Avd. Matematisk statistik

KTH Matematik

TENTAMEN I SF1901 SANNOLIKHETSTEORI OCH STATISTIK,
TISDAGEN DEN 14:E MARS 2017 KL 08.00–13.00.

Examinator: Thomas Önskog, 08 – 790 84 55.

Tillåtna hjälpmedel: Formel- och tabellsamling i Matematisk statistik, Mathematics Handbook (Beta), hjälpreda för miniräknare, miniräknare.

Införda beteckningar skall förklaras och definieras. Resonemang och uträkningar skall vara så utförliga och väl motiverade att de är lätta att följa. Numeriska svar skall anges med minst två siffrors noggrannhet. Tentamen består av 6 uppgifter. Varje korrekt lösning ger 10 poäng. Gränsen för godkänt är preliminärt 24 poäng. Möjlighet att komplettera ges för tentander med, preliminärt, 22–23 poäng. Tid och plats för komplettering kommer att anges på kursens hemsida. Det ankommer på dig själv att ta reda på om du har rätt att komplettera.

Poäng från kontrollskrivning och laborationer under innevarande kursomgång (period 3, VT2017) får tillgodoräknas under förutsättning att tentanden erhållit minst 20 poäng på denna tentamen.

Tentamen kommer att vara rättad inom tre arbetsveckor från skrivningstillfället och kommer att finnas tillgänglig på studentexpeditionen minst sju veckor efter skrivningstillfället.

Uppgift 1

Betrakta en *binär repetitionskod*: Bitar $X_1, X_2, \ldots$ sänds från en källa med lika sannolikhet för 1 och 0, och passerar genom en kanal med felsannolikhet $p < 1/2$ (med sannolikhet p ändras en 1:a till en 0:a och vice versa). För att korrigera för eventuella fel upprepas varje bit X_i ett udda antal N gånger - en 0:a sänds som N 0:or i följd, en 1:a som N 1:or - och en majoritetsomröstning bestämmer hur den mottagna följden ska tolkas (avkodas): Om $N = 3$ tolkas den mottagna följden 001 som 0, 011 som 1 osv. Fel uppstår oberoende av varandra och oberoende av vad som sänds genom kanalen.

a) Beräkna sannolikheten att en följd avkodas fel, det vill säga en ursprungliga 1:a tolkas som en 0:a eller omvänt, i fallet $N = 3$. (4 p)

b) Beräkna sannolikheten att det var en 1:a som skickades om du avkodar de tre bitarna som en 1:a. (4 p)

c) Finn ett uttryck för sannolikheten att en mottagen följd avkodas till fel bit för ett godtyckligt udda $N > 3$. (2 p)

Var god vänd!

Uppgift 2

En ikosaeders sidor är numrerade $0, 1, 2, \ldots, 9$. Man misstänker att ikosaedern är skev och vill därför undersöka vad sannolikheten p att få nio i ett enskilt kast är. Tag fram ML-skattningen av sannolikheten p i följande två fall:

a) 20 oberoende kast görs, varav en 9:a erhålls i fem kast. (5 p)

b) Oberoende kast görs tills dess att fyra nior erhållits. Dessa inträffade i kast 3, 9, 15, 18. (5 p)

Uppgift 3

En våg har inte bara mätfel utan utsätts även för en slumpmässig störning. Mätresultatet vid vägning av ett föremål med vikt μ beskrivs av en stokastisk variabel X där

$$X = \text{ vikt} + \text{mätfel} + \text{störning } = \mu + \epsilon + \delta,$$

där mätfelet ϵ är $N(0, \sigma)$ och störningen är $N(\mu_\delta, \sigma_\delta)$ och oberoende av ϵ. Vid vägningar i en ostörd miljö beskrivs mätresultaten av $\mu + \epsilon$. För att få en uppfattning om störningens storlek görs 5 mätningar på föremål med vikter $\mu_1, \ldots, \mu_5$ både i den störda miljön och en ostörd miljö:

Störd miljö	48.47	51.39	46.87	45.52	53.87
Ostörd miljö	47.85	52.07	47.47	47.50	55.10

Alla mätningar kan anses vara utfall av oberoende stokastiska variabler.

a) Bestäm ett 95% konfidensintervall för μ_{delta}. (7 p)

b) Testa nollhypotesen $H_0 : \ \mu_\delta = 0$ mot den alternativa hypotesen $H_1 : \ \mu_\delta > 0$ på nivån 5%. Slutsatsen om H_0 skall anges och motiveras tydligt. (3 p)

