

Tommy Norberg
Matematisk statistik
Chalmers & GU
9 maj 2006

Poissonprocessen och extrema laster

Denna föreläsning ska vi teoritisera lite grand på temat ovanliga händelser. Vi börjar med att införa Poissonprocessen som ofta är en bra modell för när i tiden ovanliga händelser inträffar. Sedan ska vi med hjälp av den så kallade POT-metoden skatta storleken eller styrkan av de ovanliga händelserna vi studerar.

Poissonprocessen

Tänk dig att en viss konstruktion belastas kontinuerligt och att vi är intresserade av att registrera de tillfällen då lasten överstiger ett kritiskt värde x_T . Det skulle kunna vara vågor som bryter mot en oljeborrplattform, det skulle kunna vara vinden som blåser över en skog, det skulle kunna vara koncentrationen av en viss förorening i dagvattnet, etc. Det behöver dock inte vara någon "man-made" konstruktion. Tex mäter man ju ständigt halten av olika föroreningar i Göteborgsluften. Av intresse är naturligtvis då att registrera de tillfällen då luften är så mycket förorenad att tex astmatiker riskerar att bli tvugna att besöka Sahlgrenska akutmottagning. I exemplena ovan finns naturliga säsongsvariationer. Sådana går att ta hänsyn till, men i den enkla analys vi ska göra ska vi antaga att sådana ej förekommer. Ett exempel på en ren lastprocess utan säsongsvariationer har du i figur 1.

Vi börjar med att betrakta varje dygn för sig. Så tänk dig nu att vi varje dag kl 12:00 gör försöket att observera huruvida lastprocessen någon gång under det senaste dygnet har varit över den kritiska nivån x_T . Låt p_d vara sannolikheten att detta ska inträffa.

Anta att alla dessa försök är oberoende¹ och betrakta vad som händer under ett helt år. Totala antalet dygn med överskridande, N^d , är då en binomialfördelad stokastisk variabel med parametrar $n = 365$, som är totala antalet försök som görs, och sannolikheten p_d .

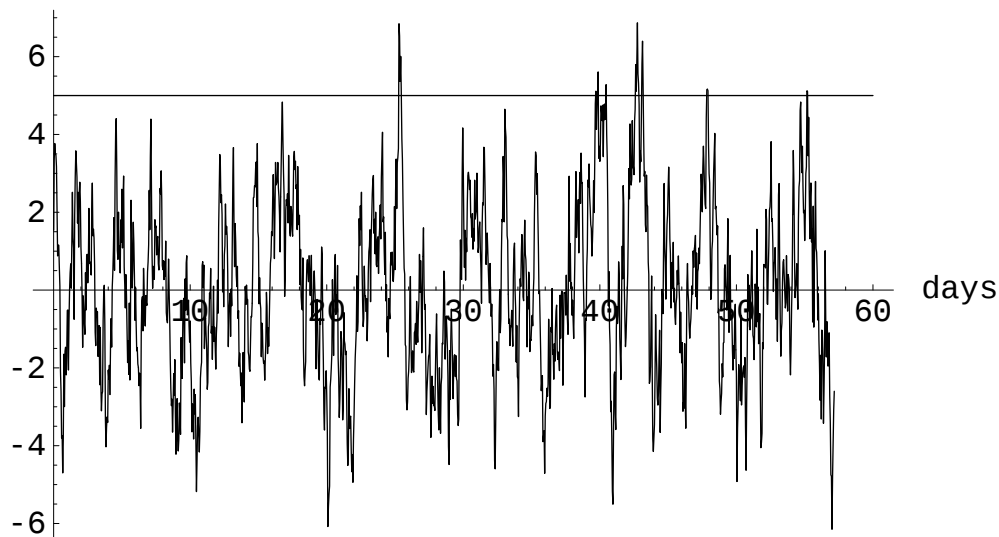
Anta nu att vi inte bara är intresserade av totala antalet lyckade försök N^d utan också av hur många som lyckas de olika månaderna under året. Då kan vi införa stokastiska variabler $N_1^d, N_2^d, \dots, N_{12}^d$, där $N_1^d \sim \text{Bin}(31, p_d)$ är antalet lyckade försök i januari, $N_2^d \sim \text{Bin}(28, p_d)$ är antalet lyckade försök i februari, etc. Observera att $N_1^d, \dots, N_{12}^d$ är oberoende och att $N^d = N_1^d + \dots + N_{12}^d$.

Om p_d är en liten sannolikhet är det inte orimligt att tänka på samtliga dessa stokastiska variabler som om de vore Poissonfördelade med parameter $\lambda = np_d$, där n är antalet dagar vi mäter över.

⁰Preliminär per den 25 mars 2008 kl 15:12

¹För många lastprocesser är detta ett hyfsat realistiskt antagande om den kritiska nivån ej är för låg.

level



Figur 1: En belastningsprocess studerad under nästan 60 dagar. Kritisk nivå är $x_T = 5.0$

