Avd. Matematisk statistik

KTH Matematik

TENTAMEN I SF1903 SANNOLIKHETSLÄRA OCH STATISTIK FÖR 3-ÅRIG Media TIMEH TORSDAGEN DEN TREDJE JUNI 2010 KL 14.00–19.00.

Examinator: Gunnar Englund, tel. 790 74 06

Tillåtna hjälpmedel: Läroboken. Formel- och tabellsamling i Matematisk statistik. 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 de tentander med 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 finnas tillgänglig på elevexpeditionen sju veckor efter skrivningstillfället.

Uppgift 1

a) X, Y och Z är oberoende normalfördelade stokastiska variabler, $E(X) = 2$, $E(Y) = 1$ och $E(Z) = 0$. Alla har varians 2. Beräkna $P(4X - 3Y > 5Z)$. (5 p)

b) I en produktionsprocess blir enheter felaktiga oberoende av varandra och alla med sannolikhet 0.001. Man tillverkar 8000 enheter. Beräkna sannolikheten att högst 10 av dessa är felaktiga. Välmotiverade approximationer får användas. (5 p)

Uppgift 2

a) Beräkna minsta-kvadrat skattningen θ^* av θ då man erhållit observationerna $x_1, x_2, \ldots, x_n$ på en stokastisk variabel med täthetsfunktion

$$f(x) = \begin{cases} \theta x^{\theta-1} & \text{om } 0 \leq x \leq 1 \\ 0 & \text{annars} \end{cases}$$

Beräkna även det numeriska värdet på θ^* då man erhållit observationerna 0.15, 0.40, 0.28 och 0.70. (5 p)

b) Man erhöll $\theta^* = 0.58$ baserat på $n = 100$ observationer. Ge ett approximativt 95 % konfidensintervall för väntevärdet på fördelningen i a) och beräkna från detta ett approximativt 95 % konfidensintervall för θ. (5 p)

Uppgift 3

Koncentrationen av en aktiv ingrediens i ett material tros påverkas av vilken katalysator som används i processen. Standardavvikelsen av den aktiva koncentrationen är känd till att vara 3.0 g/l, oberoende av katalysatorteknik. Tjugo observationer av koncentrationen togs, tio från katalysator 1 och tio från katalysator 2 med följande resultat:

Katalysator 1:	57.9	66.2	65.4	65.4	65.2	62.6	67.6	63.7	67.2	71.0
Katalysator 2:	66.4	71.7	70.3	69.3	64.8	69.6	68.6	69.4	65.3	68.8

Finns det någon anledning att tro att den aktiva koncentrationen beror på valet av katalysator? Basera ditt svar på beräkning av ett lämpligt 95%-konfidensintervall. Observationerna antas vara normalfördelade. (10 p)

Uppgift 4

Vid en tillverkningsprocess kontrolleras de tillverkade enheterna av en datorstyrd sensor. Härvid är den betingade sannolikheten för att en enhet klassificeras som defekt, givet att den är defekt, lika med 0.9 och den betingade sannolikheten för att en enhet klassificeras som korrekt, givet att den är defekt, lika med 0.1. Vidare är den betingade sannolikheten att klassificera en enhet som korrekt, givet att den är korrekt, lika med 0.85 och den betingade sannolikheten att klassificera en enhet som defekt, givet att den är korrekt, lika med 0.15. Vad är den betingade sannolikheten att en enhet är defekt givet att den klassificerats som defekt, om sannolikheten för en defekt enhet är 0.1 ? (10 p)

Uppgift 5

$X_1, \ldots, X_n, X_{n+1}$ är oberoende stokastiska variabler, $N(m, \sigma)$-fördelade och

$$\bar{X} = \frac{1}{n}\left(X_1 + \ldots + X_n\right)$$

är det aritmetiska medelvärdet av de första n av dessa. Vi vill använda $\bar{X}$ för att *prediktera* värdet på X_{n+1}, dvs. vi försöker förutsäga värdet på den nästkommande observationen. *Prediktionsfelet* är en stokastisk variabel Z_n definierad som

$$Z_n = X_{n+1} - \bar{X}.$$

a) Beräkna väntevärdet och standardavvikelsen för prediktionsfelet Z_n samt bestäm sannolikhetsfördelningen för Z_n. Noggranna motiveringar krävs. (5 p)

b) Beräkna gränsvärdet

$$\lim_{n\to\infty} P\left(|Z_n| > \sigma\right),$$

där $|Z_n|$ är absolutbeloppet och definieras av $|x| = x$, om $x > 0$ och $|x| = -x$, om $x \leq 0$. (5 p)

