LINKÖPINGS UNIVERSITET
MAI
Johan Thim

Tentamen i matematisk statistik (9MA241/9MA341/LIMAB6, STN2) 2012-01-09 kl 08-13

Hjälpmedel är: miniräknare med tömda minnen och formelbladet bifogat.
Varje uppgift är värd 6 poäng. För godkänd tentamen räcker 16 poäng. Noggrann motivering krävs där alla viktiga detaljer skall motiveras.

För lösningsskisser, se kurshemsidan `www.mai.liu.se/~jothi/kurser/9MA241-stat/` efter skrivningens slut. Lycka till!

---

1. Låt A, B och C vara händelser så att $P(A) = 0.4$, $P(B \cap A^*) = 0.2$ och $P(B \cap C) = 0.05$. Vidare så är $P(A \cap B) = P(A \cap C) = 0.10$ och $A \cap B$ samt $A \cap C$ är disjunkta (de har inget snitt). Händelserna B och C är dessutom oberoende.

   (a) Bestäm $P(B)$. (2p)

   (b) Är A och B oberoende? Oförenliga? Motivera ditt svar! (2p)

   (c) Beräkna $P(A \mid B \cup C)$. (2p)

2. HiFi-Håkan funderar på att köpa nya högtalare, så han lånar hem ett par för att testa om de ger bättre ljudupplevelse än de han har idag. Håkan bestämmer sig för att genomföra försök med åtta av sina favoritlåtar (däribland *"Breaking the Law"* och *"Beyond the Realms of Death"*). Håkan lyssnar först på en låt med sina gamla högtalare, graderar upplevelsen i en 100-gradig skala från 0 till 10, byter till de nya, och lyssnar på samma låt igen och graderar på samma sätt. Antag att $X_j \sim N(\mu_j, \sigma_1)$ är upplevelserna med de gamla högtalarna och att $Y_j \sim N(\mu_j + \Delta, \sigma_2)$ är med de nya. Håkan skrev ned följande siffror.

x_j	6.5	7.8	7.1	9.0	6.8	7.5	8.1	9.1
y_j	8.2	8.1	8.0	8.9	7.5	7.5	8.5	7.8

   Beräkna ett 90% konfidensintervall för Δ. Kan hypotesen att de gamla högtalarna passar Håkan lika bra som de nya förkastas på 10% nivån? (6p)

3. Låt A vara (den fyllda) enhetskvadraten där $0 \leq x \leq 1$ och $0 \leq y \leq 1$. Vi definierar en tvådimensionell täthetsfunktion enligt

$$f_{X,Y}(x,y) = \begin{cases} c\,xy, & \text{punkten } (x,y) \text{ ligger i } A, \\ 0, & \text{för övrigt,} \end{cases}$$

   där c är en konstant.

   (a) Bestäm konstanten c så att $f_{X,Y}(x,y)$ blir en täthetsfunktion. (1p)

   (b) Beräkna $P(X > \frac{1}{4}, Y \leq \frac{1}{4})$. (3p)

   (c) Är X och Y oberoende stokastiska variabler? Motivera ditt svar. (2p)

4. Låt X vara likformigt fördelad på intervallet $[0, 10]$. Vi definierar Y enligt följande: $Y = 0$ om $0 \leq X < 1$, $Y = 1$ om $1 \leq X < 3$, $Y = 2$ om $3 \leq X < 6$ och $Y = 3$ om $6 \leq X \leq 10$.

   (a) Beräkna Y's sannolikhetsfunktion, väntevärde och standardavvikelse. (3p)

   (b) Om vi har 500 st oberoende variabler Y_k, alla med Y's fördelning, beräkna en approximativ sannolikhet att $\sum_{k=1}^{500} Y_k > 1025$. (3p)

5. Nyfikna Nina funderar över hur stor del av Sveriges befolkning som tycker man kan ha en rosa julgran. Stelbenta Stefan hävdar bestämt att andelen är 70%, men Nina vill gärna undersöka saken närmare. Hon frågar 100 slumpmässigt utvalda personer om man kan ha en rosa julgran. Hon får 80 ja och 20 nej.

