Ime in priimek: _____ Vpisna št: | | | | | | | | |

Fakulteta za matematiko in fiziko

Oddelek za matematiko

Statistika

Pisni izpit

1. september 2017

Navodila

Pazljivo preberite besedilo naloge, preden se lotite reševanja. Nalog je 6, 5 rešenih nalog pa je že 100%. Na razpolago imate 2 uri.

Naloga	a.	b.	c.	d.	
1.			•	•	
2.				•	
3.					
4.			•	•	
5.				•	
6.			•	•	
Skupaj	•	•	•	•	

1. (20) Naj bodo $X_1, X_2, \ldots$ med sabo neodvisne slučajne spremenljivke s $P(X_i = 1) = P(X_i = -1) = 1/2$. Definiramo $S_0 = 0$ in $S_n = X_1 + \cdots + X_n$.

a. (10) Pokažite, da je za poljubno število c in $\lambda \in [0, \pi/2)$ velja

$$E\big[\cos(\lambda(S_{n+1} - c)) \mid X_0, \ldots, X_n\big] = \cos(\lambda(S_n - c)) \cos\lambda\,.$$

Namig: $S_{n+1} = S_n + X_{n+1}$ in adicijski izrek za kosinus.

Rešitev: Računamo

$$\begin{aligned}
&E\big[\cos(\lambda(S_{n+1} - c)) \mid X_0, \ldots, X_n\big] = \\
&\quad = E\big[\cos(\lambda(X_{n+1} + S_n - c)) \mid X_0, \ldots, X_n\big] \\
&\quad = E\big[\cos(\lambda X_{n+1}) \cos(\lambda(S_n - c)) - \sin(\lambda X_{n+1}) \sin(\lambda(S_n - c)) \mid X_0, \ldots, X_n\big] \\
&\quad = \cos(\lambda(S_n - c)) E(\cos(\lambda X_{n+1})) - \sin(\lambda(S_n - c)) E(\sin(\lambda X_{n+1})) \\
&\quad = \cos(\lambda(S_n - c)) \cos\lambda\,.
\end{aligned}$$

b. (10) Izračunajte $E\big[\cos(\lambda(S_n - c))\big]$.

Rešitev: Uporabimo pravilo, da je $E(Y) = E\left[E(Y \mid X_0, \ldots, X_n)\right]$. Iz prvega dela naloge sledi, da je

$$E\left[\cos(\lambda(S_{n+1} - c))\right] = \cos\lambda\, E\left[\cos(\lambda(S_n - c))\right]\,.$$

Iz te rekurzije takoj sledi, da je

$$E\left[\cos(\lambda(S_n - c))\right] = \cos(\lambda c) \cos^n \lambda\,.$$

2. (20) Populacija velikosti N naj bo za namene vzorčenja razdeljena v K stratumov velikosti $N_1, N_2, \ldots, N_K$. Označimo z μ in σ^2 populacijsko povprečje in populacijsko varianco. Označimo za $i = 1, 2, \ldots, K$ z μ_i populacijsko povprečje za i-ti stratum in z σ_i^2 populacijsko varianco za i-ti stratum. Predpostavite, da izberemo stratificirani vzorec, tako da bo za $i = 1, 2, \ldots, K$ velikost vzorca v i-tem stratumu enaka n_i. Označimo še $w_i = N_i/N$.

a. (5) Prepričajte se, da velja

$$\sigma^2 = \sum_{i=1}^{K} w_i \sigma_i^2 + \sum_{i=1}^{K} w_i (\mu_i - \mu)^2 \,.$$

Namig: Prepišite

$$N\sigma^2 = \sum_{i=1}^{K} \sum_{j=1}^{N_i} (y_{ij} - \mu)^2 = \sum_{i=1}^{K} \sum_{j=1}^{N_i} (y_{ij} - \mu_i + \mu_i - \mu)^2 \,,$$

kjer je y_{ij} vrednost spremenljivke za j-to enoto v i-tem stratumu.

