Meddelande: Nu på torsdag den 14/11 använder vi en del av föreläsningen till att titta på tentorna.

Kap 3 Simulering av stokastiska variabler **F05b**

Kapitlet handlar om hur man utifrån slumptal $u \in (0,1)$ genererar observationer på en stokastiska variabel X med given fördelning.

Vi har i del A lärt oss hur man gör då X är diskret. Läs själv kap 3.1 som handlar om hur man kan simulera slumpmässiga skeenden m.h.a mynt och tärningar. Avsnitt 3.2.1 handlar om sådana slumptal som genereras av program i datorer. Läs detta själv. Avsnitt 3.2.2 handlar om hur man m.h.a slumptal kan simulera observationer av diskreta stokastiska variabler. Detta kan alla redan, men studera ändå exempel 3.3 och 3.4, och, speciellt, exempel 3.5 som går igenom hur man simulerar realiseringar av Markovkedjor.

3.2.3 Generering av observationer på stokastiska variabler

Följande resultat skulle man kunna kalla för "alla simulerares huvudsats".

Sats 3.1 Låt $F(x)$ vara X:s fördelningsfunktion, som vi ska anta är strikt växande och kontinuerlig. Då har $F(x)$ en invers $F^{-1}(y)$ som uppfyller

$$y = F(x) \quad \Leftrightarrow \quad x = F^{-1}(y)$$

Låt $U \sim \mathrm{U}(0,1)$ (alltså likformigt fördelad på $(0,1)$). Då har den stokastiska variabeln $V = F^{-1}(U)$ har samma fördelning som X.

Satsen säger oss att om vi låter datorn generera ett slumptal u och bildar $x = F^{-1}(u)$, så är x en simulering av en observation på X.

Bevis av sats 3.1 Notera först att

$$F(x) < u \quad \Leftrightarrow \quad x < F^{-1}(u)$$

(F är ju strikt växande). Vi får att

$$\Pr\{V \le x\} = \Pr\{F^{-1}(U) \le x\} = \Pr\{U \le F(x)\} = F(x)$$

ty $\Pr\{U \le u\} = u$ för $0 \le u \le 1$, vilket skulle visas.

Om X har fördelningsfunktionen F, så är $U = F(X) \sim \mathrm{U}(0,1)$.

Omvänt, om $U \sim \mathrm{U}(0,1)$, så har $X = F^{-1}(U)$ fördelningsfunktionen F.

I recepten nedan betecknar u, u_1, u_2 slumptal.

Generering av observationer på Exp-fördelade variabler

Recept: $x = -\mu \ln u$, där μ är väntevärdet.

Simulering av Poissonprocessen

(J.f.r exempel 3.6.) Låt N_t vara antalet impulser i tidsintervallet $[0,t]$. Då är $N_t - N_s$ antalet impulser i tidsintervallet $(s,t]$. Poissonprocessen karaktäriseras av (1) att $N_t - N_s \sim \text{Poi}(\lambda(t-s))$ och (2) antalet impulser i disjunkta intervall är oberoende stokastiska variabler.

Man kan visa att detta är ekvivalent med att säga (3) att tiderna mellan impulser är oberoende och Exp-fördelade med väntevärdet $1/\lambda$.

Denna karaktärisering talar om för oss hur Poissonprocessen ska simuleras.

Recept: Låt $s_1, s_2, \ldots$ vara oberoende observationer på Exp med väntevärdet $1/\lambda$. Första impulsen kommer vid tidpunkten $t_1 = s_1$, andra vi tidpunkten $t_2 = t_1 + s_2$, tredje vid tidpunkten $t_3 = t_2 + s_3$, o s v. Om man vill simulera en Poissonprocess på tidsintervallet $[0,t]$, så drar man nya mellanimpulstider s_i så länge $t_m = \sum_{i=1}^{m} s_i \leq t$. Då man kommit till läget

$$t_m \leq t, \quad \text{men} \quad t_{m+1} = t_m + s_m > t$$

avbryts simuleringen. Då har vi fått reda på i vilka tidpunkter impulserna i intervallet $[0,t]$ inträffade. Observera att vi vet också då hur många de är, så vi har också erhållit en observation m av $N_t \sim \text{Poi}(\lambda t)$.

