Avd. Matematisk statistik

KTH Matematik

TENTAMEN I SF1901 och SF1905 SANNOLIKHETSTEORI OCH STATISTIK, ONSDAGEN DEN 9:E JANUARI 2013 KL 14.00–19.00.

Examinator: Tatjana Pavlenko, tel 790 8466. Jourlärare: Gunnar Englund, tel 073 3213745

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

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.

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

Uppgift 1

En tillverkad enhet kan uppvisa A- och B-fel och i genomsnitt har 1 av 200 tillverkade enheter både slagen av fel. Endast 94% av alla tillverkade enheter är felfria. Av enheterna med fel har 75% A-fel.

a) Bestäm andelen av de felaktiga enheterna som har B-fel. (5 p)

b) Förekommer A- och B-fel oberoende av varandra? Motivera svaret genom att göra de erforderliga kalkylerna. (5 p)

Uppgift 2

Ett hälsoinstitut påstår att deras 5-dagars träningsprogram i genomsnitt minskar en deltagarens midjemått. På tio personer mättes midjemåttet (enhet: centimeter) före och efter träningsprogrammet. Resultat:

Person	1	2	3	4	5	6	7	8	9	10
Midjemått före	91.4	96.5	99.7	116.9	105.0	86.6	93.8	100.0	104.2	93.4
Midjemått efter	93.7	95.9	99.4	114.8	103.3	86.0	91.3	98.5	103.8	94.3

Undersök, genom att göra ett lämpligt statistiskt test på nivån 5% eller motsvarande konfidensintervall, om hälsoinstitutet har belägg för sitt påstående? Ett tydligt svar bör framgå. Svaret och den tillämpade statistiska metoden bör motiveras. (10 p)

Uppgift 3

Löjtnant Tabbes steglängd kan anses som en stokastisk variabel med väntevärdet 0.8 (enhet: meter) och variansen 0.0144. Man kan anta att steglängderna är oberoende. Han ämnar stega upp en landningsbana. Vad är minsta antal steg han skall ta, för att hans uppstegade bana med sannolikheten 0.9 skall överstiga 3000 meter. Lämplig och väl motiverad approximation får göras. (10 p)

Uppgift 4

Låt x_1 och x_2 vara resultat av två mätningar av en storhet med värdet θ och x_3 är en mätning av en storhet med värdet 2θ. Alla mätningar saknar systematiska fel, men störs av oberoende slumpmässiga fel med samma standardavvikelse σ.

a) Bestäm minsta-kvadratskattningen θ^* av parametern θ och visa att den är väntevärdesriktig. (5 p)

b) En annan skattning $\hat{\theta} = (2x_1 + 2x_2 + x_3)/6$, av parametern θ föreslås. Visa att den är väntevärdesriktig. (2 p)

c) Är $\hat{\theta}$ eller minsta-kvadratskattningen θ^* effektivast? Svaret skall motiveras. (3 p)

Uppgift 5

Vid en undersökning rörande trafikskadade i Västergötland under tiden mellan 2008-09-01 och 2010-02-28 har man bland annat studerat hur utsatta olika trafikantkategorier är för skallskador. Följande material insamlades för totalt 695 skallskadade personer:

	Bilister	Cyklister, mopedister, motorcyklister	Fotgångare
Ej medvetslös	122	80	28
Medvetslös högst 30 min	179	163	62
Medvetslös mer är 30 min	25	22	14

Undersök med ett lämpligt statistisk test, på nivån 5% om de tre trafikantkategorierna uppvisar någon skillnad vad gäller skallskadornas svårighetsgrad. Ange tydligt vilka de uppställda hypoteserna och slutsatsen är. (10 p)

Uppgift 6

Sannolikheten p, att ett SJ-tåg skall vara försenat vid ankomsten med mer än fem minuter beror bland annat av väder, antalet resande och veckodag. Antag att man är intresserad av värdet på p en speciell dag. Man väljer slumpmässigt ur tidtabellen för alla tåg i Sverige n ankomster och undersöker hur mycket försenade just dessa ankomster är. Man vill pröva hypotesen $H_0 : p = 0.3$ (egentligen $p \leq 0.3$) mot $H_1 : p > 0.3$ med ett test på nivån 5%.

a) Antag att man undersöker $n = 80$ ankomster och att man finner att vid 28 av dessa är tåget mer än fem minuter försenat. Kan man förkasta H_0? Svaret och den tillämpade statistiska metoden bör motiveras. Lämplig approximation får användas och skall då motiveras. (5 p)

b) Antag att p i själva verket är 0.4. Hur stort måste n vara för att man skall ha 50% chans att förkasta H_0 vid ett test på nivån 5%. (5 p)

Lycka till!

