STOCKHOLMS UNIVERSITET
Statistiska institutionen
Hans Nyquist

TENTAMEN I STATISTISK TEORI MED TILLÄMPNINGAR II
2023-03-20
LÖSNINGAR

Uppgift 1.

En lektor i statistik har hört att mätningar på blodtrycket kan variera kraftigt mellan upprepade mätningar. Innan han går till en föreläsning gör han därför 10 mätningar med en minuts mellanrum på sitt eget systoliska blodtryck. Han antar att mätningarna kan ses som observationer på en normalfördelad stokastisk variabel med väntevärdet μ och variansen σ^2, båda parametrarna är okända. Observationerna blev:

$$128 \quad 124 \quad 137 \quad 126 \quad 128 \quad 121 \quad 122 \quad 135 \quad 134 \quad 133$$

Med hjälp av observationerna vill han nu göra inferens om σ^2 och beräknar därför en del summor han tänker kan vara användbara:
$\sum_{i=1}^{n} y_i = 1288$
$\sum_{i=1}^{n} \frac{1}{y_i} = 0.08$
$\sum_{i=1}^{n} y_i^2 = 166216$
$\sum_{i=1}^{n} \ln y_i = 48,58$
där $y_i, i = 1, 2, \ldots, 30$ är de observerade blodtrycken. Tyvärr hinner han inte fortsätta uträkningarna eftersom han behöver gå till en föreläsning så du måste hjälpa honom att slutföra beräkningarna.

a) Beräkna en skattning på σ^2.

b) Bilda ett 95%-igt konfidensintervall för σ^2.

Lösning:

a) En uppskattning av σ^2 är $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}\right) = \frac{1}{10-1}\left(166216 - \frac{1288^2}{10}\right) = \frac{1}{9}321,6 = 35,667$

b) Ett 95 %-igt konfidensintervall ges av $\left(\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{\chi^2_{0.975}(n-1)};\ \frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{\chi^2_{0.025}(n-1)}\right) = \left(\frac{321,6}{19.9};\ \frac{321,6}{2.70}\right) = (16,161; 119,11)$

Uppgift 2

När lektorn i uppgift 1 kommer tillbaka till sitt kontor efter föreläsningen vill han undersöka om föreläsningen har påverkat hans systoliska blodtryck. Han gör därför 10 nya mätningar. De observationer han har nu är alltså

Före föreläsningen	128	124	137	126	128	121	122	135	134	133
Efter föreläsningen	133	127	145	130	133	129	127	139	140	136

a) Gör nödvändiga antaganden, sätt upp hypoteser, välj signifikansnivå och pröva med ett parametriskt test om föreläsningen har påverkat blodtrycket.

b) Gör nödvändiga antaganden, sätt upp hypoteser, välj signifikansnivå och pröva med ett icke-parametriskt test om föreläsningen har påverkat blodtrycket

Lösning:

Observationerna på blodtrycket utgör två orelaterade urval.

a) Antag att observationerna före föreläsningen är $N(\mu_F, \sigma^2)$ och att observationerna efter föreläsningen är $N(\mu_E, \sigma^2)$, obs. lika varianser, och att observationerna är stokastiskt oberoende inom och mellan urvalen. Vi prövar $H_0 : \mu_F = \mu_E$ mot $H_A : \mu_F \neq \mu_E$, dvs. tvåsidigt alternativ, med ett t-test. Teststatistikan är

$$t = \frac{\overline{y}_F - \overline{y}_E}{s_p\sqrt{\left(\frac{1}{n_F} + \frac{1}{n_E}\right)}}$$

