq \frac{C\rho}{\sqrt{n}\mathrm{var}(X_1)^{3/2}},$$

kjer je C univerzalna konstanta.

Najboljša do zdaj znana ocena za univerzalno konstanto je $C < 0,4748$. [2]

Primer: Vzemimo neodvisne $X_1, X_2, \ldots$ s $P(X_i = 0) = P(X_i = 1) = \frac{1}{2}$. Vemo, da je $S_n \sim \mathrm{Bin}(n, \frac{1}{2})$. Velja, da je $E(S_n) = n/2$ in $\mathrm{var}(S_n) = n/4$. Aproksimacija z normalno porazdelitvijo nam da za $n = 10.000$, da je

$$P(4950 \leq S_n \leq 5050) = P(-1 \leq \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \leq 1) \approx \Phi(1) - \Phi(-1).$$

Za funkcijo Φ ni analitičnega izraza, zato uporabimo ali tabele ali pa statistične programe, ki imajo to funkcijo vgrajeno. Program R nam da $\Phi(1) - \Phi(-1) = 0,6827$. Program R ima vgrajeno tudi binomsko porazdelitev, tako da je točna verjetnost $0,6875$. Če uporabimo še korekcijo za zveznost, kar pomeni da spodnji meji odštejemo $\frac{1}{2}$ in zgornji prištejemo $\frac{1}{2}$, je rezultat aproksimacije $0,6875$, ker je točno na štiri decimalke!

Primer: Naj bodo spet $X_1, X_2, \ldots$ neodvisne in tokrat naj bo $P(X_1 = 1) = P(X_1 = 2) = P(X_1 = 9) = \frac{1}{3}$. Vzemimo $n = 300$. Izračunamo, da je

$$E(S_{300}) = 300 \cdot 4 \qquad \text{in} \qquad \mathrm{var}(S_{300}) = 300 \cdot \frac{38}{3}.$$

[1] Dokaz je, recimo, v Y. S. Chow, H. Teicher, Probability Theory: Independence, Interchangeability, Martingales, 3rd Edition, Springer 2003.

[2] Shevtsova, I., On the accuracy of the normal approximation for sums of independent symmetric random variables. (Russian) Dokl. Akad. Nauk 443 (2012), no. 6, 671-676.

Aproksimiramo

$$
\begin{aligned}
& P(1100 \leq S_{300} \leq 1300) \\
= \; & P\left(\frac{1100-1200}{\sqrt{3800}} \leq \frac{S_n - E(S_n)}{\sqrt{\mathrm{var}(S_n)}} \leq \frac{1300-1200}{\sqrt{3800}}\right) \\
\approx \; & \Phi(1,6222) - \Phi(-1,6222) \\
\doteq \; & 0,8952\,.
\end{aligned}
$$

Izračun s hitro Fourierovo transformacijo da rezultat $0,8970$. Če uporabimo še korekcijo za zveznost, je rezultat aproksimacije z normalno porazdelitvijo enak $0,8970$!

Centralni limitni izrek je najbolj znan primer izreka, ki govori o konvergenci v porazdelitvi. Ideja je, da zaporedje slučajnih spremenljivk konvergira v porazdelitvi, če so njihove porazdelitve vedno bolj podobne neki končni porazdelitvi. Kot iztočnico za matematično definicijo vzemimo Izrek 1.11.

DEFINICIJA: Zaporedje slučajnih spremenljivk $Y_1, Y_2, \ldots$ konvergira v porazdelitvi proti slučajni spremenljivki Y, če za vsako omejeno zvezno funkcijo $f\colon \mathbb{R} \to \mathbb{R}$ velja

$$E\left[f(Y_n)\right] \to E\left[f(Y)\right]\,,$$

ko $n \to \infty$. Matematična oznaka je

$$Y_n \xrightarrow{\mathrm{d}} Y\,,$$

ko $n \to \infty$.

V definiciji vzamemo omejene funkcije, da ni težav z obstojem pričakovanih vrednosti. Zvezne funkcije pa izberemo zato, da v porazdelitvi konvergirajo zaporedja konvergentnih konstantnih slučajnih spremenljivk. Da si nekoliko bolje predstavimo, kaj pomeni zgornja definicija, dokažimo naslednji izrek.

IZREK 1.13: Naj bo $(Y_n)_{n\geq 1}$ zaporedje slučajnih spremenljivk. Zaporedje konvergira v porazdelitvi proti slučajni spremenljivki Y, če in samo če za vsak $x \in \mathbb{R}$, v katerem je F_Y zvezna, velja

$$F_{Y_n}(x) \to F_Y(x)\,,$$

ko $n \to \infty$.

DOKAZ: Naj bo F_Y zvezna v točki x. Naj bo $\epsilon > 0$. Izberimo $\delta > 0$, tako da bo za $|x-y| < \delta$ veljalo $|F_Y(x) - F_Y(y)| < \epsilon$. Vzemimo enaki funkciji $f_{\pm\delta}$ kot v Izreku 1.11. Ker sta funkciji zvezni, iz predpostavke sledi

$$E(f_{-\delta}(Y)) \leq \liminf_{n\to\infty} P(Y_n \leq x) \leq \limsup_{n\to\infty} P(Y_n \leq x) \leq E(f_\delta(Y)).$$

Nadalje iz izbire funkcij sledi

$$P(Y \leq x - \delta) \leq \liminf_{n\to\infty} P(Y_n \leq x) \leq \limsup_{n\to\infty} P(Y_n \leq x) \leq P(Y \leq x + \delta).$$

Ker je bil ϵ poljuben, iz zveznosti porazdelitvene funkcije v x sledi $F_{Y_n}(x) \to F_Y(x)$, ko $n \to \infty$.

Obratno predpostavimo, da $F_{Y_n}(x) \to F_Y(x)$, ko $n \to \infty$ za vsako točko zveznosti F_Y. Naj bo $\epsilon > 0$. Ker je F_Y nepadajoča, so nezveznosti lahko samo skoki, ki jih je največ števno mnogo. Naj bo f omejena zvezna funkcija. Označimo supremum te funkcije z M. Obstajata točki zveznosti $a < b$, da je $F_Y(a) < \epsilon$ in $F_Y(b) > 1 - \epsilon$. Zaradi zveznosti na $[a, b]$ lahko najdemo particijo $a = x_0 < x_1 < \cdots < x_n = b$, tako da so vse točke v particiji točke zveznosti F_Y in funkcija f na $[x_k, x_{k+1}]$ variira za manj kot za ϵ. Funkcijo f aproksimirajmo s funkcijo

$$f^\epsilon(x) = \begin{cases} 0 & \text{za } x \leq a \\ f(x_k) & \text{za } x \in (x_{k-1}, x_k] \text{ za } k = 1, 2, \ldots, n. \\ 0 & \text{za } x > b. \end{cases}$$

Velja

$$\left|E(f^\epsilon(Y)) - E(f(Y))\right| \leq \epsilon + 2M\epsilon,$$

kot se hitro prepričamo. Po drugi strani je

$$E\left[f^\epsilon(Y_n)\right] = \sum_{k=1}^{n} f(x_k)\left(F_{Y_n}(x_k) - F_{Y_n}(x_{k-1})\right).$$

Podobna vsota velja tudi za $E\left[f^\epsilon(Y)\right]$. Vsota je končna, zato po predpostavki obstaja dovolj velik N, da bo za $n \geq N$

$$\left|E\left[f^\epsilon(Y_n)\right] - E\left[f^\epsilon(Y)\right]\right| \leq \epsilon.$$

Po potrebi povečamo N, da bo za $n \geq N$ veljalo $F_{Y_n}(a) < \epsilon$ in $F_{Y_n}(b) > 1-\epsilon$. Podobno kot za Y za $n \geq N$ velja

$$\left|E(f^\epsilon(Y_n)) - E(f(Y_n))\right| \leq \epsilon + 2M\epsilon\,,$$

Ko trikotniško združimo nenakosti, bo za $n \geq N$ veljalo

$$\left|E\left[f(Y_n)\right] - E\left[f(Y)\right]\right| \leq 3\epsilon + 4M\epsilon\,.$$

Ker je bil $\epsilon > 0$ poljuben, je izrek dokazan.

S tem izrekom lahko centralni limitni izrek, glede na to, da je Φ povsod zvezna, napišemo kot

$$\frac{S_n - E(S_n)}{\sqrt{\operatorname{var}(S_n)}} \xrightarrow{\mathrm{d}} Z\,,$$

kjer je $Z \sim \mathrm{N}(0,1)$. Definicija konvergence v porazdelitvi pa ima tudi druge prednosti, na samo da lahko elegantno povemo centralni limitni izrek.

Splošnejšo definicijo konvergence v porazdelitvi lahko uporabimo *mutatis mutandis* za vektorje.

Definicija: Zaporedje slučajnih vektorjev $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ dimenzije r konvergira v porazdelitvi proti slučajnemu vektorju $\mathbf{Y}$, če za vsako omejeno zvezno funkcijo $f\colon \mathbb{R}^r \to \mathbb{R}$ velja

$$E\left[f(\mathbf{Y}_n)\right] \to E\left[f(\mathbf{Y})\right]\,,$$

ko $n \to \infty$. Matematična oznaka je

$$\mathbf{Y}_n \xrightarrow{\mathrm{d}} \mathbf{Y}\,,$$

ko $n \to \infty$.

Centralni limitni izrek ima tudi vektorsko varianto, ki jo tukaj samo navajamo. Dokaz z milimi dodatnimi predpostavkami je podoben dokazu v eni dimenziji, le pisanja in ocenjevanja je nekaj več.

Izrek 1.14: Naj bodo $\mathbf{Y}_1, \mathbf{Y}_2, \ldots$ med sabo neodvisni, enako porazdeljeni slučajni vektorji. Predpostavimo, da obstajata $E(\mathbf{Y}_1)$ in $\mathbf{\Sigma} = \operatorname{var}(\mathbf{Y}_1)$. Označimo $\mathbf{S}_n = \mathbf{Y}_1 + \cdots + \mathbf{Y}_n$. Velja

$$\frac{\mathbf{S}_n - E(\mathbf{S}_n)}{\sqrt{n}} \xrightarrow{\mathrm{d}} \mathrm{N}(0, \mathbf{\Sigma})\,,$$

ko $n \to \infty$.

Za namene statistike je pomembna naslednja posledica definicij.

IZREK 1.15: Naj $\mathbf{Y}_n \xrightarrow{\mathrm{d}} \mathbf{Y}$, ko $n \to \infty$, kjer so vsi vektorji r-dimenzionalni. Naj bo $\phi\colon \mathbb{R}^r \to \mathbb{R}$ zvezna funkcija. Potem

$$\phi(\mathbf{Y}_n) \xrightarrow{\mathrm{d}} \phi(\mathbf{Y})\,,$$

ko $n \to \infty$.

DOKAZ: Dokaz sledi iz preproste opazke, da je za vsako omejeno zvezno funkcijo f kompozitum $f \circ \phi$ omejena zvezna funkcija in trditev sledi iz definicije konvergence v porazdelitvi za vektorje.

PRIMER: Multinomska porazdelitev nastane pri metanju kroglic v škatle. Recimo, da je škatel r, meti so med sabo neodvisni, i-to škatlo pa zadenemo z verjetnostjo p_i. Označimo z $N_i(n)$ število kroglic v škatli i po n metih. Definirajmo slučajni vektor $\mathbf{X}^{(k)}$ po komponentah s predpisom

$$\mathbf{X}_i^{(k)} = \begin{cases} 1 & \text{če } k\text{-ta kroglica pade v } i\text{-to škatlo} \\ 0 & \text{sicer.} \end{cases}$$

Zaradi neodvisnosti metov so tudi vektorji $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots$ med seboj neodvisni in enako porazdeljeni. S temi definicijami je

$$\begin{pmatrix} N_1(n) \\ N_2(n) \\ \vdots \\ N_r(n) \end{pmatrix} = \mathbf{X}^{(1)} + \mathbf{X}^{(2)} + \cdots + \mathbf{X}^{(n)}\,.$$

Iz definicij tudi sledi, da je

$$E\left(\mathbf{X}^{(1)}\right) = \mathbf{p}\,,$$

kjer označimo $\mathbf{p} = (p_1, \ldots, p_r)^T$. Ker je $\mathbf{X}_i^{(1)}$ indikatorska slučajna spremenljivka, je

$$\operatorname{var}\left(\mathbf{X}_i^{(1)}\right) = p_i(1 - p_i)\,.$$

Opazimo, da za $i \neq j$ velja $\mathbf{X}_i^{(1)}\mathbf{X}_j^{(1)} = 0$, zato je

$$\operatorname{cov}\left(\mathbf{X}_i^{(1)}, \mathbf{X}_j^{(1)}\right) = -E\left(\mathbf{X}_i^{(1)}\right) E\left(\mathbf{X}_j^{(1)}\right) = -p_i p_j\,.$$

Sledi

$$\mathbf{\Sigma} = \operatorname{var}\left(\mathbf{X}^{(1)}\right) = \operatorname{diag}(p_1, p_2, \ldots, p_r) - \mathbf{p}\mathbf{p}^T .$$

Zapišimo

$$\begin{pmatrix} \frac{N_1(n)-np_1}{\sqrt{np_1}} \\ \frac{N_2(n)-np_1}{\sqrt{np_2}} \\ \vdots \\ \frac{N_r(n)-np_r}{\sqrt{np_r}} \end{pmatrix} = \frac{\mathbf{A}\left(\mathbf{X}^{(1)} + \mathbf{X}^{(2)} + \cdots + \mathbf{X}^{(n)} - n\mathbf{p}\right)}{\sqrt{n}} ,$$

kjer je

$$\mathbf{A} = \operatorname{diag}\left(\frac{1}{\sqrt{p_1}}, \ldots, \frac{1}{\sqrt{p_r}}\right) .$$

Po izrekih 1.14 in 1.15 velja, da

$$\begin{pmatrix} \frac{N_1(n)-np_1}{\sqrt{np_1}} \\ \frac{N_2(n)-np_1}{\sqrt{np_2}} \\ \vdots \\ \frac{N_r(n)-np_r}{\sqrt{np_r}} \end{pmatrix} \xrightarrow{\mathrm{d}} \mathbf{A}\mathbf{X} ,$$

ko $n \to \infty$, kjer je $\mathbf{X} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{\Sigma}\right)$. Porazdelitev vektorja $\mathbf{AX}$ je tudi večrazsežna normalna in sicer $\mathrm{N}\left(\mathbf{0}, \mathbf{A}\mathbf{\Sigma}\mathbf{A}^T\right)$. Z množenjem matrik sledi

$$\mathbf{A}\mathbf{\Sigma}\mathbf{A}^T = \mathbf{I} - \mathbf{q}\mathbf{q}^T ,$$

kjer je

$$\mathbf{q} = \left(\sqrt{p_1}, \ldots, \sqrt{p_r}\right)^T .$$

Norma vektorja $\mathbf{q}$ je enaka 1, zato je $\mathbf{q}\mathbf{q}^T$ idempotentna matrika in s tem tudi $\mathbf{I} - \mathbf{q}\mathbf{q}^T$ idempotentna matrika z rangom $r-1$.

Če je $\mathbf{Z} \sim \mathrm{N}\left(0, \mathbf{I}\right)$ in je $\mathbf{H}$ idempotentna matrika, je $\mathbf{W} = \mathbf{HZ} \sim \mathrm{N}\left(\mathbf{0}, \mathbf{H}\right)$. Sledi, da je

$$W_1^2 + \cdots + W_r^2 = \mathbf{W}^T\mathbf{W} = (\mathbf{HZ})^T\mathbf{HZ} = \mathbf{Z}^T\mathbf{HZ} \sim \chi^2\left(\operatorname{rang}(\mathbf{H})\right) .$$

Sklepamo lahko, da velja

$$\chi^2(n) = \sum_{i=1}^{r} \frac{(N_i(n) - np_i)^2}{np_i} \xrightarrow{\mathrm{d}} U \sim \chi^2(r-1) ,$$

ko $n \to \infty$. Ta limitni rezultat ima široko uporabo v statistiki.

2. UZORČENJE

2.1 Uvodni primeri

~~4.1 Uvodni primeri~~

~~4.1.1~~ Plebiscit 1990

Plebiscit o neodvisnosti Slovenije decembra 1990 je bil prelomni dogodek pri osamosvajanju. Ko je takrat jeseni naraščala napetost, so mnogi nestrpno pričakovali rezultate predplebiscitnih anket. Ena od odmevnejših je bila SJM 90 (Slovensko javno mnenje 90), ki so jo izpeljali na Fakulteti za družbene vede na Univerzi v Ljubljani.

Za kaj pravzaprav gre pri taki anketi? Pred plebiscitom je nemogoče ugotoviti mnenje vsakega volivca, zato se je treba zateči k izbiranju manjšega števila enot iz populacije volivcev. Izbrani skupini v statističnem žargonu pravimo *vzorec*. Anketa SJM 90 je bila zasnovana na izbiri vzorca velikosti 2074 volivcev, od katerih se jih je 1306 nedvoumno izreklo tako za samostojnost kot za odcepitev Slovenije. Natančni rezultati ankete so v spodnji tabeli [1].

		Samostojnost		
		Da	Ne	Drugo
Odcepitev	Da	1306	11	34
	Ne	183	125	63
	Drugo	110	12	230

Tabela 4.1: Tabela razultatov SJM90

Podatki iz vzorca kažejo, da se je velika večina volivcev iz vzorca izrekla za neodvisnost Slovenije. Iz tega bi sklepali, da se bo tudi večina vseh volilcev v Sloveniji odločila za neodvisnost. Ta sklep je sedaj, ko je rezultat plebiscita že zdavnaj znan, očiten. Poskusimo pa se postaviti v čas jeseni 1990. Vzorec je zajel samo neznaten delež celotnega volilnega telesa, komajda nekaj več kot 0,1% vseh upravičencev. Skeptiki bi se gotovo vprašali, ali lahko na podlagi tako neznatnega vzorca zanesljivo sklepamo o

[1] Vir: SJM90

volji celotne populacije volivcev. Lahko si zamislimo, da bi se v vzorcu znašlo ali preveč privržencev neodvisnosti ali preveč nasprotnikov. V takem primeru bi bila napoved izida seveda napačna. Ali nam statistika lahko pomaga, da presodimo zanesljivost napovedi na podlagi vzorčnih podatkov? Odgovor je pritrdilen.

Prvi korak pri razmišljanju o zanesljivosti ocen na podlagi vzorca mora biti opis načina izbire vzorca ali, kot se temu pravi v statistiki, vzorčni načrt. Pri ustrezno izpeljanih anketah je postopek izbire vzorca natanko predpisan. Poglejmo si način vzorčenja pri anketah SJM, ki je bil uporabljen tudi za anketo o plebiscitu.

Okvir vzorčenja pri anketah SJM je centralni register prebivalstva, ki ga vzdržuje Statistični urad RS. Gre preprosto za primerno urejen seznam vseh prebivalcev Slovenije, iz katerega zlahka dobimo tudi seznam vseh volivcev. Ta seznam je za potrebe vzorčenja po geografskem ključu razdeljen na manjše dele po 4200 volivcev. Tem skupinam pravimo *primarne vzorčne enote.* Vsaka od teh manjših skupin po 4200 volivcev je nadalje razdeljena na skupine po 100, spet po takem ključu, da volivci v podskupini živijo v skupnosti, kot je naselje ali vas. Tem manjšim skupinam pravimo *sekundarne vzorčne enote.* Izbira vzorca SJM poteka v treh korakih:

1. Na prvem koraku anketarji naključno izberejo 140 primarnih vzorčnih enot, torej 140 skupin po 4200 volivcev.

2. Na drugem koraku so v vsaki na prvem koraku izbrani primarni vzorčni enoti izbrane tri sekundarne vzorčne enote, torej tri skupine po 100 volivcev.

3. Na tretjem koraku nato anketarji v vsaki izbrani sekundarni vzorčni enoti naključno izberejo 5 volivcev.

Če izračunamo, je na koncu v vzorec izbranih $140 \cdot 3 \cdot 5 = 2100$ volivcev. Nato je seveda treba z izbranimi volivci stopiti v stik in jih povprašati po njihovem mnenju. Izvajalci ankete na terenu imajo napotek, da od vsake izbrane osebe poskusijo dobiti odgovor. Če prvi poskus ni uspešen, anketarji poskušajo znova do največ petkrat, če je to potrebno. Če so vsi poskusi neuspešni, anketarji neznani odgovor obravnavajo kot manjkajoči podatek. Torej, način vzorčenja je vnaprej določen in anketarji se ga morajo držati. Drug pomemben dejavnik pri vzorčenju za SJM je, da je izbira enot na vsakem koraku naključna. Ni vnaprej določeno, katere skupine ali podskupine ali nazadnje volivci bodo izbrani, temveč so vse enote izbrane kot na "loteriji". Gotovo je na tem mestu utemeljeno vprašanje, zakaj je treba uvesti tovrstno naključno izbiro. Razlog je v odpravljanju vsakršne pristranosti pri izbiri vzorca. Nekako tako, kot da bi pred jemanjem vzorca kapljice tekočine iz steklenice le-to dobro pretresli. Vzorec iz dobro pretresene steklenice mnogo bolje pokaže njeno vsebino, kot pa če steklenice ne bi pretresli.

Kot bomo videli v nadaljevanju, način vzorčenja bistveno vpliva na zanesljivost ocen. Na podlagi opisa vzorčenja lahko domnevamo, da so bili rezultati ankete SJM 90 precej zanesljivi, saj je bila "steklenica zelo dobro pretresena". V naslednjem razdelku si bomo ogledali, kako natančneje opišemo, do kolikšne mere lahko verjamemo, da vzorčni rezultati zanesljivo odražajo stanje v celotni populaciji.

4.1.2 Indeks cen življenjskih potrebščin

Statistični urad Republike Slovenije vsak mesec objavi indeks cen življenjskih potrebščin, ki je eden od osnovnih indikatorjev inflacije. Osnovi za izračun tega indeksa sta:

1. Seznam najvažnejših predmetov in storitev gospodinjske porabe in podatki o višini stroškov za posamezne postavke. Osnova za določitev obsega te porabe je anketa o porabi v gospodinjstvih, ki jo izvaja Statistični urad RS.

2. Povprečna porast cen na drobno za posamezne postavke s seznama.

Formula, ki jo Statistični urad RS uporablja za izračun tega indeksa, je

$$I = \frac{\sum_{i=1}^{m} (p_{1i}/p_{0i}) \cdot w_{0i}}{\sum_{i=1}^{m} w_{0i}} \cdot 100\% \,.$$

Pri tem je kvocient p_{1i}/p_{0i} porast cene na drobno za predmet ali storitev i s seznama, utež w_{0i} pa je povprečni delež stroškov, ki je namenjen za dani predmet ali storitev glede na celotne izdatke v gospodinjstvu. Za primer navedimo, da gospodinjstva v Sloveniji za hrano v povprečju namenijo 23,1%[2] celotne mesečne porabe.

