LINKÖPINGS UNIVERSITET
MAI
Johan Thim

Tentamen i matematisk statistik (9MA241/9MA341/LIMAB6, STN2) 2011-08-27 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 och B vara händelser så att $P(A) = 0.4$ och $P(B \cap A^*) = 0.3$.

   (a) Bestäm $P(A \cup B)$ och $P(A^*)$. (2p)

   (b) Om händelserna är oberoende, vad är $P(B)$? (2p)

   (c) Om händelserna inte är oberoende, men man vet att $P(A \,|\, B) = 0.5$, vad är $P(B)$? (2p)

2. Vid en serie oberoende mätningar av en process erhöll man följande värden:

   $$2.49 \quad 3.03 \quad 2.73 \quad 1.70 \quad 2.29 \quad 1.21 \quad 2.89 \quad 0.85$$

   Vi antar att dessa värden utgör ett stickprov av en normalfördelad variabel $X \sim N(\mu_1, \sigma)$. Efter en tid misstänker man att något gått snett då produkterna inte längre håller samma kvalitet. Man gör nya mätningar på processen och erhåller då värderna

   $$1.32 \quad 1.31 \quad 0.14 \quad 0.97 \quad 0.84$$

   Antag att detta är ett stickprov från en normalfördelad variabel $Y \sim N(\mu_2, \sigma)$. Beräkna 99% konfidensintervall (dubbelsidiga) för μ_1, μ_2, samt $\mu_1 - \mu_2$. Kan man dra någon statistiskt säker slutsats med 1% felrisk från dessa intervall angående om det är någon skillnad mellan μ_1 och μ_2? Motivera ditt svar. (6p)

   En hjälpsam examinator har räknat ut följande:

   $$\sum_{i=1}^{8} x_i = 17.19 \qquad \text{och} \qquad \sum_{i=1}^{8} x_i^2 = 41.5067,$$

   $$\sum_{j=1}^{5} y_i = 4.58 \qquad \text{och} \qquad \sum_{i=1}^{5} y_i^2 = 5.1246.$$

3. Låt A vara (den fyllda) triangeln med hörn i $(0,0)$, $(1,0)$ och $(1,1)$. Vi definierar en tvådimensionell täthetsfunktion enligt

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

   där c är en konstant. Med andra ord är sannolikheten likformigt fördelad på triangeln.

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

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

(c) Bestäm $P(X > 1/2)$ och $P(Y > X)$. (3p)

4. Vi har en rättvis tärning med 6 sidor som vi kastar 4 gånger och räknar antalet 6:or vi får, låt X vara en stokastisk variabel som innehåller resultatet (antalet 6:or). I ett spel så satsar man 10 kronor och vinner enligt följande. Om vi får två eller färre 6:or vinner vi inget. Om vi får tre 6:or vinner vi 500 kronor, och om vi lyckas slå fyra 6:or vinner vi 2000 kronor.

   (a) Vad är $P(X \geq 3)$ och $E(X)$? (2p)

   (b) Om man spelar en gång, vad är väntevärdet för vinsten? Tycker du spelet är rättvist?(4p)

5. Skräckfilmsfantasten Sture samlar på italienska Giallos (thrillers från 60–70 talet). Sture brukar beställa dessa från en importör där han riskerar att få betala tullavgifter om värdet är för högt, så han beställer endast en film åt gången och inväntar leveransen innan han genast gör en ny beställning. Denna leverantör har 32 filmer som Sture inte äger, och leveranstiden från och med att beställningen görs uppskattar Sture är exponentialfördelad med väntevärde 8 dagar. Vad är sannolikheten att Sture måste vänta i högst 230 dagar innan han stolt kan visa upp sin bokhylla med alla dessa filmer? Ett approximativt svar duger om det motiveras. Vi antar också att leveranstiderna för olika filmer är tillräckligt oberoende av varandra. (6p)

6. Antag att vi vid följande x-värden har mätt upp motsvarande y-värden.

x	1	2	4	5	8
y	3.20	4.72	8.00	9.88	13.81

   Modellen är att y_j är en observation av en stokastisk variabel $Y_j = b_0 + b_1 x_j + \epsilon_j$, där $\epsilon_j \sim N(0, \sigma)$ är parvis oberoende, $j = 1, 2, 3, 4, 5$.