Uppgift 4

I en stor studie ville forskare undersöka huruvida klimat kan påverka förekomsten av astma. Man valde därför två stora städer, A och B, med olika klimat och i andra avseenden jämförbara populationer (ålder, etniticitet osv.). Totalt undersöktes tvåhundratusen människor i stad A och etthundratusen i stad B. Antalet personer med astma var 13800 i A och 8400 i B; städernas befolkningsmängder är tio miljoner (A) respektive fem miljoner (B).

a) Bestäm ett konfidensintervall med approximativ konfidensgrad 99% för skillnaden i förekomst av astma mellan de båda städerna. Var noga med att ange och motivera eventuella approximationer. (7 p)

b) Gör en hypotesprövning på approximativa nivån 1% för att se om klimat har en signifikant inverkan på förekomsten av astma. Var noga med att ange dina hypoteser och slutsatser. (3 p)

Uppgift 5

I amerikansk fotboll är en "fumble" när en spelare under ett "försök" tappar bollen; en match innehåller ca 60, 70 "försök" per lag per match. Följande tabell anger antal fumbles i 55 matcher (antalet angivet per lag).

2	1	2	2	3	1	3	4	3	4	5
5	2	1	3	2	5	2	4	1	2	2
1	0	4	2	4	1	2	0	2	0	3
0	1	2	0	1	2	2	3	5	1	3
2	3	4	5	4	3	6	0	3	1	2
1	2	2	1	2	1	3	2	4	2	4
4	2	0	5	4	3	6	5	3	5	1
3	1	1	3	1	4	3	1	5	1	2
1	3	4	4	4	2	7	4	2	5	3
1	3	6	2	1	1	4	1	2	3	0

Tabellen kan sammanfattas med följande frekvenstabell:

Antal fumbles	0	1	2	3	4	5	6	7 eller fler
Antal observationer	8	24	27	20	17	10	3	1

Vidare var det observerade genomsnittliga antalet fumbles i en match (per lag) 2.55. För enkelhets skull kan observationerna betraktas som utfall av oberoende och likafördelade stokastiska variabler. Vidare kan vi anta att sannolikheten för en fumble i ett "försök" är konstant.

a) Ange en lämplig modell, innehållande endast en parameter, för antalet fumbles för ett lag i en match. Motivera ditt val. *Ledning: Asymptotiska resultat för binomialfördelningen kan vara till nytta.* (2 p)

b) Formulera ett statistiskt test på nivån 5% som testar hur väl den föreslagna modellen passar observerad data. Var noga med att ange dina hypoteser och slutsatser. (8 p)

Uppgift 6

I en processor för akustiska signaler observerar man en stokastisk variabel Y som ges av absolutbeloppet av $X \in N(0, \sigma)$, $\sigma > 0$, dvs. $Y = |X|$. Vi har inte tillgång till direkta observationer av X. Standardavvikelsen σ är inte känd och bör skattas på basis av n oberoende observationer $y_1, \ldots, y_n$ av Y.

a) En intuitivt tilltalande skattning av σ^2 ges av

$$s^* = \frac{1}{n} \sum_{i=1}^{n} y_i^2,$$

där $y_1, \ldots, y_n$ är oberoende observationer av Y. Avgör om s^* är en väntevärdesriktig skattning. (4 p)

b) Härled täthetsfunktionen f_Y för Y. (6 p)

Lycka till!

Avd. Matematisk statistik

KTH Matematik

LÖSNINGSFÖRSLAG TENTAMEN I SF1901 MATEMATISK STATISTIK.
TISDAGEN DEN 14:E MARS 2017 KL 08.00–13.00.