Rešitev: Prvi način. *Računamo*

$$\begin{aligned} N\sigma^2 &= \sum_{i=1}^{K} \sum_{j=1}^{N_i} (y_{ij} - \mu)^2 \\ &= \sum_{i=1}^{K} \sum_{j=1}^{N_i} (y_{ij} - \mu_i + \mu_i - \mu)^2 \\ &= \sum_{i=1}^{K} \sum_{j=1}^{N_i} \left[(y_{ij} - \mu_i)^2 + 2(y_{ij} - \mu_i)(\mu_i - \mu) + (\mu_i - \mu)^2 \right] . \end{aligned}$$

Ker je

$$\sum_{j=1}^{N_i} (y_{ij} - \mu_i) = \sum_{j=1}^{N_i} y_{ij} - N_i \mu_i = 0 \,,$$

je

$$N\sigma^2 = \sum_{i=1}^{K} \left[N_i \sigma_i^2 + 0 + N_i (\mu_i - \mu)^2 \right] .$$

Obe strani enačbe delimo z N in enakost sledi.

Drugi način. *Definirajmo slučajno spremenljivko X kot vrednost dane spremenljivke na naključno izbranem elementu populacije, tako da so vsi elementi enako verjetni. Nadalje naj bo S stratum, ki mu pripada taisti naključno izbrani element. Tedaj je $E(X) = \mu$, $\mathrm{var}(X) = \sigma^2$ ter še $E(X|S = i) = \mu_i$, $\mathrm{var}(X|S = i) = \sigma_i^2$ in $P(S = i) = w_i$ za vse $i = 1, 2, \ldots, K$. Zdaj pa se spomnimo na razcep variance:*

$$\sigma^2 = \mathrm{var}(X) = E\big(\mathrm{var}(X|S)\big) + \mathrm{var}\big(E(X|S)\big) \,.$$

Velja

$$\begin{aligned} E\big(\mathrm{var}(X|S)\big) &= \sum_{i=1}^{K} P(S=i)\, E\big(\mathrm{var}(X|S) \;\big|\; S=i\big) \\ &= \sum_{i=1}^{K} P(S=i)\, E\big(\mathrm{var}(X|S=i)\big) \\ &= \sum_{i=1}^{K} w_i \sigma_i^2 \,. \end{aligned}$$

Slučajna spremenljivka $Y := E(X|S)$ pa je za vsak $i = 1, 2, \ldots, K$ na dogodku $\{S = i\}$, ki ima verjetnost w_i, enaka μ_i. Njena pričakovana vrednost je enaka $E\big(E(X|S)\big) = E(X) = \mu$, njena varianca pa je po definiciji enaka

$$\mathrm{var}(Y) = \mathrm{var}\big(E(X|S)\big) = \sum_{i=1}^{K} w_i (\mu_i - \mu)^2 \,.$$

Poberemo skupaj in dobimo zahtevano enakost.

b. (5) Recimo, da bi želeli na podlagi stratificiranega vzorca oceniti populacijsko varianco σ^2. Predlagajte cenilko.

Rešitev: Označimo z $\bar{y}_i$ vzorčno povprečje v i-tem stratumu in $\bar{y} = \sum_{i=1}^{K} w_i \bar{y}_i$ cenilko parametra μ. Glede na zgornjo formulo bi predlagali

$$\hat{\sigma}^2 = \sum_{i=1}^{K} w_i \hat{\sigma}_i^2 + \sum_{i=1}^{K} w_i (\bar{y}_i - \bar{y})^2 \,.$$

Pri tem je $\hat{\sigma}_i^2$ cenilka variance v i-tem stratumu.

c. (10) Je predlagana cenilka nepristranska?