Om H_0 är sann så är $t \sim t(n_F + n_E - 2)$. Eftersom $n_F = n_E = 10$ förkastar H_0 på signifikansnivån $\alpha = 0.05$ om $|t| > t_{0.975}(18) = 2.101$.
Vi får att $\overline{y}_F = 128.81$, $\overline{y}_E = 134$, $s_p^2 = \frac{(10-1)5.606^2+(10-1)6.072^2}{10+10-2} = 34.148$ så att $t_{obs} = \frac{128.81-134}{\sqrt{34.148}\sqrt{\left(\frac{1}{10}+\frac{1}{10}\right)}} = -1.9860$, dvs. H_0 kan inte förkastas på signifikansnivån 5% och vi har inga empiriska evidens för att föreläsningen höjde lektorns blodtryck.

b) Vi kan använda ett Mann-Whitney's test för att testa $H_0 : \mu_F = \mu_E$ mot $H_A : \mu_F \neq \mu_E$. Vi antar att fördelningarna för blodtrycket före respektive efter föreläsningen är symmetriska och att alla observationer är stokastiskt oberoende inom och mellan urvalen.
Vi börjar med att rangordna alla observationer

Före föreläsningen	7,5	3	17	4	7,5	1	2	15	14	12
Efter föreläsningen	12	5,5	20	10	12	9	5,5	18	19	16

Rangsumman för urval 1 (före) är $W = 83$ (vi kan kolla rangsummorna genom att beräkn den för urval 2 som är 127 så att rangsummorna tillsammans är 210 vilket jämförs med 20(20+1)/2=210.)
Testvariabeln är $U = n_F n_E + \frac{n_F(n_F+1)}{2} - W = 10 \cdot 10 + \frac{10(10+1)}{2} - 83 = 100 + 55 - 83 = 72$
Vi väljer signifikansnivån $\alpha = 0.05$ (det närmaste enl. tabell vi kan vi välja är $\alpha = 2 \cdot$

$0.0216 = 0.0432$) vilket ger kritiskt värde $U_0 = 23$, dvs. H_0 förkastas om $U < 23$ eller om $U > n_F n_E - U_0 = 10 \cdot 10 - 23 = 77$, men vårt observerade värde på U är 72, så vi kan inte förkasta H_0, dvs. testet ger inget evidens på 5% nivån att det är någon blodtryckshöjande effekt av föreläsningen.

Uppgift 3

En övningslärare i statistik har börjat spela tennis och övar flitigt på att slå bollen på ett korrekt sätt över nätet. Antag att sannolikheten för ett korrekt slag är $1-p$, så att antalet korrekta slag följt av ett felaktigt, Y, är en geometriskt fördelad stokastisk variabel med parametern p. Övningsläraren vill nu estimera p, sannolikheten för ett felaktigt slag, med hjälp av n stycken stokastiskt oberoende observationer på Y.

a) Bestäm momentestimatorn av p.

b) Bestäm maximum likelihoodestimatorn av p.

Lösning:

a) För ev geometrisk fördelning gäller det att $E(Y) = 1/p$. Momentestimatorn $\widetilde{p}$ av p definieras därför av

$$\frac{1}{\widetilde{p}} = \overline{y},$$

där $\overline{y} = n^{-1}\sum_{i=1}^{n} y_i$ är medelvärdet av observationerna. Detta gör att

$$\widetilde{p} = \frac{1}{\overline{y}}$$

b) För den geometrisk fördelningen gäller att

$$p(y;p) = (1-p)^{y-1}\, p$$

så att likelihoodfunktionen är

$$L(p) = \prod_{i=1}^{n} (1-p)^{y_i-1}\, p = (1-p)^{\sum_{i=1}^{n} y_i - n}\, p^n.$$

Logaritmering ger

$$\ln L(p) = \ln(1-p)\left(\sum_{i=1}^{n} y_i - n\right) + n\ln p.$$

Derivering ger

$$\frac{d\ln L(p)}{dp} = -\frac{\sum_{i=1}^{n} y_i - n}{1-p} + \frac{n}{p}.$$

Sätt derivatan lika med noll och lös ut $\widehat{p}$

$$\frac{\sum_{i=1}^{n} y_i - n}{1-\widehat{p}} = \frac{n}{\widehat{p}}$$

$$\widehat{p}\left(\sum_{i=1}^{n} y_i - n\right) = n(1-\widehat{p})$$

$$\widehat{p}\sum_{i=1}^{n} y_i - \widehat{p}n = n - n\widehat{p}$$

$$\widehat{p} = \frac{n}{\sum_{i=1}^{n} y_i} = \frac{1}{\overline{y}}$$

Uppgift 4 (20 poäng)

