Matematički fakultet u Beogradu

Matematička statistika

Ispit, 21. jun 2015

Ispit traje 180 minuta. Isključiti mobilne telefone. Nije dozvoljena upotreba beležaka ili drugih pomoćnih sredstava.

---

1. Neka je $\mathbf{X}$ uzorak obima n iz familije raspodela P_θ, gde je θ r-dimenzionalni parametar, $\theta \in \Theta \subset \mathbf{R}^r$.

[10] **a)** Definisati pojam dovoljne statistike. Zatim na konkretnom primeru objasniti ideju koja stoji iza ove definicije (zašto se zove dovoljna).

Rešenje: Statistika S je dovoljna za parametar θ ako uslovna raspodela za $\mathbf{X} \,|\, S$ ne zavisi od θ, ili ekvivalentno, ako verovatnoća $P(X \in B \,|\, S = s)$ ne zavisi od θ, za proizvoljni Borelov skup $B \in \mathbf{R}^n$ i proizvoljno $s \in \Theta$.

Kao mogući primer, posmatrajmo problem bacanja novčića n puta u cilju određivanja nepoznate verovatnoće p padanja pisma. Definišemo $X_i = 1$ ako u i-tom bacanju padne pismo i $X_i = 0$ ako padne grb. Tako dobijamo $\mathbf{X} = (X_1, \ldots, X_n)$, gde su X_i nezavisne slučajne promenljive sa Bernulijevom raspodelom sa parametrom p, koga treba oceniti. U ovom primeru, dovoljna statistika je $S = \sum X_i$, pri čemu je

$$(*) \qquad P(X_1 = x_1, \ldots, X_n = x_n \,|\, S = k) = \frac{1}{\binom{n}{k}} \cdot I_{\{\sum x_i = k\}}, \quad x_i \in \{0, 1\}, \quad k \in \{0, 1, \ldots, n\}.$$

Iz $(*)$ izlazi da znajući da je $S = k$, možemo Monte Karlo metodom da regenerišemo uzorak $(x_1, \ldots x_n)$ iz diskretne uniformne raspodele na skupu $x_1 + \cdots x_n = k$, $x_i \in \{0, 1\}$, što znači da ne moramo da registrujemo konkretne vrednosti x_i, nego samo njihov zbir, odnosno vrednost dovoljne statistike, koja se zato tako i zove. □

[5] **b)** Ako je statistika $T = T(\mathbf{X})$ nezavisna od dovoljne statistike $S(\mathbf{X})$, onda njena raspodela ne zavisi od θ. Dokazati.

Rešenje: Za proizvoljan merljiv skup u prostoru vrednosti statistike T imamo da je

$$P(T \in B) = P(T \in B \mid S) = P(X \in T^{-1}(B) \mid S).$$

Prva jednakost važi zato što su T i S nezavisne statistike. Kako raspodela za $X|S$ ne zavisi od θ (po definiciji dovoljne statistike), zaključujemo da ni raspodela za T ne zavisi od θ. □

[5] **c)** Napisati izraz kojim se definiše opšta eksponencijalna familija raspodela sa parametrom dimenzije r i koristeći se teoremom o faktorizaciji, naći dovoljnu statistiku. (Ne treba dokazivati teoremu o faktorizaciji, samo objasnite kako se koristi u ovom slučaju).

Rešenje: Opšti oblik eksponencijalne familije raspodela za slučaj kao u postavci zadatka na početku je da je

$$f(\mathbf{x} \,|\, \theta) = C(\theta) \cdot \exp\{\sum_{j=1}^{r} Q_j(\theta) T_j(\mathbf{x})\} h(\mathbf{x}),$$

gde je f gustina (ili zakon raspodele) za posmatranu familiju. Ovde je $\mathbf{x} \in \mathbf{R}^n$, $\theta \in \Theta \subset \mathbf{R}^r$.