Recept: Simulering av en observation på $\text{Poi}(\lambda)$ erhålls följaktligen genom att man gör som ovan för fallet $t = 1$.

Generering av observationer av N-fördelade variabler **F06a**

Recept: lös z ur

$$u = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)\, dt$$

och bilda $x = \mu + \sigma z$, eller använd Box-Mullers metod, som säger att z_1, z_2, givna av

$$\begin{aligned} z_1 &= \sqrt{-2 \ln u_1} \cos 2\pi u_2 \\ z_2 &= \sqrt{-2 \ln u_1} \sin 2\pi u_2 \end{aligned}$$

kan betraktas som oberoende observationer på $\text{N}(0,1)$. Då är $x_1 = \mu + \sigma z_1$ och $x_2 = \mu + \sigma z_2$ två oberoende observationer av $\text{N}(\mu, \sigma)$.

Simulering av bivariata observationer

(Det diskreta fallet gås igenom i exempel 3.4.) Givet är X, Y:s gemensamma täthet $f_{X,Y}(x,y)$. Ur den räknar man fram X:s marginaltäthet $f_X(x)$. Definiera den betingade tätheten för $Y|X = x$,

$$f_{Y|X=x}(y) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

(I det diskreta fallet gäller att $f_{Y|X=x}(y) = \Pr\{Y = y | X = x\}$.)

Recept: Simulera först en observation x av X med tätheten $f_X(x)$. Simulera sedan en observation y av $Y|X = x$ enl den betingade tätheten $f_{Y|X=x}(y)$. Då är x, y en observation på X, Y.

Observera att det ibland kan vara enklare att börja med att simulera en observation y enl Y:s täthet, och sedan en observation x enl tätheten för $X|Y = y$.

Tekniken kan utvigdas till en generell metod för simulering av stokastiska vektorer. Med omatematiska men suggestiva beteckningar gäller att

$$f(x, y, z) = f(x)f(y|x)f(z|x, y)$$

där

$$f(y|x) = \frac{f(x, y)}{f(x)}$$

$$f(z|x, y) = \frac{f(x, y, z)}{f(x, y)}$$

En observation av den tredimensionella stokastiska variabeln X, Y, Z erhålls alltså genom att man först genererar en observation x på X enligt tätheten $f(x)$, sedan en observation på $Y|X = x$ enligt tätheten $f(y|x)$ och till sist en observation på $Z|X = x, Y = y$ enligt tätheten $f(z|x, y)$.

Kom också ihåg att i det kontinuerliga fallet gäller

$$f(x, y) = \int_{-\infty}^{\infty} f(x, y, z)\, dz$$

$$f(x) = \int_{-\infty}^{\infty} f(x, y)\, dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y, z)\, dz\, dy$$

Läs själva exempel 3.6, som går igenom hur man kan simulera observationer av sammansatta ("compound") Poissonprocesser. Sådana processer kallas i en del sammanhang märkta Poissonprocesser.

Läs ej exempel 3.7 och 3.8.

Läs ej avsnitt 3.2.4.

Kap 6 Händelsestyrd simulering och statistisk analys av simuleringsresultat

Kapitel 6.1 handlar om händelsestyrd simulering. Det återkommer vi till på någon av de sista föreläsningarna i kursen.

Kap 6.2 Statistisk analys av simuleringsresultat **F06b**

Mätdata $x_1, \ldots, x_n$.Vi antar först att dessa är observationer på oberoende och likafördelade stokastiska variabler som har väntevärdet μ och standardavvikelsen σ.

