LINKÖPINGS UNIVERSITET
MAI
Johan Thim

Tentamen i matematisk statistik (9MA241/9MA341, STN2) 2012-06-05 kl 14-18

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. (a) Låt A och B vara oberoende händelser så att $P(A) = 0.4$ och $P(B) = 0.6$. Beräkna $P(A \cup B)$. (2p)

   (b) Antag att händelserna i föregående deluppgift inte är oberoende, men att $P(A \mid B) = 0.5$. Beräkna $P(B \mid A)$. (2p)

   (c) I en fabrik har man tre maskiner, M_1, M_2 och M_3, som tillverkar samma produkter. Felrisken för en produkt tillverkad av maskin M_1 är 0.03, felrisk för M_2 är 0.05 och felrisk för M_3 är 0.065. På en dag tillverkar man 200 produkter i maskin M_1, 150 produkter i maskin M_2 och 300 produkter i maskin M_3. Bestäm sannolikheten att en vid dagens slut slumpmässigt utvald produkt, som visar sig vara trasig, kommer från maskin M_2. (2p)

2. (a) Vid 20 upprepade mätningar av en process fann man att medelvärdet blev $\bar{x} = 30$ och stickprovsvariansen $s_x^2 = 7.2$. Antag att mätningarna är observationer av en stokastisk variabel $X \sim \mathrm{N}(\mu_1, \sigma)$. Konstruera ett 99% konfidensintervall för väntevärdet μ_1. (3p)

   (b) Ett annat företag föreslår en annan process som ger samma resultat och hävdar att deras proces ger minst två gånger så högt förväntat värde. Som bevis har de gjort 15 mätningar på den nya processen och funnit att medelvärdet blev $\bar{y} = 62$ och stickprovsvariansen $s_y^2 = 8.8$. Antag att mätningarna är observationer av en stokastisk variabel $Y \sim \mathrm{N}(\mu_2, \sigma)$, där vi har samma standardavvikelse som i den första processen. Konstruera ett 95% konfidensintervall för $\mu_2 - 2\mu_1$. Vad blir slutsatsen angående det nya företagets påstående? (3p)

3. Låt $f_{X,Y}(x, y) = c(1 + xy)$ för $0 \leq x \leq 1$ och $0 \leq y \leq 1$, där c är en lämplig konstant. Annars är $f_{X,Y}(x, y) = 0$. Låt den två-dimensionella s.v. (X, Y) ha täthetsfunktionen $f_{X,Y}$.

   (a) Bestäm ett värde på konstanten c så att $f_{X,Y}$ verkligen blir en täthetsfunktion. (1p)

   (b) Beräkna $P(X > Y)$. (2p)

   (c) Beräkna kovariansen $C(X, Y)$. Är X och Y oberoende? (3p)

Var god vänd!

4. En viss enhet har en felrisk på 0.05 (sannolikheten att enheten är trasig). Man paketerar dessa i lådor med 120 enheter i varje låda.

   (a) Beräkna sannolikheten att en låda innehåller fler än sex (≥ 7) stycken trasiga enheter.

   (b) En last med 1000 stycken lådor (med 120 enheter per låda) har köpts in. Vi väljer ut 250 av dessa lådor slumpmässigt. Vad är sannolikheten att mindre än 91 (≤ 90) stycken av dessa 250 lådor innehåller fler än sex stycken trasiga enheter?

   Lämpliga approximationer är tillåtna i både (a) och (b).

5. Nina har frågat 290 stycken **Doctor Who** fans om vilken av de två senaste doktorerna (nummer tio och elva) som är deras favorit. Av dessa svarade 116 att det var nummer elva, *Matt Smith*. Övriga tyckte att *David Tennant* hade presterat bättre.