Rešitev: Za $\hat{\sigma}_i^2$ lahko izberemo nepristranske cenilke. Vprašanje o nepristranskosti predlagane cenilke se prevede na vprašanje nepristranskosti drugega kosa formule. Opažena povprečja $\bar{y}_i$ in $\bar{y}$ nadomestimo s slučajnimi spremenljivkami $\bar{Y}_i$ in $\bar{Y}$ in računamo

$$\begin{aligned} E\left[\big(\bar{Y}_i - \bar{Y}\big)^2\right] &= E(\bar{Y}_i^2 - 2\bar{Y}_i\bar{Y} + \bar{Y}^2) \\ &= \mathrm{var}(\bar{Y}_i) + \mu_i^2 + \mathrm{var}(\bar{Y}) + \mu^2 - 2E(\bar{Y}_i\bar{Y}) \,. \end{aligned}$$

Zaradi neodvisnosti $\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_K$ *dobimo*

$$\begin{aligned} E(\bar{Y}_i \bar{Y}) &= \sum_{j=1}^{K} w_j \, E(\bar{Y}_i \bar{Y}_j) \\ &= \sum_{j=1, j\neq i}^{K} w_j \mu_i \mu_j + w_i \, E(\bar{Y}_i^2) \\ &= \sum_{j=1, j\neq i}^{K} w_j \mu_i \mu_j + w_i (\mathrm{var}(\bar{Y}_i) + \mu_i^2) \\ &= \sum_{j=1}^{K} \Big(w_j \mu_i \mu_j \Big) + w_i \, \mathrm{var}(\bar{Y}_i) \, . \end{aligned}$$

Računamo

$$\begin{aligned} &\sum_{i=1}^{K} w_i E\left[\left(\bar{Y}_i - \bar{Y} \right)^2 \right] = \\ &= \sum_{i=1}^{K} w_i \Bigg(\mathrm{var}(\bar{Y}_i) + \mu_i^2 + \mathrm{var}(\bar{Y}) + \mu^2 - 2 \Bigg(\sum_{j=1}^{K} (w_j \mu_i \mu_j) + w_i \mathrm{var}(\bar{Y}_i) \Bigg) \Bigg) \\ &= \sum_{i=1}^{K} w_i \mathrm{var}(\bar{Y}_i) + \sum_{i=1}^{K} w_i \mu_i^2 + \mathrm{var}(\bar{Y}) + \mu^2 - \\ &\qquad - 2 \, \mathrm{var}(\bar{Y}) - 2 \sum_{i=1}^{K} \sum_{j=1}^{K} w_i w_j \mu_i \mu_j \\ &= \sum_{i=1}^{K} w_i \mathrm{var}(\bar{Y}_i) + \sum_{i=1}^{K} w_i \mu_i^2 + \mathrm{var}(\bar{Y}) + \mu^2 - 2 \mathrm{var}(\bar{Y}) - 2\mu^2 \\ &= \sum_{i=1}^{K} w_i (\mu_i - \mu)^2 + \sum_{i=1}^{K} w_i \mathrm{var}(\bar{Y}_i) - \mathrm{var}(\bar{Y}) \, . \end{aligned}$$

Cenilka v splošnem ni nepristranska.

3. (20) Predpostavite, da so podatki $x_1, x_2, \ldots, x_n$ nastali kot med sabo neodvisne slučajne spremenljivke $X_1, X_2, \ldots, X_n$ z gostoto

$$f(x) = \frac{1}{\sqrt{2\pi x^3}}\, e^{-\frac{(1-\mu x)^2}{2x}}$$

za $x, \mu > 0$.

a. (5) Poiščite oceno parametra μ po metodi največjega verjetja.

Rešitev: Zapišemo logaritemsko funkcijo verjetja kot

$$\ell = \frac{n}{2}\log 2\pi - \frac{3}{2}\sum_{k=1}^{n}\log x_k - \sum_{k=1}^{n}\frac{(1-\mu x_k)^2}{2x_k}\,.$$

Odvajanje po μ nam da enačbo

$$\sum_{k=1}^{n}(1-\mu x_k) = 0\,.$$

Iskana cenilka je torej

$$\hat{\mu} = \frac{n}{x_1 + x_2 + \cdots + x_n} = \frac{1}{\bar{x}}\,.$$

b. (5) Ali lahko cenilko po metodi največjega verjetja popravite tako, da bo nepristranska? Kot znano privzemite naslednje:

- Vsota $X_1 + \cdots + X_n$ ima gostoto

$$f_n(x) = \frac{n}{\sqrt{2\pi x^3}}\, e^{-\frac{(n-\mu x)^2}{2x}}$$

za $x > 0$.