Räkna ut

$$\bar{x}, \quad s^2, \quad s = \sqrt{s^2}$$

Att göra konfidensintervall för μ och σ är standard. Notera att när frihetsgraderna är så många att χ^2-kvantiler ej finns tabellerade, kan man använda att

$$\frac{s}{1 + z_{\alpha/2}/\sqrt{2n}} \leq \sigma \leq \frac{s}{1 - z_{\alpha/2}/\sqrt{2n}}$$

gäller med den approximativa konfidensen $1 - \alpha$.

Även i fallet då mätdata $x_1, \ldots, x_n$ ej består av oberoende observationer är det naturligt att man skattar väntevärdet μ med medelvärdet $\bar{x}$ p.g.a stora talens lag, som säger OH 7

$$\mu = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} x_i$$

Den naturliga estimatorn av μ är alltså

$$\hat{\mu} = \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(eller

$$\hat{\mu} = \frac{1}{n-k} \sum_{i=k+1}^{n} x_i$$

vilken man tar till om mätdata har ett ca k tidsenheter långt insvängningsförlopp (lång "start-up-tid") som man vill ta bort effekterna utav).

Konstruktion av ett konfidensintervall för μ kräver att vi kan skatta estimatorns varians $\mathrm{Var}[\hat{\mu}]$. Om observationerna är oberoende eller approximativt oberoende är $\mathrm{Var}[\hat{\mu}] = \sigma^2/n$ och skattas följaktligen av s^2/n. Problemet är att nu är våra observationer typiskt beroende. Ett sätt att skaffa sig approximativt oberoende N-fördelade observationer är att dela in data i grupper och räkna ut medelvärdet i varje grupp ("batch means"). Om gruppstorleken ej är för liten blir dessa medelvärden approximativt oberoende och tack vare c.g.s approximativt normalfördelade. Man skattar nu μ med medelvärdet av gruppmedelvärdena, som blir lika med $\hat{\mu}$ och konfidensintervall för skattningen görs på sedvanligt sätt m.h.a den uppmätta variansen i medelvärdena.

Antag att vi p.g.a insvängningsförloppet plockar bort k observationer från mätdatas början och att grupperna är l observationer långa. Kalla gruppmedelvärdena för $y_1, \ldots, y_k$, där alltså

$$y_1 = \frac{1}{l} \sum_{i=n-k+1}^{n-k+l} x_i$$

$$y_2 = \frac{1}{l} \sum_{i=n-k+l+1}^{n-k+2l} x_i$$

o.s.v. Antag att antalet gruppmedelvärden är m och låt s_y^2 vara deras varians. Då gäller att

$$\hat{\mu} = \bar{y} = \frac{1}{m} \sum_{i=1}^{m} y_i$$

och att s_y^2/m skattar $\mathrm{Var}[\hat{\mu}]$, vilket enl c.g.s medför att

$$\frac{\hat{\mu} - \mu}{s_y/\sqrt{m}} \sim t(m-1)$$

gäller åtminstone approximativt. Det gäller här att inte välja gruppstorleken l för liten (för då blir inte gruppmedelvärdena $y_1, \ldots, y_m$ tillräckligt oberoende) och inte för stor (för då blir m onödigt litet; man vill ju ha m så stort som möjligt av noggranhetsskäl).

Kap 6.3 Jämförande statistik

OH 8

Nu tänker vi oss att vi har n_1 oberoende observationer på μ_1 med den teoretiska standardavvikelsen σ_1. Vi räknar ut deras medelvärde $\bar{x}_1$ och standardavvikelse s_1. Vi har dessutom n_2 oberoende observationer på μ_2 med den teoretiska standardavvikelsen σ_2. Vi räknar ut deras medelvärde $\bar{x}_2$ och standardavvikelse s_2. Om man inte har starka skäl att tro något annat än att $\sigma_1 \approx \sigma_2$, så väger man samman skattningarna s_1 och s_2 enl

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$$

till en skattning av den gemensamma standardavvikelsen σ.