Uppgift 1

a) Låt s beteckna den symbol som sändes och $\tilde{s}$ den som erhålls efter avkodning. En följd avkodas felaktigt, $\tilde{s} \neq s$, om två eller tre fel inträffar i överföringen. Vidare inses enkelt att antalet fel Y i en överföring, dvs. bland de tre bitar som repeterar den ursprungliga symbolen, är $Bin(3, p)$-fördelat. Därmed ges sannolikheten för ett fel av

$$\begin{aligned} P(\tilde{s} \neq s) &= P(Y = 2) + P(Y = 3) \\ &= \binom{3}{2} p^2(1-p) + p^3 \\ &= 3p^2 - 2p^3. \end{aligned}$$

Svar: Sannolikheten att en följd avkodas fel är $3p^2 - 2p^3$.

b) Låt S beteckna händelsen att en 1:a skickas och M händelsen att en 1:a mottas (efter avkodning). Vi söker $P(S|M)$, vilket med hjälp av Bayes sats kan skrivas

$$\begin{aligned} P(S|M) &= \frac{P(M|S)P(S)}{P(M)} \\ &= \frac{P(M|S)P(S)}{P(M|S)P(S) + P(M|S^c)P(S^c)}. \end{aligned}$$

Vi har antagit att $P(S) = 1/2$ (sannolikheten att sända 0:a eller 1:a densamma). Kombinerat med resultatet i (a) fås

$$P(M|S)P(S) = \frac{1 - 3p^2 + 2p^3}{2},$$

och

$$\begin{aligned} P(M|S)P(S) + P(M|S^c)P(S^c) &= \frac{1 - 3p^2 + 2p^3}{2} + \frac{3p^2 - 2p^3}{2} \\ &= \frac{1}{2}. \end{aligned}$$

Vi konstaterar att den sökta sannolikheten ges av

$$P(S|M) = 1 - 3p^2 + 2p^3.$$

Svar: Sannolikheten ges av $1 - 3p^2 + 2p^3$.

c) För ett godtyckligt $N \geq 3$ ges sannolikheten att avkoda en föld felaktigt med majoritetsröstning av sannolikheten för fler än $\lceil N/2 \rceil$ ändrade bitar i överföringen. Det senare är en summa av sannolikheter från $Bin(N,p)$-fördelningen: Med s, $\tilde{s}$ och Y som i (a),

$$\begin{aligned} P(\tilde{s} \neq s) &= P(Y = \lceil N/2 \rceil) + P(Y = \lceil N/2 \rceil + 1) + \cdots + P(Y = N) \\ &= \sum_{k=(N+1)/2}^{N} P(Y = k) \\ &= \sum_{k=(N+1)/2}^{N} \binom{N}{k} p^k (1-p)^{N-k}. \end{aligned}$$

Svar: För ett godtyckligt udda $N > 3$ ges sannolikheten av $\sum_{k=(N+1)/2}^{N} \binom{N}{k} p^k (1-p)^{N-k}$.

Uppgift 2

a) Den stokastiska varibeln X som räknar antalet 9:or i de 20 kasten är $Bin(20,p)$-fördelad. Likelihood-funktionen ges därför av

$$L(p) = \binom{20}{5} p^5 (1-p)^{15}.$$

Funktionen maximeras av $p = 5/20 = 0.25$ (generellt: x/N), vilket alltså är ML-skattningen av p.
Svar: ML-skattningen av av p är $p^*_{ML} = 0.25$.

b) Oberoendet ger att sannolikheten att "få det som man fått" som funktion av p, dvs likelihood-funktionen, blir

$$L(p) = (1-p)^2 p (1-p)^5 p (1-p)^5 p (1-p)^2 p = (1-p)^{14} p^4.$$

Denna maximeras av $p = 4/18 = 0.22$, som därmed är ML-skattningen av sannolikheten p.
Svar: ML-skattningen av av p är $p^*_{ML} = 0.22$.

Uppgift 3

a) Mätning i i den störda miljön, x_i, är ett utfall av en stokastisk variabel $X_i \in N(\mu_i + \mu_\delta, \sqrt{\sigma^2 + \sigma_\delta^2})$. Mätning i i den ostörda miljön, y_i är ett utfall av en stokastisk variabel $Y_i \in N(\mu_i, \sigma)$. De parvisa skillnaderna $Z_i = X_i - Y_i$ är således $N(\mu_\delta, \sqrt{2\sigma^2 + \sigma_\delta^2})$-fördelade stokastiska variabler.

