Sannolikhetsteori 1, del 2 - 2008

Följande inlämningsuppgift är frivillig, men kan generera upp till 2 (två) bonuspoäng på tentan i kursen Sannolikhetsteori 1, del 2 HT 2008. Den består i två olika uppgifter - en mer teoretisk och en mer praktisk. Man behöver INTE lämna in båda uppgifterna, utan det räcker med EN för att kunna få bonuspoäng.

Tanken är att man ska arbeta två och två, samt att man lämnar in två och två. Typiskt lär man sig bättre på att arbeta tillsammans med någon annan. Ett poäng får man om man genom att lämna in en raport visar att man har gjort en utav uppgifterna. För att få två bonuspoäng så krävs också att raporten är välskriven.

Deadline: tisdag 2 december (preliminärt).

Inlämningsuppgift 1: Neyman-Pearsons lemma (teoretisk)

Antag att vi vill testa hurvida ett mynt är rättvist. Mer specifikt så vill vi testa följande hypotes:

$$H_0 : p = 0.5$$
$$H_1 : p = 0.7.$$

Hur ska vi resonera för att hitta ett bra test? Beteckna utfallet vid singling av slanten med

$$X = \begin{cases} 1 & \text{om krona} \\ 0 & \text{om klave.} \end{cases}$$

Låt oss beteckna mängden av möjliga utfall av stickprovet $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ med Ω_n. Dvs.

$$\Omega_n = \{(x_1, x_2, \ldots, x_n) : x_j = 0 \text{ eller } 1, \text{ för alla } j = 1, 2, \ldots, n\}.$$

För att bestämma ett test för hypotesen så vill vi bestämma en delmängd C till Ω_n för vilka utfall på stickprovet som H_0 skall förkastas. C är våran kritiska region.

Vilka utfall $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ bör ingå i C för att testet ska kunna förkasta H_0 när H_1 är sann? Jo, sådana $\mathbf{x}$ som har stor sannolikhet när H_1 är sann i förhållande till sannolikheten när H_0 är sann. Dvs., utfall för vilka kvoten

$$L(\mathbf{x}) = \frac{P(\mathbf{X} = \mathbf{x} | H_1 \text{ sann})}{P(\mathbf{X} = \mathbf{x} | H_0 \text{ sann})}$$

är stor. Hur borde vi välja den kritiska regionen i vårt fall ovan?

$$L(\mathbf{x}) = \frac{0,7^{\#\text{krona}} 0,3^{\#\text{klave}}}{0,5^{\#\text{krona}} 0,5^{\#\text{klave}}} = 1,4^{\sum_j x_j} 0,6^{20-\sum_j x_j} = 0,6^{20} (7/3)^{\sum_j x_j},$$

så $L(\mathbf{x})$ är stor om $\sum_{j=1}^n x_j$ är stor. Vår testregel blir: Förkasta H_0 för stora värden på $\sum_{j=1}^n x_j$. Observera att detta är precis så som vi har gjort för att testa en okänd sannolikhet tidigare. Bara att nu har vi kommit fram till att detta är en lämplig metod genom samma resonemang vi hade för Maximum Likelihood skattare.

Betrakta det allmäna fallet. Låt $\mathbf{X}$ vara ett stickprov (från en kontinuerlig eller diskret fördelning) och beteckna frekvensfunktionen (densitet eller massfunktion) av $\mathbf{X}$ med $f_\theta(\mathbf{x})$. Vi vill på signifikansnivån α testa hypotesen

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta = \theta_1.$$

Definiera *Likelihood Ratio-funktionen*

$$L(\mathbf{x}) = \frac{f_{\theta_1}(\mathbf{x})}{f_{\theta_0}(\mathbf{x})}.$$

Kalla det test som förkastar H_0 när $L(\mathbf{x}) \geq k$, för någon konstant k, för *NP-testet* (Neyman-Pearson-testet). Konstanten k bestäms så att signifikansnivå α uppfylls. Den kritiska regionen för testet blir

$$C = \{\mathbf{x} \in \Omega_n : L(\mathbf{x}) \geq k\}.$$

Theorem 1 (Neyman-Pearsons lemma) *Låt $\mathcal{T}$ vara vilket annat test som helst som har signifikansnivå högst α. Då gäller*

$$P(\textit{typ II-fel med NP-testet}) \leq P(\textit{typ II-fel med } \mathcal{T}).$$

Uppgiften går nu ut på att visa denna sats, samt att uppskatta den.

1. Använd definitionerna för L och C för att visa att

a) $P(\mathbf{X} \in S|\ \theta = \theta_1) \geq kP(\mathbf{X} \in S|\ \theta = \theta_0)$ för $S \subseteq C$.

b) $P(\mathbf{X} \in S|\ \theta = \theta_1) \leq kP(\mathbf{X} \in S|\ \theta = \theta_0)$ för $S \subseteq C^c = \Omega_n \setminus C$.

Hint: Betrakta kontinuerliga och diskreta fallet separat.