Avd. Matematisk statistik

KTH Matematik

LÖSNINGSFÖRSLAG TENTAMEN I SF1901 och SF1905 MATEMATISK STATISTIK, SF1901 OCH SF1905.
ONSDAGEN DEN 9 JANUARI 2013 KL 14.00–19.00

Uppgift 1

a) Låt A och B vara händelserna att en tillverkad enhet har A- respektive B-fel. Rita ett Venndiagram för händelserna!
Givet är att $P(A \cap B) = 0.005$ och $P(A^* \cap B^*) = 0.94$, dvs $P(A \cup B) = P(\text{Något fel}) = 1 - P(A^* \cap B^*) = 0.06$.
Nu är $P(A) = P(A|A \cup B)P(A \cup B) = 0.75 \cdot 0.06 = 0.045$. Alltså är $P(A \cap B^*) = P(A) - P(A \cap B) = 0.045 - 0.005 = 0.04$ och $P(B) = P(A \cup B) - P(A \cap B^*) = 0.06 - 0.04 = 0.02$. Slutligen

$$P(B|A \cup B) = \frac{P(B \cap (A \cup B))}{P(A \cup B)} = \frac{P(B)}{P(A \cup B)} = \frac{0.02}{0.06} = \frac{1}{3}.$$

SVAR: Andelen av de felaktiga enheterna som har B-fel är $\frac{1}{3}$.

b) Eftersom

$$P(A \cap B) = 0.005 \neq 0.045 \cdot 0.02 = P(A)P(B)$$

är händelserna A och B beroende.

SVAR: Händelserna A och B är beroende.

Uppgift 2

Modell: Stickprov i par. Man ska direkt beräkna differenserna mellan midjemåttet efter behandlingen och midjemåttet före behandlingen. Om differenserna är $z_1, \ldots, z_{10}$ antas de vera uttfalen av $N(\mu_d, \sigma)$. Dessa differenser är

Person	1	2	3	4	5	6	7	8	9	10
Midjemått före	91.4	96.5	99.7	116.9	105.0	86.6	93.8	100.0	104.2	93.4
Midjemått efter	93.7	95.9	99.4	114.8	103.3	86.0	91.3	98.5	103.8	94.3
Differens (z_i)	2.3	-0.6	-0.3	-2.1	-1.7	-0.6	-2.5	-1.5	-0.4	0.9

Hälsoinstitutet påstår att det förväntade värdet på dessa differenser ska vara mindre än 0, vilket med hypoteser kan uttryckas som

$$H_0 : \mu_d \geq 0; \quad H_1 : \mu_d < 0$$

och baserat på differenserbńa i tabell ovan beräknas $\bar{z} = \frac{1}{10}\sum_{i=1}^{10} z_i = -0.65$ och $s_z = 1.4393$. Vi förkastar H_0 för små värden på teststorheten vilken är

$$t = \frac{-0.65 - 0}{1.44/\sqrt{10}} = -1.428.$$

Om H_0 är sann är detta ett utfall från en $t(n-1) = t(9)$-fördelning. Ur $t(9)$ fås att $t_{0.05}(9) = 1.833$ (ensidigt test). Alltså, förkasta H_0 om $t < -t_{0.05}(9) = -1.833$. Eftersom $-1.4428 > -1.833$ kan man inte förkasta H_0. Alternativt, beräknas ett uppåt begränsat konfidensintervall för μ_d:

$$(-\infty, \bar{z} + t_{0.05}(9)\frac{s_z}{\sqrt{10}}] = (-\infty, 0.1844].$$

Eftersom intervallet täcker över 0 är slutsatsen den samma som tidigare, H_0 kan ej förkastas.

SVAR: Ur det insamlade data har man inte det tillräckliga belägg för hälsoinstitutets påstående.

Uppgift 3

Låt de oberoende lika fördelade stokastiska variablerna X_i, $i = 1, 2, 3, \ldots$ med $E(X_i) = 0.8$ och $V(X_i) = 0.0144$ beteckna vad som stegs upp med olika steg. På n steg är längden av den uppstegade bana $X_1 + X_2 + \cdot + X_n$, som enligt Centrala Gränsvärdessatsen är approximativt $N(0.8 \cdot n, 0.12\sqrt{n})$. Således bör n uppfylla

$$0.9 \approx P\left(X_1 + X_2 + \cdot + X_n > 3000\right) \Leftrightarrow 0.9 \approx 1 - \Phi\left(\frac{3000 - 0.8n}{0.12\sqrt{n}}\right)$$

$$\Leftrightarrow \frac{3000 - 0.8n}{0.12\sqrt{n}} \approx -1.2816 \Leftrightarrow 0.8n - 0.1538\sqrt{n} - 3000 = 0,$$

vilket ger $n = 3761.8 \approx 3762$.