[2]Vir: Statistični letopis 1996, str. 234

Zgornji formuli ne bomo posvetili posebne pozornosti. Omenimo le, da so uteži potrebne zato, ker morajo imeti predmeti ali storitve s seznama, za katere gospodinjstva namenijo večji delež izdatkov, pri določanju rasti življenjskih stroškov večjo težo. Podražitev hrane ima na življenjsko raven večji vpliv kot, recimo, zvišanje cen frizerskih storitev.

Če želimo zgornjo formulo uporabiti, moramo priti na dan z dejanskimi številkami. Med drugim moramo za posamezne predmete ali storitve s seznama ugotoviti povprečne deleže porabe v gospodinjstvih. Ker bi bilo težko spremljati porabo v vseh gospodinjstvih, si Statistični urad RS pomaga z vzorčenjem. Gospodinjstva v Sloveniji so popisana in urejena v popisne okoliše. Za potrebe izbire vzorca so popisni okoliši razdeljeni v 6 podskupin glede na lokacijo in tip. Te podskupine v statistiki imenujejo *stratumi*.

Vzorčenje poteka v dveh korakih:

1. Najprej v vsakem stratumu naključno izberemo popisne okoliše.

2. Na drugem koraku v vsakem izbranem popisnem okolišu naključno izberemo 5 gospodinjstev.

V vzorec je zajetih 3270 gospodinjstev. Število popisnih okološev, izbranih na prvem koraku, je tako, da v vsakem stratumu na koncu postopka izbire zajamemo 0,5% gospodinjstev. Statistični urad RS izbranim gospodinjstvom razdeli vprašalnike za vodenje evidence o porabi. Anketarji urada zberejo izpolnjene evidenčne vprašalnike in na podlagi tako zbranih podatkov določijo deleže porabe za posamezne predmete ali storitve. Prej omenjenih 23,1% stroškov za hrano je ena od tako dobljenih ocen.

Seveda se moramo vprašati, do kolikšne mere lahko ocenam na podlagi vzorca zaupamo. Odgovor je odvisen od tega, kako dobro je bil vzorčni načrt premišljen in ali je sámo vzorčenje bilo izvedeno z nadzorom. Element naključnosti je bistvena sestavina vzorčenja, saj nam izkušnje iz preteklosti in teoretična razmišljanja kažejo, da tako najbolje dosežemo nepristranost vzorčnih ocen, poleg tega pa edino strog vzorčni načrt omogoča presojo o tem, kako zanesljiva je ocena. Brez naključne izbire vzorca taka presoja ni možna. Velikost vzorca je izbrana tako, da so dobljeni vzorčni odstotki dovolj zanesljivi za uporabo v formuli za izračun indeksa cen življenjskih potrebščin.

Rezultat zgornje formule za december 1996 je bil 1,3%.

4.4.2 TIMSS v Sloveniji

Naslednji primer, kjer zbiranje podatkov poteka z vzorčenjem, so raziskave uspešnosti šolskih sistemov. Učenci slovenskih osnovnih in srednjih šol so vključeni v mednarodne primerjalne raziskave znanja matematike, naravoslovja, računalništva, bralne pismenosti in drugih področij znanja. V Sloveniji izvaja te primerjalne raziskave Pedagoški inštitut v Ljubljani. Za opis postopka vzorčenja v teh raziskavah si izberimo zadnjo izmed raziskav TIMSS, ki je potekala od leta 1991 in je bila končana leta 1998. To raziskavo smo že srečali v prejšnjih poglavjih.

V jeziku prvega poglavja spadajo v populacijo, ki nas v tem primeru zanima, vsi učenci sedmih in osmih razredov slovenskih šol. V letu 1995, ko je potekalo dejansko vzorčenje, je bila ta populacija velika $N = 54.965$ učencev[3]. Iz praktičnih razlogov je bilo izključenih 310 učencev (večinoma učenci iz šol za slepe in slabovidne ter podobnih), tako da je bila na koncu populacija, iz katere je bil izbran vzorec, velika $N = 54.655$.

Zanima nas znanje matematike in naravoslovja pri učencih iz opisane populacije. Znanje je bilo merjeno z nalogami, ki so bile sestavljene posebej za TIMSS. Če se spet vrnemo k jeziku prvega poglavja, sta spremenljivki raven znanja matematike in raven znanja naravoslovja posameznega učenca. Ta ravni izrazimo kot število na posebnih lestvicah, ki so premišljene tako, da omogočajo smiselno primerjavo med državami, ki so bile vključene v raziskavo. Za nas je v tem trenutku pomembno to, da vsaki enoti v populaciji pripadata neki vrednosti, ki pomenita znanje matematike oziroma naravoslovja.

Preden se lotimo opisa vzorčenja, se moramo seveda vprašati, zakaj je vzorčenje sploh potrebno. Populacija je dokaj velika, zato si ni težko predstavljati, koliko dela bi bilo z izvedbo preizkušanja znanja, ki bi zajela vse učence v populaciji, da o vrtoglavih stroških tako velikega projekta sploh ne govorimo. Take raziskave za celotno populacijo učencev ni mogoče izpeljati iz povsem praktičnih razlogov. Rešitev je v vzorčenju. Iz celotne populacije izberemo manjši vzorec učencev in povprečno raven znanja slovenskih sedmošolcev in osmošolcev ocenimo na podlagi znanja učencev iz tega vzorca. V raziskavi TIMSS je bilo v vzorec zajetih 5927 učencev.

[3] Vir: Pedagoški inštitut Ljubljana

Samo izbiranje vzorca je bilo mednarodno usklajeno in vnaprej predpisano. Vzorčenje ni poteklo z neposrednim izbiranjem učencev, temveč posredno v dveh korakih. Na prvem koraku so sodelavci Pedagoškega inštituta med 455 osnovnimi šolami v Sloveniji naključno izbrali 150 šol, in sicer tako, da so imele večje šole nekaj več verjetnosti, da bodo izbrane v vzorec. Tukaj besedo "naključno" uporabljamo še v nekoliko ohlapnem pomenu, kasneje pa bomo ta pojem tudi natančneje opredelili. Ko so bile šole izbrane, je bil naslednji korak izbiranje razreda. Izbira je bila spet naključna, in sicer taka, da je bil izbran po en 7. in en 8. razred na vsaki šoli, ki je bila izbrana na prvem koraku. V končnem vzorcu so bili zajeti vsi učenci iz izbranih razredov, torej zgoraj omenjenih 5927 učencev. Takemu načinu izbiranja vzorca statistiki pravijo "vzorčenje v skupinicah", ker izbiramo celotne skupinice enot, v tem primeru razrede. Vsi izbrani učenci so reševali izbrane naloge in na podlagi njihovih odgovorov je bila ocenjena povprečna raven znanja matematike oziroma naravoslovja vseh slovenskih sedmošolcev in osmošolcev.

Tudi tukaj si moramo zastaviti vprašanje o zanesljivosti dobljenih ocen. Raziskava je zajela samo okrog 10% vseh učencev iz opisane populacije in na podlagi njihovih rezultatov je bila potem ocenjena raven znanja za celotno populacijo. Do kolikšne mere je to utemeljeno? Kako opisati zanesljivost ocen? Na to vprašanje bomo odgovorili v razdelkih, ki sledijo.

4.1.4 Predsedniške volitve v ZDA leta 1936

Kot zadnji primer vzorčenja si oglejmo znamenito anketo, kjer so šle stvari pri napovedovanju izida volitev zelo hudo narobe. Pred predsedniškimi volitvami v ZDA leta 1936 je prestižna revija *Literary Digest* na podlagi obsežne ankete napovedala zmagovalca. Danes vemo, da je bil na omenjenih volitvah s precejšnjo večino izvoljen Franklin D. Roosevelt, revija pa je takrat napovedala, da bo zmagal njegov republikanski tekmec Alfred Landon. V tabeli 4.2 so vsebovane napovedi revije *Literary Digest* in dejanski rezultati. Jasno je, da anketa ni mogla zajeti vseh volivcev v ZDA leta 1936, zato je bilo treba izbrati vzorec. Vzorec, ki so ga izbrali anketarji revije *Literary Digest*, je bil največji, kar so jih sploh kdaj izbrali, in je vseboval 2,4 milijona oseb. Kljub tako velikemu vzorcu pa se je napoved razlikovala od dejanskega rezultata za skoraj 20%!

	Napoved revije *Literary Digest*	Dejanski rezultat
F. D. Roosevelt	43%	62%
A. Landon	57%	38%

Tabela 4.2: Napovedi in rezultati predsedniških volitev v ZDA leta 1936

Zanimivo je seveda vprašanje, kaj je povzročilo tako zelo zgrešeno napoved. Če želimo dobiti odgovor na to vprašanje, si moramo ogledati, kako je bil vzorec izbran. Izvor za izbiro volivcev v vzorec so bili seznami naročnikov revije *Literary Digest*, telefonski imeniki, seznami članov elitnih klubov in podobno. Vprašalnike so poslali izbranim osebam po pošti, in sicer kar 10 milijonom, odgovorilo pa je samo 2,4 milijona naslovnikov, kar je že razlog za previdnost. Če pomislimo, da je bilo leto 1936 najbolj črno leto velike depresije in je bilo v ZDA 11 milijonov brezposelnih, nam takoj pade v oči, da od le-teh velika večina ni imela telefona in torej njihova imena niso bila v telefonskem imeniku ali na seznamu članov kakšnega elitnega kluba. Če še pomislimo, da republikansko stranko v ZDA po pravilu podpirajo bogatejši sloji, je eden od razlogov za slabe napovedi na dlani. Anketa je bila že vnaprej načrtovana tako, da so bili v vzorec zajeti tisti, ki so tudi med depresijo imeli kaj pod palcem. Velikost vzorca seveda ni pomagala, ker se je samo ponavljala ena in ista napaka. V vzorcu se je vedno znova znašlo več republikanskih volivcev kot pa podpornikov predsednika Roosevelta. V statističnem žargonu bi lahko rekli, da je bil vzorčni načrt pri tej anketi slab, z drugimi besedami, način izbire vzorca je bil slabo premišljen. Še bolj primerna izjava bi bila, da vzorčnega načrta sploh ni bilo. Revija *Literary Digest* se je kmalu po volitvah 1936 znašla v stečaju.

Kot zanimivost lahko povemo še to, da je v istem času mladi statistik George Gallup na podlagi vzorca velikosti 5000 oseb napovedal zmago F. D. Roosevelta s 56% glasov. Še bolj zanimivo je to, da je George Gallup napovedal tudi napoved revije *Literary Digest*, še preden jo je ta objavila. Izbral je vzorec velikosti 3000 izmed tistih, ki so prejeli vprašalnike, in na podlagi izbranega vzorca napovedal, da bo napoved 44% za Roosevelta. Res ne bi bilo treba izbirati vzorca velikosti 2,4 milijona!

> Pogosto je nemogoče zbrati podatke o vrednostih spremenljivke za vse enote v populaciji. Zato iz populacije izberemo vzorec, ki zajema le njen manjši del. Iz vrednosti spremenljivke za enote v vzorcu potem ocenimo želene količine, na primer povprečno vrednost spremenljivke ali odstotek enot z določeno lastnostjo, za celotno populacijo. Če nas zanima povprečna vrednost spremenljivke za celotno populacijo, kot oceno za to količino vzamemo povprečno vrednost spremenljivke za enote v vzorcu. Če nas zanima odstotek enot v populaciji z neko lastnostjo, za oceno tega odstotka vzamemo odstotek enot v vzorcu, ki imajo izbrano lastnost.

> Vzorčni načrt je vnaprej predpisan postopek izbiranja vzorca iz vnaprej določene in natančno opredeljene populacije. Če je izbira enot ali skupin enot naključna, potem takemu vzorčenju pravimo *verjetnostno vzorčenje*. Z vpeljavo naključnosti se najbolje izognemo pristranosti.

2.1. ~~4.2~~ Enostavno slučajno vzorčenje

~~4.2.1~~ Pojem enostavnega slučajnega vzorca

Iz primerov v prejšnjem razdelku, posebej iz zadnjega, je razvidno, da je način izbire vzorca zelo pomemben. V tem razdelku si bomo ogledali najpreprostejši vzorčni načrt: *enostavno slučajno vzorčenje*. Čeprav se ta tip vzorčenja v dejanskih anketah redko uporablja, je sestavni del mnogih, tudi bolj zapletenih vzorčnih načrtov, poleg tega pa je pri njem najbolj razviden pojem standardne napake. Za okvir razmišljanja vzemimo populacijo velikosti N, iz katere želimo izbrati vzorec velikosti n. Zamislimo

si, da imamo za vsako od N enot v populaciji listek, ki ga damo v škatlo.

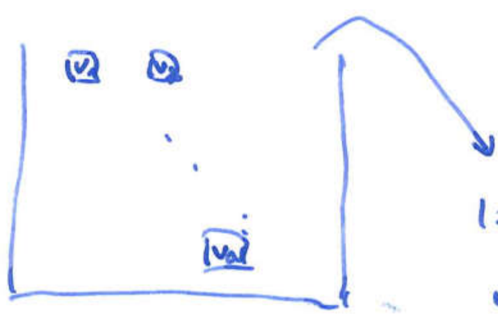
Predno izberemo enostavni slučajni vzorec, si lahko predstavljamo, da imamo v škatli listke, na katerih je zapisana vrednost. Če je listkov N, označimo vrednosti na njih z $v_1, v_2, \ldots, v_N$. Definirajmo populacijsko povprečje z

$$\mu = \sum_{k=1}^{N} v_k$$

in populacijsko varianco z

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{N} (v_k - \mu)^2 .$$

Slika:



Izberemo naključno n enot. Populacijsko povprečje ocenimo z vzorčnim povprečjem.

Glavna ideja: ko izberemo vzorec in izračunamo vzorčno povprečje, smo kreirali slučajno spremenljivko $\bar{X}$.

Definirajmo

$$I_k = \begin{cases} 1, & \text{če je } k\text{-ta enota izbrana v vzorec.} \\ 0, & \text{sicer.} \end{cases}$$

Velja $I_1 + \cdots + I_N = n$. Zapišemo lahko

$$\bar{X} = \frac{1}{n} \sum_{k=1}^{N} v_k \cdot I_k$$

Vemo

$$\begin{aligned} E(\bar{X}) &= \frac{1}{n} \sum_{k=1}^{N} v_k \cdot E(I_k) \\ &= \frac{1}{n} \sum_{k=1}^{N} v_k \, P(I_k = 1) \\ &= \frac{1}{n} \sum_{k=1}^{N} v_k \cdot \frac{n}{N} \\ &= \frac{1}{N} \sum_{k=1}^{\infty} v_k \\ &= \mu \end{aligned}$$

Zaradi simetrije je $\operatorname{cov}(I_k, I_l)$ enaka za vse $k \neq l$. Računamo

$$
\begin{aligned}
\operatorname{cov}(I_k, I_l) &= P(I_k = 1, I_l = 1) \\
&\quad - P(I_k = 1)\,P(I_l = 1) \\
&= \frac{\binom{N-2}{n-2}}{\binom{N}{n}} - \frac{n}{N} \cdot \frac{n}{N} \\
&= \frac{n(n-1)}{N(N-1)} - \frac{n}{N}^2 \\
&= \frac{n}{N}\left[\frac{n-1}{N-1} - \frac{n}{N}\right] \\
&= \frac{n}{N}\left[\frac{(n-1)N - n(N-1)}{N(N-1)}\right] \\
&= \frac{n}{N}\left[\frac{-N+n}{N(N-1)}\right] \\
&= -\frac{n}{N} \cdot \frac{(N-n)}{N(N-1)}
\end{aligned}
$$

Velja

$$
\operatorname{var}(I_k) = \frac{n}{N} \cdot \left(1 - \frac{n}{N}\right) = \frac{n}{N} \cdot \frac{(N-n)}{N}
$$

Računamo

$$\mathrm{var}\left(\overline{X}\right)$$

$$= \frac{1}{n^2}\Big[\sum_{k=1}^{N} v_k^2 \,\mathrm{var}(I_k) + \sum_{\substack{k,l=1\\k\neq l}}^{n} v_k v_l \,\mathrm{cov}(I_k, I_l) \Big]$$

$$= \frac{1}{n^2}\Big[\frac{n}{N}\cdot\frac{N-n}{N} \sum_{k=1}^{N} v_k^2 - \frac{n}{N}\cdot\frac{(N-n)}{N(N-1)} \sum_{\substack{k,l=1\\k\neq l}}^{N} v_k v_l \Big]$$

$$= \frac{1}{n}\cdot\frac{N-n}{N^2}\cdot\Big[\sum_{k=1}^{N} v_k^2 - \frac{1}{N-1} \sum_{\substack{k,l=1\\k\neq l}}^{N} v_k v_l \Big]$$

$$= \frac{1}{n}\cdot\frac{N-n}{N^2}\Big[\sum_{k=1}^{N} v_k^2 \Big(\underbrace{1 + \frac{1}{N-1}} - \frac{1}{N-1}\Big) - \frac{1}{N-1} \sum_{\substack{k,l=1\\k\neq l}}^{N} v_k v_l \Big]$$

$$= \frac{1}{n}\,\frac{N-n}{N^2}\Big[\frac{N}{N-1} \sum_{k=1}^{n} v_k^2 - \frac{1}{N-1} \sum_{k=1}^{n} v_k^2 - \frac{1}{N-1} \sum_{\substack{k,l=1\\k\neq l}}^{N} v_k v_l \Big]$$

Zaradi simetrije je $E(X_k) = \mu$ in $var(X_k) = \sigma^2$ za vse $k = 1, \dots, n$.

Potrebujemo še $cov(X_k, X_l)$.

Mislimo si lahko, da enote izbiramo do zadnje. Potem je $X_1 + X_2 + \dots + X_N = const.$ in zato

$$cov(X_1, X_1 + X_2 + \dots + X_N) = 0$$

$$\parallel$$

$$cov(X_1, X_1) + (N-1)\, cov(X_1, X_2) \quad \Rightarrow$$

$$cov(X_k, X_l) = -\frac{\sigma^2}{N-1}$$

Sledi

$$var(\bar{X}) = \frac{1}{n^2}\left[n\, var(X_1) + n(n-1)\, cov(X_1, X_2)\right]$$

$$= \frac{1}{n}\left[\sigma^2 + (n-1)\left(-\frac{\sigma^2}{N-1}\right)\right]$$

$$= \frac{\sigma^2}{n}\left[1 + \left(-\frac{n-1}{N-1}\right)\right]$$

$$= \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

$$= \frac{1}{n}\cdot\frac{N-n}{N^2}\left[\frac{N}{N-1}\sum_{k=1}^{n} v_k^2 - \frac{1}{N-1}\left(\sum_{k=1}^{n} v_k\right)^2\right]$$

$$= \frac{1}{n}\cdot\frac{N-n}{N-1}\left[\frac{1}{N}\sum_{k=1}^{n} v_k^2 - \mu^2\right]$$

$$= \frac{1}{n}\cdot\frac{N-n}{N-1}\cdot\sigma^2,$$

kjer je $\sigma^2 = \frac{1}{N}\sum_{k=1}^{N} v_k^2 - \mu^2$.

Izračunajmo še nekoliko drugače. Lahko si predstavljamo, da izbiramo enote naključno eno po eno. Vrednosti na izbranih listkih so slučajne spremenljivke $X_1, X_2, \dots, X_n$ in

$$\overline{X} = \frac{1}{n}\sum_{k=1}^{N} X_k$$

Sl. spr. X_1 ima vrednosti $v_1, \dots, v_N$ z verjetnostmi $\frac{1}{N}$, kar pomeni

$$E(X_1) = \mu \quad \text{in}$$

$$\mathrm{var}(X_1) = \sigma^2 .$$

Opomba: $\overline{X}$ smo lahko napisali na več načinov kot linearno kombinacijo enostavnejših slučajnih spremenljivk.

Najpomembnejša opazka pri tem drugem zapisu je, da je $\overline{X}$ vsota sl. spr. deljena z n. Po centralnem limitnem izreku bo ta vsota približno normalno porazdeljena. Torej porazdelitev $\overline{X}$ lahko aproksimiramo z normalno porazdelitvijo!

Definicija: Porazdelitvi $\overline{X}$ rečemo vzorčna porazdelitev.

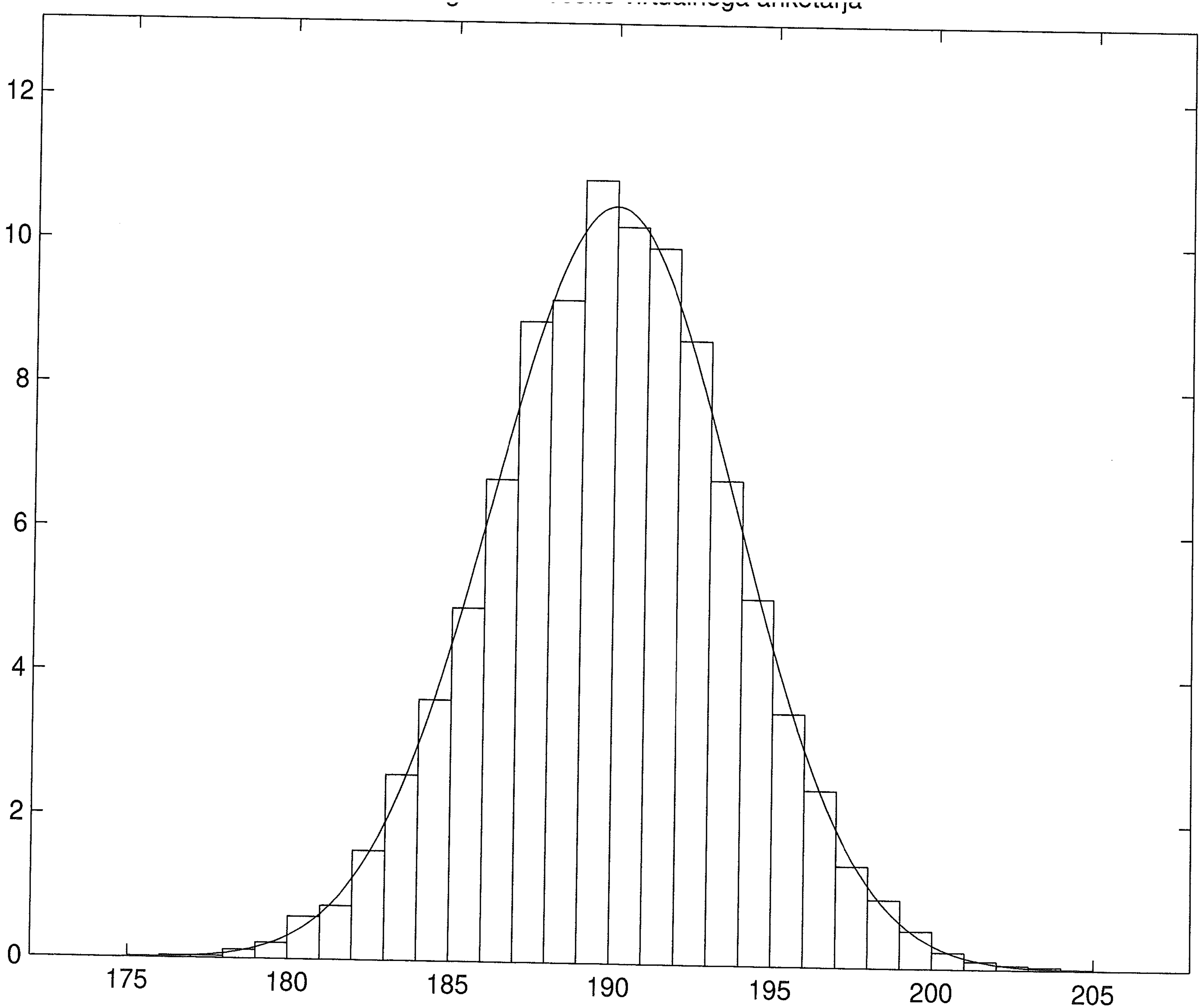
Definicija: Slučajni spremenljivki $\overline{X}$ rečemo cenilka.

