TDAB01 SANNOLIKHETSLÄRA OCH STATISTIK

TENTAMEN 2021-10-26

Lärare

Jose M. Peña. Nås via Teams (eller på `jose.m.pena@liu.se` om Teams inte fungerar).

Betyg

För full poäng i varje delfråga krävs tydliga och väl motiverade svar.
Maximalt antal poäng: 20 poäng
Betyg 5 = 18-20 poäng
Betyg 4 = 14-17 poäng
Betyg 3 = 10-13 poäng
Betyg U = 0-9 poäng

Tillåtna Hjälpmedel

Miniräknare med tomt minne. Tabell- och formelsamling (ingår i tentamen). Slides på kurswebbsidan (som pdf eller utskrivna, med eller utan egna anteckningar). Kursboken.

Instruktioner

- Ladda ner tentan från LISAM från kl 8.00. Lämna in dina svar i LISAM senaste kl 12.15. Om du har fått förlängd tid, v.v. och mejla dina svar till `jose.m.pena@liu.se`.
- Om du har frågor, v.v. och använd Teams för att kontakta mig (eller `jose.m.pena@liu.se` om Teams inte fungerar).
- Tentan är individuell, dvs hjälp från eller kommunikation med andra är inte tillåten.
- Tentan är anonym, dvs skriv ej ditt namn någonstans.
- Du rekommenderas lösa uppgifterna med papper och penna, fota lösningarna och ladda upp dem. Det är också OK att använda Word eller LaTeX.

Uppgifter

(1) (4 p = 2 + 1 + 1) X, Y och Z are binära variabler med sannolikhetsfördelningar $R(x)$, $S(y|x)$ och $T(z|x)$. Bevisa att

(a) $P(x,y,z) = R(x) \cdot S(y|x) \cdot T(z|x)$ är en giltig sannolikhetsfördelning.

(b) $P(x) = R(x)$, $P(y|x) = S(y|x)$ och $P(z|x) = T(z|x)$.

(c) Y och Z are betingat oberoende givet X.

Lösning:

(a) Först, obs. att $0 \leq P(x,y,z) \leq 1$ för alla x, y, z. Sedan, obs. att

$$\sum_{x,y,z} P(x,y,z) = \sum_x R(x) \sum_y S(y|x) \sum_z T(z|x) = 1.$$

(b) Först, obs. att

$$P(x) = \sum_{y,z} P(x,y,z) = R(x) \sum_y S(y|x) \sum_z T(z|x) = R(x).$$

Sedan, obs. att

$$P(x,y) = \sum_z P(x,y,z) = R(x) \cdot S(y|x) \sum_z T(z|x) = R(x) \cdot S(y|x).$$

Då, $P(y|x) = P(x,y)/P(x) = S(y|x)$. Likadant kan man bevisa att $P(z|x) = T(z|x)$.

(c) $P(y,z|x) = P(x,y,z)/P(x) = S(y|x) \cdot T(z|x) = P(y|x) \cdot P(z|x)$.

(2) (4 p = 2 + 1 + 1) En sjuksköterska jobbar åtta timmar varje dag. Hen testar en patient varannan minut. Hen upptäcker två smittade patienter per timme i genomsnitt.

(a) Vad är sannolikheten att hen upptäcker flera än 10 smittade patienter idag?

(b) Vad är sannolikheten att hen måste vänta mer än två timmar för att upptäcka den första smittade patienten?

(c) Vad är sannolikheten att hen upptäcker två smittade patienter i följd?

Lösning:

(a) $X \sim Binomial(n = 8 \cdot 60/2, p = 2/(60/2))$ och då $P(X > 10)$. $X \sim Binomial(n = 240, p = 1/15)$ kan approximeras som $N(\mu = n \cdot p, \sigma = \sqrt{n \cdot p \cdot (1-p)})$ eftersom $n > 30$. Då

$$P((X-\mu)/\sigma > (10.5-\mu)/\sigma) = P(Z > (10.5-\mu)/\sigma) = 1 - \Phi((10.5-\mu)/\sigma).$$

(b) $Y \sim Geometrisk(p = 2/(60/2))$ och då $P(Y > 2 \cdot 60/2)$. Eller, $W \sim Binomial(n = 2 \cdot 60/2, p = 2/(60/2))$ och då $P(W = 0) = (1-p)^n$.

(c) p^2, men man kan tolka frågan annorlunda.

(3) (2 p) Fabriker A och B producerar samma vara. I fabrik A, varans kvalitetsindikatorn följer en normal fördelning $N(\mu = 3, \sigma = 3)$. I fabrik B, kvalitetsindikatorn följer en likformig fördelning $U(a = -2, b = 2)$. I båda fabrikerna, varan godkänns om kvalitetsindikatorn tillhör intervallet väntervärdet plus/minus en standardavvikelse, dvs om kvaliteten inte avviker från väntevärdet med mer en standardavvikelse. Båda fabrikerna jobbar i samma takt. Vad är sannolikheten att en godkänd vara kommer från fabrik A?