   (a) Hitta den linjära regressionslinjen för y med avseende på x. (2p)

   (b) Visa att (MK-)skattningarna för parametrarna b_0 och b_1 är väntevärdesriktiga. (4p)

Lösningsskisser för tentamen i matematisk statistik, 9MA241, 2011-08-27

1. (a) Eftersom händelserna A och $B \cap A^*$ är oförenliga och $A \cup B = A \cup (B \cap A^*)$ (rita Venn-diagram) så får vi $P(A \cup B) = P(A) + P(B \cap A^*) = 0.7$. Vidare så är $P(A^*) = 1 - P(A) = 0.6$.

   (b) Om A och B är oberoende så gäller att $P(A \cap B) = P(A)P(B)$. Vi söker $P(B)$, och vet att

   $$0.7 = P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B)(1 - P(A)),$$

   där vi utnyttjat att A och B är oberoende. Vi löser ut $P(B)$ och erhåller

   $$P(B) = \frac{0.7 - P(A)}{1 - P(A)} = \frac{0.3}{0.6} = 0.5.$$

   (c) Ifrån uppgiften vet vi att $P(A \mid B) = 0.5$. Vidare så gäller fortfarande

   $$0.7 = P(A \cup B) = P(A) + P(B) - P(A \cap B),$$

   så

   $$0.5 = \frac{P(A \cap B)}{P(B)} = \frac{P(B) - 0.3}{P(B)}.$$

   Vi löser ut $P(B) = 0.3/0.5 = 0.6$.

   Svar: (a) $P(A \cup B) = 0.7$ och $P(A^*) = 0.6$. (b) $P(B) = 0.5$. (c) $P(B) = 0.6$.

2. Vi har två olika stickprov, ett med 8 mätningar och ett med 5. Modellen förutsätter att de kommer från två källor som har samma varians (är detta rimligt?). Vi punktskattar med medelvärdet: $\mu_1^* = \bar{X} \sim N(\mu_1, \sigma/\sqrt{8})$. Som vanligt och skattar σ med s_1, där

   $$s_1^2 = \frac{1}{7}\sum_{i=1}^{8}(x_i - \bar{x})^2 \approx 0.6528125$$

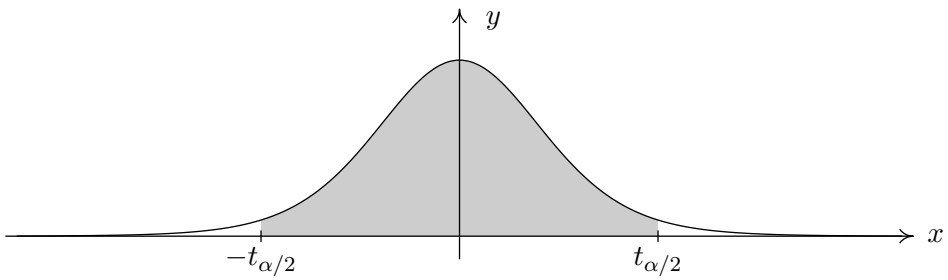
   är stickprovsvariansen. Vi skapar testvariabeln

   $$T = \frac{\mu_1^* - \mu_1}{s_1/\sqrt{8}} \sim t(7).$$

   Det följer att

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

   där $t_\alpha(7)$ är α-kvantilen till $t(7)$-fördelningen.



Figur 1: $t(7)$-fördelningen med 95% av arean markerad.

Ur tabell finner vi $t_{0.005}(7) = 3.499$. Genom att lösa ut μ_1 ur olikheten i sannolikhetsmåttet får vi uttrycket

$$\mu_1^* - \frac{3.499 \cdot 0.8080}{\sqrt{8}} < \mu_1 < \mu_1^* + \frac{3.499 \cdot 0.8080}{\sqrt{8}}.$$

Om vi ersätter μ_1^* med den observerade punktskattningen $(\mu_1)^*_{\text{obs}} = 2.149$ (medelvärdet av observationerna) så får vi ett konfidensintervall $I_{\mu_1} = [1.15, 3.15]$ med konfidensgrad 99%.

På samma sätt kan vi finna ett konfidensintervall för μ_2. Värderna vi använder är $s_2^2 = 0.2323$ och $(\mu_2)^*_{\text{obs}} = 0.9160$, och fördelningen för testvariabeln är $t(4)$. Motsvarande kvantil kan hittas i tabell och ges av $t_{0.005}(4) = 4.604$. Vi erhåller $I_{\mu_2} = [-0.077, 1.91]$.