   (a) Ange ett approximativt 95% konfidensintervall för andelen p som föredrog Matt Smith. (3p)

   (b) Ett konfidensintervall för andelen p ges av $I_p = [0, 0.4369]$. Vad har detta intervall för konfidensgrad (approximativt)? Rimliga antaganden om hur intervallet har beräknats får göras. (3p)

6. Låt X vara en binomialfördelad stokastisk variabel med parametrarna n och p, $n \geq 1$ heltal och $0 < p < 1$. Visa att väntevärdet för X ges av $E(X) = np$ och att variansen för X ges av $V(X) = np(1-p)$. Det räcker inte med att hänvisa till formelsamling! (6p)

Lösningsskisser för tentamen i matematisk statistik, 9MA241, 2012-06-05

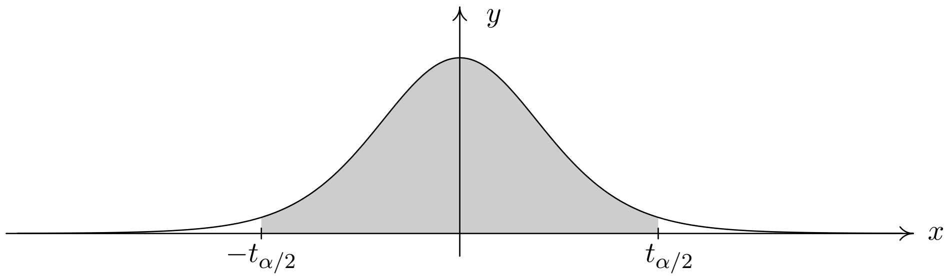
1. (a) Eftersom $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, och $P(A \cap B) = P(A)P(B)$ ty A och B är oberoende, så gäller att
$$P(A \cup B) = 0.4 + 0.6 - 0.24 = 0.76.$$

   (b) Nu är inte A och B oberoende, men eftersom vi vet att $P(A \,|\, B) = 0.5$ kan vi räkna ut att $P(A \cap B) = P(A \,|\, B)P(B) = 0.3$. Vidare gäller att
$$P(B \,|\, A) = \frac{P(A \cap B)}{P(A)} = \frac{0.3}{0.4} = 0.75.$$
Faktum är att vi nu härlett Bayes sats i specialfallet med ett B_i:
$$P(B \,|\, A) = \frac{P(B)P(A \,|\, B)}{P(A)}.$$

   (c) Det tillverkas totalt sett 650 enheter på en dag. Från detta erhåller vi sannolikheterna att en enhet producerats av en viss maskin: $P(M_1) = 200/650 = 0.308$, $P(M_2) = 150/650 = 0.231$ och $P(M_3) = 300/650 = 0.461$. Låt T vara händelsen att enheten vi väljer ut är trasig. Enligt uppgiften gäller nu att $P(T \,|\, M_1) = 0.03$, $P(T \,|\, M_2) = 0.05$ och $P(T \,|\, M_3) = 0.065$. Lagen om total sannolikhet ger att händelsen att en enhet vid dagens slut är trasig blir
$$P(T) = 0.308 \cdot 0.03 + 0.231 \cdot 0.05 + 0.461 \cdot 0.065 = 0.0508.$$
Vi söker sannolikheten $P(M_2 \,|\, T)$. Vi använder Bayes sats för ”vända” på betingningen:
$$P(M_2 \,|\, T) = \frac{P(T \,|\, M_2)P(M_2)}{P(T)} = \frac{0.05 \cdot 0.231}{0.0508} = 0.227.$$

Svar: (a) 0.76. (b) 0.75. (c) 0.227.

2. (a) Vi har $n = 20$ och bildar testvariabeln
$$T = \frac{\overline{X} - \mu_1}{S_x/\sqrt{n}} =\sim t(19)$$
och ser att
$$P(-t_{\alpha/2}(19) \leq T \leq t_{\alpha/2}(19)) = 1 - \alpha.$$
Vi löser ut μ_1 ur intervallet i sannolikhetsmåttet och får att
$$\overline{X} - \frac{S_x}{\sqrt{n}} t_{\alpha/2}(19) \leq \mu_1 \leq \overline{X} + \frac{S_x}{\sqrt{n}} t_{\alpha/2}(19).$$
Vi ersätter $\overline{X}$ med den observerade punktskattningen $X^*_{\text{obs}} = \overline{x} = 30$ och S_x med $s_x = \sqrt{7.2}$. Då erhåller vi ett konfidensintervall med konfidensgrad 99%:
$$I_{\mu_1} = [28.3, 31.7],$$
där vi använt $\alpha = 0.01$ och $t_{0.005}(19) = 2.861$.