SVAR: $\underline{n = 3762.}$

Uppgift 4

a) Vi vill minimera

$$Q(\theta) = (x_1 - \theta)^2 + (x_2 - \theta)^2 + (x_3 - 2\theta)^2.$$

Vi ser att

$$Q^{'}(\theta) = -2(x_1 - \theta) - 2(x_2 - \theta) - 2(x_3 - 2\theta) = 0$$

vilket medför att

$$x_1 + x_2 + 2x_3 - 6\theta = 0,$$

så att MK-skattningen av θ ges av $\theta^* = \frac{x_1+x_2+2x_3}{6}$.
θ^* är definitionsmässigt en väntevärdesriktig punktskattning om $E(\theta^*) = \theta$. Vi får, ty $E(X_1) = \theta$, $E(X_2) = \theta$ och $E(X_3) = 2\theta$, att

$$E(\theta^*) = \frac{1}{6}\left(E\left(X_1\right) + E\left(X_2\right) + 2E\left(X_3\right)\right) = \frac{1}{6}\left(\theta + \theta + 4\theta\right) = \theta.$$

SVAR a): MK-skattningen av θ ges av $\theta^* = \frac{x_1+x_2+2x_3}{6}$.
θ^* är en väntevärdesriktig punktskattning.

b) På samma sätt som i a) fås att

$$E\left(\hat{\theta}\right) = \frac{1}{6}\left(2E\left(X_1\right) + 2E\left(X_2\right) + E\left(X_3\right)\right) = \theta.$$

c) Eftersom X_1, X_2, X_3 är oberoende stokastiska variabler med samma standardavvikelse gäller det att

$$V\left(\hat{\theta}\right) = \frac{1}{36}\left(4V\left(X_1\right) + 4V\left(X_2\right) + V\left(X_3\right)\right) = \frac{4\sigma^2 + 4\sigma^2 + \sigma^2}{36} = \frac{1}{4}\sigma^2.$$

För θ^* gäller att

$$V\left(\theta^*\right) = \frac{1}{36}\left(V\left(X_1\right) + V\left(X_2\right) + 4V\left(X_3\right)\right) = \frac{\sigma^2 + \sigma^2 + 4\sigma^2}{36} = \frac{1}{6}\sigma^2.$$

Således är

$$V\left(\hat{\theta}\right) = \frac{1}{4}\sigma^2 > V\left(\theta^*\right) = \frac{1}{6}\sigma^2.$$

SVAR c): MK-skattningen θ^* är effektivast.

Uppgift 5

Detta är typexempel på homogenitetstestet (FS, avsn. 14.3). Vi vill testa H_0 : "Skallskadornas svårighetsgrad är likvärdig (med avseende på trafikantkategorier)" mot H_1 : "Skallskadornas svårighetsgrad är inte likvärdig". Vi använder ett approximativt χ^2-test (tillämpbarhet måste motiveras) som går ut på att jämföra de viktade och kvadrerade skillnaderna mellan de observerade frekvenserna och förväntade frekvenserna ($= 695 \cdot p_{ij}$). Vi har

	Trafikantkategorier			
	Bilister	Cyk.,moped., motorc.	Fotgångare	Total
Ej medv.	122	80	28	230
Medv, högst 30 min	179	163	62	404
Medv, mer 30 min	25	22	14	61
Total	326	265	104	695

Vår testvariabel betecknas med Q och ges av

$$Q = \sum_{ij} \frac{(x_{ij} - n_i p_j^*)^2}{n_i p_j^*},$$

där $i = 1, 2, 3$ och j=Bilister,Cyk.,moped., motorc., Fotgångare. Vidare får vi ur tabellen ovan att

$$p^*_{\text{Bilister}} = \frac{326}{695} = 0.4691, \quad p^*_{\text{Cyk.,moped., motorc.}} = \frac{256}{695} = 0.3813, \quad p^*_{\text{Fotgångare}} = \frac{104}{695} = 0.1496,$$

varefter vi får de förväntade frekvenserna

	Trafikantkategorier			
	Bilister	Cyk.,moped., motorc.	Fotgångare	Total
Ej medv.	107.89	87.69	34.42	230
Medv, högst 30 min	189.50	154.05	60.45	404
Medv, mer 30 min	28.61	23.26	9.13	61
Total	326	265	104	695

och villkoret $np_i \geq 5$ för ett approximativt χ^2 test är uppfyllt, dvs Q's fördelning kan approximeras med $\chi^2(f) = \chi^2((r-1)(s-1))$. Nu kan vi beräkna teststorheten vilken blir