Ur dessa intervall kan vi inte säga något om skillnaden $\mu_1 - \mu_2$ med konfidensgrad 99%. Istället punktskattar vi $\mu = \mu_1 - \mu_2$ med $\mu^* = \bar{X} - \bar{Y}$. Vi betraktar sedan variabeln

$$T = \frac{\mu^* - (\mu_1 - \mu_2)}{s\sqrt{1/8 + 1/5}},$$

där

$$s^2 = \frac{7s_1^2 + 4s_2^2}{11}$$

är den sammanvägda skattningen av σ^2. Det följer att $T \sim t(11)$. Nu fullföljer vi som vanligt, med $s^2 = 0.4999$, $\mu^*_{\text{obs}} = 1.2328$, $s\sqrt{1/8 + 1/5} = 0.4031$, och $t_{0.005}(11) = 3.106$, och får konfidensintervallet $I_{\mu_1 - \mu_2} = [-0.019, 2.48]$. Eftersom nollan ingår i intervallet kan vi inte med säkerhet säga att någon förändring har skett (med 1% felrisk).

Svar: a) $I_{\mu_1} = [1.15, 3.15]$, $I_{\mu_2} = [-0.08, 1.91]$, och $I_{\mu_1 - \mu_2} = [-0.019, 2.48]$.

3. Eftersom arean av triangeln är $1/2$ måste $c = 2$ för att dubbelintegralen av $f_{X,Y}$ över området skall bli ett. Eftersom området ser ut som det gör är det också tydligt att X och Y inte kan vara oberoende. Om vi skulle räkna ut de marginella täthetsfunktionerna skulle dessa bli nollskilda för argument mellan 0 och 1, så produkten av f_X och f_Y blir därmed nollskild för en kvadrat. Detta stämmer ej överens med triangeln där $f_{X,Y}$ är definierad. Sannolikheterna kan enkelt beräknas genom att betrakta arean av det sökta området, eller med hjälp av en dubbelintegral:

$$P(X > 1/2) = \int_{1/2}^{1} \int_{0}^{x} 2\, dydx = \int_{1/2}^{1} 2x\, dx = \left[2x^2/2\right]_{1/2}^{1} = 1 - 1/4 = 3/4.$$

Svaret är rimligt eftersom $3/4$ av triangelns area befinner sig till höger om $x = 1/2$. Sannolikheten att $Y > X$ är däremot noll då $f_{X,Y}(x, y) = 0$ för alla $y > x$. Det finns alltså inga punkter i triangeln där y-koordinaten är större än x-koordinaten.

Svar: (a) $c = 2$. (b) Oberoende. Se ovan. (c) $P(X > 1/2) = 3/4$ och $P(Y > X) = 0$.

4. Situationen ger att X är binomialfördelad: $X \sim \text{Bin}(4, 1/6)$. Vidare så låter vi funktionen g vara definierad enligt $g(k) = 500$ för $k = 3$ och $g(k) = 2000$ för $k = 4$. Annars är $g(k) = 0$. Vi söker nu följande: $P(X \geq 3)$, $E(X)$ samt $E(g(X))$.

Vi börjar med sannolikheten:

$$P(X \geq 3) = P(X = 3) + P(X = 4) = \binom{4}{3} \left(\frac{1}{6}\right)^3 \frac{5}{6} + \binom{4}{4} \left(\frac{1}{6}\right)^4 = \frac{20 + 1}{6^4} \approx 0.0162$$

Tydligt att det är ganska låg sannolikhet för vinst. Eftersom X är binomialfördelad blir

$$E(X) = 4 \cdot 1/6 = 2/3 \approx 0.6667.$$

Vi utnyttjar nu formel för väntevärde av en funktion av en stokastisk variabel för att beräkna $E(g(X))$:

$$\begin{aligned} E(g(X)) &= \sum_{k=0}^{4} g(k) \binom{4}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{4-k} \\ &= 500 \cdot \binom{4}{3} \left(\frac{1}{6}\right)^3 \frac{5}{6} + 2000 \cdot \binom{4}{4} \left(\frac{1}{6}\right)^4 \\ &= \frac{500 \cdot 20 + 2000}{6^4} \approx 9.2593 \end{aligned}$$

Det lönar sig alltså inte i längden att spela spelet då insatsen är högre än det förväntade värdet av vinsten i varje omgång (en följd av de stora talens lag).