Se paragrafen ”Två normalfördelade stickprov” i formelsamlingen.

Där hittar vi vad vi behöver veta för att kolla om $\sigma_1 \approx \sigma_2$. Nämligen,

$$\frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$

Välj kvantiler $F_{n_1-1,n_2-1,\alpha/2}$, $F_{n_1-1,n_2-1,1-\alpha/2}$, så att

$$\Pr\left\{ F_{n_1-1,n_2-1,1-\alpha/2} \le \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \le F_{n_1-1,n_2-1,\alpha/2} \right\} = 1-\alpha$$

Sedvanligt omstuvning av olikheten ger oss nu ett konfidensintervall för variansernas kvot,

$$\frac{1}{F_{n_1-1,n_2-1,\alpha/2}} \frac{s_1^2}{s_2^2} \le \frac{\sigma_1^2}{\sigma_2^2} \le \frac{1}{F_{n_1-1,n_2-1,1-\alpha/2}} \frac{s_1^2}{s_2^2} \quad (\approx 1-\alpha)$$

Många statistiker börjar en analys med att göra ett konfidensintervall för kvoten σ_1/σ_2 enl ovan med ganska låg konfidensgrad, t.ex ca 80%. Skälet till detta är att man vill att det inte ska vara alltför svårt att upptäcka en ev avvikelse från $\sigma_1/\sigma_2 \approx 1$. Om konfidensintervallet innehåller 1, så finns det inte starka skäl att tro något annat än att $\sigma_1 \approx \sigma_2$ och man fortsätter med att göra konfidensintervall för differensen $\mu_1 - \mu_2$ enligt ovan. Om konfidensintervallet ej innehåller 1, så gör man konfidensintervall för differensen med den metod som gås igenom i avs 10.4 i Milton & Arnold.

Om att jämföra experimentella frekvenser med teoretiska

OH 9

Här följer ett par ord om hur man t.ex testar om en tärning är osymmetrisk. Jämför med inlämningsuppgiften i del A. Metoden är generell och identisk med vad vi gjorde i inlämningsuppgift C3 i A delen av kursen. Vi ska först gå igenom mer generellt hur man jämför uppmätta frekvenser med teoretiska, och avgör om de uppmätta tyder på att de teoretiska är fel.

Antag att vi i n oberoende försök har mätt upp frekvenserna $f_1, f_2, \ldots, f_k$ för de ömsesidigt uteslutande händelserna $A_1, A_2, \ldots, A_k$. Vi ska anta att inget annat kan inträffa i våra försök, vilket är samma sak som att säga att $A_1 \cup A_2 \cup \ldots \cup A_k$ är lika med vårt utfallsrum Ω.

Antag att vi har en nollhypotes

$$H_0 : P(A_1) = p_1, P(A_2) = p_2, \ldots, P(A_k) = p_k$$

som vi vill testa mot den naturliga mothypotesen

$$H_1 : P(A_i) \neq p_i \quad \text{för något} \quad i$$

Då beräknar vi det s k χ^2-avståndet mellan teoretiska frekvenser $np_1, np_2, \ldots, np_k$ (observera att $E[f_i] = np_i$ under H_0) och uppmätta frekvenser $f_1, f_2, \ldots, f_k$, enligt

$$\chi^2 = \sum_i \frac{(f_i - np_i)^2}{np_i}$$

som om H_0 är sann approximativt har en $\chi^2(k-1)$-fördelning. Regeln som säger då denna approximation kan användas är $\min_i np_i \geq 5$. Det test av H_0 mot H_1 som förkastar H_0 då $\chi^2 \geq \chi^2_{k-1,\alpha}$ håller approximativt nivån α.