Zadržek: Centralni limitni izrek velja za neodvisne $X_1, X_2, \ldots$. V vsoti $\frac{1}{n}\sum_{k=1}^{n} X_k$ so slučajne spremenljivke korelirane! Možna sta dva odgovora:

(1) Odvisnost je šibka, dokler je vzorec majhen delež populacije. CLI še vedno dobro aproksimira porazdelitev vsote.

(2) CLI velja tudi v primerih, ko so slučajne spremenljivke odvisne, vendar so formulacije in dokazi bolj komplicirani.

SKLEP: Vzorčna porazdelitev $\bar{X}$ je približno normalna s parametroma μ in $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$. Na podlagi tega lahko dajemo izjave o zanesljivosti ocen.


12
10
8
6
4
2
0
175
180
185
190
195
200
205

Definicija: Cenilka je nepristransa, če je njena pričakovana vrednost enaka parametru, ki ga ocenjujemo.

Pri vzorčenju ne poznamo σ^2. Kako bi ta parameter ocenili? V rokah imamo samo vzorec in rečemo

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} (X_k - \overline{X})^2 .$$

V statistiki s ^ označimo cenilke. Računamo

$$E(\hat{\sigma}^2) = \frac{1}{n} \sum_{k=1}^{n} E[(X_k - \overline{X})^2]$$

$$= E[(X_1 - \overline{X})^2] \quad (\text{simetrija})$$

$$= \mathrm{var}(X_1 - \overline{X})$$

$$= \mathrm{var}(X_1) + \mathrm{var}(\overline{X}) - 2\,\mathrm{cov}(X_1, \frac{1}{n} \sum_{k=1}^{n} X_k)$$

$$= \sigma^2 + \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$$

$$- 2 \cdot \frac{1}{n} \left(\sigma^2 - \frac{\sigma^2 (n-1)}{N-1} \right)$$

$$= \sigma^2 \left[1 + \frac{N-n}{n(N-1)} - \frac{2}{n} \cdot \frac{N-n}{N-1} \right]$$

$$= \sigma^2 \left[1 - \frac{N-n}{n(N-1)} \right]$$

$$= \sigma^2 \frac{n(N-1) - N + n}{n(N-1)}$$

$$= \sigma^2 \cdot \frac{nN - N}{n(N-1)}$$

$$= \sigma^2 \cdot \frac{N(n-1)}{n(N-1)}$$

Cenilka $\hat{\sigma}^2$ ni nepristranska. Vendar jo lahko popravimo v

$$\boxed{\tilde{\sigma}^2 = \frac{N-1}{N(n-1)} \sum_{k=1}^{n} (X_k - \overline{X})^2 ,}$$ ki

je nepristranska cenilka σ^2.

Še nekaj terminologije:

Ko govorimo o cenilkah, je to načeloma pred izbiranjem vzorca. Takrat še ne vemo njihovih vrednosti. Ko je vzorec izbran, označimo konkretne vrednosti z $x_1, x_2, \ldots x_n$, povprečje z $\bar{x}$ in govorimo o <u>ocenah</u>.

Za presojo kvalitete ocen pa je merodajna vzorčna porazdelitev.

<u>Definicija</u>: Količini $\sqrt{var(\bar{x})}$ rečemo standardna napaka cenilke $\bar{x}$ in označimo z $se(\bar{x})$.

Glede natančnosti cenilke potem lahko rečemo: z verjetnostjo ~ 0.68 se zmotimo za manj kot $se(\bar{x})$, z verjetnostjo 0.95 za manj kot $1{,}96 \times se(\bar{x})$, z verjetnostjo ~ 0.99 za manj kot $2.56 \times se(\bar{x})$

Ker je $se(\bar{x}) = \frac{\sigma}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$, moramo tudi σ aposteriori nadomestiti z $\hat{\sigma}$.

<u>Definicija</u>: Faktorju $\sqrt{\frac{N-n}{N-1}}$ rečemo popravni faktor za končnost.

<u>Opomba</u>. Če bi vzorčili ~~brez~~ z vračanjem, popravnega faktorja ne bi bilo.

2.4. Intervali zaupanja

Intervali zaupanja so grafična metoda za prikaz točnosti vzorčnih ocen.
Ideja je preprosta: če ocenjujemo neko količino, za oceno dobimo število. To število lahko „razpihnemo" v interval

Če je z_α tako število, da je za

$z \sim N(0,1)$ $\quad P(-z_\alpha \le z \le z_\alpha) = \alpha$,

potem interval $\bar{x} \mp z_\alpha \times se(\bar{x})$

pokriva pravo vrednost parametra μ, če in samo če se ocena $\bar{x}$ od μ razlikuje za manj kot $z_\alpha \times se(\bar{x})$, iz razprave o vzorčni porazdelitvi pa vemo, da je ta verjetnost aproksimativno α. Izjava še vedno velja, če nadomestimo σ v izrazu za $se(\bar{x})$ z oceno $\hat{\sigma}$.

<u>Opomba</u>: O verjetnostih lahko spet govorimo <u>pred</u> izbiro vzorca. Če rečemo, da bomo izbrali vzorec, ocenili μ z $\bar{x}$, ocenili standardno napako z $\frac{\hat{\sigma}}{\sqrt{n}} \cdot \sqrt{\frac{N-n}{N-1}}$, potem <u>bo</u> <u>interval</u> <u>zaupanja</u>

$$\bar{x} \mp z_\alpha \times se(\bar{x})$$

pokril μ z verjetnostjo približno α.

Po izbiri vzorca o verjetnosti ne govorimo več in so intervali zaupanja le grafični način prikaza točnosti vzorčnih ocen.

2.5 Primeri drugih vzorčnih načrtov

Kot smo videli, v praksi enostavno slučajno vzorčenje ni v uporabi. Večinoma se uporablja večfazno vzorčenje kot v slovenskem javnem mnenju. Ogledali si bomo dva od možnih vzorčnih načrtov in izpeljali ustrezne cenilke. V realnih primerih so lahko cenilke dokaj zapletene.

Stratificirano vzorčenje

Ideja je, da populacijo razdelimo na podskupine, ki jim rečemo stratumi. Recimo, da je populacija velikosti N in jo razdelimo v k podmnožic velikosti $N_1, N_2, \dots, N_k$, tako da je $N = N_1 + \dots + N_k$. Iz vsake podmnožice = stratuma izberemo enostavni slučajni vzorec velikosti $n_1, n_2, \dots, n_k$. Predpostavljam da so postopki izbire vzorcev v stratumih neodvisni.

Oznake: μ = populacijsko povprečje

σ^2 = populacijska varianca

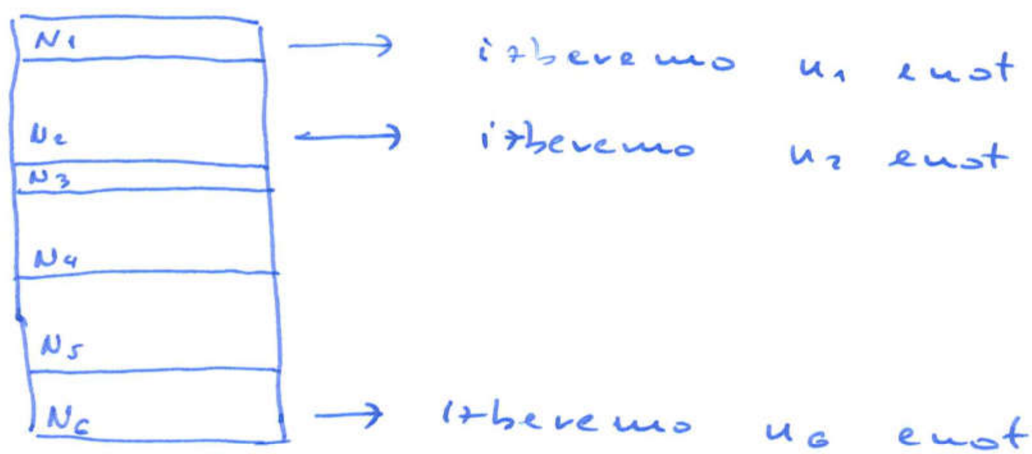
μ_k = populacijsko povprečje v k-tem stratumu

σ_k^2 = populacijska varianca v k-tem stratumu

$w_k = \frac{N_k}{N}$ = delež k-tega stratuma v celotni populaciji.

Zakaj bi stratificirali? Če je znotraj stratumov varianca občutno manjša kot je σ^2, bodo lahko končne ocene bolj točne.

Slika:



Še nekaj oznak:

$\overline{X}_k$ = vzorčno povprečje za enostavni slučajni vzorec v k-tem stratumu.

$\hat{\sigma}_k^2$ = nepristranska cenilka σ_k^2 na podlagi enostavnega slučajnega vzorca

Kako bi ocenili μ? Opazimo, da je

$$\mu = \sum_{k=1}^{K} \frac{N_k}{N} \cdot \mu_k = \sum_{k=1}^{K} w_k \cdot \mu_k$$

Razumno je reči, da bo cenilka

$$\boxed{\overline{X} = \sum_{k=1}^{K} w_k \cdot \overline{X_k}}$$

cenilka μ za stratificirano vzorčenje

Zaradi linearnosti je

$$E(\overline{X}) = \sum_{k=1}^{k} w_k \cdot E(X_k)$$

$$= \sum_{k=1}^{k} w_k \cdot \mu_k$$

$$= \mu$$

Zgornja cenilka za μ je torej nepristranska. Iz predpostavk sledi, da so $\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_K$ neodvisne.

Zaradi neodvisnosti je

$$\boxed{\begin{aligned} \mathrm{var}(\bar{X}) &= \sum_{k=1}^{K} w_k^2 \cdot \mathrm{var}(\bar{X}_k) \\ &= \sum_{k=1}^{K} w_k^2 \cdot \frac{\sigma_k^2}{n_k} \cdot \frac{N_k - n_k}{N_k - 1} \end{aligned}}$$

Ker lahko σ_k^2 ocenimo nepristransko z $\hat{\sigma}_k^2$, je

$$\sum_{k=1}^{K} w_k^2 \cdot \frac{\hat{\sigma}_k^2}{n_k} \cdot \frac{N_k - n_k}{N_k - 1}$$

nepristranska cenilka $\mathrm{var}(\bar{X})$. Standardni napako ocenimo z

$$\hat{se}(\bar{X}) = \left(\sum_{k=1}^{K} w_k^2 \frac{\hat{\sigma}_k^2}{n_k} \cdot \frac{N_k - n_k}{N_k - 1} \right)^{1/2}.$$

Kako pa je z normalnostjo vzorčne porazdelitve? $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_K$ so približno normalno porazdeljene.

Zato lahko pričakujemo, da bodo tudi njihove linearne kombinacije aproksimativno normalni

Opomba: Sledeče dejstvo se da dokazati, vendar je dokaz dokaj zahteven.

Pri stratificiranem vzorčenju se pojavi še naslednje vprašanje. Recimo, da lahko izberemo vzorec velikosti n. Če imamo stratume velikosti $N_1, \ldots, N_k$, se moramo odločiti, kako velike vzorce bomo izbirali iz posameznih stratumov. Kako izbrati, da bo standardna napaka cenilke $\bar{X}$ čim manjša? Najti moramo $n_1, \ldots, n_k$, da bo $n = n_1 + n_2 + \ldots + n_k$ in bo

$$\operatorname{var}(\bar{X}) = \sum_{k=1}^{k} \frac{\sigma_k^2}{n_k} \cdot \frac{N_k - n_k}{N_k - 1} \cdot w_k^2$$

čim manjša?

V večini praktičnih situacij so korekturni faktorji $\frac{N_k - n_k}{N_k - 1}$ praktično 1 in jih zanemarimo. Rešujemo problem vezanega ekstrema:

$$f(n_1, n_2, \dots, n_k) = \sum_{k=1}^{k} \frac{\sigma_k^2}{n_k} \cdot w_k^2$$

z vezjo $n_1 + n_2 + \dots + n_k = n$. Sestavimo Lagrangeovo funkcijo

$$F(n_1, n_2, \dots, n_k) = f(n_1, \dots, n_k) - \lambda (n_1 + \quad + n_k$$

Parcialne odvode izenačimo z 0:

$$\frac{\partial F}{\partial n_k} = -\frac{\sigma_k^2}{n_k^2} w_k^2 - \lambda = 0$$

Sledi

$$-\lambda n_k^2 = \sigma_k^2 w_k^2 .$$

Iz pogoja $n_1 + \quad + n_k = n$ potem sledi

$$\boxed{n_k = n \cdot \frac{\sigma_k \cdot w_k}{\sum_{j=1}^{k} \sigma_j w_j}}$$

Našli smo optimalno razmestitev.

Opombi:

(i) Na videz je stvar nesmisel, ker za optimalno razmestitev potrebujemo $\sigma_1^2, \ldots, \sigma_K^2$. Vendar te količine pogosto lahko vsaj približno ocenimo na podlagi bivših vzorcev (kot to recimo dela statistični urad) ali na podlagi manjših pilotskih vzorcev. Točnost lahko s tem bistveno izboljšamo

(ii) Velikosti n_k v optimalni rešitvi niso v splošnem cela števila, zato zaokrožimo. Če ne bi zanemarili korekturnega faktorja bi dobili rešitev, v kateri w_k zamenjamo z $w_k \cdot \sqrt{\frac{N_k}{N_k - 1}}$, kar je v primerjavi z zaokrožanjem zanemarljivo.

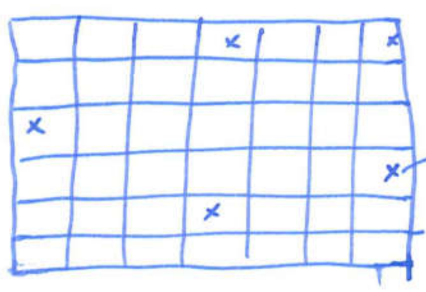
Vzorčenje v skupinicah

Recimo, da je populacija velikosti N razdeljena na K enako velikih podskupin velikosti M, tako da je $N = K \cdot M$. Vzorec izberemo na naslednji način:

(i) izberemo enostavni slučajni vzorec skupin velikosti K.

(ii) v vsaki izbrani skupini izberemo enostavni slučajni vzorec velikosti m.

Slika:



Oceniti želimo populacijsko povprečje μ in izračunati ter oceniti standardno napako.

Predpostavljamo, da so postopki izbiranja vzorca v izbranih podskupinah neodvisni.

<u>Oznake</u>:

μ_k = populacijsko povprečje v k-ti skupini.

σ_k^2 = populacijska varianca v k-ti skupini

μ = populacijsko povprečje

$$\sigma_b^2 = \frac{1}{K}\sum_{k=1}^{K}(\mu_k-\mu)^2$$

$\overline{X}_k$ = vzorčno povprečje v k-ti skupini.

$$I_k = \begin{cases} 1, & \text{če } k\text{-to skupino izberemo} \\ 0, & \text{sicer.} \end{cases}$$

Iz računov pri enostavnem slučajnem vzorčenju vemo, da je $E(I_k) = \frac{k}{K}$ in $\mathrm{cov}(I_k, I_l) =$ za $k \neq l$.

$$= -\frac{k}{K}\left(1-\frac{k}{K}\right)\cdot\frac{1}{K-1}.$$

Kandidatka za cenilko μ je

$$\overline{X} = \frac{1}{k} \sum_{j=1}^{k} \overline{X_j} \cdot I_j$$

= povprečje ocen μ_k za izbrane podskupinice.

<u>Opomba</u>: Pišemo $\overline{X_1}, \ldots, \overline{X_k}$, čeprav ne bomo jemali vzorcev iz vseh podskupinic.

Vendar kot sl. spr. te le ~~#~~ obstajajo kot matematični objekti.

Iz predpostavk v bolj matematični obliki sledi, da sta vektorja $(\overline{X_1}, \ldots, \overline{X_k})$ in $(I_1, I_2, \ldots, I_k)$ neodvisna.

Sledi, ker so $\overline{X_1}, \ldots, \overline{X_k}$ neodvisni

$$E(\overline{X} \mid I_1, \ldots, I_k) = \frac{1}{k} \sum_{j=1}^{k} \mu_j \cdot I_j$$

$$\mathrm{var}(\overline{X} \mid I_1, \ldots, I_k) = \frac{1}{k^2} \sum_{j=1}^{k} I_j \, \mathrm{var}(\overline{X_j})$$

Računamo

$$E(\overline{X}) = \frac{1}{k} \cdot \sum_{j=1}^{k} \cdot \mu_j \cdot E(I_j)$$
$$= \frac{1}{k} \sum_{j=1}^{k} \mu_j$$
$$= \mu.$$

Cenilka $\overline{X}$ je nepristranska.

Računamo

$$\operatorname{var}(\overline{X}) = E(\operatorname{var}(\overline{X} \mid I_1, \ldots, I_k))$$
$$+ \operatorname{var}(E(\overline{X} \mid I_1 \ldots I_k))$$
$$= \frac{1}{k^2} \cdot \sum_{j=1}^{k} E(I_j) \operatorname{var}(\overline{X}_j)$$
$$+ \frac{1}{k^2} \operatorname{var}\left(\sum_{j=1}^{k} \mu_j \cdot I_j\right)$$
$$= \frac{1}{k \cdot K} \sum_{j=1}^{k} \operatorname{var}(\overline{X}_j)$$
$$+ \frac{1}{k^2} \sum_{j=1}^{k} \mu_j^2 \cdot \frac{k}{K}\left(1 - \frac{k}{K}\right)$$

$$+ \frac{1}{k^2} \sum_{i \neq j} \mu_i \mu_j \frac{k}{K}\left(1 - \frac{k}{K}\right)\left(-\frac{1}{K-1}\right)$$

$$= \frac{1}{k \cdot K} \sum_{j=1}^{K} \frac{\sigma_j^2}{m} \cdot \frac{M-m}{M-1}$$

$$+ \frac{1}{k \cdot K}\left(1 - \frac{k}{K}\right)$$

$$\times \left[\sum_{j=1}^{K} \mu_j^2 - \sum_{i \neq j} \mu_i \mu_j \left(\frac{1}{K-1}\right)\right]$$

$$= \frac{1}{k \cdot K} \cdot \sum_{j=1}^{K} \frac{\sigma_j^2}{m} \cdot \frac{M-m}{M-1}$$

$$+ \frac{K-k}{k \cdot K^2} \left[\sum_{j=1}^{K} \left(1 + \frac{1}{K-1} - \frac{1}{K-1}\right) \mu_j^2 \right.$$

$$\left. - \sum_{i \neq j} \mu_i \mu_j \frac{1}{K-1} \right]$$

$$= \quad \text{— " —}$$

$$+ \frac{K-k}{k \cdot K^2} \cdot \left[\frac{K}{K-1} \sum_{j=1}^{K} \mu_j^2 \right.$$

$$\left. - \frac{1}{K-1} \left(\sum_{j=1}^{K} \mu_j \right)^2 \right]$$

$$= \frac{1}{k \cdot K} \sum_{j=1}^{K} \frac{\sigma_j^2}{m} \cdot \frac{M-m}{M-1}$$

$$+ \frac{1}{k} \cdot \frac{K-k}{K-1} \cdot \frac{1}{K} \sum_{j=1}^{K} (\mu_j - \mu)^2$$

$$= \frac{1}{k \cdot K} \sum_{j=1}^{k} \frac{\sigma_j^2}{m} \cdot \frac{M-m}{M-1}$$

$$+ \frac{1}{k} \cdot \frac{K-k}{K-1} \cdot \sigma_b^2$$

Če želimo presojati točnost cenilke $\overline{X}$ moramo znati oceniti količini

$$\gamma = \frac{1}{k} \sum_{j=1}^{k} \sigma_j^2 \quad \text{in} \quad \sigma_b^2 .$$

Za γ hitro najdemo nepristransko cenilko:

$$\hat{\gamma} = \frac{1}{k} \sum_{j=1}^{k} \hat{\sigma}_j^2 \cdot I_j , \quad \text{kjer je}$$

$\hat{\sigma}_j^2$ nepristranska cenilka σ_j^2. Za σ_b^2 pa je nekoliko bolj komplicirano.

