LINKÖPINGS UNIVERSITET
MAI
Johan Thim

Tentamen i matematisk statistik (92MA31, STN2)
2013-10-16 kl 08–12

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 efter skrivningens slut. Lycka till!

---

1. Motivera svaren på följande frågor noggrant!
   (a) Låt $f(x) = x$ för $0 \leq x < 1$ och $f(x) = 0$ för övrigt. Är f en täthetsfunktion? (1p)
   (b) Låt $f(x) = 2x/3$ för $-1 \leq x \leq 2$ och $f(x) = 0$ för övrigt. Är f en täthetsfunktion? (1p)
   (c) Låt $f(x, y) = x^2 - y^2$ för $-1 < x < 1$ och $0 \leq y \leq 1$, och $f(x, y) = 0$ för övrigt. Är f en täthetsfunktion? (1p)
   (d) Låt $P(A) = P(B) = 1$. Visa att om A och B är oberoende så är $P(A \cap B) = 1$. (1p)
   (e) Låt $P(A) = P(B) = 1$. Gäller $P(A \cap B) = 1$ även om A och B är beroende? Bevisa eller ange motexempel. (1p)
   (f) Låt $P(A \mid B) = 0.45$ och $P(B^*) = 0.32$. Beräkna $P(A)$ om $A \subset B$. (1p)
2. (a) Låt $X \sim \mathrm{N}(10, 2)$ och $Y \sim \mathrm{N}(12, 4)$ vara oberoende. Beräkna $P(X > Y)$ (2p)
   (b) Låt $X \sim \mathrm{N}(10, \sigma)$. Vad måste σ vara för att $P(X > 12) = 0.2$? (2p)
   (c) Antag att X_i, $i = 1, 2, \ldots, 7$, är oberoende och har fördelningen $X_i \sim \mathrm{N}(\mu, \sigma)$ där både μ och σ är okända. Om man vid en observationsserie erhöll följande värden, ange ett 99% konfidensintervall för μ. (2p)

i	1	2	3	4	5	6	7
x_i	10.0	9.5	9.9	9.2	10.6	11.1	10.4

   Tips: $\overline{x} = 10.1$ och $\displaystyle\sum_{i=1}^{7}(x_i - \overline{x})^2 = 2.56$.
3. Låt $f(x, y) = cx(y + 1)$ på rektangeln $0 \leq x \leq 3$ och $0 \leq y \leq 2$.
   (a) Bestäm konstanten c. (1p)
   (b) Beräkna $f_X(x)$ och $f_Y(y)$. Är variablerna X och Y oberoende (2p)
   (c) Beräkna $P(3Y > 6 - 2X)$. (3p)

Vänd!

4. Bonden Benny har tröttnat på sina mjölkkor och har istället börjat odla sydamerikansk kaktus i sin gigantiska ladugård. Enligt tidigare undersökningar har det visat sig att 25% av den sorts kaktus som Benny odlar innehåller tillräckligt mycket meskalin för att kunna extraheras effektivt. Benny har odlat 5030 stycken kaktusar. Approximationer nedan är OK om dessa motiveras.

   (a) Antag att påståendet 25% stämmer. Om Benny på måfå väljer ut precis 20 stycken kaktusar från ladugården och testar dessa, vad är sannolikheten att fyra eller fler innehåller tillräckligt mycket meskalin för effektiv extraktion? (3p)

   (b) När Benny till slut testade 200 stycken kaktusar fann han att 42 innehöll tillräckligt med meskalin. Ange ett 95% konfidensintervall för andelen av alla kaktusar i ladugården som har tillräckligt hög meskalinkoncentration. Är 25% rimligt? (3p)

5. Cigarrconnoisseuren Cesar börjar få tomt i sin humidor och har endast 3 st Cohiba Esplendidos, 2 st H. Upmann Monarchs, och 5 st Partagas Lusitanias kvar.