Figur 1: Den markerade arean innehåller 99% av sannolikheten för $t(19)$-fördelningen.

(b) Vi använder oss av testvariabeln

$$T = \frac{\overline{Y} - 2\overline{X} - (\mu_2 - 2\mu_1)}{S\sqrt{1/20 + 4/15}} \sim t(33),$$

där vi kommer att skatta S^2 med

$$s^2 = \frac{19s_x^2 + 14s_y^2}{33} = 7.879.$$

Det följer att

$$P(-t_{\alpha/2}(33) \leq T \leq t_{\alpha/2}(33)) = 1 - \alpha.$$

På samma sätt som ovan löser vi ut $\mu_2 - 2\mu_1$ ur intervallet i sannolikhetsmåttet och får att

$$\overline{Y} - 2\overline{X} - S\sqrt{4/20 + 1/15}t_{\alpha/2}(33) \leq \mu_2 - 2\mu_1 \leq \overline{Y} - 2\overline{X} + S\sqrt{4/20 + 1/15}t_{\alpha/2}(33).$$

Vi ersätter $\overline{Y} - 2\overline{X}$ med den observerade punktskattningen $\overline{y} - 2\overline{x} = 2$ och S enligt ovan. Då erhåller vi ett konfidensintervall av konfidensgrad 95%:

$$I_{\mu_2 - 2\mu_1} = [-0.95, 4.95],$$

där vi använt $\alpha = 05$ och $t_{0.025}(33) \approx 2.042 + 0.3 \cdot (2.021 - 2.042) = 2.0357$ (linjär interpolation[1]).

Svar:

(a) $I_{\mu_1} = [28.3, 31.7]$.

(b) $I_{\mu_2 - 2\mu_1} = [-0.95, 4.95]$. Nej, med signifikansnivån 5% kan man inte statistiskt säkert styrka det nya företagets påstående.

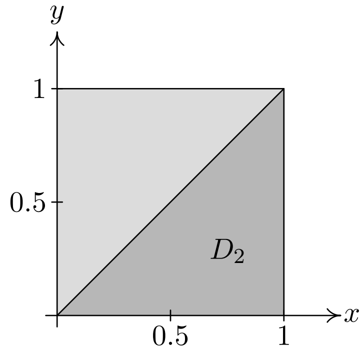
3. Vi skisserar området i Figur 2.

(a) Om $c > 0$ så är $f(x, y) \geq 0$ för alla x och y. Återstår att visa att arean blir ett. Vi itererar med y-integralen ytterst:

$$\iint_D c(1 + xy)\,dxdy = c\int_0^1 \int_0^1 \big(1 + xy\big)\,dxdy = c\int_0^1 \left[x + \frac{x^2 y}{2}\right]_{x=0}^1 dy$$
$$= c\int_0^1 \big(1 + \frac{y}{2}\big)\,dy = c\left[y + \frac{y^2}{4}\right]_{y=0}^1 = c\frac{5}{4}$$

Detta implicerar att $c = 4/5$ är nödvändigt för att f skall vara en täthetsfunktion.

[1] Det är ok att avrunda här också om man vill.



Figur 2: Området i uppgift 3. Delområdet D_2 är det vi är intresserade av i deluppgift (b).

(b) I figur 2 ser vi att området där $x > y$ ges av triangeln D_2. Detta område kan beskrivas som $0 \leq x \leq 1$ och $0 \leq y < x$. Om man undersöker $f_{X,Y}(x, y)$ lite närmare kan man inse att funktionen är symmetrisk kring linjen $y = x$ (d v s att samma funktionsvärden dyker upp på båda sidorna av linjen). Eftersom integralen vi behöver räkna ut täcker hela området under linjen $y = x$ måste således sannolikheten vara lika med en halv (hälften av volymen finns på varje sida linjen).