Teorema o faktorizaciji kaže da je S dovoljna statistika za θ ako i samo ako postoje nenegativne funkcije g i h tako da je

$$f(\mathbf{x} \,|\, \theta) = g(S(\mathbf{x}), \theta) \cdot h(\mathbf{x}),$$

pri čemu g zavisi od S i θ ali ne od $\mathbf{x}$ direktno. Poređenjem sa definicijom eksponencijalne raspodele, imamo da je S dovoljna statistika ako i samo ako je

$$g(S(\mathbf{x}), \theta) = C(\theta) \cdot \exp\{\sum_{j=1}^{r} Q_j(\theta) T_j(\mathbf{x})\},$$

što je moguće u opštem slučaju ako i samo ako je $S(\mathbf{X}) = (T_1(\mathbf{X}), \ldots, T_r(\mathbf{X}))$. Ovo je potpun odgovor na postavljeno pitanje.

Mnogi studenti su posmatrali *samo slučaj kada je* $\mathbf{X}$ *nezavisan uzorak iz jednodimenzionalne eksponencijalne familije*, mada to nigde u formulaciji zadatka nije pretpostavljeno. Korektno rešenje u tom slučaju je sledeće:

U slučaju jednodimenzionalne eksponencijalne raspodele sa parametrom θ dimenzije r, gustina $f(x\,|\,\theta)$ za X je data istim izrazom kao gore (s tim što umesto vektora $\mathbf{x}$ imamo $x \in \mathbf{R}$, a ukoliko je $\mathbf{X}$ nezavisan uzorak obima n iz raspodele, onda imamo da je $\mathbf{X} = (X_1, \ldots, X_n)$ gde $X_i \sim f(\cdot\,|\,\theta)$ i nezavisne su, tako da je gustina za $\mathbf{X}$ data sa

$$f(\mathbf{x}\,|\,\theta) = C(\theta) \cdot \exp\{\sum_{j=1}^{r} Q_j(\theta) \sum_{i=1}^{n} T_j(x_i)\} h(\mathbf{x})$$

Na isti način kao u opštem slučaju, poređenjem sa uslovom iz teoreme o faktorizaciji, dobija se da je $S(\mathbf{X}) = (\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_r(X_i))$. □

[10] **d)** Neka je $\mathbf{X}$ uzorak obima n iz familije normalnih $\mathcal{N}(m, \sigma^2)$ raspodela. Naći jedinstvenu najbolju ocenu parametra σ (standardna devijacija).

Rešenje: Najpre, prema uputstvu koga ste dobili na ispitu, treba pokazati da postoji C tako da je

$$U = C \cdot \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (X_i - \hat{\mu})^2}$$

nepristrasna ocena parametra σ. Ovo je logičan izbor jer je statistika pod kvadratnim korenom u gornjem izrazu nepristrasna ocena parametra σ^2.

Polazeći od dobro poznate činjenice da slučajna promenljiva

$$Y = \frac{1}{\sigma^2} \sum_{k=1}^{n} (X_i - \hat{\mu})^2$$

ima raspodelu $\chi^2(n-1)$, sa konačnim matematematičkim očekivanjem $\mathrm{E}\, Y \ (= n)$, zaključujemo da je

$$\mathrm{E}\, U = \frac{\mathrm{E}\sqrt{Y}}{\sqrt{n-1}} \sigma,$$

tako da je $\mathrm{E}\, U = \sigma$ ako i samo ako je $C = \frac{\sqrt{n-1}}{\mathrm{E}\sqrt{Y}}$. Primetimo da $\mathrm{E}\sqrt{Y}$ postoji zbog toga što postoji $\mathrm{E}\, Y$, osim toga, $\mathrm{E}\sqrt{Y}$ nije nula, jer je $P(Y > 0) > 0$, tako da je C realan broj. Nije bilo potrebno dalje izračunavati C.

Dalje, možemo da iskoristimo poznatu činjenicu da je u ovom primeru dovoljna statistika za σ dvodimenzionalna (bez obzira da li je m poznato ili nije):

$$S = \left(\hat{\mu}, \sum_{k=1}^{n} (X_i - \hat{\mu})^2 \right),$$

ili neka od ekvivalentnih verzija, na primer $(\sum X_i, \sum X_i^2)$. S obzirom da normalna familija raspodela čini eksponencijalnu familiju i da su ispunjeni uslovi regularnosti, statistika S je kompletna dovoljna statistika.