- Za $a, b > 0$ velja

$$\int_0^\infty x^{-5/2} e^{-ax-\frac{b}{x}}\,\mathrm{d}x = \frac{\sqrt{\pi}\big(1+2\sqrt{ab}\,\big)}{2b^{3/2}}\, e^{-2\sqrt{ab}}\,.$$

Rešitev: Naj ima X gostoto $f_n(x)$. Računamo

$$\begin{aligned}
E\left(\frac{n}{X}\right) &= n\int_0^\infty \frac{1}{x}\, f_n(x)\,\mathrm{d}x \\
&= n^2 \frac{e^{n\mu}}{\sqrt{2\pi}} \int_0^\infty x^{-5/2} e^{-\frac{\mu^2}{2}x - \frac{n^2}{2x}}\,\mathrm{d}x \\
&= n^2 \frac{e^{n\mu}}{\sqrt{2\pi}}\, \sqrt{2\pi}\, \frac{1+n\mu}{n^3}\, e^{-n\mu} \\
&= \mu + \frac{1}{n}\,.
\end{aligned}$$

Nepristranska cenilka bi bila torej

$$\tilde{\mu} = \frac{1}{\bar{X}} - \frac{1}{n}\,.$$

c. (5) Izračunajte varianco cenilke po metodi največjega verjetja. Kot znano privzemite, da za $a, b > 0$ velja

$$\int_0^\infty x^{-7/2} e^{-ax-\frac{b}{x}}\,\mathrm{d}x = \frac{\sqrt{\pi}\big(3+6\sqrt{ab}+4ab\big)}{4b^{5/2}}\, e^{-2\sqrt{ab}}\,.$$

Rešitev: Za slučajno spremenljivko X z gostoto $f_n(x)$ računamo

$$\begin{aligned} E\left(\frac{n^2}{X^2}\right) &= \int_0^\infty \frac{n^2}{x^2}\, f_n(x)\,\mathrm{d}x \\ &= n^3 \frac{e^{n\mu}}{\sqrt{2\pi}} \int_0^\infty x^{-7/2} e^{-\frac{\mu^2}{2}x-\frac{n^2}{2x}}\,\mathrm{d}x \\ &= n^3 \frac{e^{n\mu}}{\sqrt{2\pi}} \frac{\sqrt{2\pi}(3+3n\mu+n^2\mu^2)}{n^5}\, e^{-n\mu} \\ &= \frac{3}{n^2}+\frac{3\mu}{n}+\mu^2\,. \end{aligned}$$

Varianca je torej enaka

$$\mathrm{var}(\hat{\mu}) = E(\hat{\mu}^2) - \big(E(\hat{\mu})\big)^2 = \frac{\mu}{n}+\frac{2}{n^2}\,.$$

d. (5) Kakšno aproksimacijo za varianco dobimo za $\hat{\mu}$ z uporabo Fisherjeve informacije? Kot znano privzemite, da je

$$\int_0^\infty x^{-1/2} e^{-ax-\frac{b}{x}}\,\mathrm{d}x = \frac{\sqrt{\pi}}{\sqrt{a}}\, e^{-2\sqrt{ab}}\,.$$

Rešitev: Z odvajanjem dobimo, da za $n = 1$ velja

$$\ell'' = -x\,.$$

Sledi, da je

$$\begin{aligned} I(\mu) &= E(X) \\ &= \frac{e^{\mu}}{\sqrt{2\pi}} \int_0^\infty \frac{1}{\sqrt{x}}\, e^{-\frac{\mu^2 x}{2}-\frac{1}{2x}}\,\mathrm{d}x \\ &= \frac{e^{\mu}}{\sqrt{2\pi}} \cdot \sqrt{\frac{2\pi}{\mu^2}}\, e^{-\mu} \\ &= \frac{1}{\mu}\,. \end{aligned}$$

Aproksimacija variance po Fisherju je torej

$$\frac{\mu}{n}\,,$$

kar je vodilni člen v izrazu za pravo varianco.