Tycker man inte om det argumentet kan man helt sonika räkna ut integralen. Vi itererar med x-integralen ytterst:

$$\begin{aligned}\int_0^1 \int_0^x f_{X,Y}(x, y)\, dydx &= \frac{4}{5} \int_0^1 \int_0^x (1 + xy)\, dydx = \frac{4}{5} \int_0^1 \left[y + \frac{xy^2}{2} \right]_{y=0}^{x} dx \\ &= \frac{4}{5} \int_0^1 (x + \frac{x^3}{2})\, dx = \frac{4}{5} \left[\frac{x^2}{2} + \frac{x^4}{8} \right]_{x=0}^{1} dx \\ &= \frac{4}{5}\frac{5}{8} = \frac{1}{2}.\end{aligned}$$

(c) Kovariansen kan, t ex, beräknas med formeln $C(X, Y) = E(XY) - E(X)E(Y)$. Vi börjar med att ta fram $E(XY)$:

$$\begin{aligned}E(XY) &= \int_0^1 \int_0^1 xy \cdot c(1 + xy)\, dxdy = c \int_0^1 \int_0^1 (xy + x^2y^2)\, dxdy \\ &= c \int_0^1 \left[\frac{x^2y}{2} + \frac{x^3y^2}{3} \right]_{x=0}^{1} dy = c \int_0^1 \left(\frac{y}{2} + \frac{y^2}{3} \right) dy = c \left[\frac{y^2}{4} + \frac{y^3}{9} \right]_{y=0}^{1} = \frac{4}{5}\frac{13}{36}.\end{aligned}$$

När det gäller $E(X)$ och $E(Y)$ kan man av symmetriskäl se att $E(X) = E(Y)$ (ser man inte det får man räkna ut båda integralerna). Vi undersöker $E(X)$ närmare:

$$\begin{aligned}E(X) &= \int_0^1 \int_0^1 x \cdot c(1 + xy)\, dxdy = c \int_0^1 \int_0^1 (x + x^2y)\, dxdy \\ &= c \int_0^1 \left[\frac{x^2}{2} + \frac{x^3y}{3} \right]_{x=0}^{1} dy = c \int_0^1 \left(\frac{1}{2} + \frac{y}{3} \right) dy = c \left[\frac{y}{2} + \frac{y^2}{6} \right]_{y=0}^{1} = \frac{4}{5}\frac{2}{3}.\end{aligned}$$

Total sett erhåller vi nu att

$$C(X, Y) = E(XY) - E(X)E(Y) = \frac{1}{225}.$$

Eftersom $C(X, Y) \neq 0$ så kan inte X och Y vara oberoende. Man kan också beräkna de de marginella täthetsfunktionerna och kontrollera att $f_X(x)f_Y(y) \neq f_{X,Y}(x, y)$ för vissa punkter (x, y).

Svar: (a) $c = 4/5$. (b) $P(X > Y) = 1/2$. (c) $C(X, Y) = 1/225$, beroende. Se ovan.

4. (a) Låt X vara antalet trasiga enheter i en låda. Vi vet att $X \sim \text{Bin}(120, 0.05)$ och söker sannolikheten $P(X \geq 7)$. Vi approximerar med Poissonfördelning (normalapproximation fungerar inte), vilket är OK ty $n = 120 \geq 10$ och $p = 0.05 \leq 0.1$. Således är $X \overset{\text{appr.}}{\sim} \text{Po}(6)$. Vi erhåller nu att

$$P(X \geq 7) = 1 - P(X \leq 6) \approx 1 - 0.6063 = 0.3937$$

ur Poissontabell.