2. Använd resultatet i förra uppgiften för att visa att om $S \subseteq \Omega_n$, så gäller

$$P(\mathbf{X} \in C|\ \theta = \theta_1) - P(\mathbf{X} \in S|\ \theta = \theta_1) \geq k\big(P(\mathbf{X} \in C|\ \theta = \theta_0) - P(\mathbf{X} \in S|\ \theta = \theta_0)\big).$$

Hint: För mängder A och B gäller $A = (A \cap B) \cup (A \setminus B)$.

3. Förkastningsregionen för NP-testet har vi betecknat med C. Låt nu S beteckna förkastningsregionen för det alternativa testet $\mathcal{T}$. Använd uppgift 2 för att visa att

$$P(\mathbf{X} \in C|\ \theta = \theta_1) \geq P(\mathbf{X} \in S|\ \theta = \theta_1),$$

dvs. att NP-testet är minst lika starkt som testet $\mathcal{T}$. Motivera varför Neyman-Pearsons lemma följer av detta.

Det kan tyckas att Neyman-Pearsons lemma inte är så intressant, eftersom det behandlar ett väldigt enkelt fall (vår hypotes H_0 och H_1 innehåller bara två värden). Det visar sig dock att resultatet kan användas i fler situationer än så. Återgå till det exempel som vi hade om slantsingling. Det följer av Neyman-Pearsons lemma att det inte finns något starkare test än NP-testet för att testa

$$\begin{aligned} H_0 &: p = 0.5 \\ H_1 &: p = p_1, \end{aligned} \tag{1}$$

där $p_1 > 0.5$.

4. Låt $n = 100$ och ta fram förkastningssregion C för NP-testet för att testa (1) på signifikansnivå $\alpha = 0.05$.

Att hitta ett test som är likformigt starkast är inte alltid möjligt.

Extra. Förklara varför det följer att NP-testet faktiskt är *likformigt starkast* för att testa

$$H_0 : p = 0.5$$
$$H_1 : p > 0.5.$$

Inlämningsuppgift 2: Centrala gränsvärdessatsen (praktisk)

Denna uppgift går ut på att förstå sig på den viktiga centrala gränsvärdessatsen. För att göra detta kommer ni att genomföra vissa uppgifter med dator. Till exempel kan man använda sig av MatLab för att göra uppgifterna. Det vi ska undersöka är hur fördelningen för

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

ändras när n ökar.

1. Generera värden från en likformig fördelning där de möjliga värdena är 1,2,3,4,5 och 6 (som om ni kastade en rättvis tärning). Ni ska simulera värden för 50 tärningskast, och upprepa detta 250 gånger. Om ni använder MatLab så gör man det enklast genom att spara dom i en 250×50-matris som man kallar M. Totalt har ni data föreställande 12500 tärningskast.

2. Beräkna i tur och ordning $\bar{X}$ för de 2, 10, 30 och 50 första kasten i vardera rad. Spara de värden ni får som tex. $my2$, $my10$, $my30$ och $my50$ (var och en av dom är en 250×1-matris).

3. Plotta histogram av vardera av kolon 1 i M, $my2$, $my10$, $my30$ och $my50$. Det är instruktivt att ändra samtliga skalor så att de stämmer överens.

4. Vi vet sedan tidigare att

$$\mathrm{E}[\bar{X}] = \mu$$
$$\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

I detta fall är $\mu = 3.5$ och $\sigma = 1.71$. Normalisera värdena i $my2$, $my10$, $my30$ och $my50$ genom att subtrahera μ och dividera med σ/n (för respektive n). Plotta histogram av de normaliserade kolonvektorerna. Rita in i vardera plot densiteten för en N(0,1)-fördelning. Vad för slutsatser kan ni dra av detta?

En stokastisk variabel med densitet

$$f(x) = \frac{1}{\pi(1+y^2)}, \quad \text{för } y \in \mathbb{R}$$

kallas *Cauchyfördelad*.

5. Visa att om Y är Cauchyfördelad, så är $\mathrm{E}[|Y|] = \infty$. Visa ytterligare att $\mathrm{Var}(Y) = \infty$.

6. Upprepa steg 1-3 fast denna gång med data genererade från en Cauchyfördelning. Vad händer? Skiljer det sig från fallet då data bestod av tärningskast? Förväntade du dig detta? Förklara vad skillnaden från fallet med tärningskast beror på.

Ett sätt att undersöka hurvida data man har kommer från en viss fördelning eller inte är att plotta data i en så kallad *qq-plot* (quantile-quantile plot). Man definierar *q-kvantilen* av en fördelning med fördelningsfunktion F som det minsta värde x så att $F(x) \geq q$. Om vi inte vet fördelningen vårt stickprov kommer ifrån, så kan vi skatta kvantilerna. Vi får

$$\frac{k}{n}\text{-kvantilen för stickprovet } X_1, X_2, \ldots, X_n = X_{(k)}.$$

För att undersöka om vårat stickprov är normalfördelat kan vi plotta dess kvantiler mot motsvarande kvantiler för en normalfördelning.

Extra. Plotta kvantilerna för $my2$, $my10$, $my30$ och $my50$ mot kvantilerna för en normalfördelning. När de plottas mot en normalfördelning så kallas en sådan plot också för *normal probability plot*. Vad kan man dra för slutsats från detta? Hur borde grafen se ut om data är normalfördelade?

Hint: I MatLab kan göra detta automatiskt med komandot *normplot(X)*.