Avd. Matematisk statistik

KTH Matematik

TENTAMEN I SF1911, STATISTIK FÖR BIOTEKNIK
Torsdag den 13e april 08:00-13:00.

Examinator: Timo Koski, 70 − 237 00 47.
Kursledare: Timo Koski, − 790 71 34 .

Tillåtna hjälpmedel: Formel- och tabellsamling för SF911, 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 8 uppgifter. Varje korrekt lösning ger 4 poäng. Gränsen för godkänt är preliminärt 16 poäng (ev. bonuspoäng inräknade). Möjlighet att komplettera ges för tentander med, preliminärt, 14–15 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

Infektionsmultiplicitet (MOI) är kvoten av antalet bakteriofager mot antalet celler i ett experiment med fager, se t.ex. Frank H. Stephenson: *Calculations for Molecular Biology and Biotechnology. Third Edition*, Academic Press, 2016.
Antag att en kultur av bakterier är infekterad av bakteriofager med $MOI = 0.2$. Betrakta $MOI = 0.2$ som sannolikheten för att en bakteriecell är infekterad av en fag. En alikvot (för statistiker ung. ett stickprov) av 20 celler dras ur kulturen. Talet 20 är givet på förhand.

i) Låt X = antalet celler som är infekterade. En eller flera av av följande utsagor är sanna. Vilka ? (2 p)

a) $X \sim \mathcal{G}e(0.2)$.

b) $X \sim \mathcal{B}in(20, 0.2)$.

c) $X \sim \mathcal{G}eom(0.2)$.

d) $E(X) = 5$.

e) $E(X) = 4$.

ii) Låt X = antalet bakterier i en cell. Vi antar att bakteriekulturen är mycket stor och att en alikvot av 20 celler är mycket liten i jämförelse.

Vi gör nu gällande att $X \sim \mathcal{P}oi(MOI \cdot 20)$. En av av följande utsagor är en korrekt motivering till detta. Vilken? (2 p)

a) Centrala gränsvärdessatsen.

b) Rare Event Principle.

c) De små talens lag.

d) De stora talens lag.

Uppgift 2

Enligt CDC (= Centers for Disease Control and Prevention, USA) var andelen gonorréfall resistenta mot antibiotika lika med 0.27 för år 2010.
Du tar ett slumpmässigt stickprov på 50 vårdjournaler av patienter som diagnosticerats med gonorré och kontrollerar antalet fall med resistens mot antibiotika.

i) En av utsagorna **a)-c)** är riktig. Svara med det rätta alternativet.

a) Talet 0.27 är en empirisk fördelning för stickprovet.

b) Talet 0.27 är en populationsparameter för stickprovet.

b) Talet 0.27 är en statistika för stickprovet.

ii) Du tar upprepade gånger nya slumpmässiga stickprov på 50 vårdjournaler oberoende av varandra och med återläggning. Låt X = antalet gonorréfall resistenta mot antibiotika i Ditt stickprov. Din skattning av antalet antibiotikaresistenta är $\widehat{p} = X/50$. Vad är medelfelet för $\widehat{p}$ lika med? Svara med det rätta alternativet ur **a)-c)**.

a) 0.27.

b) $0.27 \cdot 0.73/\sqrt{50}$.

c) $\sqrt{0.27 \cdot 0.73/50}$.

iii) Du tar upprepade gånger nya slumpmässiga stickprov på 50 vårdjournaler obereonded av varandra och med återläggning. Låt X = antalet gonorréfall resistenta mot antibiotika i Ditt stickprov. Din skattning av antalet antibiotikaresistenta är $\widehat{p} = X/50$. Vad är sannolikheten för att få $\widehat{p} \geq 25\%$ *approximativt* lika med?

a) 0.530.

b) 0.625.

c) 0.762.