4. (20) Gaussova gama porazdelitev je dana z gostoto

$$f(x,y) = \sqrt{\frac{2\lambda}{\pi}}\, y\, e^{-y}\, e^{-\frac{\lambda y(x-\mu)^2}{2}}\,.$$

za $-\infty < x < \infty$ in $y > 0$ ter $(\mu, \lambda) \in \mathbb{R} \times (0, \infty)$. Predpostavite, da so opaženi pari $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ nastali kot med seboj neodvisni pari slučajnih spremenljivk $(X_1, Y_1), \ldots, (X_n, Y_n)$ z gostoto $f(x, y)$. Preizkusiti želimo domnevo

$$H_0\colon \mu = 0 \qquad \text{proti} \qquad H_1\colon \mu \neq 0\,.$$

a. (10) Izračunajte oceni $\hat{\mu}$ in $\hat{\lambda}$ po metodi največjega verjetja v splošnem in še cenilko $\tilde{\lambda}$ za primer, ko je $\mu = 0$.

Rešitev: Logaritemska funkcija verjetja je

$$\ell = \frac{n}{2}\log\left(\frac{2\lambda}{\pi}\right) + \sum_{k=1}^{n}(\log y_k - y_k) - \frac{\lambda}{2}\sum_{k=1}^{n} y_k(x_k - \mu)^2\,.$$

Parcialne odvode izenačimo z 0 *in dobimo enačbi*

$$\frac{n}{2\lambda} - \frac{1}{2}\sum_{k=1}^{n} y_k(x_k - \mu)^2 = 0$$

in

$$\lambda \sum_{k=1}^{n} y_k(x_k - \mu) = 0\,.$$

Iz druge enačbe sledi, da je

$$\hat{\mu} = \frac{\sum_{k=1}^{n} x_k y_k}{\sum_{k=1}^{n} y_k}\,.$$

Vstavimo lahko v prvo enačbo in sledi

$$\hat{\lambda} = \frac{n}{\sum_{k=1}^{n} y_k(x_k - \hat{\mu})^2}\,.$$

V primeru, ko privzamemo $\mu = 0$*, je ustrezna cenilka določena prav s prej omenjeno prvo enačbo. Vanjo torej vstavimo* $\mu = 0$ *in dobimo cenilko*

$$\tilde{\lambda} = \frac{n}{\sum_{k=1}^{n} x_k^2 y_k}\,.$$

b. (10) Poiščite testno statistiko po metodi kvocienta verjetij in navedite njeno aproksimativno porazdelitev pri veljavnosti ničelne domneve H_0.

Rešitev: Testna statistika je enaka

$$\begin{aligned}\lambda &= 2\Big[\ell(\hat{\lambda}, \hat{\mu}|\mathbf{x}, \mathbf{y}) - \ell(\tilde{\lambda}, 0|\mathbf{x}, \mathbf{y})\Big] \\ &= n(\log\hat{\lambda} - \log\tilde{\lambda}) - \hat{\lambda}\sum_{k=1}^{n} y_k(x_k - \hat{\mu})^2 + \tilde{\lambda}\sum_{k=1}^{n} x_k^2 y_k\,.\end{aligned}$$

Toda iz enačb, iz katerih smo dobili cenilke, sledi

$$\hat{\lambda} \sum_{k=1}^{n} y_k (x_k - \hat{\mu})^2 = \tilde{\lambda} \sum_{k=1}^{n} x_k^2 y_k = n \,,$$

torej je kar

$$\lambda = n \log \frac{\hat{\lambda}}{\tilde{\lambda}} \,.$$

Po Wilksovem izreku je aproksimativna porazdelitev te statistike, če H_0 drži, enaka $\chi^2(1)$.