År 2001 simulerade vi gemensamt en osymetrisk tärning. Vi gjorde tillsammans $n = 2800$ ”kast”, vilket gav oss följande frekvenser:

f_1	f_2	f_3	f_4	f_5	f_6
359	459	468	447	487	530

Totalt har vi alltså $n = \sum_i f_i = 2800$ observationer, och de teoretiska frekvenserna för en symmetrisk tärning är $np_1 = \cdots = np_6 = 2800/6 \approx 466.67$. Här är ju nollhypotesen

$$H_0 : p_1 = \cdots = p_6 = 1/6$$

χ^2-avståndet mellan uppmätta och teoretiska frekvenser är därför

$$\chi^2 = \frac{(359-466.67)^2}{466.67} + \frac{(459-466.67)^2}{466.67} + \frac{(468-466.67)^2}{466.67}$$
$$+ \frac{(447-466.67)^2}{466.67} + \frac{(487-466.67)^2}{466.67} + \frac{(530-466.67)^2}{466.67}$$
$$= 21.53 + 0.00 + 0.20 + 0.28 + 1.79 + 11.21 = 35.01$$

I tabell ser vi att $\chi^2_{5,0.01} = 15.08$, så utfallets P-värde är alltså betydligt lägre än 1%. Inga tvivel, alltså, om att tärningen är osymmetrisk. Vi ser också genom att titta på termerna i χ^2-avståndet och motsvarande frekvenser, att tärningen ger för få ettor och för många sexor.

Samma slutsats hade vi kunnat dra från konfidensintervallet

$$p_6 = 0.193 \pm 0.015 \quad (\approx 95\%)$$

(ty $1/6 \notin$ det) och ett analogt för p_1.

Så när ska man använda χ^2-testet och när ska man göra konfidensintervall för en sannolikhet? En bra regel är att när man har starka skäl att tro att osymmetrin yttrar sig på så sätt att en viss sannolikhet är "fel", så gör man ett konfidensintervall (eller test) för denna sannolikhet. Annars χ^2-testar man likheten mellan teoretiska och uppmätta frekvenser.

Den här typen av statistiska metoder diskuteras i kap 15 i M & A.

Om att punktskatta en kvantil OH 10

Låt X vara kontinuerligt fördelad och låt x_p vara p-kvantilen i X:s fördelning. Då gäller

$$p = P(X \le x_p)$$

Antag att du har n oberoende observationer $x_1, \ldots, x_n$ av X och vill skatta x_p.
Börja med att sortera observationerna i storleksordning. Kalla de sorterade observationerna för $x_{(1)}, \ldots, x_{(n)}$. Då

$$x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$$

(T.ex om observationerna är 3.2, 1.9, 1.6, 2.6, så är det ordnade stickprovet 1.6, 1.9, 2.6, 3.2 och vi ser att $x_{(1)} = 1.6, x_{(2)} = 1.9$, o.s.v.) Vi kopplar ihop den i:te observationen i storleksordning med sannolikheten

$$p_i = \frac{i - \frac{1}{2}}{n} = \frac{2i-1}{2n}$$

Kravet $p = p_i$ ger nu att

$$i = pn + \frac{1}{2}$$

Har man möjlighet att välja n kan man ju försöka att välja n dels stort och dels så att $pn + \frac{1}{2}$ blir ett heltal, och helt enkelt låta $x_{(i)}$ vara punktskattningen av kvantilen x_p.

I annat fall (alltså då $pn + \frac{1}{2}$ ej är ett heltal) är det rimligt att man interpolerar linjärt mellan de två närmaste observationerna i storleksordning. Om t.ex $n = 45$, $p = 0.75$, så blir $np + \frac{1}{2} = 34.25$ och som skattning av $x_p = x_{0.75}$ väljes då $\hat{x}_{0.75} = 0.75x_{(34)} + 0.25x_{(35)}$. Det finns metoder med vilka man kan beräkna konfidensintervall för skattningen av x_p, som vi inte går in på här. Notera dock att om man t.ex simulerar observationer av X, så kan man ju tänka sig att medelst upprepning skaffa sig flera oberoende skattningar av x_p och därefter skatta x_p med medelvärdet och använda den erhållna standardavvikelsen till att skatta felet.