Zapišimo

$$\sigma_b^2 = \frac{1}{K}\sum_{j=1}^{k} \mu_j^2 - \mu^2$$

Ideja:

$$\hat{\sigma}_b^2 = \frac{1}{k}\sum_{j=1}^{k} \overline{X}_{j\cdot}^2 \cdot I_j - \overline{X}^2$$

Vemo: $$E(\overline{X}_{j\cdot}^2) = var(\overline{X}_{j\cdot}) + \mu_j^2$$

$$E(\overline{X}^2) = var(\overline{X}) + \mu^2$$

Računamo

$$E(\hat{\sigma}_b^2) = \sigma_b^2 + \frac{1}{K}\sum var(\overline{X}_{j\cdot})$$

$$- var(\overline{X})$$

$$= \sigma_b^2 + \gamma \cdot \frac{1}{m} \cdot \frac{M-m}{M-1}$$

$$- \frac{1}{k}\gamma \frac{M-m}{m(M-1)}$$

$$- \frac{1}{k}\frac{K-k}{K-1}\sigma_b^2$$

Preoblikujemo v

$$E(\hat{\sigma}_b^2) = \sigma_b^2 \cdot \frac{K(k-1)}{k(K-1)} + \gamma \cdot \frac{k-1}{k} \cdot \frac{M-m}{m(M-1)}$$

Tovej lahko $\hat{\sigma}_b^2$ popravimo v nepristransko cenilko:

$$\tilde{\sigma}_b^2 = \left[\hat{\sigma}_b^2 - \hat{\gamma} \cdot \frac{k-1}{k} \cdot \frac{M-m}{m(M-1)}\right] \cdot \frac{k(K-1)}{K(k-1)}$$

$$= \frac{k(K-1)}{K(k-1)} \hat{\sigma}_b^2 - \hat{\gamma} \cdot \frac{(M-m)(K-1)}{m(M-1) \cdot K}$$

Ostane še vprašanje aproksimativne normalnosti. Lahko si mislimo, da skupine izbiramo eno po eno dokler jih ne izberemo k. V vsaki tudi izberemo enostavni slučajni vzorec in dobimo slučajne spr. $\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$.
Cenilka je $\frac{1}{k}(\bar{X}_1 + \cdots + \bar{X}_k)$ in vse $\bar{X}_1, \ldots, \bar{X}_k$ so enako porazdeljene. Kot pri enostavnem sl. vzorčenju je potem vzorčna porazdelitev aproksimativno normalna

POVZETEK

1. Osnova za vzorčenje je vzorčni načrt. Vzorčenje je verjetnostno, če lahko za vsak možen vzorec vnaprej povemo verjetnost, da bo izbran.

2. Ocena katerekoli količine (μ, σ^2, Y, σ_p^2) je neka funkcija vrednosti statistične spremenljivke na izbranih enotah.

3. Na postopek ocenjevanja lahko matematično gledamo kot na „ustvarjanje" slučajne spremenljivke cenilke. Vsa informacija o točnosti je vsebovana v tem, kar znamo povedati o porazdelitvi cenilke.

3. Ocenjevanje parametrov

3.1. Pojem statističnega modela

Statistika se ukvarja z analizo podatkov. Eno od bistvenih orodij za ta namen so statistični modeli. Statistični model je opis mehanizma, za katerega menimo, da je generiral podatke. Tak mehanizem praviloma vsebuje naključnost, zato so statistični modeli formulirani v jeziku slučajnih spremenljivk, porazdelitev, neodvisnosti ali v jeziku pogojnih porazdelitev. Oglejmo si nekaj primerov.

<u>Primer</u>: Linearna regresija

Podatki so oblike $y_i, x_{i1}, x_{i2}, \ldots, x_{im}$ za $i = 1, 2, \ldots, n$;

Tipično rečemo, da je y_i odziv i-te enote, vrednosti $x_{i1}, x_{i2}, \ldots, x_{im}$ pa kovariate. Tipično bi radi „pojasnili" spremenljivko y_i kot funkcijo kovariat $x_{i1}, x_{i2}, \ldots, x_{im}$, vendar funkcijska odvisnost ni „točna", tako da nekaj vlogo igra še naključje.

Prvi primer regresije je zgodovinsko ta, da je F. Galton želel pojasniti telesno višino sina (Y) ~~z~~ telesno višino očeta (x). Imel je podatke:

$$\begin{array}{cc} Y_1 & x_1 \\ Y_2 & x_2 \\ \vdots & \vdots \\ Y_n & x_n \end{array}$$

Statistični model pravi: pari $(x_1, y_1), \ldots, (x_n, y_n)$ „nastanejo" kot med sabo neodvisni slučajni vektorji iz neke porazdelitve $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$.

Predpostavljamo, da velja

$$Y_i = \alpha + \beta X_i + \varepsilon_i,$$

kjer so $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ neodvisne in enako porazdeljene slučajne spremenljivke z $E(\varepsilon_i) = 0$ in $var(\varepsilon_i) = \sigma^2$ za $i = 1, 2, \dots, n$.
V grobem torej rečemo, da je telesna višina sina linearna funkcija telesne višine očeta + slučajni člen. Na drugačen način bi lahko zapisali:

$$E(Y_i | X_i) = \alpha + \beta X_i$$

$$var(Y_i | X_i) = \sigma^2$$

Torej, statistični model je predpostavka da so (X_i, Y_i) generirani iz neke porazdelitve, o kateri nekaj predpostavljamo.

Primer: LOGISTIČNA REGRESIJA

Banke tipično morajo razvrščati kvaliteto kreditojemalcev v razrede. To zahtevajo bančna zakonodaja. Podatki iz preteklosti so oblike

$$
\begin{array}{ll}
y_1 & x_{11}\ x_{12} \dots x_{1m} \\
y_2 & x_{21}\ x_{22} \dots x_{2m} \\
\vdots & \\
y_n & x_{n1}\ x_{n2} \dots x_{nm}\,.
\end{array}
$$

Pri tem je $y_i \in \{0,1\}$ indikator, ali je kreditojemalec i nehal odplačevati dolg, $x_{i1}, x_{i2} \dots x_{im}$ pa „kovariate", ki so starost, delovna doba, dohodki, zakonski stan, tip zaposlitve, ... Vprašanje: ali lahko iz kovariat napovemo verjetnost, da kreditojemalec ne bo vrnil kredita?

Logistična regresija je ena od možnosti. Predpostavka je: ~~vektorji~~ vektorji $(Y_i, X_{i1}, \ldots, X_{im})$ „nastanejo" kot med sabo neodvisni vektorji $(Y_i, X_{i1}, \ldots, X_{im})$ z enako porazdelitvijo. Predpostavimo, da je

$$P(Y_i = 1 \mid X_{i1}, X_{i2}, \ldots, X_{im}) = \frac{e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^{m} \beta_j x_{ij}}}$$

za $i = 1, 2, \ldots, n$. Konstantam $\beta_0, \beta_1, \ldots, \beta_m$ rečemo konstante v modelu ali parametri.

Parametre ocenimo na podlagi znanih podatkov. Ko imamo konkretnega kreditojemalca s kovariatami $x_1, x_2, \dots, x_m$ in ko imamo ocenjene parametre $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_m$ lahko ocenimo verjetnost, da kreditojemalec ne bo vrnil kredita kot

$$\hat{p}_i = \frac{e^{\hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_j}}{1 + e^{\hat{\beta}_0 + \sum_{j=1}^{m} \hat{\beta}_j x_j}}.$$

Torej, predpostavljamo, da so podatki nastali na določen način naključno in predpostavljam določene lastnosti teh porazdelitev. Ali lahko trdimo, da je model točen? Tipično ne, lahko pa izvedemo določena preverjanja.

V tem poglavju bomo obravnavali kako oceniti parametre in kaj lahko rečemo o točnosti ocen.

3.2. Cenilke, vzorčne porazdelitve, intervali zaupanja

Primer: Recimo, da imamo podatke $x_1, x_2, \ldots, x_n$. Podatkom pogosto rečemo opazovane vrednosti. Predpostavimo, da so podatki nastali iz slučajne spremenljivke $X_1, X_2, \ldots, X_n$, za katere predpostavljamo neodvisnost in enako $N(\mu, \sigma^2)$ porazdeljenost. Parametrov μ in σ^2 ne poznamo, predpostavljamo le $(\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$.

Kako bi ocenili μ in σ^2 ?

Ideja:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Oceni $\hat{\mu}$ in $\hat{\sigma}^2$ smo dobili tako, da smo „sprožili" slučajni spremenljivki

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \qquad \text{in}$$

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2 .$$

Razprava je podobna, kot pri vzorčenju. Preden podatki „nastanejo" so slučajne spremenljivke, le da pri vzorčenju vemo, da jih sprožimo z izvedbo vzorčnega načrta, tukaj pa je „sprožanje" nekoliko bolj abstraktno. Kaj znamo povedati o porazdelitvah $\overline{X}$ in S^2?

$$E(\overline{X}) = \frac{1}{n} \sum_{i=1}^{n} E(X_i) = \mu$$

$$\mathrm{var}(\overline{X}) = \frac{1}{n^2}\,\mathrm{var}(X_1 + \dots + X_n)$$

$$= \frac{\sigma^2}{n}$$

Kot pri vzorčenju bomo slučajni spremenljivki $\overline{X}$ rekli <u>cenilka</u>. Številu $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ rečemo <u>ocena</u>. Oznako $\hat{\mu}$ bomo uporabljali za eno ali drugo, raba pa bo jasna iz konteksta.

Cenilka $\hat{\mu} = \overline{X}$ je <u>nepristranska</u>.

Poleg tega je $se(\overline{X}) = \frac{\sigma}{\sqrt{n}}$ in $\overline{X}$ je normalna.

Računamo še

$$E(S^2) = \frac{1}{n} \sum_{i=1}^{n} E[(X_i - \overline{X})^2]$$

$$= E[(X_1 - \overline{X})^2] \quad (\text{simetrija})$$

$$= \operatorname{var}(X_1 - \overline{X})$$

$$= \operatorname{var}(X_1) + \operatorname{var}(\overline{X})$$

$$- 2 \operatorname{cov}\left(X_1, \frac{1}{n} \sum_{i=1}^{n} X_i\right)$$

$$= \sigma^2 + \frac{\sigma^2}{n} - \frac{2}{n} \cdot \sigma^2$$

$$= \frac{n-1}{n} \cdot \sigma^2$$

Cenilka S^2 ni nepristranska, lahko pa jo popravimo v nepristransko ~~[illegible]~~ cenilko

$$\boxed{\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Kako pa je s porazdelitvijo $\hat{\sigma}^2$?

Iz poglavja o večrazsežni normalni porazdelitvi vemo, da je

$$\sum_{i=1}^{n} (x_i - \overline{x})^2 \sim \sigma^2 \chi^2(n-1)$$

$\chi^2(n)$ porazdelitev je vsota med seboj neodvisnih slučajnih spremenljivk $\Rightarrow$ po centralnem limitnem izreku aproksimativno normalna. Rečemo lahko torej, da je $\hat{\sigma}^2$ približno normalno porazdeljena s parametroma σ^2 in $2\sigma^4(n-1)$, ker je varianca $\chi^2(n)$ porazdelitve $2n$. Z drugimi besedami

$$\sqrt{n}\,(\hat{\sigma}^2 - \sigma^2) \overset{\text{aprox.}}{\sim} N\left(0, 2\sigma^4 \frac{(n-1)}{n}\right)$$

Potrebujemo nekaj terminologije..

Privzeli bomo, da so podatki ali opazovane vrednosti $x_1, x_2, \dots x_n$ „nastale" kot slučajni vektor $\underline{X} = (X_1, X_2, \dots, X_n)$ z gostoto $f_{\underline{X}}(\underline{x}, \underline{\theta})$, kjer je $\underline{\theta} \in \Theta \subseteq \mathbb{R}^m$ vektor možnih parametrov.

Opomba: Zaenkrat ne privzamemo, da so $X_1, X_2, \dots, X_n$ neodvisne.

Opomba: Za diskretne vektorje $\underline{X}$ bodo od parametra $\underline{\theta}$ odvisne verjetnosti $p(\underline{x}, \underline{\theta}) := P_{\underline{\theta}}(\underline{X} = \underline{x})$.

Definicije:

(i) Funkciji $\hat{\underline{\theta}} = \hat{\underline{\theta}}(x_1, x_2, \dots, x_n)$, ki „ocenjuje" $\underline{\theta}$ rečemo ocena.

(ii) Slučajnemu vektorju

$$\underline{\hat{\theta}} = \underline{\hat{\theta}}(X_1, X_2, \ldots, X_n) = \hat{\theta}(\underline{x})$$

rečemo cenilka parametra $\underline{\theta}$.

<u>Opomba</u>: Za komponente pišemo $\hat{\theta}_j$.

(iii) Cenilka $\underline{\hat{\theta}}$ je nepristranska, če je

$$E_{\underline{\theta}}(\underline{\hat{\theta}}) = \underline{\theta}.$$

(iv) Srednja kvadratična napaka cenilke $\hat{\theta}$ je količina

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2],$$

če je parameter enodimenzionalen.

(v) Za enodimenzionalno $\hat{\theta}$ je standardna napaka količina

$$se(\hat{\theta}) = \sqrt{var_{\theta}(\hat{\theta})}.$$

Primer: Recimo, da so $X_1, X_2, \ldots, X_n$ vzorec iz Poissonove porazdelitve $Po(\lambda)$. Povprečje

$$\hat{\lambda} = \overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

je nepristranska cenilka λ, ker je $E(X_i) = \lambda$. Velja tudi

$$var(\overline{X}) = \frac{1}{n} \cdot var(X_1) = \frac{\lambda}{n},$$

torej je $se(\overline{X}) = \frac{\sqrt{\lambda}}{\sqrt{n}}$. Po centralnem limitnem izreku je $\overline{X}$ aproksimativno normalna, varianco pa lahko ocenimo z $\frac{\hat{\lambda}}{n}$. Spet lahko rečemo

$$\sqrt{n}(\hat{\lambda} - \lambda) \overset{\text{aprox}}{\sim} N(0, \lambda)$$

Primer: Opazovane vrednosti $\underline{x}_1, \underline{x}_2 \cdots \underline{x}_n$ so lahko tudi vektorji. Privzemimo da so nastali kot neodvisni, enako porazdeljeni normalni vektorji, torej $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$ $\ni$ $\underline{x}_k \sim N(\underline{\mu}, \underline{\Sigma})$.

Recimo

$$\hat{\underline{\mu}} = \frac{1}{n} \sum_{k=1}^{n} \underline{x}_k$$

$$\hat{\underline{\Sigma}} = \frac{1}{n-1} \sum_{k=1}^{n} (\underline{x}_k - \overline{\underline{x}})(\underline{x}_k - \overline{\underline{x}})^T$$

Računamo

$$E(\hat{\underline{\mu}}) = \underline{\mu} \quad (\text{nepristranska})$$

$$E(\hat{\underline{\Sigma}}) = \frac{1}{n-1} \sum_{k=1}^{n} E\left[(\underline{x}_k - \overline{\underline{x}})(\underline{x}_k - \overline{\underline{x}})^T\right]$$

$$= \frac{n}{n-1} E\left[(\underline{x}_1 - \overline{\underline{x}})(\underline{x}_1 - \overline{\underline{x}})^T\right]$$

$$= \frac{n}{n-1} E\left[\underline{x}_1 \underline{x}_1^T - \overline{\underline{x}}\, \underline{x}_1^T - \underline{x}_1 \overline{\underline{x}}^T + \overline{\underline{x}}\, \overline{\underline{x}}^T\right]$$

$$= \frac{n}{n-1} \Big[\underline{\Sigma} + \underline{\mu} \cdot \underline{\mu}^T$$

$$- \operatorname{cov}(\overline{\underline{X}}, \underline{X}_1) - \underline{\mu}\underline{\mu}^T$$

$$- \operatorname{cov}(\underline{X}_1, \overline{X}) - \underline{\mu}\underline{\mu}^T$$

$$+ \operatorname{cov}(\overline{\underline{X}}, \overline{\underline{X}}) + \underline{\mu} \cdot \underline{\mu}^T \Big]$$

$$= \frac{n}{n-1} \Big[\underline{\Sigma} - \frac{1}{n} \underline{\Sigma} - \frac{1}{n} \underline{\Sigma} + \frac{1}{n} \underline{\Sigma} \Big]$$

$$= \frac{n}{n-1} \cdot \frac{n-1}{n} \underline{\Sigma}$$

$$= \underline{\Sigma}$$

Predlagana $\hat{\underline{\Sigma}}$ je nepristranska.

Po vektorski verziji centralnega limitnega izreka je

$$\sqrt{n} (\hat{\underline{\mu}} - \underline{\mu}) \overset{\text{aproks.}}{\sim} N(0, \underline{\Sigma})$$

3.3. Metoda največjega verjetja

Primer: Predpostavimo, da so opazovane vrednosti $x_1, x_2, \ldots, x_n$ vzorec iz Weibullove gostote. To pomeni, da so „nastale" kot med sabo neodvisne, enako porazdeljene slučajne spremenljivke z gostoto

$$f(x, \alpha, \sigma) = \begin{cases} \frac{\alpha}{\sigma}\left(\frac{x}{\sigma}\right)^{\alpha-1} e^{-\left(\frac{x}{\sigma}\right)^{\alpha}}, & x> \\ 0. & \text{sicer.} \end{cases}$$

Pri tem je $\alpha > 0$ in $\sigma > 0$.

Opomba: Abstraktno je $\underline{\theta} = (\alpha, \sigma)$ in $\Theta = (0,\infty)^2$.

Kako bi ocenili α in σ !

V tem primeru ni očitno, kako oceniti parametre. Ugotovimo, da je

$$F_X(x) = 1 - e^{-\left(\frac{x}{\sigma}\right)^\alpha}.$$

Če je $Y = \left(\frac{X}{\sigma}\right)^\alpha$, je

$$P(Y \leq y) = P\left(\left(\frac{X}{\sigma}\right)^\alpha \leq y\right)$$
$$= P\left(X \leq \sigma \cdot y^{1/\alpha}\right)$$
$$= 1 - e^{-\left(\frac{\sigma y^{1/\alpha}}{\sigma}\right)^\alpha}$$
$$= 1 - e^{-y},$$

torej je $Y \sim \exp(1)$. Sledi

$$E(X) = E\left[\sigma \cdot Y^{1/\alpha}\right]$$
$$= \sigma \cdot \int_0^\infty y^{1/\alpha} e^{-y}\, dy$$
$$= \sigma \cdot \Gamma\left(1 + \frac{1}{\alpha}\right)$$

$$E(X^2) = \sigma^2 \, \Gamma\left(1 + \frac{2}{\alpha}\right)$$

V tem primeru $\bar{X}$ in izvedenke ne pomagajo. Morda bi lahko rekli

$$\bar{x} \approx \sigma \, \Gamma(1 + 1/\alpha) \quad \text{in}$$

$$\frac{1}{n}\sum_{k=1}^{n} x_k^2 \approx E(X^2) = \sigma^2 \, \Gamma(1 + 1/\alpha)$$

in izrazili α in σ, vendar se ta t. i. metoda momentov izkaže za slabotno.

<u>Ideja</u>: Vektor podatkov $\underline{x} = (x_1, x_2, \dots, x_n)^T$ je nastal kot naključno izbrana točka $\underline{X} = (X_1, X_2, \dots, X_n)^T$ z gostoto

$$f_{\underline{X}}(\underline{x}, \alpha, \sigma) = \prod_{k=1}^{n} f_{X_k}(x_k, \alpha, \sigma).$$

kje smo bolj verjetno izbrali $\underline{x}$, tam kjer je gostota velika ali tam, kjer je gostota majhna?

Če že moramo izbirati, tam kjer je gostota velika. Obrnimo razmislek: za fiksne opazovane vrednosti $x_1, x_2, \ldots, x_n$ si oglejmo funkcijo

$$L(a, b \mid \underline{x}) = \prod_{k=1}^{n} f_{X_k}(x_k, a, b),$$

tovej funkcijo

$$(a, b) \xrightarrow{L} \prod_{k=1}^{n} f_{X_k}(x_k, a, b).$$

Če ima ta funkcija maksimum, je argument maksimuma dober kandidat za $(\hat{a}, \hat{b})$.

Ideji, da fiksiramo $\underline{x}$ in iščemo maksimum po $\underline{\theta}$ se imenuje <u>metoda</u> <u>največjega</u> <u>verjetja</u>.

(angl. maximum likelihood method)

Avtor ideje je angleški statistik Sir Ronald A. Fischer (1890–1962).

V primeru Weibulove porazdelitve dobimo

$$L(\alpha, \sigma \mid \underline{x}) = \frac{\alpha^n}{\sigma^n} \prod_{k=1}^{n} \left(\frac{x_k}{\sigma}\right)^{\alpha-1} e^{-\left(\frac{x_k}{\sigma}\right)^{\alpha}}$$

Ker je $x_k > 0$ lahko logaritmiramo in dobimo

$$\log L(\alpha, \sigma \mid \underline{x}) = n \log \alpha - n \log \sigma + (\alpha - 1) \sum_{k=1}^{n} \log x_k - \sum_{k=1}^{n} \left(\frac{x_k}{\sigma}\right)^{\alpha} - (\alpha-1)\, n \log \alpha$$

Poenostavimo v

$$\log L(\alpha, b \mid \underline{x}) = n \log \alpha - n\alpha \log b + (\alpha - 1) \sum_{k=1}^{n} \log x_k - \sum_{k=1}^{n} \left(\frac{x_k}{b}\right)^{\alpha}.$$

Parcialno odvajamo po α in b.

$$\frac{\partial \log L}{\partial \alpha} = \frac{n}{\alpha} - n \log b + \sum_{k=1}^{n} \log x_k - \sum_{k=1}^{n} \left(\frac{x_k}{b}\right)^{\alpha} \log\left(\frac{x_k}{b}\right) = 0$$

$$\frac{\partial \log L}{\partial b} = -\frac{n\alpha}{b} - \sum_{k=1}^{n} \frac{x_k^{\alpha}}{b^{\alpha+1}} (-\alpha) = 0$$

Iz druge enačbe sledi:

$$\frac{1}{n} \sum_{k=1}^{n} \left(\frac{x_k}{b}\right)^{\alpha} = 1.$$

Prvo enačbo delimo z n in preuredimo:

$$\frac{1}{\alpha} = \frac{1}{n} \sum_{k=1}^{n} \log x_k - \frac{1}{n} \sum_{k=1}^{n} \left(\frac{x_k}{b}\right)^{\alpha} \log x_k$$

Enoličnost rešitev enačb po metodi največjega verjetja.