Kako je $\mathrm{E}\, U = \sigma$, primenom teoreme Leman-Šefea nalazimo da je

$$V = \mathrm{E}\,(U\,|\,S)$$

jedinstvena nepristrasna ocena za σ najmanje varijanse. S druge strane, kako je U funkcija od S, imamo da je $V = U$, što je i odgovor na postavljeno pitanje. □

---

[10] **2.** Dat je slučajni vektor $\mathbf{X} = (X_1, \ldots, X_n)$ sa raspodelom iz familije raspodela sa gustinom $f(\mathbf{x}\,|\,\theta)$, gde je θ skalarni parametar, $\theta \in \Theta$. Neka je $\mathbf{Q}(\theta) = \frac{\partial}{\partial\theta} \log f(\mathbf{X}\,|\,\theta)$ i $I(\theta) = \mathrm{Var}\,(\mathbf{Q}(\theta))$ - informaciona funkcija. Pretpostavljamo da su ispunjeni uslovi regularnosti.

a) Dokazati da je $\mathrm{E}\,(\mathbf{Q}(\theta)) = 0$ i da je

[10]
$$I(\theta) = -\mathrm{E}_\theta \left(\frac{\partial^2}{\partial\theta^2} \log f(\mathbf{X}\,|\,\theta) \right) = -\mathrm{E}_\theta \frac{\partial}{\partial\theta} Q(\mathbf{X}\,|\,\theta)$$

Rešenje: Prva jednakost se dobija iz

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta} f(\mathbf{x}\,|\,\theta)\, \mathrm{d}\mathbf{x} = \int \frac{\frac{\partial}{\partial\theta} f(\mathbf{x}\,|\,\theta)}{f(\mathbf{x}\,|\,\theta)} f(\mathbf{x}\,|\,\theta)\, \mathrm{d}\mathbf{x} = \int \frac{\partial}{\partial\theta} \log f(\mathbf{x}\,|\,\theta) f(x\,|\,\theta)\, \mathrm{d}\mathbf{x} \\ &= \int Q(\mathbf{x}\,|\,\theta) f(\mathbf{x}\,|\,\theta)\, \mathrm{d}\mathbf{x} = \mathrm{E}_\theta Q(\mathbf{X}\,|\,\theta), \end{aligned}$$

pri čemu se integrali po oblasti u kojoj je $f(\mathbf{x}\,|\,\theta) > 0$.

Druga jednakost (uz skraćene oznake):

$$\begin{aligned} \mathrm{E}_\theta\left((\log f_\theta)'^2 + (\log f_\theta)''\right) &= \mathrm{E}\left(\frac{f_\theta'^2}{f_\theta^2} + \frac{f_\theta'' f_\theta - f_\theta'^2}{f_\theta^2}\right) \\ = \mathrm{E}\left(\frac{f_\theta'' f_\theta}{f_\theta^2}\right) = \int \frac{f_\theta'' f_\theta}{f_\theta^2} f_\theta \,\mathrm{d}\mathbf{x} &= \int f_\theta'' \,\mathrm{d}\mathbf{x} = \left(\int f_\theta \,\mathrm{d}\mathbf{x}\right)'' = 0, \end{aligned}$$

[5] **b)** Ako se umesto parametra θ familija raspodela parametrizuje pomoću parametra $\eta = h(\theta)$, gde je h monotona funkcija, dobija se familija gustina $g(\mathbf{x}\,|\,\theta)$. (Ovo često koristimo kod varijanse, kad uvodimo smenu $\eta = \theta^2$.) Ako odgovarajuće informacione funkcije označimo sa I_θ i I_η respektivno, izraziti $I_\theta(\theta)$ preko I_η.

Rešenje: $I_\theta(\theta) = (h'(\theta))^2 I_\eta(h(\theta))$. *Videti kompletno rešenje u fotokopiranom tekstu beležaka sa predavanja.* □

[5] **c)** Definisati Jeffreys-ovu apriornu raspodelu i objasniti ideju koja stoji iza definicije.

Rešenje: Jeffreys-ova apriorna raspodela definisana je gustinom $\pi(\theta) = C \cdot \sqrt{I(\theta)}$, gde je $C = \left(\int_\Theta \sqrt{I(\theta)}\,\mathrm{d}\theta\right)^{-1}$. Ako integral ne postoji ili je beskonačan, uzimamo da je $C = 1$, pri čemu se dobija neprava apriorna raspodela.