Uppgift 3

Du utvärderar ett nytt diagnostiskt test för en viss infektionssjukdom. Vi har gjort testet på 200 individer med känt sjukdomstillstånd. $D+$ betyder att personen är infekterad och $D-$ betyder att personen inte är infekterad. I tabellen nedan har du de observerade frekvenserna av de fyra möjliga utfallen av utvärderingen, där $T+$ är ett positivt testresultat i den bemärkelsen att en infektion diagnosticeras.

	$D+$	$D-$
Positivt test $T+$	95	10
Negativt test $T-$	5	90

Två av de följande utsagorna är sanna. Vilka är de? (4 p)

a) Sensitivitet = 90%.

b) Sensitivitet = 10%.

c) Sannolikheten för falsk positiv = 5%.

d) Positivt prediktivt värde = $P(D+|T+) = 90.05$ %

e) Specificitet = 95%

f) Sensitivitet = 95%

Uppgift 4

En logistisk regression av formen

$$p_x = P(Y = 1 \mid x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

betraktas. Variablerna Y och x är binära, och har värdena 0 och 1.

i) Låt OR beteckna oddskvoten för p_x när $x = 1$ och $x = 0$ (t.ex. oddskvoten för död i hjärtsjukdom för rökare v.s icke-rökare). En av de följande utsagorna är sann. Vilken? (2 p)

a) $OR = \beta_1$.

b) $OR = e^{\beta_1}$.

d) $OR = e^{-\beta_1}$.

e) $OR = e^{\beta_0}$.

ii) En av de följande formlerna är sann. Vilken ? (2 p)

a) $P(Y = 0 \mid x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$.

b) $P(Y = 0 \mid x) = e^{(\beta_0+\beta_1 x)}$.

d) $\text{logit}(p_0) = \beta_0$.

e) $\text{logit}(p_0) = \frac{1}{1+e^{-\beta_0}}$.

Uppgift 5

Låt A och B beteckna två händelser. Givet är $P(A) = 3/5$, $P(B|A) = 2/3$ och $P(B|A^c) = 1/3$, där A^c betecknar komplementet till händelsen A.

Avgör om A och B är beroende händelser. (4 p)

Uppgift 6

Progeri eller progeria är en ovanlig sjukdom som gör att kroppen åldras i förtid. En grupp av kliniska forskare mätte pulsvågshastigheten (PWV) hos 18 barn som diagnosticerats med en aggressiv form av progeria. PWV är ett standardmått på arterial styvhet som tilltar med åldern. De erhöll följande data (i meter per sekund m/s)

18.8 17.6 17.5 16.0 14.8 14.1 13.76 13.1 12.9

12.9 12.4 10.1 9.3 9.1 8.3 8.3 7.9 7.2

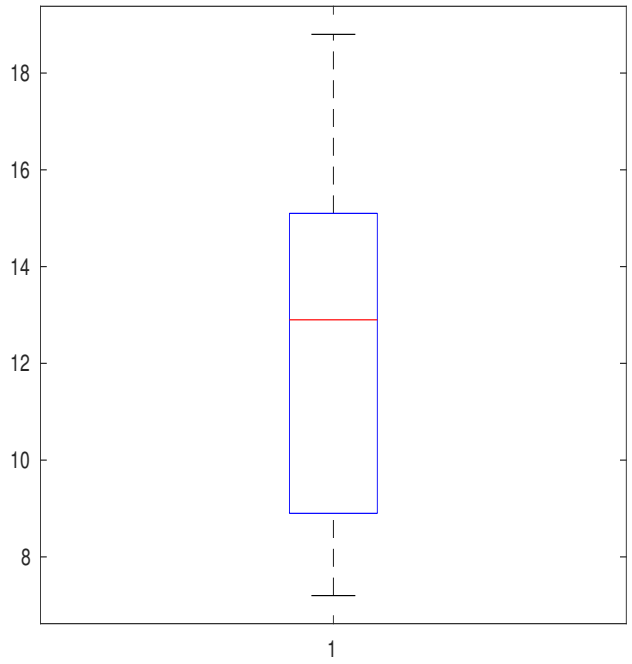
För friska barn anses ett värde på PWV över 6.6 vara onormalt.

a) Vi är intresserade av μ, medelvärdet för populationen av alla barn som diagnosticerats med progeria. Vår nollhypotes är

$$H_o : \mu = 6.6$$

vilket säger att det inte finns någon skillnad i PWV mellan friska barn och barn med progeria. Vilken mothypotes/alternativ hypotes måste vi rimligen ta? (1 p)

b) I figuren ser Du boxplotten för dessa data.