$$Q = \frac{(122 - 107.89)^2}{107.89} + \frac{(80 - 87.69)^2}{87.69} + \cdots + \frac{(14 - 9.13)^2}{9.13} = 7.9813,$$

och $f = (r-1)(s-1) = 4$, vilket vidare ger kvcantilen för $\chi^2(4)$, som är $\chi^2_{0.05}(4) = 9.48$. Om H_0 är sann så är 7.9813 ett utfall från en stokastisk variabel som approximativt har en χ^2 fördelning med 4 frihetsgrader. Eftersom $\chi^2_{0.05}(4) = 9.48 > 7.9813$ så kan H_0 inte förkastas på nivån 5%. Alternativt kan vi beräkna sannolikheten att en $\chi^2(4)$-variabel är större än eller lika med 7.9813 (`X2cdf` på en TI-räknare). Denna sannolikhet, dvs p-värdet för testet, är 0.092. Detta p-värde är inte så lågt att vi förkastar H_0. Både teststorheten och p-värdet fås direkt med funktionen `X2-Test` på en TI-räknare.

SVAR: Data ger inte belägg för att svårhetsgrad av skallskador skiljer sig åt i de undersökta trafikantkategorierna.

Uppgift 6

a) Låt X betäkna antal försenade ankomster. Då $X \in Bin(n, p)$ där $n = 80$ och $x = 28$. Om H_0 är sann så har vi $np(1-p) = 80 \cdot 0.3 \cdot 0.7 = 16.8 > 10$ vilket innebär att normalapproximation kan användas. Då gäller att under H_0 $X \in N(np, \sqrt{np(1-p)}$ approximativt, eller $X \in N(24, \sqrt{16.8})$ approximativt. Vi testar H_0 mot H_1 genom att beränka teststorheten

$$u = \frac{x - 24}{\sqrt{16.8}} = \frac{28 - 24}{\sqrt{16.8}} = 0.98.$$

Vi skall förkasta H_0 för stora värden på u (ensidigt mothypotes) och vet att under H_0 så är u en observation från en stokastisk variabel som approximatit har $N(0, 1)$ fördelning. Om vi gör testet på nivån 5% så använder vi 5%-kvantilen för $N(0, 1)$ fördelning, som är 1.64 (tabell). H_0 ska förkastas om $u > \lambda_\alpha = \lambda_{0.05} = 1.64$. Eftersom $0.98 < 1.64$ så förkastar vi *inte* H_0 på nivån 5%. Alternativt kan vi beräkna sannolikheten att en $N(0, 1)$-fördelad variabel är större än 0.98 (`tcdf` på en TI-räknare); denna sannolikhet, dvs p-värdet för testet, är 0.1635. Detta p-värde är inte tillräkligt litet för att förkasta H_0.
Vill vi besvara frågan med ett konfidensintervall, skall vi göra ett nedåt begränsat ensidigt intervall för p med den ungefärliga konfidensgraden $1 - \alpha$ (jfr. FS 12.3). Detta blir

$$[p^*_{obs} - \lambda_\alpha \cdot \sqrt{\frac{p^*_{obs}(1 - p^*_{obs})}{n}}, \infty) = [0.2626, \infty).$$

Då detta intervall innehåller $p = 0.3$, kan vi inte förkasta H_0 på nivån 5%.

SVAR: Slutsatsen är alltså att det inte finns något statistist stöd i mätdata för att sannolikheten att ett SJ-tåg skall vara försenat vid ankomsen med mer än fem minuter är större än 0.3.

b) Vi har att $X \in Bin(n, 0.4)$, och vidare $X \in N(0.4n, \sqrt{n \cdot 0.4 \cdot 0.6}) = N(0.4n, \sqrt{0.24n})$, approximativt. Vårt test är som ovan, dvs förkasta H_0 om

$$\frac{x - n \cdot 0.3}{\sqrt{n \cdot 0.3 \cdot 0.7}} > 1.64 \Rightarrow x > 0.3n + 1.64\sqrt{0.21n}.$$

Styrkefunktionen (FS 14.1) är

$$h(0.4) = P\left(X > 0.3n + 1.64\sqrt{0.21n}\right) = 1 - \Phi\left(\frac{0.3n + 1.64\sqrt{0.21n} - 0.4n}{\sqrt{0.24n}}\right) \approx 0.5.$$

Detta innebär att

$$\frac{-0.1n + 1.64\sqrt{0.21n}}{\sqrt{0.24n}} = 0 \Rightarrow \sqrt{n} = 16.4\sqrt{0.21} \Rightarrow n = 56.5 \approx 57.$$

SVAR: n måste vara 57 för att man skall ha 50% chans att förkasta H_0 vid ett test på nivån 5%.