Motivacija za ovu definiciju je sledeća: Traži se univerzalno pravilo po kome bi se dodeljivale gustine koje su invarijantne u odnosu na monotone transformacije, u sledećem smislu: Ako je $\pi_1(\theta)$ apriorna gustina za θ i $\pi_2(\eta)$ apriorna gustina za $\eta = h(\theta)$ gde je h striktno monotona i diferencijabilna funkcija, onda bi trebalo da važi da je $\pi_2(\eta) = \pi_1(h^{-1}(\eta)) \cdot |u'(\eta)|$, gde je $u = h^{-1}$, ili, ekvivalentno, $\pi_1(\theta) = \pi_2(h(\theta) \cdot |h'(\theta)|$. Jedno očigledno rešenje je da se za apriornu gustinu uzima uniformna raspodela, ali to nije moguće za slučaj kad parametar uzima vrednosti na beskonačnim intervalima. Ako se uzme gustina definisana pod c), traženi uslov biće ispunjen, tako da je Jeffreys-ova raspodela ustvari neka vrsta generalizacije uniformne raspodele. To i jeste primarni motiv, jer je uniformna raspodela neinformativna. Za više detalja oko izvođenja navedenih osobina videti u fotokopiranom tekstu beležaka sa predavanja. □

[10] **d)** Neka X ima gustinu raspodele $f(x\,|\,\theta) = ax^{a-1}\theta^{-a}$, $x \in (0,\theta)$, $\theta > 0$, i neka θ ima nepravu apriornu raspodelu sa (nepravom) gustinom $\pi(\theta) = \theta^{-1}$. Naći Bajesovu ocenu parametra θ pri funkciji gubitka $L(\theta, \hat\theta) = |\theta - \hat\theta|$.

Rešenje: U ovom zadatku možemo da pretpostavimo da imamo nezavisan uzorak $\mathbf{X}$ obima n iz raspodele sa zadatom gustinom. S obzirom da se služimo istim metodama za $n > 1$ kao i za $n = 1$, dajemo detaljno rešenje samo za $n = 1$. U tom slučaju imamo da je zajednička "gustina" za (X, θ) data sa

$$f(x,\theta) = f(x\,|\,\theta)\pi(\theta) = a\frac{x^{a-1}}{\theta^{a+1}} I_{\{0<x<\theta\}} I_{\{\theta>0\}}.$$

Bezuslovna "gustina" za X je data sa (za $x > 0$):

$$m(x) = \int f(x,\theta)\,\mathrm{d}\theta = ax^{a-1}\int_x^{+\infty} \frac{1}{\theta^{a+1}}\,\mathrm{d}\theta = \frac{1}{x}, \quad 0 < x < \theta$$

i jednaka je nuli za $x \notin (0,\theta)$. Ovo nije prava gustina, jer je

$$\int m(x)\,\mathrm{d}x = \int_0^\theta \frac{\mathrm{d}x}{x} = +\infty$$

Međutim, aposteriorna gustina može da se definiše, takođe u nepravom smislu - formalnom primenom formule

$$(*) \qquad f(\theta\,|\,x) = \frac{f(x,\theta)}{m(x)} = \frac{ax^a}{\theta^{a+1}} I_{\{0<x<\theta\}} I_{\{\theta>0\}},$$

i za dato $x > 0$ ovo je prava gustina, tj. $\int_0^{+\infty} f(\theta\,|\,x)\,\mathrm{d}\theta = 1$.

Bajesova ocena je ona koja minimizira matematičko očekivanje funkcije gubitka u odnosu na aposteriornu raspodelu $f(\theta\,|\,x)$, odnosno u ovom slučaju to znači da se $\hat\theta = \hat\theta(X)$ dobija kao ona vrednost c za koju je

$$\varphi(c) = \int |\theta - c| f(\theta\,|\,X)\,\mathrm{d}\theta \to \max,$$

gde je X slučajna promenljiva iz date familije raspodela sa nepoznatim θ. Pretpostavljajući da je $X > 0$, znamo sa predavanja da funkcija $\varphi(c)$ dostiže maksimum za $c = \mathrm{Med}\,(\theta \,|\, X)$, odnosno za ono c za koje je

$$aX^a \int_X^c \frac{\mathrm{d}\theta}{\theta^{a+1}} = \frac{1}{2},$$

odakle je $\hat{\theta} = c = 2^{\frac{1}{a}} X$, i ovo je rešenje u slučaju da ocenu tražimo na bazi posmatranja jedne vrednosti za X.