   (a) Ställ upp ett dubbelsidigt konfidensintervall, med approximativ konfidensgrad 99%, för andelen p som tycker man kan ha en rosa julgran. Kan man statistiskt säkert säga något om Stefans påstående? (3p)

   (b) Vid närmare eftertanke tycker Nina att det räcker med att begränsa andelen p nedåt, alltså att hitta ett enkelsidigt konfidensintervall $I_p = [a, 1]$ för något a. Ställ upp ett sådant konfidensintervall med approximativ konfidensgrad 99%. Kan man säga något annat om Stefans kommentar nu? (3p)

6. (a) Visa att $V(X+Y) = V(X) + V(Y) + 2C(X, Y)$ för två stokastiska variabler X och Y. (2p)

   (b) Låt $A_1, A_2, \ldots, A_n$ vara händelser. Visa att $P(A_1 \cup A_2 \cup \cdots \cup A_n) \leq \sum_{k=1}^{n} P(A_k)$. (2p)

   (c) Låt Y vara en kontinuerlig stokastisk variabel så att $Y \geq 0$. Visa att

$$P(Y \geq a) \leq \frac{E(Y)}{a},$$

   för $a > 0$. (2p)

   Ledning: skriv ut integralen $E(Y)$ och dela upp på lämpligt sätt.

Lösningsskisser för tentamen i matematisk statistik, 9MA241, 2012-01-09

1. (a) Vi vet att $0.20 = P(B \cap A^*) = P(B) - P(A \cap B)$, och då $P(A \cap B) = 0.10$ följer det att $P(B) = 0.30$.

   (b) Om A och B är oberoende så gäller att $P(A \cap B) = P(A)P(B)$. Vi vet att $P(A \cap B) = 0.10$, men $P(A)P(B) = 0.40 \cdot 0.30 = 0.12$. Händelserna är alltså inte oberoende. Då $P(A \cap B) \neq 0$ kan händelserna inte heller vara oförenliga (att vara oförenliga skulle innebära att $A \cap B = \emptyset$, så $P(A \cap B) = 0$ i det fallet).

   (c) Enligt definitionen av betingad sannolikhet så gäller att

   $$P(A \mid B \cup C) = \frac{P(A \cap (B \cup C))}{P(B \cup C)}.$$

   Eftersom B och C är oberoende så är $P(B \cap C) = P(B)P(C)$. Detta ger att

   $$P(C) = P(B \cap C)/P(B) = 0.05/0.3 = 1/6.$$

   Vi erhåller nu att

   $$P(B \cup C) = P(B) + P(C) - P(B \cap C) = 0.30 + 0.1667 - 0.05 = 0.4167$$

   och att

   $$P(A \cap (B \cup C)) = P((A \cap B) \cup (A \cap C)) = P(A \cap B) + P(A \cap C) = 0.20.$$

   Vi kan nu beräkna att

   $$P(A \mid B \cup C) \approx \frac{0.20}{0.4167} \approx 0.48.$$

   Svar: (a) $P(B) = 0.30$. (b) Händelserna är varken oberoende eller oförenliga. (c) ≈ 0.48.

2. Modellen är stickprov i par. Vi betraktar skillnaden som ett enda stickprov och skattar väntevärdet med

   $$\bar{\Delta} = \sum_{i=1}^{8} y_i - x_i = 0.325$$

   och variansen med

   $$s = \left(\frac{1}{8-1} \sum_{i=1}^{10} (y_i - x_i - \bar{\Delta})^2 \right)^{1/2} = 0.8697.$$

   Vi bildar testvariabeln

   $$T = \frac{\Delta^* - \Delta}{S/\sqrt{n}} \sim t(7)$$

   där vi skattar S med s, så $s/\sqrt{n} = 0.3075$ och

   $$P(-t_{\alpha/2}(7) < T < t_{\alpha/2}(7)) = 1 - \alpha.$$

   Vi löser ut Δ ur intervallet i sannolikhetsmåttet och får att

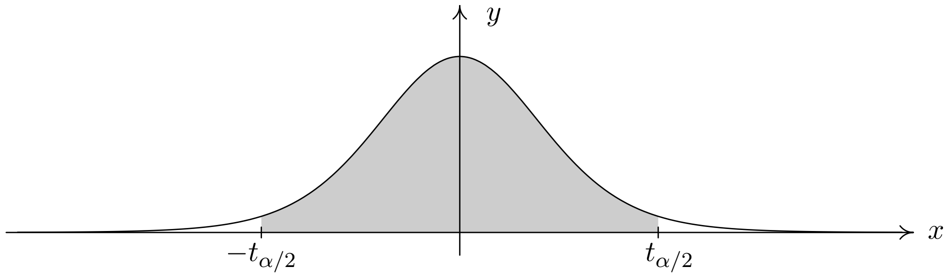
   $$\Delta^* - 0.3075 t_{\alpha/2}(7) < \Delta < \Delta^* + 0.3075 t_{\alpha/2}(7).$$