5. (20) Predpostavite regresijski model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}\,,$$

kjer je $E(\boldsymbol{\epsilon}) = 0$ in $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$, pri čemer je $\mathbf{V}$ obrnljiva matrika. Predpostavite, da je $\mathbf{X}$ oblike

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$$

z $\sum_{k=1}^n x_k = 0$, znana matrika $\mathbf{V}$ pa oblike

$$\mathbf{V} = (1-\rho)\mathbf{I} + \rho\mathbf{1}\mathbf{1}^T$$

z $\rho \neq -1/(n-1)$ in $\rho \neq 1$.

a. (5) Utemeljite, da je

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}$$

najboljša nepristranska linearna cenilka parametra $\boldsymbol{\beta}$.

Rešitev: Regresijsko enačbo pomnožimo na obeh straneh z $\mathbf{V}^{-1/2}$ ter označimo $\tilde{\mathbf{Y}} = \mathbf{V}^{-1/2}\mathbf{Y}$ in podobno za $\tilde{\mathbf{X}}$ in $\tilde{\boldsymbol{\epsilon}}$. S tem regresijski model prevedemo na standardno obliko in po izreku Gauss–Markova lahko zapišemo zgornjo cenilko, ki je najboljša.

b. (5) Pokažite, da je cenilka iz prvega dela naloge kar običajna cenilka, kot da bi bila $\mathbf{V}$ identiteta.

Namig: $\mathbf{V}^{-1}$ je oblike $a\mathbf{I} + b\mathbf{1}\mathbf{1}^T$.

Rešitev: Z množenjem dobimo

$$\mathbf{V}\left(a\mathbf{I} + b\mathbf{1}\mathbf{1}^T\right) = a(1-\rho)\mathbf{I} + a\rho\mathbf{1}\mathbf{1}^T + b(1-\rho)\mathbf{1}\mathbf{1}^T + b\rho\mathbf{1}\mathbf{1}^T\mathbf{1}\mathbf{1}^T\,.$$

Upoštevajoč, da je $\mathbf{1}^T\mathbf{1} = n$, poenostavimo v

$$a(1-\rho)\mathbf{I} + (a\rho + b(1-\rho) + bn\rho)\mathbf{1}\mathbf{1}^T\,.$$

Sledi, da je

$$\mathbf{V}^{-1} = \frac{1}{1-\rho}\,\mathbf{I} - \frac{\rho}{(1-\rho)(1+(n-1)\rho)}\,\mathbf{1}\mathbf{1}^T$$

(inverz obstaja, brž ko so izpolnjeni dani pogoji). Računamo

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} = \begin{pmatrix} an + bn^2 & 0 \\ 0 & a\sum_{k=1}^n x_k^2 \end{pmatrix}$$

in podobno

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y} = \begin{pmatrix} (a+nb)\sum_{k=1}^n Y_k \\ a\sum_{k=1}^n x_k Y_k \end{pmatrix}\,.$$

Sledi

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \bar{Y} \\ \sum_{k=1}^n x_k Y_k / \sum_{k=1}^n x_k^2 \end{pmatrix}.$$

Prav to cenilko bi dobili tudi v primeru, če bi bilo $\mathbf{V} = \mathbf{I}$*: cenilka pride neodvisna od* ρ*.*