Låt $\tilde{\sigma} = \sqrt{2\sigma^2 + \sigma_\delta^2}$. Vi skattar nu μ_δ med $\bar{z} = -0.7740$ och $\tilde{\sigma}$ med $s_z = 0.9546$. Skattningen $\bar{z}$ kan ses som ett utfall av en $N(\mu_\delta, \frac{\tilde{\sigma}}{\sqrt{5}})$-fördelad stokastisk variabel. Ett tvåsidigt 95%-igt konfidensintervall för parameter μ_δ ges därför av

$$\bar{z} \pm t_{0.025}(4) \frac{s_z}{\sqrt{5}} = -0.7740 \pm 2.78 \times \frac{0.9546}{\sqrt{5}} = -0.77 \pm 1.19,$$

eller

$$I_\delta = (-1.96, 0.41).$$

Vi kan även välja att göra ett enkelsidigt intervall; de två möjliga intervallen ges av

$$(\bar{z} - t_{0.05}(4)\frac{s_z}{\sqrt{5}}, \infty) = (-0.77740 - 2.13 \times \frac{0.9546}{\sqrt{5}}, \infty) = (-1.68, \infty),$$
$$(-\infty, \bar{z} - t_{0.05}(4)\frac{s_z}{\sqrt{5}}) = (-\infty, -0.77740 + 2.13 \times \frac{0.9546}{\sqrt{5}}) = (-\infty, 0.14),$$

Samtliga intervall är godtagbara svar.

b) Vi använder oss av konfidensmetoden för att pröva hypotesen H_0 mot H_1. Då 0 ligger i motsvarande *ensidiga* konfidensintervall kan vi på nivån 5% **ej** förkasta $H_0: \ \mu_\delta = 0$.
Svar: Vi kan ej förkasta H_0 på nivån 5%.

Uppgift 4

a) För att ta fram ett konfidensintervall används normalapproximation till binomialfördelningen. Låt X och Y beteckna de stokastiska variabler som svarar mot antalet tillfrågade personer som har astma. Vi noterar att om p_A respektive p_B betecknar sannolikheten att en slumpmässigt utvald person i respektive stad har astma så gäller

$$X \in Bin(n_A, p_A), \quad Y \in Bin(n_B, p_B),$$

där $n_A = 2 \times 10^5$ och $n_B = 10^5$ är antalet tillfrågade personer i respektive stad. Notera att vi kan använda binomialfördelningen snarare än en hypergeometrisk fördelning då befolkningsmängderna är så pass stora jämfört med antalet tillfrågade i de två städerna.

Med hjälp av normalapproximationen vet vi att ett konfidensintervall med approximativ konfidensgrad 99% för skillnaden $p_A - p_B$ ges av

$$p_A^* - p_B^* \pm \lambda_{0.005}\sqrt{\frac{p_A^*(1-p_A^*)}{n_A} + \frac{p_B^*(1-p_B^*)}{n_B}},$$

där $p_A^* = 13800/n_A = 0.069$ och $p_B^* = 8400/n_B = 0.084$ är de erhållna punktskattningarna av p_A och p_B. Konfidensintervallet ges av $I_{p_A-p_B}(-0.018, -0.012)$.
Svar: Ett konfidensintervall med approximativ konfidensnivå är $I_{p_A-p_B}(-0.018, -0.012)$.

Notera att normalapproximationen är applicerbar då antalet försök, n_A och n_B, i de två binomialfördelningarna är tillrčckligt stort för att tumregeln $np(1-p) \geq 10$ ska vara uppfylld.

b) Vi vill pröva hypotesen att klimatet ej har inverkan på förekomsten av astma,

$$H_0: \ p_A - p_B = 0,$$

mot

$$H_1: \ p_A - p_B \neq 0.$$

Vi använder oss av konfidensmetoden för att pröva hypotesen H_0 mot H_1. Då ett konfidensintervall med approximativ konfidensgrad 99% för $p_A - p_B$ inte innehåller 0 kan vi på approximativa nivån 1% förkasta H_0.
Svar: På signifikansnivån 1% förkastar vi hypotesen att klimatet ej har inverkan på förekomsten av astma.

Uppgift 5

a) En lämplig modell är att antalet fumbles är Poisson-fördelat med parameter $\lambda > 0$. Som motivering kan vi använda antagandet att sannolikheten för en fumble är konstant i alla försök - det leder till att antalet fumbles i en match kan ses som $Bin(n, p)$-fördelat med n antalet försök (60 till 70 st). Då p kan antas litet är Poisson-approximationen till binomialfördleningen lämplig.
Svar: En lämplig statistisk modell är att antalet fumbles är Poisson-fördelat med parameter $\mu > 0$.

b) Med X som antalet fumbles i en match önskar vi nu pröva hypotesen

$$H_0: \ X \text{ följer en Poisson-fördelning,}$$

mot

$$H_0: \ X \text{ följer ej en Poisson-fördelning.}$$

Detta görs lämpligen med ett χ^2-test med skattad parameter. Vi har givet att en skattning av parametern ges av $\hat{\lambda} = 2.55$ (genomsnittliga antalet fumbles per match). Med det kan vi beräkna sannolikheten, p_i^*, av $0, 1, \ldots$ upp till "7 eller fler" fumbles samt motsvarande np_i^* för $n = 110$ matcher. Resultatet ses i tabellen nedan.