Za logaritemsko funkcijo verjetja smo dobili

$$\ell(\alpha, \sigma \mid \underline{x})$$

$$= n \log \alpha - n\alpha \log \sigma + (\alpha - 1) \sum_{k=1}^{n} \log x_k - \sum_{k=1}^{n} \left(\frac{x_k}{\sigma}\right)^{\alpha}$$

Parcialno odvajanje da enačbi

$$\frac{\partial \ell}{\partial \alpha} = \frac{n}{\alpha} - n \log \sigma + \sum_{k=1}^{n} \log x_k - \sum_{k=1}^{n} \left(\frac{x_k}{\sigma}\right)^{\alpha} \log\left(\frac{x_k}{\sigma}\right) = 0$$

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n\alpha}{\sigma} + \frac{\alpha}{\sigma} \sum_{k=1}^{n} \left(\frac{x_k}{\sigma}\right)^{\alpha} = 0$$

Iz druge enačbe sledi, da je

$$\sigma^{\alpha} = \Big(\sum_{k=1}^{n} x_k^{\alpha}\Big)/n \quad \text{in} \quad \sum_{k=1}^{n} \left(\frac{x_k}{\sigma}\right)^{\alpha} = \underline{1}.$$

Vstavimo v prvo enačbo in dobimo po deljenju z n

$$\frac{1}{\alpha} = -\frac{1}{n} \sum_{k=1}^{n} \log x_k + \frac{1}{n} \frac{\sum_{k=1}^{n} x_k^{\alpha} \log x_k}{\frac{1}{n} \sum_{k=1}^{n} x_k^{\alpha}} = 0$$

Prepišemo v

$$\frac{1}{\alpha} = \underbrace{\frac{\sum_{k=1}^{n} x_k^{\alpha} \log x_k}{\sum_{k=1}^{n} x_k^{\alpha}} - \frac{1}{n} \sum_{k=1}^{n} \log x_k}_{g(\alpha)} = g(\alpha)$$

Računamo

$$g'(\alpha) = \frac{\sum_{k=1}^{n} x_k^{\alpha} \log^2 x_k \cdot \sum_{k=1}^{n} x_k^{\alpha} - \left(\sum_{k=1}^{n} x_k^{\alpha} \log x_k\right)^2}{\left(\sum_{k=1}^{n} x_k^{\alpha}\right)^2}$$

Po Cauchy-Schwartzovi neenačbi je

$$\left(\sum_{k=1}^{n} x_k^{\alpha} \log x_k\right)^2 \leq \left(\sum_{k=1}^{n} x_k^{\alpha}\right) \cdot \left(\sum_{k=1}^{n} x_k^{\alpha} \log^2 x_k\right)$$

Opomba: Pišemo $x_k^{\alpha} = x_k^{\alpha/2} \cdot x_k^{\alpha/2}$.

Enakost dobimo le, če sta vektorja

$(x_1^{\alpha/2}, \ldots, x_n^{\alpha/2})$ in $(x_1^{\alpha/2} \log x_1, \ldots, x_n^{\alpha/2} \log x_n)$

kolinearna. To je možno le, če je $x_1 = x_2 = \ldots = x_n$, kar lahko izključimo.

Funkcija $g(\alpha)$ je na $(0, \infty)$ strogo naraščajoča.

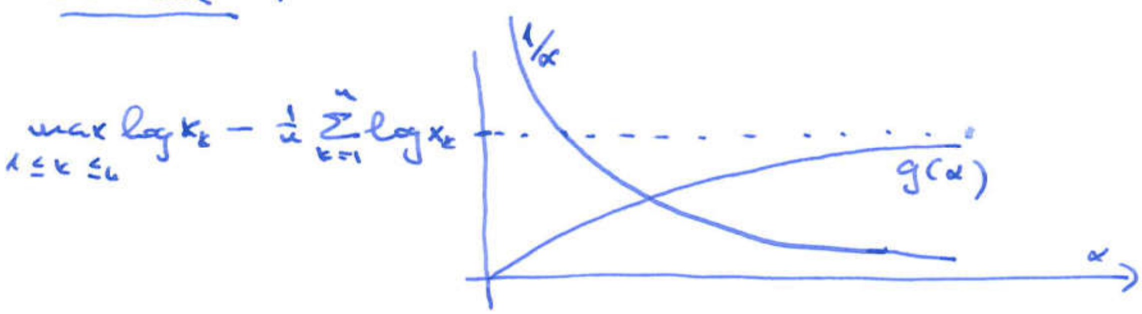
Velja $\lim_{\alpha \to 0} g(\alpha) = 0$ in

$$\lim_{\alpha \to \infty} \frac{\sum_{k=1}^{n} x_k^\alpha \log x_k}{\sum_{k=1}^{n} x_k^\alpha} = \max_{1 \le k \le n} \log x_k$$

Ker niso vsi $x_1, x_2, \dots, x_n$ enaki, je

$$\max_{1 \le k \le n} \log x_k > \frac{1}{n} \sum_{k=1}^{n} \log x_k .$$

Slika:



Enačba $1/\alpha = g(\alpha)$ ima natanko eno rešitev > 0. Iz tega sledi tudi enoličnost 6.

Enačbe nimajo analitične rešitve. Z nekaj truda lahko dokažemo, da je rešitev enolična in obstaja, vendar eksplicitne oblike $\hat{\alpha}$ in $\hat{b}$ ne poznamo. Kaj pa zdaj?

Oglejmo si simulacije.

3.4. Asimptotske lastnosti cenilk po metodi največjega verjetja

Iz simulacij izgleda, da utegneta biti $\hat{\alpha}$ in $\hat{b}$ aproksimativno normalno porazdeljeni. Za izpeljavo tega dejstva potrebujemo nekaj pripravljalnih izrekov. Naposled bo trditev sledila iz centralnega limitnega izreka.

Izrek 3.1: Naj ima slučajna spremenljivka gostoto $f(x,\theta)$, kjer je $\Theta \subseteq \mathbb{R}$ odprta množica. Definirajmo

$$\log_+ f(x,\theta) = \begin{cases} \log f(x,\theta) & \text{za } f(x,\theta) > 0; \\ 0 & \text{sicer.} \end{cases}$$

Predpostavimo, da povsod kjer je treba lahko integrale s parametrom odvajamo pod integralskim znakom. Predpostavimo, da je $\theta \mapsto f(x,\theta)$ zvezno odvedljiva po θ za vsak fiksen x. Definirajmo

$$Z = \frac{\partial}{\partial\theta} \log_+ f(X,\theta)\Big|_{\theta=\theta_0}$$

$$W = \frac{\partial^2}{\partial\theta^2} \log_+ f(X,\theta)\Big|_{\theta=\theta_0}.$$

Velja

$$E(Z) = 0 \quad \text{in} \quad E(W) = -\operatorname{var}(Z)$$

Dokaz: Računamo

$$E(Z) = \int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} \log f(x,\theta)\Big|_{\theta=\theta_0} \cdot f(x,\theta_0)\, dx$$

$$= \int_{-\infty \atop \{f(x,\theta_0)>0\}}^{\infty} \frac{\partial}{\partial \theta} \log f(x,\theta)\Big|_{\theta=\theta_0} f(x,\theta)\, dx$$

$$= \int_{\{f(x,\theta_0)>0\}} \frac{\frac{\partial}{\partial \theta} f(x,\theta)\big|_{\theta=\theta_0}}{f(x,\theta_0)} f(x,\theta_0)\, dx$$

$$= \int_{\{f(x,\theta_0)>0\}} \frac{\partial}{\partial \theta} f(x,\theta)\Big|_{\theta=\theta_0} dx$$

$$= \left(\frac{\partial}{\partial \theta} \int_{-\infty}^{\infty} f(x,\theta)\, dx \right)\Big|_{\theta=\theta_0}$$

$$= \left(\frac{\partial}{\partial \theta} 1 \right)_{\theta=\theta_0}$$

$$= 0 .$$

Podobno računamo za w.

$$E(w) = \int_{\{f(x,\theta_0) > 0\}} \frac{\partial^2}{\partial \theta^2} \log f(x,\theta) \Big|_{\theta=\theta_0} f(x,\theta_0)$$

$$= \int_{\{f(x,\theta_0)>0\}} \frac{\frac{\partial^2 f}{\partial \theta^2}(x,\theta_0) f(x,\theta_0) - \left(\frac{\partial}{\partial \theta} f(x,\theta_0)\right)^2}{f^2(x,\theta_0)} \times f(x,\theta_0)\, dx$$

$$= \int_{-\infty}^{\infty} \frac{\partial^2}{\partial \theta^2} f(x,\theta_0)\, dx - \int_{\{f(x,\theta_0)>0\}} \left(\frac{\partial}{\partial \theta} \log f(x,\theta)\right)^2 \Big|_{\theta=\theta_0} \cdot f(x,\theta_0)\, dx$$

$$= \left(\frac{\partial^2}{\partial \theta^2} 1\right)\Big|_{\theta=\theta_0} - E(z^2)$$

$$= -E(z^2)$$

Ker je $E(z) = 0$, je drugi del trditve dokazan.

Opomba: Podoben račun velja za parcialne odvode z več pisanja. Če je $\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ lahko definiramo

$$z_i = \frac{\partial}{\partial \theta_i} \log_+ f(\underline{x}, \underline{\theta})\Big|_{\underline{\theta} = \underline{\theta}_0}$$

$$w_{i,j} = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \log_+ f(\underline{x}, \underline{\theta})\Big|_{\underline{\theta} = \underline{\theta}_0}$$

Velja (pod podobnimi predpostavkami)

$$\boxed{\begin{aligned} E(z_i) &= 0 \quad \text{in} \\ E(w_{i,j}) &= -\operatorname{cov}(z_i, z_j) \end{aligned}}$$

Primer: Če je X Weibullova je

$$\log f(x, \alpha, \sigma) = \log \alpha - \log \sigma + (\alpha - 1)\log \frac{x}{\sigma} - \left(\frac{x}{\sigma}\right)^{\alpha}$$

Odvajamo

$$\frac{\partial}{\partial \alpha} \log f(x, \alpha, \sigma)$$

$$= \frac{1}{\alpha} + \log \frac{x}{\sigma} - \left(\frac{x}{\sigma}\right)^{\alpha} \cdot \log \frac{x}{\sigma}$$

Nadomestimo x z X in izračunamo

$$Z_1 = \frac{1}{\alpha} + \log \frac{X}{\sigma} - \left(\frac{X}{\sigma}\right)^{\alpha} \cdot \log \frac{X}{\sigma}$$

$$= \frac{1}{\alpha} + \frac{1}{\alpha} \log \left(\frac{X}{\sigma}\right)^{\alpha} - \frac{1}{\alpha} \left(\frac{X}{\sigma}\right)^{\alpha} \log \left(\frac{X}{\sigma}\right)^{\alpha}$$

$$= \frac{1}{\alpha} + \frac{1}{\alpha} \log Y - \frac{1}{\alpha} Y \log Y$$

$$E(\log Y) = \int_0^{\infty} \log y \cdot e^{-y}\, dy$$

Vemo:

$$\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\, du \quad \Rightarrow$$

$$\Gamma'(x) = \int_0^\infty u^{x-1} \log u \cdot e^{-u}\, du$$

$$\Gamma''(x) = \int_0^\infty u^{x-1} (\log u)^2 e^{-u}\, du$$

Slodi

$$E(\log Y) = \Gamma'(1) \qquad E[\log Y \cdot Y^2]$$

$$E(Y \log Y) = \Gamma'(2) \qquad = \Gamma'(3)$$

$$\Gamma(x+1) = x\,\Gamma(x) \quad \Rightarrow$$

$$\Gamma'(x+1) = \Gamma(x) + x\,\Gamma'(x) \quad \Rightarrow$$

$$\Gamma'(2) = \Gamma(1) + \Gamma'(1) = 1 + \Gamma'(1)$$

Dobimo

$$E(Z_1) = \frac{1}{\alpha} + \frac{1}{\alpha}\Gamma'(1) - \frac{1}{\alpha}(1 + \Gamma'(1))$$

$$= 0 \;!$$

Podobno je

$$\frac{\partial}{\partial \sigma} \log (x, \alpha, \sigma)$$

$$= - \frac{\alpha}{\sigma} + \frac{\alpha}{\sigma} \left(\frac{x}{\sigma} \right)^{\alpha} \text{, torej}$$

$$z_2 = \frac{\alpha}{\sigma} \left(-1 + \left(\frac{x}{\sigma} \right)^{\alpha} \right)$$

$$= \frac{\alpha}{\sigma} (-1 + Y) ,$$

torej $E(z_2) = 0$.

Računamo še

$$\frac{\partial^2}{\partial \alpha \partial \sigma} \log f(x, \alpha, \sigma))$$

$$= - \frac{1}{\sigma} + \frac{1}{\sigma} \left(\frac{x}{\sigma} \right)^{\alpha} \left(1 + \log \left(\frac{x}{\sigma} \right)^{\alpha} \right)$$

Zamenjamo x z X in izračunamo pričakovano vrednost:

$$E(W_{1,2}) = E \left[- 1/\sigma + \frac{1}{\sigma} Y (1 + \log Y) \right]$$

$$= \frac{1}{6} E[Y \log Y]$$

$$= \frac{1}{6} \Gamma'(2)$$

Prevenimos:

$$\operatorname{cov}(Z_1, Z_2)$$

$$= \operatorname{cov}\left(\frac{1}{\alpha} \log Y - \frac{1}{\alpha} Y \log Y, \frac{\alpha}{6} Y\right)$$

$$= \frac{1}{6} \operatorname{cov}(\log Y - Y \log Y, Y)$$

$$= \frac{1}{6} \Big[E(Y \log Y - Y^2 \log Y)$$

$$- E[\log Y - Y \log Y] E(Y) \Big]$$

$$= \frac{1}{6} \Big[\Gamma'(2) - \Gamma'(3)$$

$$- \Gamma'(1) + \Gamma'(2) \Big]$$

$$\Gamma'(3) = 1 + 2\Gamma'(2)$$

$$= \frac{1}{6}\left[-1 - \Gamma'(1)\right] = -\frac{1}{6} \Gamma'(2)$$

Za izpeljavo bomo uporabili Taylorjevo vrsto. Aproksimacija bo potekala v več korakih.

Najprej uvedimo terminologijo in predpostavke:

<u>Definicija</u>:

(i) Funkcijo $\underline{\theta} \longmapsto f(\underline{x}, \underline{\theta})$ imenujemo funkcija verjetja in označimo z $L(\underline{\theta} \mid \underline{x})$ (angl. likelihood function)

(ii) Funkcijo $\underline{\theta} \longmapsto \log_+(L(\underline{\theta} \mid \underline{x}))$ imenujemo logaritemsko funkcijo verjetja. Oznaka: $\ell(\underline{\theta} \mid \underline{x})$.

Pri izpeljavi se bomo omejili na eno dimenzijo z $\theta \in (a, b)$.

Korak 1: Predpostavimo, da so $x_1, x_2, \dots, x_n$ vzorec iz gostote $f(x, \theta)$ in da vedno obstaja enolično določen maksimum $\hat{\theta}$.

Opomba: „Vedno" pomeni z verjetnostjo 1. Fiksirajmo $\theta_0 \in (a,b)$ in predpostavimo, da so $x_1, x_2 \dots x_n$ generirani iz $f(x, \theta_0)$. Velja

$$l'(\hat{\theta} \mid \underline{x}) = 0$$

Po Taylorjevi formuli je

$$l'(\theta_0 \mid \underline{x}) = \underbrace{l'(\hat{\theta} \mid x)}_{=0} + (\theta_0 - \hat{\theta}) l''(\hat{\theta} \mid \underline{x}) + R(\underline{x}, \underline{\theta}_0)$$

Korak 2 : Ignoriramo R in zapišemo

$$\hat{\theta} - \theta_0 \simeq -\frac{l'(\theta_0 | x)}{l''(\hat{\theta} | \underline{x})}$$

Ker so $x_1, x_2, \dots x_n$ neodvisne, je

$$l(\theta_0 | \underline{x}) = \sum_{k=1}^{n} \log_+ f(x_k, \theta)$$

$$l'(\theta_0 | \underline{x}) = \sum_{k=1}^{n} \frac{d}{d\theta} \log_+ f(x_k, \theta) \Big|_{\theta=\theta}$$

Nadomestimo male x z velikimi X:

$$l'(\theta_0 | \underline{X}) = \sum_{k=1}^{n} \underbrace{\frac{d}{d\theta} \log_+ f(X_k, \theta) \Big|_{\theta=}}_{Z_k}$$

$$= \sum_{k=1}^{n} Z_k$$

Izrek 3.1 nam pove, da je $E(Z_k) = 0$. Poleg tega so $Z_1, Z_2, \ldots, Z_n$ enako porazdeljene in neodvisne.

Korak 3: Ker je $\hat{\theta}$ „blizu" θ_0 v imenovalcu $\hat{\theta}$ nadomestimo z θ_0. Zapišemo

$$l''(\theta_0 | \underline{x}) = \sum_{k=1}^{n} \underbrace{\frac{d^2}{d\theta^2} \log f(x_k, \theta)\Big|_{\theta=\theta_0}}_{w_k}$$

Nadomestimo x z X.

$$l''(\theta_0 | \underline{X}) = \sum_{k=0}^{n} W_k .$$

Izrek 3.1 pove, da je $E(W_k) = - \mathrm{var}(Z_k)$

Korak 4 : Izpeljali smo

$$\sqrt{n}\,(\hat{\theta}-\theta_0) \doteq \frac{\frac{1}{\sqrt{n}}\sum_{k=1}^{n} z_k}{\frac{1}{n}\sum_{k=1}^{n} w_k}$$

Ker je $E(z_k) = 0$, je po centralnem limitnem izreku števec približno normalno porazdeljen. Imenovalec pa je „približno" konstanta $E(w_1)$. Če normalno porazdeljeno slučajno spremenljivko delimo s konstanto, je porazdelitev še vedno normalna. Tukaj „približno" normalno delimo s „skoraj" konstanto.

Varianca števca je $var(Z_1)$, konstanta v imenovalcu pa $E(W_1) = -var(Z_1)$. Torej je

$$\sqrt{n}\,(\hat{\theta} - \theta_0) \mathrel{\dot\sim} N\left(0, -\frac{1}{E(W_1)}\right)$$

<u>Sklep</u>: Karkoli je θ_0, je $\sqrt{n}\,(\hat{\theta} - \theta_0)$ približno normalno porazdeljena (kot cenilka!).

<u>Komentar</u>: $\hat{\theta}$ ni nujno nepristranska. Izpeljava nam v resnici da srednjo kvadratično napako.

Definicija: Naj bo $f(\underline{x}, \underline{\theta})$ gostota. Definiramo

$(\underline{\theta} = (\theta_1, \theta_2, \ldots, \theta_m))$

$$W_{i,j} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log_+ f(\underline{X}, \underline{\theta})$$

Matriko

$$\underline{I}(\underline{\theta}) = \left(-E(W_{i,j})\right)_{i,j=1}^{m}$$

imenujemo Fisherjeva matrika informacije.

Komentarji:

(i) V eni dimenziji je $I(\underline{\theta})$ število odvisno od $\underline{\theta}$.

(ii) Če v primeru

največjeverjetnega $\underline{\theta}$ definiramo

$$Z_i = \frac{\partial}{\partial \theta_i} \log_+ f(\underline{X}, \underline{\theta})$$

dobimo, da je

$$E(W_{i,j}) = -\operatorname{cov}(Z_i, Z_j)$$

V splošnem imamo več parametrov. Enaki koraki kot za eno dimenzijo z več pisanja nam dajo

$$\sqrt{n}\,(\hat{\underline{\theta}} - \underline{\theta}_0) \xrightarrow{d} N(0, \underline{I}(\theta_0)^{-1})$$

<u>komentar</u>: Strogo formulacijo bom objavil na spletni strani.

Primer: Za Weibullovo gostoto je

$$\ell(\alpha, b \mid x) = \log \alpha - \alpha \log b + \alpha \log x - \left(\frac{x}{b}\right)^{\alpha}$$

Odvajamo

$$\frac{\partial^2 \ell}{\partial \alpha^2} = -\frac{1}{\alpha^2} - \left(\frac{x}{b}\right)^{\alpha} \log^2 \frac{x}{b}$$

$$= -\frac{1}{\alpha^2} - \frac{1}{\alpha^2}\left(\frac{x}{b}\right)^{\alpha} \log^2 \left(\frac{x}{b}\right)^{\alpha}$$

$$\frac{\partial^2 \ell}{\partial \alpha \partial b} = -\frac{1}{b} + \frac{1}{b}\left(\frac{x}{b}\right)^{\alpha}\left(1 + \log\left(\frac{x}{b}\right)^{\alpha}\right)$$

$$\frac{\partial^2 \ell}{\partial b^2} = -\frac{\alpha(1+\alpha)}{b^2}\left(\frac{x}{b}\right)^{\alpha} + \frac{\alpha}{b^2}$$

upoštevamo $Y = \left(\frac{X}{b}\right)^\alpha \sim \exp(1)$

Dobimo

$$I_{11}(\alpha, b) = \frac{1}{\alpha^2}\left(1 + E[Y \log^2 Y]\right)$$

$$I_{12}(\alpha, b) = -\frac{1}{b}\left(-1 + E[Y(1 + \log Y)]\right)$$

$$I_{22}(\alpha, b) = \frac{\alpha(\alpha+1)}{b^2} E(Y) - \frac{\alpha}{b^2}$$

Sledi

$$\underline{I}(\alpha, b) = \begin{pmatrix} 1/\alpha^2(1 + \Gamma''(2)) , & * \\ -\frac{1}{b}\Gamma'(2) , & \frac{\alpha^2}{b^2} \end{pmatrix}$$

<u>Sklep</u>:

$$\sqrt{n}\left((\hat{\alpha}, \hat{b}) - (\alpha_0, b_0)\right)$$

$$\sim N(0, I^{-1}(\alpha_0, b_0))$$

V končni fazi zamenjamo neznana α_0, b_0 z ocenama $\hat{\alpha}_0$ in $\hat{b}_0$.

S tem lahko rečemo

$$\widehat{MSE(\hat{\alpha})} = \frac{I_{11}^{-1}(\hat{\alpha}, \hat{b})}{n}$$

in

$$\widehat{MSE(\hat{\alpha})} = \frac{I_{22}^{-1}(\hat{\alpha}, \hat{b})}{n}.$$

Ti oceni in dejstvo, da sta $\hat{\alpha}$ in $\hat{b}$ aproksimativno normalni nam omogoča izjave o točnost vzorčnih cenilk.

Zaključne opombe

(i) Metoda največjega verjetja je najpogostejša, ne pa edina. Je pa najbolj uporabna in prožna metoda.

(ii) Cenilke po metodi največjega verjetja niso nujno nepristranske. Če je $\underline{\hat{\theta}}_n$ cenilka na osnovi n opazovanih vrednosti, potem

$$P(|\underline{\hat{\theta}}_n - \underline{\theta}_0| > \varepsilon) \to 0 \text{ , ko } n \to \infty$$

za vsak $\varepsilon > 0$. V oznakah $\underline{\hat{\theta}}_n \xrightarrow{P} \underline{\theta}_0$. Rečemo, da je cenilka dosledna.