Varför kan denna plott tas som motivering för att vi kan anta normalfördelning för våra data? (1 p)

c) Som teststorhet använder vi ($\bar{x}$ är medelvärdet och s är standardavvikelsen för datat)

$$t = \frac{\bar{x} - 6.6}{s/\sqrt{18}}.$$

Under antagandet om oberoende normalfördelade data x, vilken fördelning har t? (1 p)

d) Kommer H_o att förkastas eller ej på signifikansnivån 0.001? *Ledning:* Jämför det numeriska värdet på t mot en lämplig kvantil för fördelningen i **c)**. (1 p)

Uppgift 7

Vi studerar farmakokinetiken hos en diabetesmedicin. Innan denna medicin kan godtas av läkemedelsmyndigheterna, måste man veta, hur kroppen absorberar och utsöndrar den. Diabetespatienter gavs en dos på antingen 1 mg eller 2mg av denna medicin och medicinens maximala plasmakoncentration (i nanogram per milliliter, ng/ml) utvärderades. Resultatet sammanfattas i tabellen nedan.

$$\begin{array}{ll} 1 \text{ mg} & n_1 = 32 \quad \overline{x} = 76, s_1 = 13 \\ 2 \text{ mg} & n_2 = 32 \quad \overline{y} = 156, s_2 = 42 \end{array}$$

Frågan lyder: är den maximala plasmakoncentrationen beroende av dosen? Låt μ_1 vara det okända populationsmedelvärdet för den maximala plasmakoncentrationen för dos 1 och μ_2 vara det okända populationsmedelvärdet för den maximala plasmakoncentrationen för dos 2. Vi ställer upp följande hypoteser.

$$H_o : \mu_1 = \mu_2$$

och

$$H_a : \mu_1 < \mu_2.$$

Vi antar att det handlar om oberoende data x från normalfördelningen $\mathcal{N}(\mu_1, \sigma^2)$ och oberoende data y från normalfördelningen $\mathcal{N}(\mu_2, \sigma^2)$, där medelvärdena och den gemensamma variansen är okända.

Avgör på signifikansnivån 1 % om det finns dosberoende i den maximala plasmakoncentrationen, d.v.s, om högre dos leder till större högre koncentration. (4 p)

Uppgift 8

I polymeraskedjereaktion (PCR) förekommer, vad som kalls på engelska **misincorporation rate**, varvid en bas annan än den komplementära mot DNA -mallen (templatet) adderas till $3'$-änden av den sträng som utvidgas.
En data-analytiker testar en polymeraskedjereaktion under vissa villkor. Efter ett antal cykler i fem försök kan hen observera följande andelar av PCR fragment fria från nämnda misincorporation.