   (a) Om Cesar på måfå (utan att titta) plockar tre cigarrer, vad är sannolikheten att han finner minst en H. Upmann Monarch bland de tre utvalda? (2p)

   (b) Antag att Cesar på måfå (utan att titta) plockar en cigarr i taget. Låt X vara antalet cigarrer till och med att Cesar får en Cohiba Esplendidos för första gången. Beräkna sannolikhetsfunktionen p_X samt väntevärdet $E(X)$. (4p)

6. Låt $X \sim \text{Bin}(n, p)$ där $np = 2950$. För vilka p gäller att $P(\ln X > 8) \leq 0.05$? Svaret får ej innehålla n. Approximationer går bra om dessa motiveras. (6p)

Lösningsskisser för matematisk statistik 2013-10-16

1. (a) Nej. Vi ser att $\int_0^1 x\,dx = 1/2$, så arean är inte ett.
   (b) Nej. Funktionen är < 0 då $-1 < x < 0$.
   (c) Nej. Funktionen är < 0 till exempel då $x = 0$ och $0 < y < 1$.
   (d) Detta gäller eftersom $P(A \cap B) = P(A)P(B) = 1 \cdot 1 = 1$.
   (e) Fortfarande sant. Detta följer ur

   $$1 = P(A \cup B) = P(A) + P(B) - P(A \cap B) = 2 - P(A \cap B).$$

   Enda möjligheten är att $P(A \cap B) = 1$.
   (f) $P(A) = P(A \cap B) = P(A \mid B)P(B) = 0.45 \cdot (1 - 0.32) = 0.306$.

   Svar: (a) Nej. (b) Nej. (c) Nej. (d)-(e) Se ovan. (f) 0.306.

2. (a) Låt $Z = Y - X$. Då är $Z \sim \mathrm{N}(12 - 10, \sqrt{4^2 + 2^2}) = \mathrm{N}(2, \sqrt{20})$, så

   $$P(X > Y) = P(Z < 0) = P((Z-2)/\sqrt{20}) = \Phi(-2/\sqrt{20}) = 1 - \Phi(2/\sqrt{20}) \approx 0.33.$$

   (b) Vi vill ha:

   $$0.2 = P(X > 12) = 1 - P(X \leq 12) = 1 - P((X - 10)/\sigma \leq 2/\sigma) \quad \Leftrightarrow \quad \Phi(2/\sigma) = 0.8.$$

   Ur tabell finner vi nu att $2/\sigma = 0.8416$, så $\sigma = 2.38$.
   (c) Vi skattar σ med s som ges enligt

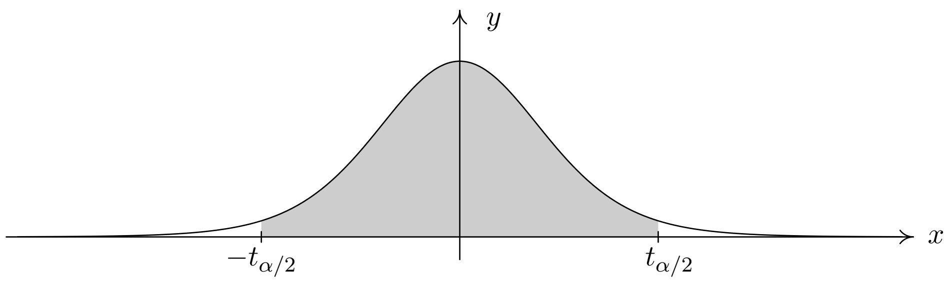
   $$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{2.56}{6} = 0.4267.$$

   Då vet vi att

   $$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim \mathrm{t}(6),$$

   så

   $$P(-t_{\alpha/2}(6) \leq T \leq t_{\alpha/2}(6)) = 1 - \alpha.$$



Figur 1: Den markerade arean innehåller 99% av sannolikheten för $t(6)$-fördelningen.

   Vi löser ut μ ur intervallet i sannolikhetsmåttet och får att

   $$\overline{X} - \frac{S}{\sqrt{n}}t_{\alpha/2}(6) \leq \mu \leq \overline{X} + \frac{S}{\sqrt{n}}t_{\alpha/2}(6).$$