4. Preizkušanje domnev

4.1. Primeri in terminologija

Primer: Recimo, da želimo ugotoviti ali je ruletni cilinder „pošten". To bi pomenilo, da so vsi izidi enako verjetni, zaporedni izidi pa med sabo neodvisni. Za preizkus potrebujemo podatke. Recimo, da imamo n izidov $x_1, x_2, \ldots, x_n$. Matematično je predpostavka, da so $x_1, x_2, \ldots, x_n$ nastale kot slučajne spremenljivke. Privzeli bomo, da so te slučajne spremenljivke $X_1, X_2, \ldots, X_n$ med sabo neodvisne, sicer pa imajo vse enako porazdelitev z možnimi vrednostmi v $\{0, 1, 2, \ldots, 36\}$.

Predpostavljamo $P(X_i = l) =: p_l$ za $l = 0, 1, \ldots, 36$;

Kako ugotoviti ali je ruletni cilinder pošten?

<u>Vprašanje</u>: Ali bomo lahko trdili, da je, z gotovostjo na podlagi naključnih izidov? Ne!

Označimo z $N_i(n)$ število pojavitev izida i v n igrah. Vemo, da je

$$N_i(n) \sim Bin(n, p_i)$$

za $i = 0, 1, \ldots, 36$;

Po tem, kar vemo o binomski porazdelitvi je

$$E[N_i(n)] = np_i$$

in

$$var(N_i(n)) = np_i(1-p_i)$$

Centralni limitni izrek pove, da lahko pričakujemo, da bo $N_i(n) \approx np_i \mp 2.56 \times \sqrt{np_i(1-p_i)}$ z verjetnostjo 0.99. To velja za fiksen i, možnih izidov pa je 37. V grobem lahko rečemo, da za vsak izid i pričakujemo, da bo $N_i(n)$ „nekje okrog" pričakovane vrednosti, vendar le za nekaj standardnih odklonov stran.

Ideja: Na nek način moramo „izmeriti", koliko se dejanski podatki razlikujejo od tega, kar bi pričakovali od poštenega cilindra.

Tipično v statistiki velikost odklonov merimo s standardnimi odkloni in odklone merimo s kvadrati. To je vodilo angleškega statistika Karla Pearsona (1857-1936), da je definiral slučajno spremenljivko

$$\chi^2(n) = \sum_{i=0}^{36} \frac{(N_i(n) - np_i)^2}{n \cdot p_i}$$

Ta količina nekako „meri",
kako daleč so izidi od
„pričakovanih", če privzamemo,
da so verjetnosti $p_0, p_1, \ldots, p_{36}$.
V konkretnem primeru je
$p_0 = p_1 = \quad = p_{36}$.

<u>Komentar</u>: V imenovalcu niso
variance sl. spr. $N_i(n)$. Spustimo
jih iz matematičnih razlogov, da
ima $\chi^2(n)$ porazdelitev, ki jo
znamo opisati in aproksimirati.

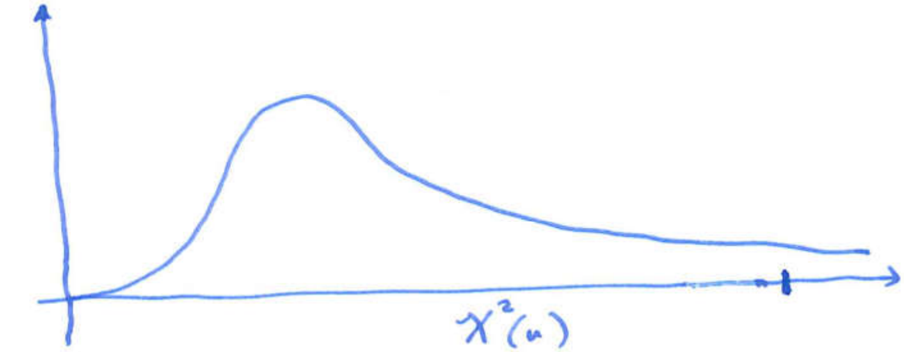
Manjka nam še naslednje:
kdaj je ta mera razhajanja
„prevelika", da bi še verjeli,
da je $p_0 = p_1 = \quad = p_{36}$.

Oglejmo si simulacije. Zavedati se moramo, da je $\chi^2(n)$ slučajna spremenljivka. Iz poglavja o centralnem limitnem izreku vemo, da

$$\chi^2(n) \xrightarrow{d} \chi^2(36) = \Gamma(18, \tfrac{1}{2})$$

To pomeni, da ima za „končne" n $\chi^2(n)$ aproksimativno $\chi^2(36)$ porazdelitev.

Slika:



Statistični software da

$$P(\chi^2(n) \leq 50.99) \doteq 0.95$$

$$P(\chi^2(n) \leq 58.62) \doteq 0.99$$

$$P(\chi^2(n) \leq 67{,}99) \doteq 0.999$$

$$P(\chi^2(n) \leq 76.36) \doteq 0.9999$$

Sklep: Aproksimativna porazdelitev nam pomaga, da presenimo ali in koliko je „mera razhajanja" prevelika. Kdaj bomo smatrali, da je verjetnost tako velike $\chi^2(n)$ prevelika, pa je stvar presoje. Tipično se bomo odločili za nek prag c_α in rekli, da je razhajanje preveliko, če $\chi^2(n) \geq c_\alpha$

Nekaj terminologije in definicij

(i) Vedno bomo privzeli statistični model. Rekli bomo, da so podatki $\underline{x}$ nastali kot slučajni vektor $\underline{X}$ s gostoto $f_{\underline{X}}(\underline{x}, \underline{\theta})$ ali v diskretnem primeru porazdelitvijo $P_{\underline{\theta}}(\underline{X} = \underline{x})$. Parameter $\underline{\theta}$ bo iz neke množice $\underline{\theta} \in \Theta$.

<u>Primer</u>: V primeru ruletnega cilindra je $\underline{X} = (X_1, X_2, \ldots, X_n)$, kjer so komponente neodvisne in je parameter $\underline{\theta} = \underline{p} \in \Delta$ z $\Delta = \{\underline{p} : p_i \geq 0, \sum_i p_i = 1\}$.

(ii) Radi bi preizkusili trditev o parametru $\underline{\theta}$. Tipično ko to trditev $\underline{\theta} \in \Theta_0 \subseteq \Theta$. Tej trditvi bomo rekli ničelna domneva in označili

$$H_0 : \underline{\theta} \in \Theta_0 .$$

Alternativa je alternativna domneva

$$H_1 : \underline{\theta} \in \Theta \setminus \Theta_0$$

Primer: V primeru ruletnega cilindra je Θ_0 singleton $\{(1/37, \ldots, 1/37)\}$.

(iii) Meri razkoraja, torej slučajni spremenljivki, ki meri razkoraj med tem, kaj pričakujemo in kar smo videli v podatkih, rečemo testna statistika.

Tipično je to enorazsežna slučajna spremenljivka, ki jo bomo označili s T.

Primer: Za vrelni cilinder je testna statistika enaka $\chi^2(n)$.

(iv) Izbrati si moramo, kaj bo za nas „majhna verjetnost". Tej izbiri rečemo stopnja tveganja $\alpha \in (0,1)$.

Komentar: Zgodovinsko so tipične izbire $\alpha = 0.05$, 0.01 in 0.001.

(v) Mera razhajanja T bo „prevelika", če bo padla v kritično območje C_α.

Pri tem poskrbimo, da bo za $\underline{\theta} \in \Theta_0$

$$P_{\underline{\theta}}(T \in C_\alpha) \leq \alpha.$$

<u>Komentar</u>: če je Θ_0 singelton, lahko zahtevamo tudi enakost

$$P_{\underline{\theta}_0}(T = C_\alpha) = \alpha.$$

<u>Primer</u>: Za vaterni cilinder in $\chi^2(n)$ je $C_\alpha = [c_\alpha, \infty)$ z

$$P(\chi^2(36) \geq c_\alpha) = \alpha.$$

(vi) Če je $T \in C_\alpha$ rečemo, da ničelno domnevo zavrnemo pri stopnji tveganja α.

Komentar: z zavrnitvijo nismo dokončno potrdili H_0, le po najboljših močeh smo se odločili.

(vii) Funkciji

$$F(\underline{\theta}) = P_{\underline{\theta}}(T \in C_\alpha)$$

rečemo moč testa. Za $\theta \notin \Theta_0$ nam pove, s kolikšno verjetnostjo se bomo odločili prav v tem primeru

(viii) Če zavrnemo H_0 pa ta drži zagrešimo napako I vrste. Če ne zavrnemo H_0 pa ta ne drži, smo zagrešili napako II vrste. Prvo verjetnost poskušamo nadzorovati, da je $\approx \alpha$, drugo verjetnost, torej verjetnost napake II vrste pa je moč testa.

<u>Primer</u>: Recimo, da so podatki $x_1, x_2, \ldots, x_n$ nastali kot neodvisne, enako porazdeljene slučajne spremenljivke $X_1, X_2, \ldots, X_n$

z $X_i \sim N(\mu, \sigma^2)$. Vprašamo se

$$H_0: \quad \mu = \mu_0$$

$$H_1: \quad \mu \neq \mu_0$$

V tem primeru je

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 > 0\}$$

$$\Theta_0 = \{(\mu_0, \sigma^2) : \sigma^2 > 0\}.$$

Vemo, da je $\overline{X}$ nepristranska cenilka μ in

$$S^2 = \frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X})^2$$

nepristranska cenilka σ^2.

Po ideji, da razdalje v statistiki merimo s standardnimi odkloni definiramo T tako, da

razdaljo med $\overline{X}$ in μ_0

izmerimo s standardnim

odklonom $\overline{X}$, ki je $\frac{\sigma}{\sqrt{n}}$.

Vendar σ ne poznamo, zato

vzamemo cenilko $\hat{\sigma}^2$.

Torej bi definirali

$$T = \frac{\overline{X} - \mu_0}{\sqrt{\hat{\sigma}^2/n}}$$

$$= \frac{\sqrt{n}\,(\overline{X} - \mu_0)}{\sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (X_k - \overline{X})^2}}$$

Vemo, da sta $\overline{X}$ in S^2

neodvisni in je T

porazdeljena po t_{n-1} -

porazdelitvi, če je $\mu = \mu_0$.

Še en koTček terminologije

(ix) Če je Θ_0 singelton $\underline{\theta}_0$, potem rečemo porazdelitvi T <u>testna porazdelitev</u>.

Enako velja, če je porazdelitev T enaka za vse $\underline{\theta} \in \Theta_0$.

V zadnjem primeru porazdelitev T ni odvisna od σ^2 in je za vsak (μ_0, σ^2) enaka.

Statistični software nam za vsak $\alpha \in (0,1)$ da c_α tako da je $P_{\mu_0, \sigma^2}(|T| \geq c_\alpha) = \alpha$.

Kritično območje je torej oblike $C_\alpha = (-\infty, -c_\alpha] \cup [c_\alpha, \infty)$.

Komentar: Trditev o porazdelitvi velja le, če je $\mu = \mu_0$, ker je v tem primeru

$$\sqrt{n}(\overline{X} - \mu_0) \sim N(0, \sigma^2)$$

Moč testa je odvisna od μ in σ kot funkcija

$$F(\mu, \sigma) = P\left(\left|\frac{\sqrt{n}(\overline{X} - \mu_0)}{\left(\frac{1}{n-1}\sum_{k=1}^{n}(X_k - \overline{X})^2\right)^{1/2}}\right| \geq c_\alpha\right)$$

za $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Moč je težno izračunljiva.

Komentar: Zgornji preizkus domneve je izumil William Sealy Gossett (1876-1937) znan pod vzdevkom Student.

(xi) Ko imamo podatke $x_1, x_2, \ldots, x_n$, statistični model in testno statistiko T, lahko izračunamo dejansko vrednost t. Če H_0 zavrnemo, če je t prevelika ali $|t|$ prevelika, lahko izračunamo

$$P_{H_0}(T \geq t)$$

ali

$$P_{H_0}(|T| \geq t).$$

Tej verjetnosti rečemo p-vrednost. H_0 zavrnemo, če je p-vrednost manjša od α.

<u>Komentar</u>: p-vrednost je eden od najbolj zlorabljanih konceptov v statistiki.

Kako pa v splošnem pridemo do testnih statistik T. Kakšna splošna metoda?

4.2 Test s kvocientom verjetij in Wilksov izrek

Kot smo videli, je tipična situacija

$$H_0 : \underline{\theta} \in \Theta_0 \quad \text{proti} \quad H_1 : \underline{\theta} \in \Theta \setminus \Theta_0 .$$

Recimo, da imajo podatki gostoto $f(\underline{x}, \underline{\theta})$. Funkcijo

$$\underline{\theta} \longmapsto f(\underline{x}, \underline{\theta})$$

za fiksen $\underline{x}$ interpretiramo kot „verjetje". Ideja:

Lahko izračunamo pri fiksnem $\underline{x}$

$$\max_{\underline{\theta} \in \Theta_0} f(\underline{x}, \underline{\theta}) \quad \text{ali} \quad \max_{\underline{\theta} \in \Theta} f(\underline{x}, \theta).$$

Zapišemo drugače:

$$\max_{\underline{\theta} \in \Theta_0} L(\underline{\theta} \mid \underline{x}) \quad \text{in} \quad \max_{\underline{\theta} \in \Theta} L(\underline{\theta} \mid \underline{x}).$$

Če je drugi maksimum „dosti večji" od prvega bi to pomenilo boljši $\underline{\theta} \in \Theta \setminus \Theta_0$, torej dokazno gradivo proti ničelni domnevi.

Definicija: Kvocient

$$\Lambda = \frac{\max_{\underline{\theta} \in \Theta} L(\underline{\theta} \mid \underline{x})}{\max_{\theta \in \Theta_0} L(\underline{\theta} \mid \underline{x})}$$

imenujemo Wilksova lambda statistika.

Komentarja:

(i) pisali smo max, čeprav vnaprej ni očitno, da bo tudi dosežen. Preverjali bomo v konkretnih situacijah.

(ii) Statistika Λ je odvisna samo od podatkov (v okviru predpostavljenega statističnega modela). To dejstvo je pomembno, saj lahko odločamo o H_0 samo na podlagi podatkov.

Primer: Naj bodo $X_1, X_2, \ldots, X_n$ normalne z porazdelitvijo $N(\mu, \sigma^2)$ in neodvisne in $x_1, x_2, \ldots, x_n$ vzorec. V tem primeru je

$$L(\mu, \sigma^2 \mid \underline{x}) = \prod_{k=1}^{n} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2\sigma^2}(x_k-\mu)^2}$$

Testiramo $H_0: \mu = 0$ proti $H_1: \mu \neq 0$. Torej je

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma > 0\}$$

in

$$\Theta_0 = \{(0, \sigma^2) : \sigma > 0\}$$

Za izračun maksimuma logaritmiramo

$$\ell(\mu, \sigma^2 \mid \underline{x}) = \frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2\sigma^2} \sum_{k=1}^{n} (x_k - \mu)^2$$

Izračunamo parcialna odvoda

$$\frac{\partial \ell}{\partial \mu} = \frac{1}{2\sigma^2} \cdot 2 \cdot \sum_{k=1}^{n} (x_k - \mu) = 0$$

$$\Rightarrow \mu = \overline{x}$$

in

$$\frac{\partial \ell}{\partial \sigma} = -\frac{n}{2} + \frac{1}{\sigma^3} \sum_{k=1}^{n} (x_k - \mu)^2 = 0 \Rightarrow$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} (x_k - \overline{x})^2$$

Če je $\mu = 0$, potem maksimiziramo

$$l(0, \sigma^2 \mid \underline{x}) = \frac{n}{2} \log 2\pi - n \log \sigma - \frac{1}{2\sigma^2} \sum_{k=1}^{n} x_k^2$$

po σ. Odvajamo in sledi

$$-\frac{n}{\sigma} + \frac{1}{\sigma^3} \sum_{k=1}^{n} x_k^2 = 0 \quad \Rightarrow$$

$$\tilde{\sigma}^2 = \frac{1}{n} \sum_{k=1}^{n} x_k^2$$

Po ideji bomo zavrnili H_0, če bo Λ „prevelika" oziroma, če bo $\log \Lambda$ preveliki...

<u>Definicija</u>: Količini

$$\lambda = 2 \log \Lambda$$

rečemo Wilksova logaritemska lambda statistika.

Izračunamo

$$\ell(\hat{\mu}, \hat{\sigma}^2 \mid \underline{x})$$

$$= \frac{n}{2}\log 2\pi - n \log \hat{\sigma} - \frac{1}{2\hat{\sigma}^2} \sum_{k=1}^{n} (\underline{x_k} - \overline{x})^2$$

$$= \frac{n}{2} \log 2\pi - n \log \hat{\sigma} - \frac{n}{2}$$

i u

$$\ell(0, \tilde{\sigma}^2 \mid \underline{x}) = \frac{n}{2} \log 2\pi - n \log \tilde{\sigma} - \frac{n}{2}$$

Slijedi

$$\lambda = -2 \log \Lambda$$

$$= 2\Big\{ \frac{n}{2} \log 2\pi - n \log \hat{\sigma} - \frac{n}{2} - \frac{n}{2} \log 2\pi + n \log \tilde{\sigma} + \frac{n}{2}$$

$$= n \log \tilde{\sigma} - n \log \hat{\sigma} \Big\}$$

Ničelno domnevo zavrnemo, če bo $\lambda \geq \lambda_0$ za nek λ_0. To pomeni

$$2n \log \frac{\hat{\sigma}^2}{\tilde{\sigma}^2} \geq \lambda_0 \quad \text{ali}$$

$$\log \frac{\hat{\sigma}^1}{\tilde{\sigma}} \leq -\frac{2n}{\lambda_0}$$

$$\frac{\hat{\sigma}^1}{\tilde{\sigma}} = \frac{\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^2}{\frac{1}{n}\sum_{k=1}^{n}x_k^2}$$

$$= \frac{\frac{1}{n}\sum_{k=1}^{n}x_k^2 - \bar{x}^2}{\frac{1}{n}\sum_{k=1}^{n}x_k^2}$$

$$= 1 - \frac{n\bar{x}^2}{\sum_{k=1}^{n}x_k^2}$$

Ničelno domnevo zavrnemo, če je

$$\frac{n\bar{x}^2}{\sum_{k=1}^{n}x_k^2} \geq c_\alpha$$

Po Cauchy-Schwarzu je

$$\left(\sum_{k=1}^{n} x_k \cdot 1\right)^2 \leq \left(\sum_{k=1}^{n} x_k^2\right) \cdot n \qquad \text{ali}$$

$$n\bar{x}^2 \leq \sum_{k=1}^{n} x_k^2 ,$$

zato bo $c_\alpha < 1$. Še enkrat prepišemo

$$n\bar{x}^2 \geq c_\alpha \sum_{k=1}^{n} x_k^2 \qquad \text{ali}$$

$$(1-c_\alpha)\,\bar{x}^2 \cdot n \geq c_\alpha\left(\sum_{k=1}^{n} x_k - n\bar{x}^2\right)$$

$$\geq c_\alpha \sum_{k=1}^{n} (x_k - \bar{x})^2$$

ali

$$\frac{\bar{x}^2 \cdot n}{\sum_{k=1}^{n} (x_k - \bar{x})^2} \geq \frac{c_\alpha}{1-c_\alpha} .$$

ali

$$\frac{|\bar{x}| \cdot \sqrt{n}}{\sqrt{\sum_{k=1}^{n} (x_k - \bar{x})^2}} \geq \sqrt{\frac{c_\alpha}{1-c_\alpha}} .$$

Torej je $\lambda \geq \lambda_0$ enako kot $|T| \geq t_\alpha$ za T statistiko iz prejšnjega razdelka. Poleg tega je porazdelitev λ neodvisna od σ. Ponovno smo predelali t-test!

<u>Komentar</u>: t_n-porazdelitev je definirana kot

$$T = \frac{Z}{\sqrt{\frac{1}{n} U}},$$ kjer sta

Z, U neodvisni in $U \sim \chi^2(n)$, $Z \sim N(0,1)$

Za velike n je $\frac{1}{n} \cdot U \approx \underline{1}$, torej bo $T^2 \sim Z^2 \sim \chi^2(1)$.

<u>Primer</u>: Vrnimo se k primeru rulete. Opazovane vrednosti $x_1, x_2, \ldots x_n$ so neodvisne in enako porazdeljene na $\{0, 1, \ldots, 36\}$.

Vemo, da je

$$\Theta = \{(p_0, \dots, p_{36}) : p_i \geq 0, \ \sum_{i=0}^{36} p_i = 1\}.$$

Recimo, da je $\Theta_0 = \{\underline{p}^0\}$. Vemo, da je

$$\ell(\underline{p} \mid \underline{x}) = \sum_{k=0}^{36} n_k \log p_k,$$

kjer je n_k število pojavitev izida k. Maksimum po Θ najdemo z Lagrangeovo metodo: maksimiziramo

$$\sum_{k=0}^{36} n_k \log p_k \quad \text{pri pogoju} \quad \sum_{k=0}^{36} p_k = 1.$$

Definiramo

$$F(\underline{p}) = \sum_{k=0}^{36} n_k \log p_k - \lambda \left(\sum_{k=0}^{36} p_k - 1 \right).$$

$$\frac{\partial F}{\partial p_k} = \frac{n_k}{p_k} - \lambda = 0$$

Sledi, da je $p_k = c \cdot n_k$ za neko konstanto. Ker se p_k seštejejo v 1, je

$$\hat{p}_k = \frac{n_k}{n}$$

Torej je

$$\ell(\hat{\underline{p}} \mid \underline{x}) = \sum_{k=0}^{36} n_k \cdot \log \frac{n_k}{n} .$$

Ker je Θ_0 singelton, nam ni treba računati maksimuma.

$$\ell(\underline{p}^0 \mid \underline{x}) = \sum_{k=0}^{36} n_k \log p_k^0 .$$

Sledi, da je

$$\lambda = 2 \sum_{k=0}^{36} n_k \cdot \log \frac{n_k}{n} - 2 \sum_{k=0}^{36} n_k \cdot \log p_k^0$$

Ali znamo kaj reči o porazdelitvi λ, če H_0 drži?

Funkcijo $f(x) = x \log(x/x_0)$ razvijmo okrog točke x_0.

$$f(x) \approx f(x_0) + f'(x_0)(x-x_0) + \frac{1}{2} f''(x_0)(x-x_0)^2$$

Računamo

$$f'(x_0) = 1$$

$$f''(x_0) = \frac{2}{x_0} - \frac{1}{x_0} = \frac{1}{x_0}$$

Prepišemo

$$\lambda = 2n \sum_{k=0}^{36} \frac{n_k}{n} \cdot \log \frac{n_k/n}{p_k^0}$$

$$= 2n \cdot \sum_{k=0}^{36} \hat{p}_k \log \hat{p}_k / p_k^0$$

Če H_0 drži, bo $\hat{p}_k$ blizu p_k^0 in bo Taylorjeva aproksimacija dobra.