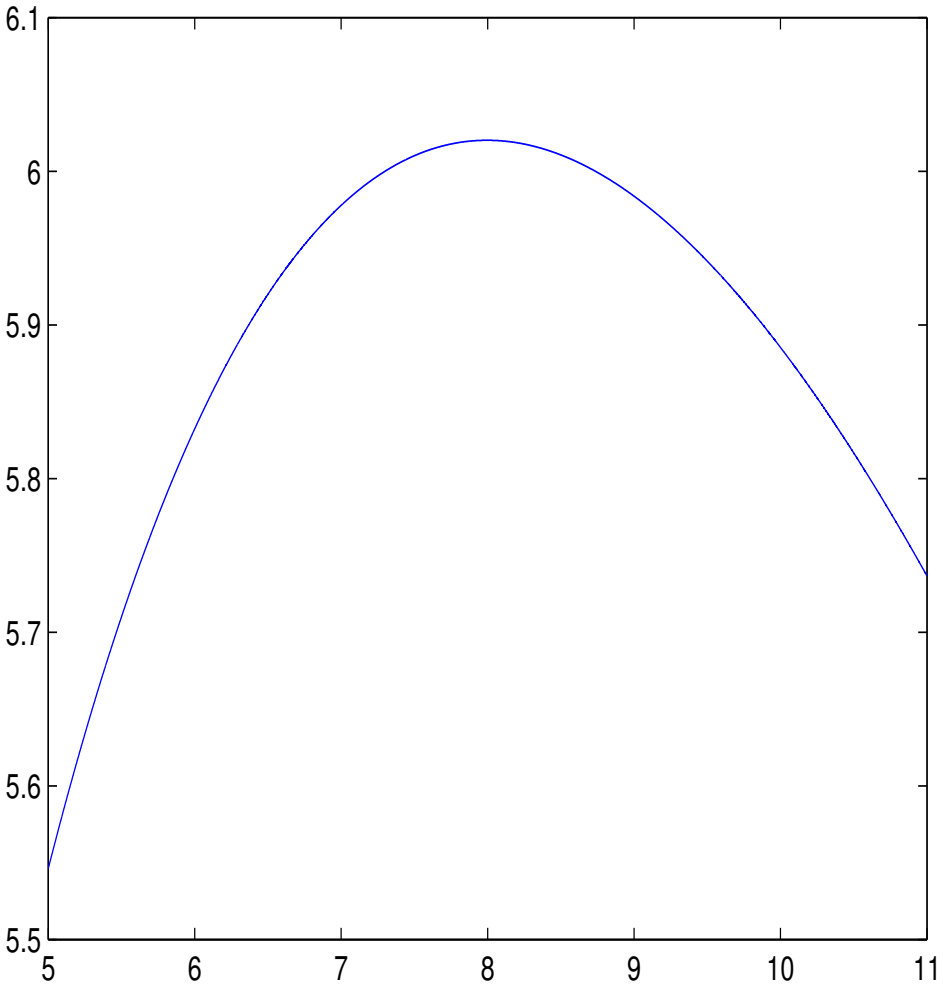
$$x_1 = 0.77, x_2 = 0.82, x_3 = 0.92, x_4 = 0.94, x_5 = 0.98.$$

Data-analytikern modellerar dessa mätvärden som oberoende utfall av en stokastisk variabel X med täthetsfunktionen,

$$f_X(x) = \begin{cases} \theta x^{\theta-1} & \text{om } 0 \leq x \leq 1, \\ 0 & \text{för övrigt,} \end{cases}$$

där $\theta > 0$.

Härled Maximum Likelihood-skattningen (ML-skattningen) av θ och beräkna den numeriskt för de givna mätvärdena. I figuren nedan har data-analytikern plottat för dessa mätvärden den naturliga logaritmen av likelihoodfunktionen som funktion av θ i ett visst intervall.



(4 p)

Lycka till!

Avd. Matematisk statistik

KTH Matematik

LÖSNINGSFÖRSLAG
TENTAMEN I SF1911 STATISTIK FÖR BIOTEKNIK.
Torsdag de 13e april 08:00-13:00.

Uppgift 1

i) Låt X = antalet celler som är infekterade. Följande utsagor är sanna.

b) $X \sim \mathcal{B}in(20, 0.2)$: samma två alternativ vid varje försök, samma sannolkhet för lyckat försök vid varje försök, et på förhand givet antal försök.

e) $E(X) = 4$, ty $E(X) = 20 \cdot 0.2 = 4$ när X är binomialfördelad $\mathcal{B}in(20, 0.2)$.

ii) c) De små talens lag.