Opomba. *To se je zgodilo zato, ker smo matriko* $\mathbf{I}$ *(poleg množenja z* $1-\rho$*) popravili za matriko, ki ima zalogo vrednosti matrike* $\mathbf{X}$ *oziroma njen ortogonalni komplement za invariantni podprostor (pri simetričnih matrikah je dovolj preveriti samo enega od omenjenih dveh).*

c. (10) Izračunajte

$$E\left(\sum_{k=1}^n \hat{\epsilon}_k^2\right).$$

Namig: za poljuben vektor $\mathbf{a}$ *je* $\mathbf{a}^T\mathbf{a} = \mathrm{Sl}(\mathbf{a}\mathbf{a}^T)$*, za matriki* $\mathbf{A}$ *in* $\mathbf{B}$ *pa velja* $\mathrm{Sl}(\mathbf{AB}) = \mathrm{Sl}(\mathbf{BA})$*.*

Rešitev: Računamo

$$\begin{aligned} E\left(\sum_{k=1}^n \hat{\epsilon}_k^2\right) &= E\left(\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}\right) \\ &= E\left(\mathrm{Sl}\left(\hat{\boldsymbol{\epsilon}}\hat{\boldsymbol{\epsilon}}^T\right)\right) \\ &= \mathrm{Sl}\left(\mathrm{var}(\hat{\boldsymbol{\epsilon}})\right) \\ &= \mathrm{Sl}\left(\sigma^2(\mathbf{I}-\mathbf{H})\mathbf{V}(\mathbf{I}-\mathbf{H})\right) \\ &= \mathrm{Sl}\left(\sigma^2\mathbf{V}(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})\right) \\ &= \mathrm{Sl}\left(\sigma^2\mathbf{V}(\mathbf{I}-\mathbf{H})\right) \\ &= \sigma^2\big(\mathrm{Sl}\left(\mathbf{V}\right) - \mathrm{Sl}(\mathbf{VH})\big), \end{aligned}$$

kjer je

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T.$$

Preprosto je videti, da je $\mathrm{Sl}\left(\mathbf{V}\right) = n$*. Nadalje je*

$$\begin{aligned} \mathrm{Sl}\left(\mathbf{VH}\right) &= \mathrm{Sl}\left((1-\rho)\mathbf{H} + \rho\mathbf{1}\mathbf{1}^T\mathbf{H}\right) \\ &= 2(1-\rho) + \rho\,\mathbf{1}^T\mathbf{H}\mathbf{1} \\ &= 2(1-\rho) + \rho\,\mathbf{1}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{1} \\ &= 2(1-\rho) + \rho\begin{pmatrix} n & 0 \end{pmatrix}\begin{pmatrix} 1/n & 0 \\ 0 & 1/\sum_{k=1}^n x_k^2 \end{pmatrix}\begin{pmatrix} n \\ 0 \end{pmatrix} \\ &= 2(1-\rho) + n\rho. \end{aligned}$$

Sledi

$$E\left(\sum_{k=1}^n \hat{\epsilon}_k^2\right) = (n-2)(1-\rho)\sigma^2.$$

6. (20) *Wilcoxonov*[1] *test* služi za preizkus domneve, da je porazdelitev določene statistične spremenljivke X simetrična okoli izhodišča (kar pomeni, da za poljubna $0 < a < b$ velja $P(-b < X < -a) = P(a < X < b)$).

Če so $X_1, X_2, \ldots, X_n$ opažene vrednosti, najprej po velikosti razvrstimo njihove absolutne vrednosti: naj bo $Y_1, Y_2, \ldots, Y_n$ taka preureditev opaženih vrednosti, da je $|Y_1| \leq |Y_2| \leq \cdots \leq |Y_n|$. Wilcoxonova statistika je enaka:

$$\operatorname{sgn}(Y_1) + 2\operatorname{sgn}(Y_2) + \cdots + n\operatorname{sgn}(Y_n)\,,$$

kjer je

$$\operatorname{sgn}(y) = \begin{cases} -1 & ;\ y < 0 \\ 0 & ;\ y = 0 \\ 1 & ;\ y > 0\,. \end{cases}$$

Privzemimo, da so opažene vrednosti $X_1, \ldots, X_n$ neodvisne in enako porazdeljene, njihova porazdelitev pa je zvezna. Tedaj so z verjetnostjo ena vse opažene vrednosti različne, kar pomeni, da je Wilcoxonova statistika nedvoumno definirana.