(b) Vi väljer ut 250 lådor från 1000 stycken slumpmässigt. Låt Y vara antalet av dessa 250 lådor som innehåller fler än 6 stycken trasiga enheter. Vi söker sannolikheten att mindre än 91 stycken innehåller fler än 6 stycken trasiga enheter. Det följer att $Y \sim \text{Hyp}(N, n, p)$ där $N = 1000$, $n = 250$ och $p \approx 0.3937$. Vi söker $P(Y \leq 90)$. Vi beräknar variansen för Y:

$$V(Y) = \frac{1000 - 250}{1000 - 1} 250 \cdot 0.3937 \cdot 0.6063 = 44.80 \gg 10.$$

Således kan vi använda normalapproximation, och finner att $Y \overset{\text{appr.}}{\sim} \text{N}(98.425, 6.69)$. Sannolikheten kan nu beräknas enligt

$$P(Y \leq 90) \approx \Phi\left(\frac{90 - 98.425}{6.69}\right) = \Phi(-1.26) = 1 - \Phi(1.26) = 0.1038.$$

Svar: (a) ≈ 0.40. (b) ≈ 0.10.

5. (a) Vi söker konfidensintervall för andelen som svarat Matt Smith. Nina har frågat $n = 290$ stycken personer, och fått $X = 116$ stycken Matt Smith. En naturlig skattning på den verkliga andelen ges av

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

och vårt observerade värde är $\widehat{p} = 116/290 = 0.40$. Fördelningen för X ges av $X \sim \text{Bin}(n, p)$. Vidare ser vi att

$$np(1 - p) \approx 289 \cdot 0.4 \cdot 0.6 = 69.6,$$

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

$$d = \sqrt{\widehat{p}(1 - \widehat{p})/n} = \sqrt{69.6/290^2} = 0.0288.$$

Vår testvariabel blir alltså

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

Vi stänger in Z:

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

där λ_α är normalfördelningens α-kvantil. Vi har $\alpha = 0.05$ och $\lambda_{\alpha/2} = \lambda_{0.025} = 1.96$. Vi löser ut p ur olikheten inuti sannolikhetsmåttet, och erhåller då

$$\widehat{P} - \lambda_{0.025} d \leq p \leq \widehat{P} + \lambda_{0.025} d.$$

Detta ger oss konfidensintervallet $I_p = [0.34, 0.46]$ om vi ersätter $\widehat{P}$ med det observerade värdet $\widehat{p} = 0.40$.

(b) I uppgift (b) har vi fått intervallet $I_p = [0, 0.4369]$. Vi antar att man har använt samma testvariabel som vi gjort ovan. Det innebär att I_p kommer från sambandet

$$P(-\lambda_\alpha \leq Z) = 1 - \alpha,$$

vilket ger att $p \leq \widehat{P} + \lambda_\alpha d$. Vi skattar $\widehat{P}$ med $\widehat{p} = 0.40$ och $d = 0.0288$ som i uppgift (a). Vi jämnför med siffrorna i I_p och ser att $0.4369 - 0.40 = \lambda_\alpha d$, så $\lambda_\alpha = 1.28$. Eftersom $\Phi(1.28) = 0.90$ följer det att $\alpha = 0.10$ så konfidensgraden är 90%.

Svar: (a) $I_p = [0.34, 0.46]$. (b) ca 90%.

6. Låt $X \sim \mathrm{Bin}(n, p)$. Vi betraktar n stycken olika försök där varje försök har chansen p att lyckas. Låt $X_i = 0$ om försök nummer i misslyckas och $X_i = 1$ om försök nummer i lyckas. Det följer att vi måste ha $X_1 + X_2 + \cdots X_n = X$. Varje X_i är oberoende av övriga X_j och sannolikhetsfunktionen ges av $p_{X_i}(k) = 1 - p$ om $k = 0$ och $p_{X_i}(k) = p$ om $k = 1$. Följaktligen kan vi beräkna

$$E(X_i) = 0 \cdot (1 - p) + 1 \cdot p = p$$

och

$$V(X_i) = E(X_i^2) - E(X_i)^2 = 0^2 \cdot (1 - p) + 1^2 \cdot p - p^2 = p - p^2.$$

Vidare ser vi att

$$E(X) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} p = np,$$

och eftersom variablerna i summan är oberoende, även

$$V(X) = V\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} V(X_i) = \sum_{i=1}^{n} (p - p^2) = n(p - p^2) = np(1 - p).$$

Svar: Se ovan.