Torej bo z $x_0 = p_k^\circ$

$$\lambda \approx 2n \left\{ \sum_{k=0}^{36} (\hat{p}_k - p_k^\circ) \right.$$

$$\left. + \frac{1}{2} \sum_{k=0}^{36} (\hat{p}_k - p_k^\circ)^2 / p_k^\circ \right\}$$

$$= \sum_{k=0}^{36} \frac{(n\hat{p}_k - np_k^\circ)^2}{n \cdot p_k^\circ}$$

$$= \sum_{k=0}^{36} \frac{(n_k - np_k^\circ)^2}{n \cdot p_k^\circ}$$

$$= \chi^2 \; !$$

Torej nam da kvocient verjetij χ^2 statistiko! Vemo tudi, da ima v primeru, ko H_0 drži, χ^2 aproksimativno $\chi^2(36)$ porazdelitev.

V obeh primerih smo ugotovili, da ima λ aproksimativno χ^2 porazdelitev. V obeh primerih nam je ideja vrnila statistike, ki jih potrebujemo, torej deluje dobro.

Kaj pa v splošnem? Oglejmo si preprost primer, ko je $\Theta = (a,b)$ in $\Theta_0 = \{\theta_0\}$ z $\theta_0 \in (a,b)$. Predpostavimo, da so opazovane vrednosti $x_1, x_2, \ldots, x_n$ vzorec iz porazdelitve z gostoto $f(x,\theta)$. Naj bo $\hat{\theta}$ ocena θ po metodi največjega verjetja. Po Taylorju bo

$$2\ell(\theta_0 \mid \underline{x}) - 2\ell(\hat\theta \mid \underline{x})$$

$$\doteq \underbrace{2\ell'(\hat\theta \mid \underline{x})}_{=0 \text{ po def.}} + 2\cdot\frac{1}{2}\,\ell''(\hat\theta \mid \underline{x})\,(\theta_0-\hat\theta)^2$$

$$= \frac{1}{n}\,\ell''(\hat\theta \mid \underline{x})\cdot\left[\sqrt{n}\,(\theta_0-\hat\theta)\right]^2$$

Vemo, da je v primeru, ko H_0 drži

$$\frac{1}{n}\,\ell''(\hat\theta \mid \underline{X})$$

$$\approx \frac{1}{n}\,\ell''(\theta_0 \mid \underline{X})$$

$$\approx -\operatorname{var}(Z_1) = E(W_1)$$

z oznakami z 3. poglavja.

Ampak $\sqrt{n}\,(\hat\theta-\theta_0)$ je približno normalno porazdeljena s parametroma 0 in $\frac{1}{\operatorname{var}(Z_1)}$.

To pomeni

$$\lambda \approx \left[\sqrt{n}\,(\hat{\theta} - \theta_0) \cdot \sqrt{\mathrm{var}(Z_1)}\right]^2$$

Ampak izraz v oglatih oklepajih ima približno standardizirano normalno porazdelitev. To pomeni, da ima λ aproksimativno $\chi^2(1)$ porazdelitev.

Ta grob račun nakazuje, da bi utegnila imeti λ v splošnem χ^2 porazdelitev.

Izrek 4.1 (Wilks) Naj bodo $x_1, x_2, \ldots x_n$ vzorec iz porazdelitve z gostoto $f(\underline{x}, \theta)$ z $\theta \in \Theta \subseteq \mathbb{R}^p$ odprta množica. Naj bo $g : \Theta \to \mathbb{R}^k$ zvezno parcialno odvedljiva z rang $Dg = k$ na Ω. Naj bo

$$H_0 : \; g(\underline{\theta}) = \underline{\psi}_0 \quad \text{in}$$

$$H_1 : \; g(\underline{\theta}) \neq \underline{\psi}_0 .$$

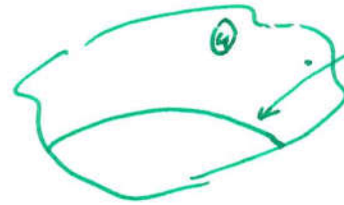
Naj bodo izpolnjeni vsi pogoji izreka o asimptotski normalnosti cenilk po metodi največjega verjetja. Potem velja

$$\lambda \xrightarrow{d} \chi^2(p-k), \text{ ko } n \to \infty .$$

Komentarji:

(i) $g(\underline{\theta}) = \underline{\psi}_0$ pomeni, da $\underline{\theta}$ leži na gladki ploskvi v Θ.

<u>Slika</u> :



$\{g(\underline{\theta}) = \underline{\psi}_0\} = \Theta_0$

(ii) Tipično lahko H_0 in H_1 preformuliramo v

$$H_0 : \underline{\theta} \in \Theta_0 \quad \text{in} \quad H_1 : \underline{\theta} \in \Theta \setminus \Theta_0 .$$

Tipično je Θ_0 neka podmnogoterost. Rečemo lahko, da je λ porazdeljena približno $\chi^2(r)$, kjer je $r = \dim \Theta - \dim \Theta_0$.

(iii) Implikacija izreka je, da aproks. porazdelitev λ ni odvisna od $\underline{\theta}_0 \in \Theta_0$ oziroma od $\underline{\theta}_0$, za katerega je $g(\underline{\theta}_0) = \underline{\psi}_0$. Zato je λ ustrezna statistika.

REGRESIJA

4.1. Uvodni primeri

Uvodni primeri so v elementarnih gradivih objavljenih na spletni strani predmeta.

4.2 Ocenjevanje parametrov, izrek Gauss-Markova

Iz primerov razberemo, da bomo privzeli, da so podatki $y_1, y_2, \ldots y_n$ nastali kot

$$Y_k = \alpha + \beta x_k + \varepsilon_k,$$

kjer so $x_1, x_2, \ldots, x_n$ konstante, za katere privzamemo, da so znane. Parametra α in β sta neznana, za slučajne spremenljivke $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ pa privzamemo, da je

$$E(\varepsilon_k) = 0, \quad \mathrm{var}(\varepsilon_k) = \sigma^2 \quad \text{za}$$

vse $k = 1, 2, \ldots, n$ in

$$\mathrm{cov}(\varepsilon_k, \varepsilon_l) = 0 \quad \text{za} \quad k \neq l.$$

<u>Komentar</u>: Zgornjim predpostavkam rečemo predpostavke standardne linearne regresije.

V matrični obliki lahko zapišemo:

$$\underline{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ & \\ 1 & x_n \end{pmatrix} \quad \underline{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

$$\underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$ S tem oznakami so

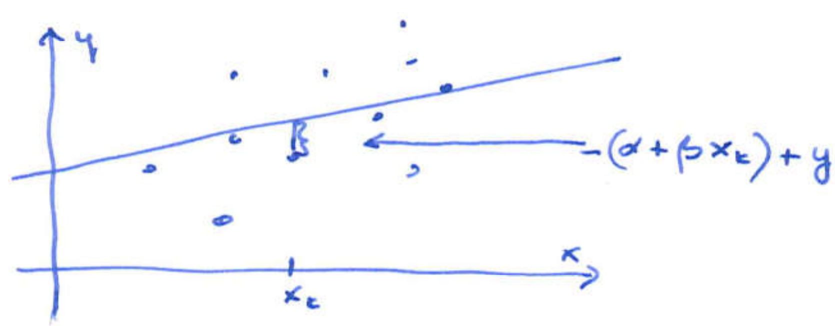
predpostavke $E(\underline{\varepsilon}) = 0$, $\mathrm{var}(\underline{\varepsilon}) = \sigma^2 \underline{I}$.

Predpostavljamo

$$\underline{Y} = \underline{X} \cdot \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \underline{\varepsilon}.$$

<u>Ideja ocenjevanja</u>:

<u>Slika</u>:



Vsaka točka ima navpično razdaljo do premice $\alpha + \beta x$.

Iščemo premico, ki se bo čim bolje „prilegala" točkam. Sledimo Gaussu in razdalje kvadriramo in seštejemo. Iščemo premico, za katero je vsota kvadratov razdalj najmanjša možna. Definiramo

$$s(\alpha,\beta) = \sum_{k=1}^{n} (y_k - \alpha - \beta x_k)^2.$$

Parcialno odvajamo po α in β in izenačimo z 0.

$$\frac{\partial s}{\partial \alpha} = -2 \sum_{k=1}^{n} (y_k - \alpha - \beta x_k) = 0$$

$$\frac{\partial s}{\partial \beta} = -2 \sum_{k=1}^{n} (y_k - \alpha - \beta x_k) x_k = 0$$

Dobimo linearni enačbi za α in β. Zapišimo v matrični obliki.

$$n\alpha + \beta \sum_{k=1}^{n} x_k = \sum_{k=1}^{n} y_k$$

$$\alpha \cdot \sum_{k=1}^{n} x_k + \beta \sum_{k=1}^{n} x_k^2 = \sum_{k=1}^{n} x_k y_k$$

ali

$$\underbrace{\begin{pmatrix} n & \sum_{k=1}^{n} x_k \\ \sum_{k=1}^{n} x_k & \sum_{k=1}^{n} x_k^2 \end{pmatrix}}_{\underline{X}^T \cdot \underline{X}} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \underbrace{\begin{pmatrix} \sum_{k=1}^{n} y_k \\ \sum_{k=1}^{n} x_k y_k \end{pmatrix}}_{\underline{X}^T \cdot y}$$

Predpostavimo, da ima $\underline{X}$ polni rang. Potem bo rešitev zgornje enačbe

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (\underline{X}^T \underline{X})^{-1} \cdot \underline{X}^T \cdot y$$

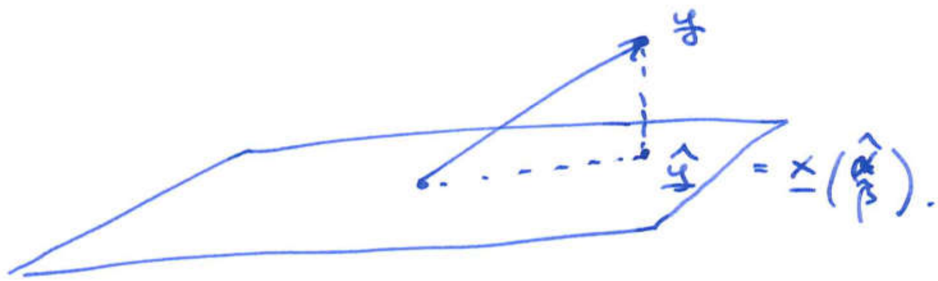
Rečemo, da sta $\hat{\alpha}$ in $\hat{\beta}$ oceni α in β po <u>metodi najmanjših kvadratov</u>. Opazimo še, da je

$$\underline{X} \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \underline{X} (\underline{X}^T \underline{X})^{-1} \underline{X}^T \cdot y$$

Matrika $\underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T$ je ortogonalna projekcija na podprostor v $\mathbb{R}^n$, ki ga napenjata stolpca $\underline{X}$. Torej sta $\hat{\alpha}$ in $\hat{\beta}$ enolični rešitvi enačbe

$$\underline{X}\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} = \underline{X}(\underline{X}^T\cdot\underline{X})^{-1}\cdot\underline{X}^T\cdot y = \hat{y}$$

Slika:



Če zamenjamo y z $\underline{Y}$, dobimo slučajni vektor $\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix}$, ki mu rečemo cenilka. Iz matričnega zapisa hitro sledi nekaj osnovnih trditev.

Če je $\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix} = (\underline{X}^T\underline{X})^{-1}\underline{X}^T\cdot\underline{Y}$ je

$$E\left[\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix}\right] = (\underline{X}^T\underline{X})^{-1}\underline{X}^T\cdot\underline{X}\cdot\begin{pmatrix}\alpha\\ \beta\end{pmatrix}$$
$$= \begin{pmatrix}\alpha\\ \beta\end{pmatrix}$$

Cenilka po metodi najmanjših kvadratov je nepristranska.

$$\mathrm{var}\left(\begin{pmatrix}\hat{\alpha}\\ \hat{\beta}\end{pmatrix}\right) = (\underline{X}^T\underline{X})^{-1}\underline{X}^T\underbrace{\mathrm{var}(\underline{Y})}_{\sigma^2 I}\underline{X}(\underline{X}^T\underline{X})^{-1}$$
$$= \sigma^2\cdot(\underline{X}^T\underline{X})^{-1}$$

Sledi torej

$$\mathrm{var}(\hat{\alpha}) = \sigma^2\cdot\frac{\sum_{k=1}^n x_k^2}{n\sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k\right)^2}$$

$$\mathrm{var}(\hat{\beta}) = \sigma^2\cdot\frac{n}{n\sum_{k=1}^n x_k^2 - \left(\sum_{k=1}^n x_k\right)^2}$$

Ostane še vprašanje, kako oceniti σ^2.

V splošnem dopuščamo več stolpcev v $\underline{X}$. Podatki so oblike

$$
\begin{matrix}
y_1 & x_{11} \cdots & x_{1m} \\
y_2 & x_{21} & \cdots x_{2m} \\
\vdots & & \\
y_n & x_{n1} & \cdots x_{nm}
\end{matrix}
$$

Privzamemo, da so x_{kl} fiksne konstante, $y_1, y_2 \dots y_n$ pa so „nastali" kot

$$Y_k = \beta_1 x_{k1} + \beta_2 x_{k2} + \dots + \beta_m x_{km} + \varepsilon_k .$$

<u>Opomba</u>: x_{k1} so lahko 1, ne zahtevamo pa tega nujno.

Pišemo matrično:

$$\underline{X} = \begin{pmatrix} x_{11} \cdots & x_{1m} \\ x_{21} & \cdots x_{2m} \\ \vdots & \vdots \\ x_{n1} \cdots & x_{nm} \end{pmatrix}, \quad \underline{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \underline{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix}, \quad \underline{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Predpostavke standardnega regresijskega modela zapišemo kot

$$\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$$

z

$$E(\underline{\varepsilon}) = 0 \quad \text{in} \quad \mathrm{var}(\underline{\varepsilon}) = \sigma^2 \cdot \underline{I}$$

<u>Opomba</u>: ves čas predpostavljamo, da je $\underline{X}$ matrika s polnim rangom.

Logika za ocenjevanje $\underline{\beta}$ je enaka kot prej. Iščemo $\beta_1, \beta_2, \ldots \beta_m$, da bo

$$S(\beta) = \sum_{k=1}^{n} (y_k - \beta_1 x_{k1} - \beta_2 x_{k2} - \cdots - \beta_m x_{km})^2$$

najmanjša možna.

Parcialno odvajamo in dobimo

$$\frac{\partial s}{\partial \beta_i} = -2 \sum_{k=1}^{n} (y_k - \beta_1 x_{k1} - \dots - \beta_m x_{km}) x_{ki}$$

Izenačimo z 0 in zapišemo v matrični obliki (pokrajšamo -2):

$$(\underline{X}^T \cdot \underline{X}) \cdot \underline{\beta} = \underline{X}^T \cdot \underline{y} .$$

Ocena β bo

$$\hat{\beta} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \cdot \underline{y} .$$

Kot cenilka bo

$$\hat{\beta} = (\underline{X}^T \underline{X})^{-1} \cdot \underline{X}^T \cdot \underline{Y}$$

<u>Izrek 5.1</u>: Naj bodo izpolnjene predpostavke standardnega linearnega regresijskega modela.

Cenilka po metodi najmanjših kvadratov je nepristranska z $\operatorname{var}(\hat{\beta}) = \sigma^2 \cdot (\underline{X}^T \underline{X})^{-1}$.

Dokaz: Enak kot prej.

Definicija: Vektorju $\hat{y} = \underline{X} \cdot \hat{\beta}$ rečemo vektor prilagojenih vrednosti. Vektorju $\hat{\underline{\varepsilon}} = y - \hat{y}$ rečemo vektor ostankov ali residualov.

Opomba: Tudi v več dimenzijah je $\hat{y}$ ortogonalna projekcija y na podprostor, ki ga napenjajo stolpci $\underline{X}$.

Oceniti moramo še σ^2.
Iz predpostavk bi načeloma sledilo $\frac{1}{n}\sum_{k=1}^{n} \varepsilon_k^2 \approx \sigma^2$.

Vendar ε_k ne poznamo, poznamo pa $\hat{\varepsilon}_k$. Izračunajmo

$$E\left(\sum_{k=1}^{n} \hat{\varepsilon}_k^2\right) = E\left[\underline{\hat{\varepsilon}}^T \cdot \hat{\varepsilon}\right]$$

$$= E\left[(\underline{Y} - \underline{X}\cdot\hat{\beta})^T(\underline{Y} - \underline{X}\cdot\hat{\beta})\right]$$

$$= E\left[(\underline{Y} - \underline{X}\cdot(\underline{X}^T\underline{X})^{-1}\cdot\underline{X}^T Y)^T(Y - \underline{X}\cdot(\underline{X}^T\underline{X})^{-1}\underline{X}^T Y)\right]$$

$$= E\left[\underline{Y}^T\cdot\underbrace{(\underline{I} - \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T)^T}_{\text{Idempotentna matrika}}(\underline{I} - \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T)Y\right]$$

$$= E\left[\underline{Y}^T(\underline{I} - \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T)\underline{Y}\right]$$

$$= E\left[(\underline{Y} - \underline{X}\beta)^T(\underline{I} - \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T)(\underline{Y} - \underline{X}\beta)\right]$$

ker je $(\underline{I} - \underline{X}(\underline{X}^T X)^{-1}\underline{X}^T)\underline{X} = 0$

$$= E\left[\underline{\varepsilon}^T(\underline{I} - \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T)\underline{\varepsilon}\right]$$

$$= E\left[\mathrm{Sl}((\underline{I} - \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T)\underline{\varepsilon}\cdot\underline{\varepsilon}^T\right]$$

$$= \operatorname{sl}\left(\left(\underline{I} - \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T\right) \underbrace{E(\underline{\varepsilon}\,\underline{\varepsilon}^T)}_{\operatorname{var}(\underline{\varepsilon}) = \sigma^2 \cdot \underline{I}}\right)$$

(linearnost)

$$= \operatorname{sl}\left(\sigma^2 \cdot \left(\underline{I} - \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T\right)\right)$$

$$= \sigma^2 \operatorname{sl}\left(\left(\underline{I} - \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}\right)\right)$$

Za idempotentne matrike je sled enaka rangu, $\underline{I} - \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}$ pa je ortogonalna projekcija na komplement podprostora, ki ga napenjajo stolpci $\underline{X}$. Dimenzija tega podprostora je $n-m$.

Torej je

$$E\left(\sum_{k=1}^{n} \hat{\varepsilon}_k^2\right) = \sigma^2 (n-m)$$

<u>Izrek 4.2</u>: Cenilka

$$\hat{\sigma}^2 = \frac{1}{n-m} \sum_{k=1}^{n} \hat{\varepsilon}_k^2$$

je nepristranska cenilka σ^2.

Dokaz: Smo že.

Ali lahko kaj rečemo o tem, koliko je „dobra" cenilka $\hat{\beta}$? Ali morda obstaja boljša cenilka? Pri takih vprašanjih moramo definirati razred „tekmic". Glede na to, da je cenilka po metodi najmanjših kvadratov linearna funkcija $\underline{Y}$ se zdi smiselno, da najprej pogledamo linearne tekmice, potem pa še zahtevamo nepristranskost. Primerjali bomo $\hat{\beta} = (\underline{X}^T\underline{X})^{-1}\cdot\underline{X}^T\cdot\underline{Y}$ in tekmice oblike $\tilde{\beta} = \underline{L}\cdot\underline{Y}$, kjer je $\underline{L}\,(n\times m)$ matrika (ki je lahko odvisna od $\underline{X}$. Ker zahtevamo nepristranskost, mora veljati

$$E(\tilde{\beta}) = E(\underline{L}\cdot\underline{Y}) = \underline{L}\cdot\underline{X}\,\beta = \beta$$

To pomeni, da je $\underline{L} \cdot \underline{X} = \underline{I}$.

<u>Izrek 4.3</u> (Gauss-Markov) Cenilka po metodi najmanjših kvadratov je najboljša med vsemi linearnimi nepristranskimi cenilkami $\underline{\beta}$.

<u>Komentar</u>: Kaj pomeni najboljša? Pomen bomo razjasnili po dokazu.

<u>Dokaz</u>: Naj bo $\tilde{\underline{\beta}} = \underline{L} \cdot \underline{Y}$ nepristranska.

Računamo

$$
\begin{aligned}
\operatorname{var}(\tilde{\beta}) &= \operatorname{var}(\tilde{\underline{\beta}} - \hat{\underline{\beta}} + \hat{\underline{\beta}}) \\
&= \operatorname{var}(\hat{\beta}) + \operatorname{var}(\tilde{\beta} - \hat{\beta}) \\
&\qquad + \operatorname{cov}(\tilde{\beta} - \hat{\beta}, \hat{\beta}) \\
&\qquad\qquad + \operatorname{cov}(\hat{\beta}, \tilde{\beta} - \hat{\beta})
\end{aligned}
$$

Posebej računamo

$$\operatorname{cov}(\tilde\beta - \hat\beta, \hat\beta)$$

$$= \operatorname{cov}\left(\underline{L}\,\underline{Y} - (\underline{X}^T\underline{X})^{-1}\underline{X}^T Y,\ (\underline{X}^T\underline{X})^{-1}\underline{X}^T Y\right)$$

$$= \left(\underline{L} - (\underline{X}^T\underline{X})^{-1}\underline{X}^T\right)\underbrace{\operatorname{var}(\underline{Y})}_{\sigma^2 \cdot \underline{I}}\,\underline{X}\cdot(\underline{X}^T\underline{X})^{-1}$$

$$= \sigma^2\left(\underline{L} - (\underline{X}^T\underline{X})^{-1}\underline{X}^T\right)\underline{X}\,(\underline{X}^T\underline{X})^{-1}$$

$$= \sigma^2\left(\underline{L}\,\underline{X} - \underline{I}\right)(\underline{X}^T\underline{X})^{-1}$$

$$= \underline{0}$$

Podobno je $\operatorname{cov}(\hat\beta, \tilde\beta - \hat\beta) = 0$.

Sledi

$$\operatorname{var}(\tilde\beta) = \operatorname{var}(\hat\beta) + \operatorname{var}(\tilde\beta - \hat\beta).$$

Vse matrike so pozitivno semi-definitne.

<u>Definicija</u>: Za pozitivno semidefinitne matriki $\underline{A}$ in $\underline{B}$ rečemo, da je $\underline{A} \geq \underline{B}$, če je $\underline{A} - \underline{B}$ pozitivno semidefinitna matrika.