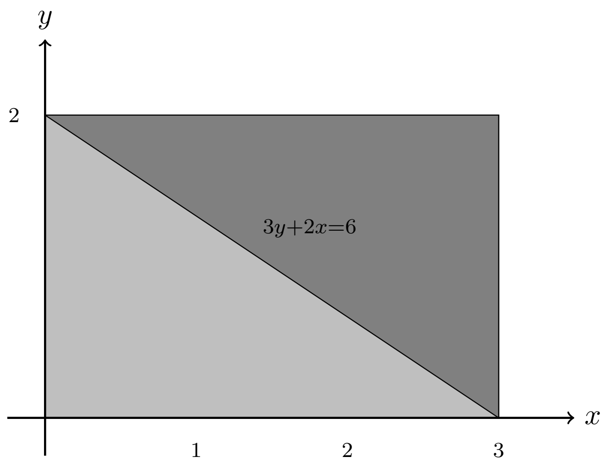
Vi ersätter $\overline{X}$ med den observerade punktskattningen $\widehat{x} = \overline{x} = 10.1$ och S med $s = 0.6532$. Då erhåller vi ett konfidensintervall med konfidensgrad 99%:

$$I_{\mu_1} = [9.18, 11.02],$$

där vi använt $\alpha = 0.01$ och $t_{0.005}(6) = 3.707$.

Svar: (a) 0.33. (b) $\sigma = 2.34$. (c) $[9.496, 10.704]$.

3. Vi skisserar de intressanta områdena i Figur 2.



Figur 2: Den skuggade rektangeln är området där $f_{X,Y}$ inte är noll. Den mörkare skuggade triangeln är det område vi är intresserade av i (c).

(a) Konstanten c fås ur sambandet

$$c^{-1} = \int_0^2 \int_0^3 x(y+1)\, dxdy = \cdots = 18,$$

så $c = 1/18$.

(b) Vi räknar ut de marginella täthetsfunktionerna:

$$f_X(x) = \int_0^2 x(y+1)\, dy = \frac{2x}{9}, \quad 0 \le x \le 3,$$

och

$$f_X(x) = \int_0^3 x(y+1)\, dy = \frac{y+1}{4}, \quad 0 \le y \le 2,$$

Vi ser direkt att $f_X(x) f_Y(y) = f_{X,Y}(x, y)$ för $0 \le x \le 3$ och $0 \le y \le 2$, så X och Y är oberoende ($f_{X,Y}$, f_X och f_Y är noll annars). Vi kan även direkt se att $f_{X,Y}$ är en produkt av en funktion av x och en av y. Återstår då bara att bestämma konstanterna. Men ovanstående argument fungerar alltid.

(c) Den del av det skuggade området där $3Y + 2X > 6$ ses i figuren. Vi ställer upp volymen och finner att

$$P(3Y > 6 - 2X) = \frac{1}{18} \int_0^3 \int_{2(1-x/3)}^2 x(y+1)\, dydx = \cdots = \frac{3}{4}.$$

Svar: (a) $c = 1/18$ (b) Se ovan. Variablerna är beroende. (c) 3/4.

4. (a) Vi har här att $X \sim \text{Hyp}(N, n, p)$ som modell. Med $p = 0.25$, $N = 3060$ och $n = 20$ ser vi att vi kan betrakta $X \sim \text{Bin}(20, 0.25)$ som lämplig approximation. Således erhåller vi

$$P(X \geq 4) = 1 - P(X \leq 3) = 1 - 0.2252 = 0.7748.$$

Verkligt värde (via hypergeometrisk fördelning) är 0.7757.

(b) Vi söker konfidensintervall för andelen som har denna gensammansättning. Totalt sett $n = 200$ stycken personer, och uppmätt är $X = 42$. En naturlig skattning på den verkliga andelen ges av

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

och vårt observerade värde är $\widehat{p} = 42/200 = 0.21$. Fördelningen för X ges approximativt av $X \sim \text{Bin}(n, p)$. Vidare ser vi att

$$np(1-p) \approx 200 \cdot 0.25 \cdot 0.75 = 37.5,$$

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

$$d = \sqrt{\widehat{p}(1-\widehat{p})/n} = \sqrt{37.5/200^2} = 0.0306.$$

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 $I_p = [0.15, 0.27]$ om vi ersätter $\widehat{P}$ med det observerade värdet $\widehat{p} = 0.21$.

Svar: (a) Ca 77%. (b) $[0.15, 0.27]$. Andelen $p = 0.25$ förefaller inte orimlig.