En väg i ett visst område har en smal bro där bilar inte kan mötas. Om två bilar kommer från vardera hållet måste en av bilarna vänta tills den andra har kört över. Lyckligtvis inträffar det relativt sällan att bilar kommer samtidigt och behöver vänta. En ekonometriker modellerar antalet gånger per dag bilar behöver vänta vid bron med en Poissonfördelning med parametern λ.

a) Hjälp ekonometrikern att konstruera ett likformigt starkaste test av hypotesen $H_0 : \lambda = 1$ mot $H_A : \lambda > 1$.

b) Under en tiodagarsperiod observerades följande antal händelser per dag då bilar behövde vänta vid bron:

$$2 \quad 0 \quad 1 \quad 2 \quad 1 \quad 1 \quad 3 \quad 2 \quad 1 \quad 3$$

Ange p-värdet för testet i uppgift a.

Lösning:

a) Vi har ett slumpmässigt urval från $Po(\lambda)$ och börjar med att konstruera ett starkaste test av den enkla nollhypotesen $H_0 : \lambda = 1$ den enkla alternativhypotesen $H_A : \lambda = \lambda_A$ för något fixt med godtyckligt värde $\lambda_A > 1$. Likelihoodfunktionen för $Po(\lambda)$ är

$$L(\lambda) = \prod_{i=1}^{n} \frac{\lambda^{y_i}}{y_i!} e^{-\lambda} = \frac{\lambda^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} e^{-n\lambda}$$

Neyman-Pearsons lemma ger att H_0 förkastas då

$$\frac{L(\lambda = 1)}{L(\lambda = \lambda_A)} < k$$

för något tal k, men detta är ekvivalent med

$$\frac{\frac{1}{\prod_{i=1}^{n} y_i!} e^{-n}}{\frac{\lambda_A^{\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} e^{-n\lambda_A}} = \frac{e^{n(\lambda_A - 1)}}{\lambda_A^{\sum_{i=1}^{n} y_i}} < k$$

Logaritmering ger

$$n(\lambda_A - 1) - \ln \lambda_A \sum_{i=1}^{n} y_i < \ln k$$

Eftersom $\lambda_A > 1$ är $\ln \lambda_A > 0$.Detta ger

$$\sum_{i=1}^{n} y_i > \frac{n(\lambda_A - 1) - \ln k}{\ln \lambda_A} = k'$$

för något tal k. H_0 förkastas alltså för stora värden på $\sum_{i=1}^{n} y_i$. Eftersom $\lambda_A > 1$ valdes godtyckligt är detta det likformigt starkaste testet.

b) Vi förkastar H_0 om $S = \sum_{i=1}^{n} y_i$ är stor. Eftersom observationerna är oberoende och Poissonfördelade är $S \sim Po(n\lambda)$. Om H_0 är sann är alltså $S \sim Po(10)$. Vi har att observerat värde på S är 16. Detta ger $p = P(S \geq 16) = 1 - P(S \leq 16) = 1 - 0.951 = 0.049$.

Uppgift 5

a) Vilka fördelar respektive nackdelar har parametriska metoder jämfört med icke-parametriska metoder?

b) Vad innebär det att en estimator är konsistent?

c) Två dataanalytiker estimerar en parameter i en modell. Båda har samma observationer, men de får olika kredibilitetsintervall. Förklara varför de kan få olika resultat förutsatt att båda inte gör några matematiska fel.

d) Vad är a posteriori fördelning?

e) Hur påverkas styrkan hos ett test om signifikansnivån minskas och antale observationer är konstant? Hur påverkas styrkan om dessutom antalet observationer ökar?