   Vi ersätter Δ^* med den observerade punktskattningen $\Delta^*_{\text{obs}} = 0.325$, d v s med medelvärdet av våra $y_j - x_j$, och erhåller ett konfidensintervall av konfidensgrad 90%:

   $$I_\Delta = [-0.258, 0.908],$$

   där vi använt $\alpha = 0.10$ och $t_{0.05}(7) = 1.895$. Eftersom $0 \in I_\Delta$ kan vi inte förkasta hypotesen.

   Svar: Hypotesen kan inte förkastas på 10%-nivån.



Figur 1: Den markerade arean innehåller 90% av sannolikheten för $t(7)$-fördelningen.

3. (a) För att $f_{X,Y}$ skall vara en täthetsfunktion krävs att dubbelintegralen av funktionen blir ett. Vi beräknar:

$$\int_0^1 \int_0^1 f_{X,Y}(x,y)\, dx\, dy = c \int_0^1 \int_0^1 xy\, dx\, dy = c \int_0^1 y \left[\frac{x^2}{2}\right]_{x=0}^{x=1} dy = \frac{c}{2} \int_0^1 y\, dy = \frac{c}{4}.$$

Vi ser här att $c = 4$ är nödvändigt.

(b) Vi söker sannolikheten att $X > 1/4$ och att $Y \leq 1/4$ samtidigt. Detta kan beräknas enligt

$$\int_{1/4}^1 \int_0^{1/4} 4xy\, dy\, dx = \cdots = \frac{15}{256}.$$

(c) Vi beräknar de marginella täthetsfunktionerna:

$$f_X(x) = \int_0^1 4xy\, dy = 2x\,, \quad 0 \leq x \leq 1$$
$$f_Y(y) = \int_0^1 4xy\, dy = 2y\,, \quad 0 \leq y \leq 1$$

Det följer att $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ för alla x och y.

Svar: (a) $c = 4$. (b) $P(X > 1/4, Y < 1/4) = 15/256$. (c) Oberoende.

4. Vi börjar med att ställa upp Y's sannolikhetsfunktion. Möjliga utfall är $Y = 0, 1, 2, 3$, och

$$p_Y(k) = P(Y = k) = \begin{cases} 1/10, & k = 0, \\ 1/5, & k = 1, \\ 3/10, & k = 2, \\ 2/5, & k = 3. \end{cases}$$

Vi kan nu räkna ut väntevärde och varians:

$$E(Y) = \sum_{k=0}^{3} k p_Y(k) = \frac{2 + 6 + 12}{10} = 2, \quad V(Y) = \sum_{k=0}^{3} k^2 p_Y(k) - E(Y)^2 = \frac{2 + 12 + 36}{10} - 4 = 1.$$

Låt $Z = \sum_{k=1}^{500} Y_k$. Då är $E(Z) = 500 \cdot 2 = 1000$ och $V(Z) = 500 \cdot 1$. Enligt centrala gränsvärdessatsen så kommer

$$P(Z > 1025) = 1 - P(Z \leq 1025) \approx 1 - \Phi\left(\frac{1025 - 1000}{\sqrt{500}}\right) = 1 - \Phi\left(\frac{\sqrt{5}}{2}\right) \approx 1 - \Phi(1.12) = 0.13.$$

Svar: (a) Se ovan. (b) $P(Z > 1025) \approx 0.13$.

5. Vi söker konfidensintervall för andel. Nina har frågat $n = 100$ stycken personer, och fått $X = 80$ stycken ja. En naturlig skattning på den verkliga andelen ges av

$$\widehat{P} = \frac{X}{n},$$

och vårt observerade värde är $\widehat{p} = 80/100 = 0.80$. Då $X \sim \mathrm{Hyp}(N, n, p)$, där p är den verkliga (okända) andelen och N är hela Sveriges befolkning, kan vi med gott samvete approximera med att $X \sim \mathrm{Bin}(n, p)$ om vi antar att Nina verkligen gjort ett helt slumpmässigt urval. Vidare ser vi att