Uppgift 6

Den stokastiska variabeln X är $N(0, 1)$–fördelad. Beräkna fördelningsfunktionen för den stokastiska variabeln $Y = \Phi(X)$, där $\Phi(\cdot)$ är fördelningsfunktionen för X. Motivera noggrant ! (10 p)

Avd. Matematisk statistik

KTH Matematik

LÖSNINGAR TILL
TENTAMEN I SF1903 (f d 5B2501) SANNOLIKHETSLÄRA OCH STATISTIK FÖR 3-ÅRIG MEDIA TIMEH TORSDAGEN DEN TREDJE JUNI 2010 KL 14.00–19.00.

Uppgift 1

a) Vi har att $P(4X-3Y>5Z)=P(5Z-4X+3Y\leq 0)$. Sätt $U=5Z-4X+3Y$. Vi får att $E(5Z-4X+3Y)=5\cdot 0-4\cdot 2+3\cdot 1=-5$ och $V(U)=V(5Z-4X+3Y)=25V(Z)+16V(X)+9V(Y)=100$, dvs $D(Z)=10$ Alltså får vi att $U\in$N(-5,10) och

$$P(U\leq 0)=\Phi\left(\frac{0-(-5)}{10}\right)=\Phi(0.5)=0.6915$$

b) Låt X vara antalet felaktiga enheter. Då är X antalet gånger händelsen felaktig enhet tillverkad har uppkommit i 8000 oberoende försök och således är $X\in$Bin(8000,0.001). Vi kan inte använda tabell och därför approximerar vi. Sannolikheten $p<0.1$ och poissonapproximation är tillåten, X är approximativt Po($8000\cdot 0.001$)=Po(8). Vi erhåller

$$P(X\leq 10)=0.816$$

i tabell 7.

Uppgift 2

Sätt fördelningens väntevärde till m. Kvadratsumman $Q=\sum_{i=1}^{n}(x_i-m)^2$ skall minimeras. Derivatan $Q'=-2\sum_{i=1}^{n}(x_i-m)$ är 0 för $m=m^*=\bar{x}$, vilket ger minimum. Eftersom $m=\int_0^1 x\cdot\theta x^{\theta-1}dx=\int_0^1\theta x^{\theta}dx=\theta[\frac{x^{\theta+1}}{\theta+1}]_0^1=\frac{\theta}{\theta+1}$ erhåller vi att MK-skattningen av θ uppfyller $\frac{\theta^*}{\theta^*+1}=\bar{x}$, vilket innebär att $\theta^*=\frac{\bar{x}}{1-\bar{x}}$. Med siffror insatta erhålls $\bar{x}=0.3825$ och alltså $\theta^*=0.6194$.

b) Om X har fördelningen i a) är $E(X^2)=\int_0^1 x^2\theta x^{\theta-1}dx=\frac{\theta}{\theta+2}$ och variansen är $\sigma^2=E(X^2)-m^2=(\frac{\theta}{\theta+2})-(\frac{\theta}{\theta+1})^2=\frac{\theta}{(\theta+1)^2(\theta+2)}$. Enligt centrala gränsvärdessatsen är $\bar{X}$ approximativt $N(m,\sigma/\sqrt{n})$ och ett konfidensintervall för m ges av $\bar{x}\pm\lambda_{\alpha/2}d$ där d är en skattning av standardavvikelsen $\sigma/\sqrt{n}$. Men skattningen av σ ges av $\sigma^*=\sqrt{\frac{\theta^*}{(\theta^*+1)^2(\theta^*+2)}}$. Eftersom $\bar{x}=\frac{\theta^*}{\theta^*+1}$ erhålls numeriskt att $\bar{x}=0.3671$ och $\sigma^*=0.3001$. Antalet observationer $n=100$ och $\lambda_{0.025}=1.96$. Detta ger att intervallet $0.3671\pm 1.96\cdot 0.3001/10=0.3671\pm 0.0588=(0.3083,\ 0.4259)$ är ett approximativt 95 % konfidensintervallet för m. Men att $0.3083\leq m\leq 0.4259$ är ekvivalent med att $0.3083\leq\frac{\theta}{\theta+1}\leq 0.4259$, d.v.s. $0.3083(1+\theta)\leq\theta\leq 0.4259(1+\theta)$ vilket i sin tur är ekvivalent med att $0.4457\leq\theta\leq 0.7419$ varför detta sista intervall är ett approximativt 95 % konfidensintervall för θ.