Uppgift 2

i) Det rätta alternativet är

b) Talet 0.27 är en populationsparameter för stickprovet.

ii) Det rätta alternativet är

c) $\sqrt{0.27 \cdot 0.73/50}$.

iii) b) $X \sim \mathcal{B}in(50, 0.27)$. $50 \cdot 0.27 \cdot 0.73 = 9.8 \approx 10$. Vi använder en appproximation med en normalfördelning, och $\widehat{p} = X/50$ är approximativt $\sim \mathcal{N}(0.27, 0.27 \cdot 0.73/50)$.

$$P(X/50 \geq 0.25) = 1 - P(X/50 \leq 0.25) =$$

$$= 1 - 1 - P\left(\frac{X/50 - 0.27}{\sqrt{0.27 \cdot 0.73/50}} \leq \frac{0.25 - 0.27}{\sqrt{0.27 \cdot 0.73/50}}\right) = 1 - \Phi\left(\frac{0.25 - 0.27}{\sqrt{0.27 \cdot 0.73/50}}\right)$$

$$= 1 - \Phi\left(\frac{0.25 - 0.27}{\sqrt{0.27 \cdot 0.73/50}}\right)$$

$$= 1 - \Phi\left(-0.3185\right) = 1 - \left(1 - \Phi\left(0.3185\right)\right) = \Phi\left(0.3185\right) = 0.6249 \approx 0.625.$$

Uppgift 3

Sensitivitet $= P(T+|D+) = \frac{95}{95+5} = 0.95$. Specificitet $= P(T-|D-) = \frac{90}{95+5} = 0.90$. Sannolikheten för falsk positiv $= P(T+|D-) = \frac{10}{95+5} = 0.10$. Positivt prediktivt värde

$$= P(D+|T+) = \frac{P(T+|D+)}{P(T+)} = 95/(95+10) = 0.905.$$

Detta ger

d) Positivt prediktivt värde $= P(D+|T+) = 90.05$. Sant. %

f) Sensitivitet $= 95\%$. Sant.

Uppgift 4

En logistisk regression av formen

$$p_x = P(Y = 1 \mid x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

betraktas.

$$P(Y = 0 \mid x) = 1 - P(Y = 1 \mid x) = \frac{1}{1 + e^{\beta_0+\beta_1 x}}$$

Odds med $x = 1$ är

$$\frac{P(Y = 1 \mid 1)}{P(Y = 0 \mid 1)} = e^{\beta_0+\beta_1}.$$

Odds med $x = 0$ är

$$\frac{P(Y = 1 \mid 0)}{P(Y = 0 \mid 0)} = e^{\beta_0}.$$

Detta ger

$$OR = \frac{\frac{P(Y=1|1)}{P(Y=0|1)}}{\frac{P(Y=1|0)}{P(Y=0|0)}} = e^{\beta_1}.$$

Dessutom

$$\begin{aligned}\text{logit}(p_0) = \ln \frac{p_0}{1 - p_0} &= \ln \frac{P(Y = 1 \mid 0))}{P(Y = 0 \mid 0)} \\ &= \ln e^{\beta_0} = \beta_0.\end{aligned}$$

i) b) $OR = e^{\beta_1}$ är en sann utsaga.

ii) Den sanna formeln är

d) $\text{logit}(p_0) = \beta_0$.

Uppgift 5

Först bestäms

$$P\left(A \cap B\right) = P\left(B|A\right) P\left(A\right) = \frac{2}{3} \cdot \frac{3}{5} = \frac{2}{5}.$$

På samma sätt fås

$$P\left(A^c \cap B\right) = P\left(B|A^c\right) P\left(A^c\right) = \frac{1}{3} \cdot (1 - \frac{3}{5}) = \frac{1}{3} \cdot \frac{2}{5} = \frac{2}{15}.$$

Sålunda får vi att

$$P\left(B\right) = P\left(A \cap B\right) + P\left(A^c \cap B\right) = \frac{2}{5} + \frac{2}{15} = \frac{8}{15}.$$

Eftersom $P(A \cap B) = 2/5$, medan

$$P(A)\,P(B) = \frac{3}{5} \cdot \frac{8}{15} = \frac{8}{25} \neq \frac{2}{5},$$

så är händelserna A och B beroende.