$$np(1-p) \approx 100 \cdot 0.8 \cdot 0.2 = 16,$$

så vi borde även kunna göra en normalapproximation: $X \overset{\text{appr.}}{\sim} N(np, \sqrt{np(1-p)})$. Alltså blir $\widehat{P} \overset{\text{appr.}}{\sim} N(p, d)$, där vi använder "medelfelet"

$$d = \sqrt{\widehat{p}(1-\widehat{p})/n} = \sqrt{0.16/100} = 0.04.$$

Vår testvariabel blir alltså

$$Z = \frac{\widehat{P} - p}{d} \overset{\text{appr.}}{\sim} N(0, 1).$$

Vi stänger in Z:

$$P(-\lambda_{\alpha/2} < Z < \lambda_{\alpha/2}) = 1 - \alpha,$$

där λ_α är normalfördelningens α-kvantil. Vi har $\alpha = 0.01$ och $\lambda_{\alpha/2} = \lambda_{0.005} = 2.57$. Vi löser ut p ur olikheten inuti sannolikhetsmåttet, och erhåller då

$$\widehat{P} - \lambda_{0.005} d \leq p \leq \widehat{P} + \lambda_{0.005} d.$$

Detta ger oss konfidensintervallet $I_p = [0.697, 0.903]$ om vi ersätter $\widehat{P}$ med det observerade värdet $\widehat{p} = 0.80$.

Om vi istället bara vill ha ett nedåt begränsat intervall så kan vi utnyttja att

$$P(Z < \lambda_\alpha) = 1 - \alpha$$

och återigen lösa ut p ur olikheten inuti sannolikhetsmåttet. Detta ger oss $p > \widehat{P} - \lambda_{0.01} d$. Vi ersätter $\widehat{P}$ och använder oss av $\lambda_{0.01} \approx 2.325$ och får konfidensintervallet $I_p = [0.707, 1]$ (varför 1 som höger ändpunkt?).

Svar: (a) $I_p = [0.697, 0.903]$. Då $p = 0.70 \in I_p$ kan vi inte säga att Stefan hade fel.
(b) $I_p = [0.707, 1]$. Nu är $p = 0.70 \notin I_p$, så vi kan på nivån 1% säga att Stefan hade fel.

Anm.: vi använder approximationer, och andelen ligger nära gränsen så man bör vara lite försiktig med att dra slutsatser.

6. (a) Låt $E(X) = \mu_X$ och $E(Y) = \mu_Y$. Enligt definition gäller att

$$\begin{aligned}
V(X+Y) &= E\big((X + Y - \mu_X - \mu_Y)^2\big) \\
&= E\big((X-\mu_X)^2 + 2(X-\mu_X)(Y-\mu_Y) + (Y-\mu_Y)^2\big) \\
&= E\big((X-\mu_X)^2\big) + E\big((Y-\mu_Y)^2\big) + 2E\big((X-\mu_X)(Y-\mu_Y)\big) \\
&= V(X) + V(Y) + 2C(X, Y)
\end{aligned}$$

(b) Vi utnyttjar att $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (rita Venn-diagram) och ser följande:

$$\begin{aligned}
P\big(A_1 \cup A_2 \cup \cdots \cup A_n\big) &= P\big(A_1 \cup (A_2 \cup \cdots \cup A_n)\big) \\
&= P(A_1) + P\big(A_2 \cup A_3 \cup \cdots \cup A_n\big) - P\big(A_1 \cap \big(A_2 \cup A_3 \cup \cdots \cup A_n\big)\big) \\
&\leq P(A_1) + P\big(A_2 \cup A_3 \cup \cdots \cup A_n\big) \\
&\leq \cdots \leq \sum_{k=1}^{n} P(A_k)
\end{aligned}$$

(c) Vi ser att

$$\begin{aligned} E(Y) &= \int_0^\infty y f_Y(y)\,dy = \int_0^a y f_Y(y)\,dy + \int_a^\infty y f_Y(y)\,dy \\ &\geq \int_a^\infty y f_Y(y)\,dy \\ &\geq a\int_a^\infty f_Y(y)\,dy = aP(Y \geq a), \end{aligned}$$

så division med a ger den sökta olikheten: $P(Y \geq a) \leq \frac{1}{a}E(Y)$. Olikheterna i beviset gäller eftersom integranden hela tiden är positiv, så vi kan ta bort delar och få ett uttryck som är mindre.

Svar: Se ovan.