

Bonusuppgift 3

Uppgiften ska vara inlämnad senast i samband med föreläsningen tisdagen den 13/5-03. Den består av 2 deluppgifter. Den första (värd 3 p) syftar till att du ska förstå centrala gränsvärdessatsen lite bättre. I den andra (värd 2 p) ska du simulera vad som kan hända vid mätning av förorenad mark. Bonusuppgiften ger alltså maximalt 5 p. De som får 2-3 p sänker gränserna för fyra och femma med 1 p till 17 p resp. 23 p. De som får minst 4 p sänker dessa gränser med ytterligare 1 p.

Deluppgift 1) Centrala gränsvärdessatsen (cgs)

Inom statistik är centrala gränsvärdessatsen ett av de viktigaste hjälpmedeln. Dock är den endast tillämpbar under specifika förutsättningar. Denna bonusuppgift går ut på att belysa problematiken kring cgs och påvisa att den inte alltid är tillämpbar.

Devore & Farnums beskrivning av cgs på s. 222 är ungefär som följer: Samplingsfördelningen av $\bar{X}$ kan bli approximerad av en normalfördelning då stickprovets storlek, n , är tillräckligt stor. Detta oberoende av utseendet på populationens fördelning. Ju större värde på n , desstå bättre approximation.

Denna beskrivning är inte korrekt och för att se detta formulerar vi cgs matematiskt: Låt $X_1, X_2, \dots$ vara oberoende och likafördelade med definierat väntevärde, μ_X , och definierad varians, σ_X^2 . För medelvärdet $\bar{X} = (X_1 + \dots + X_n)/n$ gäller

$$P\left(\frac{\bar{X} - \mu_X}{\sigma_X/\sqrt{n}} \leq z\right) \rightarrow \Phi(z)$$

då $n \rightarrow \infty$, där $\Phi(z)$ är den standardiserade normalfördelningsfunktionen. (Obs. att μ_X respektive σ_X^2 är definierade om, och endast om, $\mu_{|X|} < \infty$ respektive $\sigma_{|X|}^2 < \infty$.)

Ni skall i deluppgiften undersöka om cgs är tillämpbar på Cauchyfördelningen, vars täthet är

$$f(x) = \frac{1}{\pi(1+x^2)} \quad -\infty < x < \infty$$

(a) Beräkna Cauchyfördelningens väntevärde μ_X genom att använda definition överst på s. 65 i Devore & Farnum. Är $\mu_{|X|} < \infty$?

(b) Ur Beta fås att $\tan \phi$ är Cauchyfördelad om ϕ är likformigt fördelad på intervallet $(-\pi/2, \pi/2)$. Skriv Matlab-kod som genererar $n = 100$ Cauchyobservationer och räknar ut deras medelvärde. Stoppa in denna kod i en for-loop som upprepas 1000 gånger och

spara de 1000 genererade medelvärdena i vektorn `cm`. Exekvera sedan nedanstående kod, som ritar en qq-plot.

```
% Normal qq-plot
xx=sort(cm);
pp=((1:n)-0.5)/n;
yy=norminv(pp,mean(xx),std(xx));
qqplot(xx,yy);
xlabel('Data');
ylabel('Theoretical quantiles');
title(['Normal QQ-plot']);
grid on;
```

Upprepa för t.ex. $n = 50$ och $n = 200$. Upprepa alla plottningar några gånger, så att du ser vad som är typiskt för Cauchydata. Jämför med vad som händer om du istället genererar $\text{Exp}(1)$ eller $U(0, 1)$ -observationer.

(c) Tolka vad du ser i qq-plottarna och svara på om cgs verkar fungera för Cauchydata.

Deluppgift 2) En eventuellt förorenad tomt

Koncentrationen x av Pb i en punkt, vilken som helst, antas vara lognormalfördelad. Modell: $X \sim \text{LN}(\mu_C, \sigma_C)$. Man vet dessutom att mätvärdet y , givet att den sanna koncentrationen Pb är x , är lognormalfördelat. Modell: $Y|X = x \sim \text{LN}(\ln x, \sigma_M)$. Fundera gärna några minuter på om denna mätmetod är väntevärdesriktig eller ej.

Om man samtidigt betraktar två punkter $i = 1, 2$ på ett visst avstånd från varandra, gäller att koncentrationerna x_1, x_2 i dessa punkter typiskt är beroende. Enligt vad som sagts i stycket ovan gäller att $X_i \sim \text{LN}(\mu_C, \sigma_C)$ för $i = 1, 2$. Vi låter ρ beteckna korrelationen mellan de normalfördelade stokastiska variablerna $\ln X_1$ och $\ln X_2$.

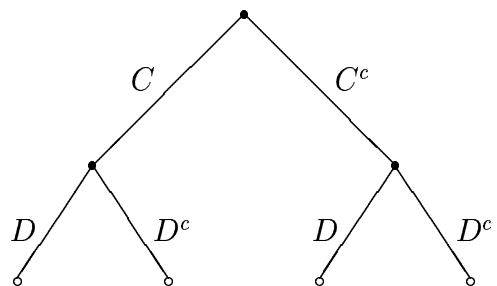
Antag att myndigheterna har fastställt ett gränsvärde g , sådant att marken ska betraktas som förorenad om koncentrationen x är minst g . Låt, för $i = 1, 2$, $C_i = \{X_i \geq g\}$. Observera att händelserna C_i ej går att observera eftersom vi inte kan mäta den sanna koncentrationen x_i utan fel. Vi betraktar tomten som förorenad om någon av C_1 eller C_2 inträffar, och definierar därför $C = C_1 \cup C_2$. Låt dessutom, för $i = 1, 2$, $D_i = \{Y_i \geq g\}$ och definiera $D = D_1 \cup D_2$. Om C är händelsen att tomten är förorenad, så är D händelsen att förorening detekteras.

Följande parametervärden ska du använda i simuleringarna av vad som kan hända i denna mätsituation: $\mu_C = 1.0$, $\sigma_C = 1.1$, $\rho = 0.9$, $\sigma_M = 0.5$ och $g = 2.0$.

(a) Skriv ett Matlab-program som simulerar $n = 1000$ (beroende) observationer x_1, x_2 av X_1, X_2 .

(b) Skriv ett Matlab-program som tar de $n = 1000$ sanna koncentrationsparen x_1, x_2 , och simulerar fram motsvarande $n = 1000$ uppmätta koncentrationer y_1, y_2 .

(c) Du har nu $n = 1000$ observerade utfall av formen x_1, x_2, y_1, y_2 . För vart och ett av dessa utfall kan du nu avgöra huruvida C resp. D har inträffat eller ej. Gör så med ett Matlab-program, samt punkt- och intervallskatta sannolikheterna $P(C)$, $P(D|C)$ och $P(D|C^c)$ i träd-diagrammet nedan, samt $P(C|D)$.



Räkna m.h.a. Bayes formel ut $P(C|D)$ och kontrollera att ditt svar överensstämmer med skattningen.

P.S. Det är tillåtet att lägga ihop de tre Matlab-programmen från (a), (b) och (c) till ett.