Projekt 8: Test av fördelningsantagande med goodness-of-fit-test

17 maj 2017

Ett mycket vanligt antagande i statistiska test-situationer är att data $X_1, X_2, \ldots, X_N$ är normalfördelade, t.ex. ett stickprov på någon storhet där antaganden om väntevärde eller varians ska testas, eller residualerna i en linjär regression. Det är vanligt att detta är ett approximativt korrekt antagande och man kontrollerar ofta detta genom att göra en normalfördelningsplot av data och bedöma om de följer en rät linje hyfsat bra. Detta är förstås ingen matematiskt rigorös metod. För att vara mer noggrann kan man använda ett s.k. *goodness-of-fit-test* för att kontrollera om data följer en viss fördelning med en fördelningsfunktion F (inte nödvändigtvis just en normalfördelning).

Tillvägagångssättet är följande. Dela in reella linjen i K intervall, kallade *celler*, $(x_{k-1}, x_k]$, $k = 1, \ldots, K$, där $-\infty = x_0 < x_1 < \ldots < x_K = \infty$. Låt N_k vara antalet observationer i data som antar ett värde i cell k och låt $E_k = N(F(x_k) - F(x_{k-1}))$ vara väntevärdet av antalet observationer i den cellen. Låt nu

$$T = \sum_{k=1}^{K} \frac{(N_k - E_k)^2}{E_k}.$$

Man kan visa att då $N \to \infty$ på ett sådant sätt att $E_k \to \infty$ för alla k, så gäller

$$F_T(x) \to F_{\chi^2_{K-1}}(x)$$

för alla $x \in (0, \infty)$. Med andra ord; när alla E_k är stora gäller att T är approximativt χ^2-fördelad med $K - 1$ frihetsgrader. Er första uppgift är att bevisa detta i fallet då $K = 2$ och $E_1 = E_2 = N/2$. Kom ihåg att en χ^2_1-fördelad stokastisk variabel per definition har samma fördelning som Z^2 där Z är $N(0, 1)$-fördelad.

En vanlig tumregel för att den givna approximationen av T's fördelning ska vara tillräckligt god för tillämpning är att $E_k > 5$ för alla k. Det lättaste sättet att åstadkomma detta är förstås att om möjligt dela upp reella linjen på att sådant sätt att $E_k = N/k$ för alla k och välja k litet nog så att $N/k > 5$. Å andra sidan vill man dela upp i så många intervall som möjligt för att kunna skilja på den fördelning man vill testa från en annan fördelning med samma E_k:n.

När man testar nollhypotesen att data verkligen kommer från fördelningen F mot alternativet att den kommer från en annan fördelning, beräknar man T och förkastar om $T \geq F^{-1}_{\chi^2_{K-1}}(1 - \alpha)$ där α är en lämplig signifikansnivå, t.ex. $\alpha = 0.05$. Er uppgift är nu först att låta F vara standard-normalfördelning, välja ett lämpligt k, säg $k = 10$, dela in i lämpliga intervall. Sedan simulerar ni data enligt just en standard-normalfördelning, beräknar T och ser om $T \geq c$ där $c = F^{-1}_{\chi^2_{K-1}}(0.95)$. Upprepa detta ett stort antal gånger. Om approximationen med χ^2-fördelningen stämmer bör man få

att $T \geq c$ nära 5% av gångerna. Får ni detta? Pröva olika N. (Om N är riktigt stort bör svaret bli ja, men N litet, säg $N < 100$ kan man vänta sig en avvikelse.) Nästa uppgift blir att simulera data från en $N(0.1, 1)$-fördelning och testa om data är standard-normalfördelade. Eftersom vi vet att detta inte är fallet, förväntar vi oss att vi ser $T \geq c$ oftare. Hur ofta beror på hur stort N är. Om N är mycket stort bör detta ske nästan jämt. Pröva olika N. Vad får ni för resultat?

Normalt är fallet förstås att kontrollera om data följer en viss fördelning med ospecificerade parametrar. Dessa måste skattas och man förlorar då lika många frihetsgrader i den approximativa χ^2-fördelningen som antalet parametrar. Ni ska tillämpa detta när ni vill testa om data är normalfördelade, dvs om data är fördelade enligt någon normalfördelning $N(\mu, \sigma^2)$. Om vi sätter $Z_i = (X_i - \mu)/\sigma$ är detta ekvivalent med att testa om Z_i:a är standard-normalfördelade. Men detta kan vi ju inte göra eftersom vi inte vet vad μ och σ^2 är. Dessa måste skattas av $\hat{\mu}$ och $\hat{\sigma}^2$ enligt någon skattningsmetod. Låt sedan $\hat{Z}_i = (X_i - \hat{\mu})/\hat{\sigma}$. Definera nu T som ovan, men med Z_i:na ersatta av $\hat{Z}_i$:na. Eftersom två parametrar skattas, hoppas vi att den asymptotiska fördelningen är χ^2_{K-3}. Uppgiften här är först att simulera $X_1, \ldots, X_N$ enligt en $N(100, 10^2)$-fördelning, skatta μ och σ^2 med medelvärdet och stickprovsvariansen och definera $\hat{Z}_i$:na som ovan. Upprepa detta ett stort antal gånger för ett stort N. Hur ofta blir $T \geq F^{-1}_{\chi^2_{K-3}}(0.95)$. Det borde bli 5% av gångerna. Tycks det stämma? Om ni gjort rätt kommer det att visa sig att svaret blir nej. Det visar sig att för den här tillämpningen är valet av medelvärde och stickprovsvarians som skattningar olämpliga. Man ska istället använda de skattningar som *minimerar* T. Detta kallas för att använda sig av minimum-χ^2. Nu är minimeringen inte så lätt att göra, men Matlab har en funktion för goodness-of-fit-test som gör det automatiskt: **chi2gof**. Gör samma simuleringar igen och se om $T \geq F^{-1}_{\chi^2_{K-3}}(0.95)$ rätt andel gånger.

Slutligen är uppgiften nu att använda minimum-χ^2 till att testa om födelsevikterna i **birth.dat** kommer från en normalfördelning. Låt testet också ge er testets p-värde. Testa också för graviditetslängd. Gör först en normalfördelningsplot för att bilda en uppfattning om vad ni tror att resultaten kommer att bli. Filen **birth.dat** och mer information om vad den innehåller finns på kurshemsidan och i filen **birth.txt**. Om det skulle visa sig att testet tyder på att en normalfördelning inte föreligger, vad är det för mekanism bakom data som inte stämmer överens med normalfördelning? Kan man ”rensa” data från den? Kan ni finna någon fördelning som bättre stämmer med data? Funktionen **chi2gof** har en möjlighet att testa fördelningsfunktion som man själv anger och som kan innehålla så många parametrar som behövs. Använd helpfunktionen.