Lösning:

Först, $P(Fabrik = A) = P(Fabrik = B) = 0.5$. För fabrik A,

$$P(|Indikator - \mu| \leq \sigma) = P(|Indikator - \mu|/\sigma \leq 1) = P(|Z| \leq 1) = \Phi(1) - \Phi(-1) = 0.68.$$

För fabrik B, $\mu = (a+b)/2 = 0$ och $\sigma = (b-a)/\sqrt{12} = 1.16$. Då,

$$P(|Indikator - \mu| \leq \sigma) = P(-1.16 \leq Indikator \leq 1.16) = \int_{-1.16}^{1.16} \frac{1}{b-a} dx = 2.26/4 = 0.58$$

Då, $P(Fabrik = A|Kvalitet = G) = 0.5 \cdot 0.68/(0.5 \cdot 0.68 + 0.5 \cdot 0.58)$.

(4) (4 p) Vi har mättat avvikelserna från den utlovade bredbandshästigheten de senaste 100 timmarna, och det blev så här:

Avvikelse	$[-\infty, -1.5]$	$[-1.5, -0.5]$	$[-0.5, +0.5]$	$[+0.5, +1.5]$	$[+1.5, +\infty]$
Antal timmar	10	20	40	15	15

Bestäm p-värdet för följande hypotestest: H_0 : Avvikelserna följer en normal fördelning $N(0, 1)$, mot H_1 : De följer en annan fördelning. Bestäm om du förkastar H_0 eller ej. **För att kunna lösa denna uppgift, måste du först läsa avsnitt 10.1 och 10.1.1 i kursboken**.

Lösning: Jag löste den i R, men ni borde lösa den på hand eftersom den är en papper-och-penna tenta.

```
e<-NULL
e[1]<-100*pnorm(-1.5)
e[2]<-100*(pnorm(-0.5)-pnorm(-1.5))
e[3]<-100*(pnorm(0.5)-pnorm(-0.5))
e[4]<-e[2]
e[5]<-e[1]
o<-c(10,20,40,15,15)
t<-sum((e-o)^2/e)
pchisq(t,df=4,lower.tail = FALSE)
```

(5) (3 p) Låt $X_1, \ldots, X_n$ vara oberoende och likafördelade variabler med fördelning $f(x; \theta) \propto (\theta^x e^{-\theta})/x!$ där θ är en parameter. Anta att apriorifördelningen för θ är $\pi(\theta) \propto \theta^{-1/2} e^{-\theta/2}$. Härled aposteriori fördelningen för θ. Tycker du att den apriori- och aposteriorifördelningarna är konjugerade? Förklara hur du skulle bygga ett 95% Bayesiansk konfidensintervall för θ. Kan du bygga flera såna intervall? Är alla lika bra?

Lösning:

Den aposteriori fördelningen är

$$\pi(\theta|x_1, \ldots, x_n) \propto \theta^{-1/2} e^{-\theta/2} \prod_{i=1}^{n} (\theta^{x_i} e^{-\theta})/x_i! \propto \theta^{\sum_{i=1}^{n} x_i - 1/2} e^{-(n+1)\theta/2}$$

Obs. att $f(x; \theta)$ är egentligen en Poisson fördelning och $\pi(\theta)$ är egentligen en Gamma fördelning. Då, den aposteriorifördelningen är också Gamma (v.v. och se boken). De är då konjugerade. För att bygga ett 95% Bayesiansk konfidensintervall $[a, b]$ räcker det med att lista ut a och b så att

$$\int_a^b \pi(\theta|x_1, \ldots, x_n) = 0.95$$

och det brukas finnas flera sådana intervall, men man brukar föredra den som innehåller de värdena för θ som har höst aposterioritäthet, också kallad highest posterior density intervall.

(6) (3 p) Bygg en regression modell från det följande strickprovet. Välj mellan linjär, polynom eller logistik regression.

X	-2	-1	0	1	2
Y	4.1	1.1	0	0.9	4.2

Ge din prediktion för populationens väntevärde för $X^* = 3$. Din regression modell bygger på vissa antagande. Identifiera en antagande som du kan faktiskt testa med hjälp av metoden i uppgift 4.

Lösning:

```
mindata<-matrix(data=c(-2,-1,0,1,2,4.1,1.1,0,0.9,4.2),ncol = 2,nrow = 5)

x<-mindata[,1]
y<-mindata[,2]
```

```
x<-x^2 # Det syns lätt att Y är en kvadratisk funktion av X.

b1<-sum((x-mean(x))*(y-mean(y)))/sum((x-mean(x))^2)
b0<-mean(y)-mean(x)*b1
prediktion<-b0+b1*3^2
```

Lycka till!