Uppgift 3

Två oberoende stickprov. Observationerna från katalysator 1 antas vara N(μ_1, 3) och de från katalysator 2 N(μ_2, 3). Konfidensintervall för $\mu_1-\mu_2$ ges av $\bar{x}-\bar{y}\pm\lambda_{0.025}\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$ där $\bar{x}$ är medelvärdet av data från katalysator 1, $\bar{y}$ är medelvärdet för observationerna från katalysator 2, $\sigma^2 = 3^2 = 9$ och $n_1 = n_2 = 10$ (kombinera FS 11.1 och 10.2 a)). Eftersom $\lambda_{0.025} = 1.96$ erhåller vi intervallet $65.22 - 68.48 \pm 1.96 \cdot 3 \cdot \sqrt{1/10 + 1/10} = \underline{-3.2 \ \pm 2.63}$. Eftersom 0 inte tillhör intervallet kan vi på signifikansnivån 5% dra slutsatsen att katalysatorerna skiljer sig åt.

Uppgift 4

Inför följande beteckningar: K = en enhet är korrekt. D = en enhet är defekt. $\widehat{K}$ = en enhet klassificeras som korrekt. $\widehat{D}$ = en enhet klassificeras som defekt. Enligt uppgiften har man att $P(D) = 0.1$ och

$$P\left(\widehat{D} \mid D\right) = 0.9, \quad P\left(\widehat{K} \mid D\right) = 0.1$$

och

$$P\left(\widehat{K} \mid K\right) = 0.85, \quad P\left(\widehat{D} \mid K\right) = 0.15.$$

SÖKT är den betingade sannolikheten för att en enhet är defekt givet att den klassificerats som defekt, dvs. med beteckningarna ovan, $P\left(D \mid \widehat{D}\right)$. För att bestämma denna sannolikhet utnyttjas *Bayes' formel*, som ger

$$P\left(D \mid \widehat{D}\right) = \frac{P\left(\widehat{D} \mid D\right) \cdot P(D)}{P\left(\widehat{D}\right)}.$$

Enligt *lagen om total sannolikhet* kan nämnaren i högra ledet uttryckas som

$$P\left(\widehat{D}\right) = P\left(\widehat{D} \mid D\right) P(D) + P\left(\widehat{D} \mid K\right) P(K)$$

som är lika med

$$= 0.9 \cdot 0.1 + 0.15 \cdot 0.9 = 0.225,$$

ty $P(K) = 1 - P(D) = 1 - 0.1$. Således får man

$$P\left(D \mid \widehat{D}\right) = \frac{0.9 \cdot 0.1}{0.225} = 0.4.$$

SVAR: <u>Sannolikheten för att en enhet är defekt givet att den klassificerats som defekt är lika med = 0.</u>

Uppgift 5

(a)

$$E(Z_n) = E\left(X_{n+1} - \bar{X}\right) = E(X_{n+1}) - E\left(\bar{X}\right) = m - m = 0,$$

ty $E\left(\bar{X}\right) = \frac{1}{n}\left(E(X_1) + \ldots + E(X_n)\right) = \frac{1}{n} \cdot n \cdot m = m$. Eftersom $X_1, \ldots, X_n$ är oberoende $N(m, \sigma)$-fördelade variabler, gäller att

$$\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n) \in N\left(m, \sigma/\sqrt{n}\right)$$