V tem smislu lahko zapišemo

$$\mathrm{var}(\tilde{\beta}) \geq \mathrm{var}(\hat{\beta})$$

Med drugim to pomeni:

(i) $\hat{\beta}_i$ je najboljša nepristranska linearna cenilka β_i.

(ii) če hočemo nepristransko oceniti $\gamma = \underline{a}^T \cdot \underline{\beta}$ za nek $\underline{a}$, je

$$\mathrm{var}(\underline{a}^T \hat{\beta}) = \underline{a}^T \cdot \mathrm{var}(\hat{\underline{\beta}})\, \underline{a}$$

$$\geq \underline{a}^T \mathrm{var}(\tilde{\beta}) \cdot \underline{a}$$

Torej je $\hat{\gamma} = \underline{a}^T \hat{\beta}$ najboljša cenilka $\gamma = \underline{a}^T \cdot \underline{\beta}$.

(iii) Velja tudi, da $\hat{\beta}$ minimizira

$$E[(\tilde{\beta}-\beta)^T(\tilde{\beta}-\beta)]$$

med vsemi nepristranskimi linearnimi $\tilde{\beta}$. Računamo

$$E[(\tilde{\beta}-\beta)^T(\tilde{\beta}-\beta)]$$

$$= E[(\tilde{\beta}-\beta+\hat{\beta}-\hat{\beta})^T(\tilde{\beta}-\beta+\hat{\beta}-\hat{\beta})]$$

$$= E[(\tilde{\beta}-\hat{\beta})^T(\tilde{\beta}-\hat{\beta})]$$
$$+ E[(\hat{\beta}-\beta)^T(\hat{\beta}-\beta)]$$
$$+ E[(\tilde{\beta}-\hat{\beta})^T(\hat{\beta}-\beta)]$$
$$+ E[(\hat{\beta}-\beta)^T(\tilde{\beta}-\hat{\beta})]$$

Ampak

$$E[(\tilde{\beta}-\hat{\beta})^T(\hat{\beta}-\beta)]$$

$$= E[\mathrm{Sl}((\hat{\beta}-\beta)(\tilde{\beta}-\hat{\beta})^T)]$$

$$= \mathrm{Sl}[\mathrm{cov}(\hat{\beta},\tilde{\beta}-\hat{\beta})]$$

$$= 0$$

Podobno dobimo za drugo pričakovano vrednost, zato je

$$E[(\tilde{\underline{\beta}}-\underline{\beta})^T(\tilde{\underline{\beta}}-\underline{\beta})]$$

$$\geq E[(\hat{\underline{\beta}}-\underline{\beta})^T(\hat{\underline{\beta}}-\underline{\beta})].$$

<u>Definicija</u>: Vektorju $\hat{\underline{y}} = \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T\hat{\underline{\beta}}$ rečemo vektor prilagojenih vrednosti (angl. fitted values).

Izrek Gauss-Markova lahko nekoliko posplošimo. Predpostavimo

$$\underline{Y} = \underline{X}\beta + \underline{\varepsilon}$$

z $\quad E(\underline{\varepsilon}) = 0 \quad , \quad var(\underline{\varepsilon}) = \sigma^2 \cdot \underline{\Sigma} \quad ,$

kjer je $\underline{\Sigma}$ znana pozitivno-definitna matrika. Velja

$$\underbrace{\underline{\Sigma}^{-1/2}\underline{Y}}_{\tilde{Y}} = \underbrace{\underline{\Sigma}^{-1/2}\underline{X}}_{\tilde{\underline{X}}}\beta + \underbrace{\underline{\Sigma}^{-1/2}\underline{\varepsilon}}_{\tilde{\underline{\varepsilon}}}$$

$$\begin{aligned} var(\tilde{\underline{\varepsilon}}) &= var(\underline{\Sigma}^{-1/2}\underline{\varepsilon}) \\ &= \underline{\Sigma}^{-1/2}\sigma^2 \cdot \underline{\Sigma}\,\underline{\Sigma}^{-1/2} \\ &= \sigma^2 \underline{I}. \end{aligned}$$

Ocena β je $\quad \hat{\beta} = (\tilde{\underline{X}}^T\tilde{\underline{X}})^{-1}\tilde{X}^T\tilde{Y}$

ali zapisano drugače

$$\hat{\beta} = (\underline{X}^T\underline{\Sigma}^{-1}\underline{X})^{-1}\underline{X}\,\underline{\Sigma}^{-1}\underline{Y}$$

Izrek 4.3a (Gauss-Markov splošen)

Naj bo $\underline{Y} = \underline{X}\beta + \underline{\varepsilon}$ z $E(\underline{\varepsilon}) = 0$ in $\mathrm{var}(\underline{\varepsilon}) = \sigma^2 \underline{\Sigma}$, kjer je $\underline{\Sigma}$ znana matrika in σ^2 neznan parameter.

Cenilka $\underline{\hat{\beta}} = (\underline{X}^T \underline{\Sigma}^{-1} \underline{X})^{-1} \underline{X}^T \underline{\Sigma}^{-1} \underline{Y}$ je najboljša nepristranska cenilka β, ki je linearna.

Dokaz: Cenilka $\hat{\beta}$ je linearna in nepristranska. Recimo, da bi obstajala boljša cenilka $\underline{\tilde{\beta}} = \underline{L} \cdot \underline{Y}$. Potem bi $\underline{\tilde{\beta}}$ bila boljša cenilka tudi za $\underline{\tilde{Y}} = \underline{\tilde{X}}\beta + \underline{\tilde{\varepsilon}}$, kar pa vemo da ni.

Opomba: Lahko tudi na enak način kot v dokazu izreka 4.3 dokažemo, da je $\mathrm{var}(\tilde{\beta}) \geq \mathrm{var}(\hat{\beta})$.

Primer: Recimo, da je

$$Y_{kl} = \alpha + \beta x_{kl} + \varepsilon_k + \varepsilon_{k,l}$$

za $k = 1, 2, \ldots, K$ in za vsak k je $l = 1, 2, \ldots, L_k$. Privzemamo, da so $\varepsilon_k, \varepsilon_{k,l}$ vse nekorelirane z $E(\varepsilon_k) = 0$, $E(\varepsilon_{k,l}) = 0$, $\mathrm{var}(\varepsilon_k) = \sigma^2$ in $\mathrm{var}(\varepsilon_{k,l}) = \tau^2$. Privzamemo, da je $\rho = \tau^2/\sigma^2$ znano število.

Zapišemo

$$\begin{pmatrix} Y_{11} \\ Y_{12} \\ \vdots \\ Y_{1L_1} \\ Y_{21} \\ \vdots \\ Y_{2L_2} \\ \vdots \\ Y_{K1} \\ Y_{KL_K} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} \\ 1 & \\ \vdots & \\ 1 & x_{1L_1} \\ & \vdots \\ & \vdots \\ 1 & x_{K1} \\ \vdots & \vdots \\ 1 & x_{KL_K} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \underbrace{\begin{pmatrix} \varepsilon_1 + \varepsilon_{11} \\ \vdots \\ \varepsilon_1 + \varepsilon_{1L_1} \\ \vdots \\ \vdots \\ \varepsilon_K + \varepsilon_{K1} \\ \vdots \\ \varepsilon_K + \varepsilon_{KL_K} \end{pmatrix}}_{\eta}$$

Ugotovimo

$$\operatorname{var}(\eta) = \begin{pmatrix} \begin{matrix} \sigma^2+\tau^2 & \sigma^2 & \cdots & \sigma^2 \\ \sigma^2 & \sigma^2+\tau^2 & & \\ & & \ddots & \\ \sigma^2 & & & \sigma^2+\tau^2 \end{matrix} \Big\} L_1 & & & 0 \\ & \boxed{} \} L_2 & & \\ & & \boxed{} & \\ 0 & & & \boxed{} \} L_k \end{pmatrix}$$

$$= \sigma^2 \underline{\Sigma}$$

Izpostavimo σ^2 in škatlice postanejo oblike

$$\sigma^2 \begin{pmatrix} 1+\rho & 1 & \cdots & 1 \\ 1 & 1+\rho & \cdots & 1 \\ \vdots & & & \\ 1 & 1 & \cdots & 1+\rho \end{pmatrix} = \sigma^2 (\rho \underline{I} + \underline{1} \cdot \underline{1}^T)$$

V smislu izreka 4.3 a bo

$$\underline{\Sigma} = \begin{pmatrix} \boxed{} & & & \\ & \boxed{} & & \\ & & \boxed{} & \\ & & & \boxed{} \end{pmatrix}$$

s škatlicami oblike $\rho \underline{I} + \underline{1} \cdot \underline{1}^T$

Izpostavimo σ^2 in škatlice postanejo oblike

$$\sigma^2 \begin{pmatrix} 1+\rho & 1 & 1 & & 1 \\ 1 & 1+\rho & & & 1 \\ & & & & \\ 1 & 1 & & & 1+\rho \end{pmatrix} = \sigma^2(\rho I + \underline{1}\cdot\underline{1}^T)$$

V smislu izreka 4.3 a bo

$$\underline{\Sigma} = \begin{pmatrix} \boxed{\Sigma_1} & & & & \\ & \boxed{} & & & \\ & & \boxed{} & & \\ & & & \ddots & \\ & & & & \boxed{} \end{pmatrix}$$

s škatlicami oblike $\rho I + \underline{1}\underline{1}^T$.

Potrebovali bomo inverze škatlic po diagonali. Predpostavimo, da je inverz oblike $\left(\frac{1}{\rho} I + c\,\underline{1}\cdot\underline{1}^T\right)$.

Zmnožimo in sledi

$$(\rho I + \underline{1}\cdot\underline{1}^T)\left(\frac{1}{\rho} I + c\,\underline{1}\cdot\underline{1}^T\right)$$

$$= I + \frac{1}{\rho}\underline{1}\underline{1}^T + c\rho\,\underline{1}\cdot\underline{1}^T + c\cdot n\,\underline{1}\cdot\underline{1}^T$$

Če želimo, da bo rezultat I, mora biti

$$\frac{1}{\rho} + c\rho + c \cdot n = 0 \Rightarrow$$

$$c = -\frac{1}{\rho(n+\rho)}$$

S tem lahko izračunamo

$$\underline{X}^T \underline{\Sigma}^{-1} \underline{X} \quad \text{in} \quad \underline{X}^T \underline{\Sigma}^{-1} y.$$

Označimo $\underline{x}_k = \begin{pmatrix} x_{k1} \\ x_{k2} \\ \vdots \\ x_{kL_k} \end{pmatrix}$, $\underline{y}_k = \begin{pmatrix} y_{k1} \\ \vdots \\ y_{kL_k} \end{pmatrix}$

Označimo $c_k = -\frac{1}{\rho(L_k+\rho)}$

$$\underline{X}^T \underline{\Sigma}^{-1} \underline{X} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

Velja

$$a_{11} = \underline{1}^T \underline{\Sigma}^{-1} \underline{1}$$

$$= \text{vsota vseh elementov } \underline{\Sigma}^{-1}$$

$$= \frac{1}{\rho} \sum_{k=1}^{K} L_k + \sum_{k=1}^{K} c_k L_k^2$$

$$a_{12} = a_{21} = \underline{1}^T \underline{\Sigma}^{-1} \underline{x}$$

$$= \sum_{k=1}^{K} \underline{1}^T \underline{\Sigma}_k^{-1} \underline{x}_k$$

$$= \sum_{k=1}^{k} \underline{1}^T \left(\frac{1}{\rho} I + c_k \underline{1} \cdot \underline{1}^T \right) \underline{x}_k$$

$$= \sum_{k=1}^{k} \left(\frac{1}{\rho} \sum_{\ell=1}^{L_k} x_{k\ell} \right)$$

$$+ \sum_{k=1}^{k} c_k L_k \cdot \sum_{\ell=1}^{L_k} x_{k\ell}$$

$$a_{22} = \sum_{k=1}^{K} \underline{x}_k^T \underline{\Sigma}_k^{-1} \underline{x}_k$$

$$= \sum_{k=1}^{k} \underline{x}_k^T \left(\frac{1}{\rho} I + c_k \cdot \underline{1} \cdot 1^T \right) \underline{x}_k$$

$$= \sum_{k=1}^{k} \frac{1}{\rho} \sum_{\ell=1}^{L_k} x_{k\ell}^2$$

$$+ \sum_{k=1}^{K} c_k \left(\sum_{\ell=1}^{L_k} x_{k\ell} \right)^2$$

Označimo

$$\underline{x}^T \underline{\Sigma}^{-1} y = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

Velja

$$b_1 = \underline{1}^T \underline{\Sigma}^{-1} y$$

$$= \sum_{k=1}^{K} \underline{1}^T \left(\frac{1}{\rho} I + c_k \underline{1} \cdot \underline{1}^T \right) \underline{y_k}$$

$$= \frac{1}{\rho} \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} y_{k\ell} + \sum_{k=1}^{K} c_k L_k \cdot \sum_{\ell=1}^{L_k} y_{k\ell}$$

$$b_2 = \underline{x}^T \underline{\Sigma}^{-1} y$$

$$= \sum_{k=1}^{K} \underline{x}_k^T \left(\frac{1}{\rho} I + c_k \underline{1} \underline{1}^T \right) \underline{y_k}$$

$$= \frac{1}{\rho} \sum_{k=1}^{K} \sum_{\ell=1}^{L_k} x_{k\ell} y_{k\ell}$$

$$+ \sum_{k=1}^{K} c_k \left(\sum_{\ell=1}^{L_k} x_{k\ell} \right) \left(\sum_{\ell=1}^{L_k} y_{k\ell} \right)$$

S tem je

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}^{-1} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}.$$

4.3 <u>Normalni $\underline{\varepsilon}$</u>

V standardnem modelu linearne regresije

$$\underline{Y} = \underline{X}\beta + \underline{\varepsilon}$$

lahko predpostavimo $\underline{\varepsilon} \sim N(\underline{0}, \sigma^2 \underline{I})$.

<u>Definicija</u>: Matriko $\underline{H} = \underline{X}(X^T \cdot \underline{X})^{-1}\underline{X}^T$ imenujemo projekcijska matrika.

Vemo, da je $\underline{H}$ idempotentna. Potem je tudi $\underline{I} - \underline{H}$ idempotentna.

Vemo, da je vektor

$$\begin{pmatrix} \hat{\beta} \\ \hat{\varepsilon} \end{pmatrix} = \begin{pmatrix} (\underline{X}^T\underline{X})^{-1}\underline{X}^T Y \\ (\underline{I} - \underline{H}) Y \end{pmatrix}$$

večrazsežen normalen. Za neodvisnost preverimo (vemo $(\underline{I} - \underline{H})\underline{X} = 0$)

$$\mathrm{cov}(\hat{\beta}, \hat{\varepsilon}) = \sigma^2 (\underline{X}^T\underline{X})^{-1}\underline{X}^T(\underline{I} - \underline{H}) = 0.$$

Posledično je $\hat{\beta}$ neodvisen od $\hat{\varepsilon}$. Poleg tega je

$$\hat{\sigma}^2 = \frac{1}{n-m} \sum_{k=1}^{n} \hat{\varepsilon}_k^2$$

$$= \frac{1}{n-m} \hat{\underline{\varepsilon}}^T \underline{\varepsilon}$$

$$= \frac{1}{n-m} \underline{\varepsilon}^T (\underline{I} - \underline{H}) \underline{\varepsilon}$$ (dokazali že prej)

$$\sim \frac{\sigma^2}{n-m} \cdot \chi^2(n-m)$$

Sledi, da je

$$T_i = \frac{\hat{\beta}_i - \beta_i}{\sqrt{c_{ii}}\ \hat{\sigma}} \sim t_{n-m}.$$

Pri tem je $\underline{C} = (\underline{X}^T \underline{X})^{-1}$ in je c_{ii} i-ti diagonalni element $\underline{C}$.

<u>Zakaj?</u> $$\frac{\hat{\beta}_i - \beta_i}{\sqrt{c_{ii}}} \sim N(0, \sigma^2)$$

Na podlagi tega lahko testiramo, recimo, ničelno domnevo

$$H_0: \beta_i = 0 \quad \text{proti} \quad H_1: \beta_i \neq 0.$$

Testna statistika $T_i = \dfrac{\hat{\beta}_i}{\sqrt{c_{ii}}\ \hat{\sigma}}$

ima v primeru, ko H_0 drži, t_{n-m} porazdelitev.

<u>Interpretacija</u>: Če H_0 ne zavrnemo, lahko rečemo, da i-ti stolpec v $\underline{X}$ ne prispeva k $E(\underline{Y})$, torej i-ta neodvisna spremenljivka „nima vpliva" na $\underline{Y}$.

Kako pa bi v splošnem preizkusili

$$H_0: \beta_{p+1} = \beta_{p+2} = \quad = \beta_n = 0.$$

Ideja: Uporabimo kvocient verjetij.

$$L(\beta, \sigma \mid \underline{X}, y)$$

$$= \frac{1}{(2\pi)^{n/2}\,\sigma^n}\, e^{-\frac{1}{2}\sum_{k=1}^{n}(y_k-(\underline{X}\beta)_k)^2/\sigma^2}$$

ali

$$\ell(\beta, \sigma \mid \underline{X}, y)$$

$$= \frac{n}{2}\log 2\pi - n\log\sigma$$

$$- \frac{1}{2\sigma^2}\,(y-\underline{X}\beta)^T(y-\underline{X}\beta).$$

Maksimum bo očitno dosežen za

$$\hat{\beta} = (\underline{X}^T\underline{X})^{-1}\underline{X}^T y \quad \text{in}$$

$$\hat{\sigma}^2 = \frac{1}{n}(y-\underline{X}\hat{\beta})^T(y-\underline{X}\hat{\beta}).$$

Označimo $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \\ \beta_{p+1} \\ \vdots \\ \beta_m \end{pmatrix}$

in podobno $\underline{X} = (\underbrace{\underline{X}_1}_{p} ; \underbrace{\underline{X}_2}_{m-p})$

Če se omejimo na $\beta_{p+1} = \dots = \beta_m = 0$

dobimo oceni

$$\tilde{\beta} = (\underline{X}_1^T \underline{X}_1)^{-1} \underline{X}_1^T \underline{y}$$

$$\tilde{\sigma}^2 = \frac{1}{n} (y - \underline{X}_1 \tilde{\beta})^T (y - \underline{X}_1 \tilde{\beta})$$

Vstavimo in dobimo

$$\lambda = -2n \log \hat{\sigma} + 2n \log \tilde{\sigma}$$

$$= n \log\left(\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} \right)$$

Ničelno domnevo zavrnemo, če je λ prevelik. Ampak to je isto kot izjava, da je kvocient $\frac{\tilde{\sigma}^2}{\hat{\sigma}^2}$ prevelik. Označimo

$$\underline{H} = \underline{X}(\underline{X}^T\underline{X})^{-1}\underline{X}^T, \quad \underline{H}_1 = \underline{X}_1(\underline{X}_1^T\underline{X}_1)^{-1}\underline{X}_1^T.$$

Velja, če H_0 drži,

$$\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} = \frac{\underline{\varepsilon}^T(\underline{I}-\underline{H}_1)\underline{\varepsilon}}{\underline{\varepsilon}^T(\underline{I}-\underline{H})\underline{\varepsilon}}$$

$$= \frac{\underline{\varepsilon}^T(\underline{I}-\underline{H})\underline{\varepsilon} + \underline{\varepsilon}^T(\underline{H}-\underline{H}_1)\underline{\varepsilon}}{\underline{\varepsilon}^T(\underline{I}-\underline{H})\underline{\varepsilon}}$$

Velja $\underline{H}_1\underline{H} = \underline{H}\cdot\underline{H}_1 = \underline{H}_1$, torej je

$$(\underline{H}-\underline{H}_1)^2 = \underline{H}^2 - 2\underline{H}\underline{H}_1 + \underline{H}_1^2$$

$$= \underline{H} - \underline{H}_1.$$

Matrika $\underline{H}-\underline{H}_1$ je idempotentna

Poleg tega je

$$\operatorname{cov}\left((\mathbf{I}-\mathbf{H})\mathbf{Y}, (\mathbf{H}-\mathbf{H}_1)\mathbf{Y}\right)$$

$$= (\mathbf{I}-\mathbf{H})\,\sigma^2\cdot\mathbf{I}\,(\mathbf{H}-\mathbf{H}_1)$$

$$= \sigma^2\left(\mathbf{H}-\mathbf{H}^2-\mathbf{H}_1+\mathbf{H}\cdot\mathbf{H}_1\right)$$

$$= 0$$

Torej je

$$\frac{\tilde{\sigma}^2}{\hat{\sigma}^2} = 1 + \frac{\boldsymbol{\varepsilon}^T(\mathbf{H}-\mathbf{H}_1)\boldsymbol{\varepsilon}}{\boldsymbol{\varepsilon}^T(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon}}$$

prevelik, če je kvocient na desni prevelik. Če je $\mathbf{X}\beta \in R(\mathbf{H}_1)$, je $\boldsymbol{\varepsilon}^T(\mathbf{H}-\mathbf{H}_1)\boldsymbol{\varepsilon} \sim \sigma^2\cdot\chi^2(m-p)$ in $\boldsymbol{\varepsilon}^T(\mathbf{I}-\mathbf{H})\boldsymbol{\varepsilon} \sim \sigma^2\chi^2(n-m)$, poleg tega pa sta števec in imenovalec neodvisna.

Kvocient lahko še množimo s konstanto. Definiramo

$$F = \frac{\underline{Y}^T(\underline{H}-\underline{H}_1)\underline{Y}/(m-p)}{\underline{Y}^T(\underline{I}-\underline{H})\underline{Y}/(n-m)}$$

$$\sim F_{m-p,\,n-m}$$

H_0 zavrnemo, če je F prevelik, kritično vrednost pa razberemo iz F-porazdelitve.

Opomba: Zgornjemu rečemo preizkus splošne linearne domneve.

Opomba. Če je prvi stolpec v $\underline{X}$ enak $\underline{1}$ in rečemo $\underline{X} = (\underline{1}; \underline{X}_1)$ potem količini

$$R^2 = \frac{\sum_{k=1}^{n} (\hat{y}_k - \bar{y})^2}{\sum_{k=1}^{n} (y_k - \bar{y})^2}$$

$$= \frac{\underline{y}^T (\underline{H} - \frac{1}{n} \underline{1}\,\underline{1}^T) \underline{y}}{\underline{y}^T (\underline{I} - \frac{1}{n} \underline{1}\,\underline{1}^T) \underline{y}}$$

imenujemo R^2 koeficient in ga interpretiramo kot mero napovedne moči modela.

Večji R^2 pomeni večjo napovedno moč modela.