a. (10) Če alternativna domneva trdi, da za vse $0 < a < b$ velja $P(-b < X < -a) \leq P(a < X < b)$ in za vsaj eno izbiro velja $P(-b < X < -a) < P(a < X < b)$, je ničelno domnevo smiselno zavrniti, če je $W \geq w_\alpha$. Poiščite približno kritično vrednost w_α, ki pri dovolj velikem vzorcu ustreza dani stopnji tveganja α. Kot znano lahko privzamete, da za ta primer velja ustrezna posplošitev centralnega limitnega izreka.

   Pomoč:

$$1 + 2 + \cdots + n = \frac{n(n+1)}{2}\,, \quad 1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}\,.$$

   Rešitev: Ker gre za zvezno porazdelitev, so z verjetnostjo ena vse slučajne spremenljivke $X_1, \ldots, X_n$ in z njimi $Y_1, \ldots, Y_n$ različne od nič. Če velja ničelna domneva, so zaradi simetrije vse slučajne spremenljivke $\operatorname{sgn}(Y_1), \ldots, \operatorname{sgn}(Y_n)$ enake -1 z verjetnostjo $1/2$ in 1 z verjetnostjo $1/2$. Sledi, da za vse $k = 1, 2, \ldots, n$ velja $E\big(\operatorname{sgn}(Y_k)\big) = 0$ in $\operatorname{var}\big(\operatorname{sgn}(Y_k)\big) = 1$. Po centralnem limitnem izreku je Wilcoxonova testna statistika porazdeljena približno normalno s pričakovano vrednostjo 0 in varianco

$$1^2 + 2^2 + \cdots + n^2 = \frac{n(n+1)(2n+1)}{6}\,,$$

   torej je

$$P(W \geq w_\alpha) \approx 1 - \Phi\left(w_\alpha \sqrt{\frac{6}{n(n+1)(2n+1)}}\right)\,.$$

   Za kritično vrednost je torej smiselno postaviti

$$w_\alpha = \sqrt{\frac{n(n+1)(2n+1)}{6}}\,\Phi^{-1}(1-\alpha)\,.$$

[1] Frank Wilcoxon (1892–1965), ameriški kemik in statistik

b. (10) Naj bo $0 \leq \theta \leq 1$ in naj za vse $0 < a < b$ velja $P(-b < X < -a) : P(a < X < b) = (1-\theta) : \theta$. Približno določite moč testa za ta primer, če je vzorec dovolj velik. Ustrezne porazdelitve lahko (poleg posplošitve centralnega limitnega izreka) privzamete kot znane. V katerih primerih je sploh smiselno govoriti o moči testa?

Rešitev: Pri danih predpostavkah je plavzibilno, da je $P(\operatorname{sgn}(Y_k) = -1) = 1 - \theta$ *in* $P(\operatorname{sgn}(Y_k) = 1) = \theta$ *za vse* $k = 1, 2, \ldots, n$ *Torej za vse* $k = 1, 2, \ldots, n$ *velja*

$$E\big(\operatorname{sgn}(Y_i)\big) = 2\theta - 1 \quad \textit{in} \quad \operatorname{var}\big(\operatorname{sgn}(Y_i)\big) = 4\theta(1-\theta)\,,$$

torej je

$$E(W) = (2\theta - 1)\,\frac{n(n+1)}{2} \quad \textit{in} \quad \operatorname{var}(W) = 2\theta(1-\theta)\,\frac{n(n+1)(2n+1)}{3}$$

in iz centralnega limitnega izreka sledi, da je moč testa približno enaka

$$P(W \geq w_\alpha) \approx 1 - \Phi\left[\big(2w_\alpha - (2\theta-1)n(n+1)\big)\sqrt{\frac{3}{8\theta(1-\theta)n(n+1)(2n+1)}}\right].$$

O moči testa je smiselno govoriti pri $\theta > 1/2$*: takrat namreč velja alternativna domneva.*