U slučaju da imamo $\mathbf{X} = (X_1, \ldots, X_n)$ gde su X_i nezavisne slučajne promenljive iz familije raspodela sa gustinom $f(x \,|\, \theta)$ koja je data u postavci zadatka, dobijamo da je

$$f(\mathbf{X} \,|\, \theta) = C(\mathbf{X})\theta^{-an} I_{\{X_{(n)} \le \theta\}}, \qquad C(\mathbf{X}) = a^n \prod_{k=1}^{n} X_k.$$

Funkcija $C(X)$ se dalje ne pojavljuje (skrati se) i dobija se da je

$$f(\theta \,|\, X) = anX_{(n)}^{an}\theta^{-an-1}$$

Poređenjem sa slučajem $n = 1$ vidimo da se dobija ista aposteriorna raspodela, ako se umesto a stavi an, i X se zameni sa $X_{(n)}$. Alternativno, ako nastavimo dalje, medijana m dobija se kao rešenje jednačine

$$\int_{X_{(n)}}^{m} f(\theta \,|\, X)\, \mathrm{d}\theta = \frac{1}{2},$$

odnosno $\hat{\theta} = m = 2^{\frac{1}{na}} X_{(n)}$. Primetimo da je $\hat{\theta} \sim X_{(n)}$ kad $n \to +\infty$ kao što bi se dobilo i primenom klasične teorije. □

---

[10] **3.** Neka je $\mathbf{X}$ nezavisan uzorak iz familije raspodela sa gustinom $f(x \,|\, \theta) = 2e^{-2(x-\theta)}$, $x \ge \theta$. Naći kritičnu oblast za testiranje hipoteze $H_0 :\ \theta = 0$ protiv alternativne hipoteze $H_1 :\ \theta \ne 0$ primenom testa količnika maksimalne verodostojnosti. Koju odluku donosimo ako je u uzorku obima 10 minimalna vrednost 2.6, a uzoračka sredina 3.4? Primeniti prag značajnosti 0.05.

Rešenje: S obzirom da u zadatku nije data oblast u kojoj se nalazi parametar θ, pretpostavićemo da je $\theta \in \mathbf{R}$. Funkcija verodostojnosti u ovom zadatku je

$$f(\mathbf{X} \,|\, \theta) = 2^n \cdot e^{-2n(\hat{\mu}-\theta)} \cdot I_{\{X_{(1)} \ge \theta\}}, \qquad n = 10,$$

i dostiže maksimum po $\theta \in \mathbf{R}$ za $\theta = X_{(1)}$. Za $\theta = 0$, funkcija verodostojnosti je

$$f_0(\mathbf{X}) = 2^n \cdot e^{-2n\hat{\mu}} \cdot I_{\{X_{(1)} \ge 0\}}.$$

Statistika količnika verodostojnosti je

$$R = \frac{f_0(\mathbf{X})}{f(\mathbf{X} \,|\, \hat{\theta})} = e^{-2nX_{(1)}} \cdot I_{\{X_{(1)} \ge 0\}}$$

Kritična oblast je oblika $R \le C$. Pod nultom hipotezom, tj, ako je $\theta = 0$, imamo da je $X_{(1)} \ge 0$ sa verovatnoćom 1, tako da je

$$P(R \le C \,|\, \theta = 0) = P(-2nX_{(1)} \le \log C) = P(X_{(1)} \ge \log C^{-\frac{1}{2n}}).$$

Iz date raspodele za $\theta = 0$ dobijamo da je $P(X_{(1)} \ge t) = P((\forall i = 1, \ldots, n),\ X_i \ge t) = e^{-2nt}$, odakle je

$$P(R \le C \,|\, \theta = 0) = C$$

i odbacujemo H_0 sa nivoom značajnosti 0.05 i $n = 10$ ako je $X_{(1)} \ge \log(0.05)^{-1/20} = 0.1498$. Kako je u posmatranom slučaju $X_{(1)} = 2.6$, odbacujemo H_0.