Svar: (a) $P(X \geq 3) = 0.0162$ och $E(X) = 0.6667$.
(b) Väntevärdet för vinsten vid ett spel är 9.26.

5. Vi låter X_i beteckna leveranstiden för film i, $i = 1, 2, \ldots, 32$. De stokastiska variablerna X_i antas vara parvis oberoende och likafördelade: $X_i \sim \text{Exp}(8)$. Vi betraktar summan

$$Y = X_1 + X_2 + \cdots + X_{32}.$$

Fördelningen för Y är bökig att räkna ut direkt (i själva verket blir Y gammafördelad med parametrarna 32 och 8), så vi försöker med en normalapproximation istället. Väntevärde och varians beräknas enkelt till $E(Y) = 32 \cdot 8 = 256$ och $V(Y) = 32 \cdot V(X_1) = 32 \cdot 8^2 = 2048$. Enligt centrala gränsvärdessatsen så blir summan Y approximativt normalfördelad, så vi får

$$P(Y \leq 230) \approx \Phi((230 - 256)/\sqrt{2048}) = 1 - \Phi(26/\sqrt{2048}) = 0.2828.$$

Vi kan jämnföra med `matlab` och ser att `gamcdf(230,32,8)` $= 0.2961$ ligger ganska nära.

Svar: Sannolikheten är ca 28%.

6. Ifrån siffrorna kan vi räkna ut

$$\bar{x} = 4, \quad \bar{y} = 7.922, \quad \sum_{j=1}^{5} x_j y_j = 204.52, \quad \sum_{j=1}^{5} x_j^2 = 110.$$

Vi använder dessa siffror för att beräkna koefficienterna:

$$b_1 = \frac{\sum_{j=1}^{5} x_j y_j - n\bar{x}\bar{y}}{\sum_{j=1}^{5} x_j^2 - n\bar{x}^2} = \frac{204.52 - 5 \cdot 4 \cdot 7.922}{110 - 5 \cdot 4^2} = 1.536$$

och

$$b_0 = \bar{y} - b_1\bar{x} = 7.922 - 1.536 \cdot 4 = 1.778.$$

Regressionslinjen ges nu av $y = b_0 + b_1 x = 1.778 + 1.536x$.

För att visa att b_0 och b_1 är väntevärdesriktiga skattningar betraktar vi motsvarande stokastiska variabler:

$$B_1 = \frac{\sum x_j Y_j - n\bar{x}\bar{Y}}{\sum x_j^2 - n\bar{x}^2} \qquad (*)$$

och

$$B_0 = \bar{Y} - B_1\bar{x}.$$

Nämnaren i B_1 är deterministisk (inget som varierar när x-värderna är fixa), så vi beräknar väntevärdet av täljaren:

$$E\Big(\sum x_j Y_j - n\bar{x}\bar{Y}\Big) = \sum x_j E(Y_j) - n\bar{x}E(\bar{Y}).$$

Då $Y_j = b_0 + b_1 x_j + \epsilon_j$ följer att $E(Y_j) = b_0 + b_1 x_j$ (ϵ_j har väntevärde noll), så högerledet ovan kan skrivas om som

$$\sum x_j(b_0 + b_1 x_j) - n\bar{x}\frac{1}{n}\sum(b_0 + b_1 x_j) = b_0 n\bar{x} + b_1 \sum x_j^2 - \big(n\bar{x}b_0 + b_1 n\bar{x}^2\big) = b_1\big(\sum x_j^2 - n\bar{x}^2\big),$$

där vi känner igen den andra faktorn som nämnaren i ekvation $(*)$. Följdaktligen blir

$$E(B_1) = \frac{1}{\sum x_j^2 - n\bar{x}^2} E\Big(\sum x_j Y_j - n\bar{x}\bar{Y}\Big) = b_1,$$

så B_1 är en väntevärdesriktig skattning.

För B_0 utnyttjar vi att B_1 är väntevärdesriktig:

$$E(B_0) = b_0 + b_1\bar{x} - E(B_1)\bar{x} = b_0.$$

Svar: (a) $y = 1.78 + 1.54x$. (b) Se ovan.