Antal fumbles	0	1	2	3	4	5	6	7 eller fler
np_i^*	8.58	21.90	27.92	23.74	15.13	7.72	3.28	1.72

Då tumregeln om $np_i^* \geq 5$ inte uppfylls slår vi ihop de två sista kategorierna och får för "6 eller fler" $np_7^* = 5.00$. Motsvarande testvariabel är

$$Q_{obs} = \sum_{i=1}^{7} \frac{(x_i - np_i^*)^2}{np_i^*} = 1.97.$$

Detta kan ses som ett utfall frå en s.v. Q som är approximativt $\chi^2(5)$-fördelad och testvariabeln ska alltså jämföras med $\chi^2_{0.05}(5) = 11.07$. Eftersom $Q_{obs} < \chi^2_{0.05}(5)$ kan vi **ej** förkasta H_0 på nivån 5%.
Svar: På signifikansnivån 5% kan vi ej förkasta att antalet fumbles är Poisson-fördelat.

Uppgift 6

a) Låt $X_1, X_2, \ldots$ vara oberoende $N(0,\sigma)$ och $Y_i = |X_i|$, $i = 1, 2, \ldots$. Beteckna med S^* den stickprovsvariabel som svarar mot s^*:

$$S^* = \frac{1}{n}\sum_{i=1}^{n} Y_i^2.$$

Linjäritet av väntevärden ger att

$$E[S^*] = \frac{1}{n}\sum_{i=1}^{n} E[Y_i^2].$$

Vidare har vi

$$E[Y_i^2] = E[|X_i|^2] = E[X_i^2] = \sigma^2.$$

Det följer att

$$E[S^*] = \frac{1}{n}\sum_{i=1}^{n} \sigma^2 = \sigma^2,$$

det vill säga skattningen är väntevärdesriktig.
Svar: Skattningen s^* är väntevärdesriktig.

b) Vi inleder med att härleda fördelningsfunktionen $F_Y(y) = P(Y \le y)$. Från definitionen av Y_i står det klart att $F_Y(y) = 0$ för $y < 0$. För $y \ge 0$ fås

$$\begin{aligned} F_Y(y) &= P(Y \le y) \\ &= P(|X| \le y) \\ &= P(-y \le X \le y) \\ &= P(\frac{-y}{\sigma} \le \frac{X}{\sigma} \le \frac{y}{\sigma}) \\ &= \Phi\left(\frac{y}{\sigma}\right) - \Phi\left(\frac{-y}{\sigma}\right). \end{aligned}$$

I det sista steget har vi utnyttjat att $X/\sigma \in N(0,1)$. Symmetri ger sedan $\Phi\left(\frac{-y}{\sigma}\right) = 1 - \Phi\left(\frac{y}{\sigma}\right)$, vilket insatt ovan ger

$$F_Y(y) = 2\Phi\left(\frac{y}{\sigma}\right) - 1.$$

Derivering med avseende på y leder till

$$f_Y(y) = \frac{d}{dy}F_Y(y) = 2\left[\Phi'(x)\right]_{x=\frac{y}{\sigma}} \times \frac{1}{\sigma} = \frac{2}{\sigma}\frac{1}{\sqrt{2\pi}}e^{-y^2/2\sigma^2}.$$

För $y \ge 0$ fås alltså

$$f_Y(y) = \sqrt{\frac{2}{\pi\sigma^2}}e^{-y^2/2\sigma^2},$$

och $f_Y(y) = 0$ då $y < 0$.

Svar: Täthetsfunktionen för Y ges av $f_Y(y) = \begin{cases} \sqrt{\frac{2}{\pi\sigma^2}e^{y^2/2\sigma^2}}, & \text{då } y \ge 0, \\ 0, & \text{då } y < 0 \end{cases}$.