Eftersom $X_1, \ldots, X_n$, X_{n+1} är oberoende $N(m,\sigma)$-fördelade variabler, är även $\bar{X}$ och X_{n+1} oberoende stokastiska variabler och vi får

$$V(Z_n) = V\left(X_{n+1} - \bar{X}\right) = V(X_{n+1}) + V\left(\bar{X}\right) = \sigma^2 + \frac{\sigma^2}{n} = \sigma^2\left(1 + \frac{1}{n}\right).$$

Detta ger

$$D(Z_n) = \sigma \cdot \sqrt{1 + \frac{1}{n}}.$$

Då $\bar{X}$ och X_{n+1} är oberoende stokastiska variabler är $Z_n = X_{n+1} - \bar{X}$ en linjär kombination av oberoende normalfördelade stokastiska variabler och vi får

SVAR a): $\underline{Z_n \in N(E(Z_n), D(Z_n)) = N\left(0, \sigma \cdot \sqrt{1 + \frac{1}{n}}\right).}$

(b) Definitionen på absolutbeloppet ger

$$P(|Z_n| > \sigma) = P(Z_n < -\sigma) + P(Z_n > \sigma).$$

Enligt upgiftens del (a) gäller, att $\frac{Z_n}{\sigma \cdot \sqrt{1+\frac{1}{n}}} \in N(0,1)$. Eftersom Z_n är en kontinuerlig stokastisk variabel fås härav att

$$P(Z_n < -\sigma) = P\left(\frac{Z_n}{\sigma \cdot \sqrt{1 + \frac{1}{n}}} < \frac{-\sigma}{\sigma \cdot \sqrt{1 + \frac{1}{n}}}\right) = \Phi\left(\frac{-1}{\sqrt{1 + \frac{1}{n}}}\right),$$

där $\Phi(x)$ är fördelningsfunktionen för $N(0,1)$. Komplementsatsen och samma standardisering ger att

$$P(Z_n > \sigma) = 1 - P(Z_n \le \sigma) = 1 - \Phi\left(\frac{1}{\sqrt{1 + \frac{1}{n}}}\right).$$

Symmetriegenskapen ger $\Phi\left(\frac{-1}{\sqrt{1+\frac{1}{n}}}\right) = 1 - \Phi\left(\frac{1}{\sqrt{1+\frac{1}{n}}}\right)$. Därmed är den sökta sannolikheten lika med

$$P(|Z_n| > \sigma) = 2 \cdot \left(1 - \Phi\left(\frac{1}{\sqrt{1 + \frac{1}{n}}}\right)\right).$$

Eftersom $\frac{1}{\sqrt{1+\frac{1}{n}}} \to 1$, då n växer mot oändligheten, får vi

$$\lim_{n\to\infty} P(|Z_n| > \sigma) = 2 \cdot (1 - \Phi(1)).$$

Här använder vi räknaren (alternativt tabell 1. i Formel- och tabellsamlingen), vilket ger att $1 - \Phi(1) = 1 - 0.8413 = 0.159$.

SVAR b): $\underline{\lim_{n\to\infty} P(|Z_n| > \sigma) = 0.318.}$

Uppgift 6

Y kan bara anta värden mellan 0 och 1 eftersom $\Phi(\cdot)$ är en fördelningsfunktion, dvs en sannolikhet (och dessa ligger som bekant mellan 0 och 1). Eftersom

$$\Phi(x) = \frac{1}{2\sqrt{\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt,$$

faas att
$$\frac{d}{dx}\Phi(x) = \frac{1}{2\sqrt{\pi}} e^{-x^2/2} > 0.$$
Således är $\Phi(x)$ monotont växande, och dess invers Φ^{-1} existerar. Lå t $0 \leq y \leq 1$.
$$P(Y \leq y) = P(\Phi(X) \leq y) = P(X \leq \Phi^{-1}(y)) = \{\text{ty } X \in N(0,1)\} = \Phi\big(\Phi^{-1}(y)\big) = y$$
Alltså få s
$$P(Y \leq y) = \begin{cases} 0, & y < 0, \\ y, & 0 \leq y < 1, \\ 1, & y \geq 1. \end{cases}$$
Denna är fördelningsfunktionen för den likformiga fördelningen på $[0, 1]$.
SVAR : $\underline{Y \in U(0, 1).}$