Uppgift 6

a) Vi är intresserade av μ, medelvärdet för populationen av alla barn som diagnosticerats med progeria. Vår nollhypotes är

$$H_o : \mu = 6.6$$

Mothypotesen är

$$H_a : \mu > 6.6$$

för att vi vill veta huruvida barn som diagnosticerats med en aggressiv form av progeria har ett onormalt högt värde på PWV.

b) Inga utliggare förekommer. Data är någorlunda symmetriskt belägna i förhållande till medianen i lådagrammet.

c)

$$t = \frac{\bar{x} - 6.6}{s/\sqrt{18}} \sim t(18-1) = t(17).$$

d) Vi har från uppgiften att $\bar{x} = 12.44$ och $s = 3.638$.

$$t = \frac{\bar{x} - 6.6}{s/\sqrt{18}} = \frac{12.44 - 6.6}{3.638/\sqrt{18}} = 6.8106.$$

Det ensidiga kritiska området ges av $t_{0.001}(17) = 3.65$. Vi ser att $t = 6.8106 > 3.65$, och **nollhypotesen förkastas på signifikansnivån** 0.001.

Uppgift 7

Detta är stickprov i par. Vi bildar ett konfidensintervall för skillnaden $\mu_2 - \mu_1$. Formlerna i formelsamlingen ger skattningen av den gemensamma variansen som

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} = \frac{31 \cdot 13^2 + 31 \cdot 42^2}{32 + 32 - 2} = 966.5.$$

och

$$s = 31.0886.$$

Konfidensintervallet blir

$$I_{\mu_2-\mu_1} = \overline{y} - \overline{x} + t_{0.01}(32+32-2) \cdot s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Från tabellen $t_{0.01}(32+32-2) \approx t_{0.01}(60) = 2.39$.

$$s \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 31.0886 \cdot \sqrt{\frac{2}{32}} = 31.0886 \cdot \sqrt{\frac{1}{16}} = 31.0886 \cdot \frac{1}{4} = 7.7721.$$

Detta ger konfodensintervallet

$$I_{\mu_2-\mu_1} = (156 - 76 + 2.39 \cdot 7.7721, +\infty) = (98.5753, +\infty).$$

Nollhypotesen förkastas på signifikansnivån 1 % , ty värdet $\mu_2 - \mu_1 = 0$ ingår inte i det observerade intervallet $I_{\mu_2-\mu_1}$. De tillgängliga data ger den statistiskt säkerställda slutsatsen att högre dos leder till större högre koncentration.

Uppgift 8

Helt generellt hår vi Likelihoodfunktionen

$$L(\theta) = f_{X_1,X_2,\cdots,X_n}(x_1, x_2, \cdots, x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\cdots f_{X_n}(x_n) = \theta^n(x_1 \cdot x_2 \cdot \ldots \cdot x_n)^{\theta-1}.$$

Som vanligt lönar det sig att ta den naturliga logaritmen och detta ger

$$\ln L(\theta) = n\ln(\theta) + (\theta - 1)\sum_{j=1}^{n} \ln(x_j).$$

Derivering ger

$$\frac{d\ln(L(\theta)}{d\theta}) = \frac{n}{\theta} + \sum_{j=1}^{n} \ln(x_j)$$

och lösning av $d\ln(L(\theta))/d\theta = 0$ m.a.p. θ ger ML-skattningen

$$\widehat{\theta}_{mle} = -\frac{1}{\frac{1}{n}\sum_{j=1}^{n} \ln(x_j)}$$

Insättning av siffrorna (n=5)

$$x_1 = 0.77, x_2 = 0.82, x_3 = 0.92, x_4 = 0.94, x_5 = 0.98.$$

ger

$$\widehat{\theta}_{mle} = -\frac{1}{\frac{1}{5}(\ln(0.77) + \ln(0.82) + \ln(0.92) + \ln(0.94) + \ln(0.98))}$$
$$= 7.9965 \approx 8.0,$$

vilket även bekräftas med en visuell granskning av den plottade loglikelihoodfunktionen.