5. (a) Denna situation är hypergeometriskt fördelad med $N = 10$, $n = 3$ och $p = 2/10$ (sannolikheten att en slumpmässigt utvald cigarr är av rätt sort). Vi låter $Y \sim \text{Hyp}(10, 3, 1/5)$. Det som efterfrågas är med andra ord $P(Y \geq 1)$. Ur formelsamling vet vi vad sannolikhetsfunktionen är, så vi kan direkt räkna ut att

$$P(Y \geq 1) = p_Y(1) + p_Y(2) + p_Y(3) = \frac{\binom{2}{1}\binom{8}{2}}{\binom{10}{3}} + \frac{\binom{2}{2}\binom{8}{1}}{\binom{10}{3}} + 0 = \frac{2 \cdot 28 + 1 \cdot 8}{120} \approx 0.533$$

(b) Vi börjar med att ställa upp sannolikhetsfunktionen. Detta är en slags för-första-gången-fördelning, så för att lyckas på försök k måste man misslyckas på de $k-1$ föregående. Multiplikationsprincipen ger tabellen nedan.

k	$P(X = k) = p_X(k)$	(num)
$k = 1$	$\frac{3}{10}$	0.3000
$k = 2$	$\frac{7}{10} \cdot \frac{3}{9}$	0.2333
$k = 3$	$\frac{7}{10} \cdot \frac{6}{9} \cdot \frac{3}{8}$	0.1750
$k = 4$	$\frac{7}{10} \cdot \frac{6}{9} \cdot \frac{5}{8} \cdot \frac{3}{7}$	0.1250
$k = 5$	$\frac{7}{10} \cdot \frac{6}{9} \cdot \frac{5}{8} \cdot \frac{4}{7} \cdot \frac{3}{6}$	0.0833
$k = 6$	$\frac{7}{10} \cdot \frac{6}{9} \cdot \frac{5}{8} \cdot \frac{4}{7} \cdot \frac{3}{6} \cdot \frac{3}{5}$	0.050
$k = 7$	$\frac{7}{10} \cdot \frac{6}{9} \cdot \frac{5}{8} \cdot \frac{4}{7} \cdot \frac{3}{6} \cdot \frac{2}{5} \cdot \frac{3}{4}$	0.0250
$k = 8$	$\frac{7}{10} \cdot \frac{6}{9} \cdot \frac{5}{8} \cdot \frac{4}{7} \cdot \frac{3}{6} \cdot \frac{2}{5} \cdot \frac{1}{4} \cdot \frac{3}{3}$	0.0083

För övriga k måste $p_X(k) = 0$ (varför?). Vi kan även samla ihop till en lite snyggare formel enligt

$$p_X(k) = 3\frac{7! \cdot (10-k)!}{10! \cdot (8-k)!}, \quad k = 1, 2, \ldots, 8.$$

För att beräkna $E(X)$ använder vi definitionen och finner att

$$E(X) = \sum_{k=1}^{7} k \cdot p_X(k) = \cdots = 2.75.$$

Svar: (a) 0.533. (b) p_X se ovan, väntevärdet är 2.75.

6. Om vi antar att $np(1-p) \geq 10$ så kan vi normalapproximera (vi återkommer till detta villkor). Då är $X \overset{\text{appr.}}{\sim} \mathrm{N}(2950, \sqrt{2950(1-p)})$. Vi har då

$$P(\log X > 8) = P(X > e^8) = 1 - P(X \leq e^8) = 1 - \Phi((e^8 - np)/\sqrt{np(1-p)}) \leq 0.05,$$

vilket ger att

$$\Phi((e^8 - np)/\sqrt{np(1-p)}) \geq 0.95 \quad \Leftrightarrow \quad (e^8 - np)/\sqrt{np(1-p)}) \geq 1.645.$$

Här är det viktigt att observera att $e^8 - np > 0$, annars blir det problem med att plocka bort Φ (kräver extra steg). Det följer därmed att

$$p \geq 1 - \frac{(e^3 - np)^2}{1.645^2 np} = 0.8799$$

där vi utnyttjat att $np = 2950$. Det följer då också att $n = 3353$, vilket vi kan använda för att testa approximationsantagandet via simulering om vi har en dator. Annars räcker det att konstatera att $np(1-p) \approx 2600$ är betydligt större än 10. Testa även att räkna framlänges med $n = 3353$ och $p = 0.8799$:

$$P(X > e^8) \approx 1 - \Phi((e^8 - 2950)/\sqrt{2950 \cdot (1 - 0.8799)}) = 0.05.$$

Exakt värde (med binomialfördelning) blir 0.0531.

Man bör även övertyga sig själv om att $p = 1$ går bra även om detta in direkt följer av approximationsargumentet ovan.

Svar: $0.8799 \leq p \leq 1$.