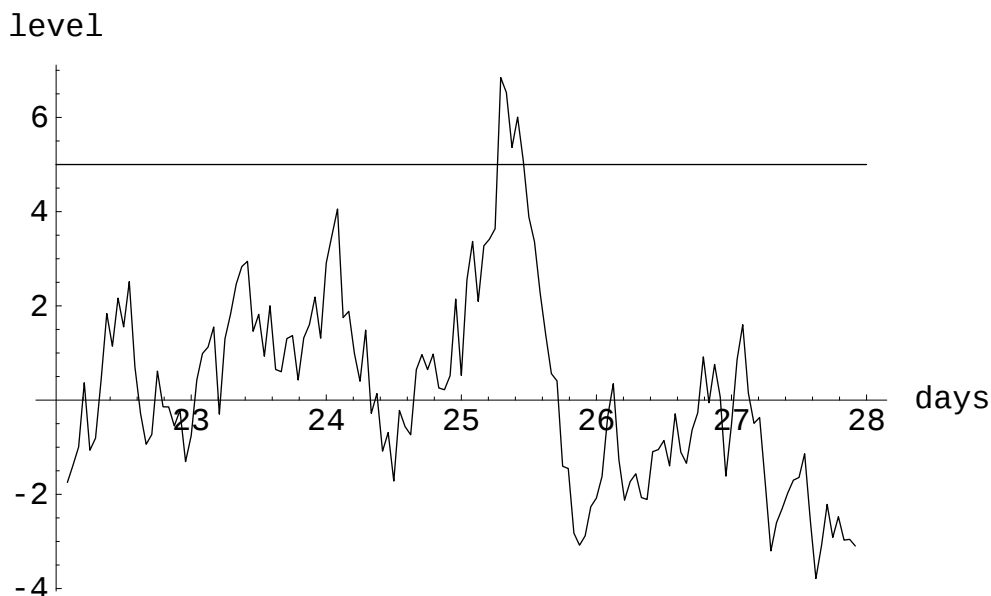
Observera också att N^d underskattar det totala antalet gånger som den kritiska nivån överskridits. Vi registrerar ju bara huruvida lasten någon gång under det senaste dygnet överskridit den kritiska nivån.

Låt oss därför tänka oss att vi ökar upplösningen och istället noterar varje hel timma huruvida lasten de senaste 60 minuterna överskridit den kritiska nivån eller ej. Under ett helt år görs nu $n = 365 \times 24 = 8760$ oberoende försök och vi låter $N^h \sim \text{Bin}(8760, p_h)$, där p_h är sannolikheten för ett överskridande under en viss timma. Vi noterar att $\text{Bin}(8760, p_h) \approx \text{Poi}(8760p_h)$ och att motsvarande approximation fungerar bra också för de stokastiska variablerna $N_{01}^h, \dots, N_{12}^h$ som anger antalet överskridanden de olika månaderna. Vi tänker ju oss att p_h är väldigt liten (den kritiska nivån är ju hög).

När vi fortsätter att öka upplösningen först till minuter och sedan till sekunder, etc, för att i gräns då tidsenheten går mot 0 få ett resultat som gäller i kontinuerlig tid, måste vi hålla i huvudet att det egentligen inte är antalet tidsenheter med lastöverskridande som vi är intresserade utav, utan totala antalet gånger lasten överskrider den kritiska nivån. Därför ska vi tänka oss att vi bara räknar s uppkorsningar av den kritiska nivån, se figur 2.

För många lastprocesser fungerar Poissonapproximationen ovan bra ändå och i gräns då tidsenheten blir oändligt liten erhålls den s k Poissonprocessen.

Det som behöver vara uppfyllt är att uppkorsningar i disjunkta tidsintervall av den kritiska nivån, kan ses som oberoende händelser och att produkterna



Figur 2: En period då belastningsprocessen i figur 1 var relativt hög

$365p_d$, $8760p_h$, $525600p_m$,² etc, konvergerar mot det förväntade antalet uppkorsningar under ett år. Obs att antalet uppkorsningar sammanfaller med antalet överskridanden. Låt oss kalla detta väntevärde för 365λ . Skälet till faktorn 365 är att det är praktiskt i det som följer att tidsenheten är dygn; λ är alltså förväntat antal överskridanden av den kritiska nivån under ett dygn.

Låt nu för $t \geq 0$, N_t vara antalet gånger den kritiska nivån överskrids under tiden från årets start och t dagar framåt. Obs att t ej behöver vara ett heltal. Vi arbetar ju nu i kontinuerlig tid. Vi behöver inte heller förutsätta att $t \leq 365$. Av diskussionen ovan förstår vi att antagandet $N_t \sim \text{Poi}(\lambda t)$ är rimligt och att det dessutom är rimligt att differensen $N_s - N_t$, som ger oss antalet lastöverskridanden i tidsintervallet $(t, s]$, är oberoende av N_t och Poissonfördelad med parameter $\lambda(s - t)$. Här tänker jag mig att $t < s$.

Detta är *Poissonprocessens* definierande egenskaper. En händelse i en Poissonprocess kallas ofta för *impuls*. Vad man ska komma ihåg om Poissonprocessen, är att antalet impulser i disjunkta tidsintervall är oberoende Poissonvariabler med parametrar proportionella mot tidsintervallens längder. Proportionalitetskonstanten λ brukar kallas Poissonprocessens *intensitet*. I en Poissonprocess är tiderna mellan impulserna oberoende och $\text{Exp}(\lambda)$ -fördelade. Den minnesgode kanske kommer ihåg att vi, under en föreläsning, räknade ut att för $\text{Exp}(\lambda)$ -fördelningen är felintensiteten $h(t) = f(t)/R(t) = \lambda$.

Poissonprocessen är många gånger en bra modell för när i tiden ovanliga händelser inträffar. Den är också, som vi just indikerat, ofta en vettig modell

² p_d, p_h, p_m är nu sannolikheten för en uppkorsning under 1 dygn, 1 timma, 1 minut

för när olika kontinuerliga processer överskrider (våldigt) höga nivåer. Men kom ihåg att modifieringar måste göras om den kontinuerliga processen uppvisar säsongsvariationer.

Övning 1) Antag att olyckor sker enligt en Poissonprocess med intensitet $\lambda = 0.025$ olyckor per dag.

- (a) Bestäm väntevärdet av antalet olyckor under en vecka.
- (b) Bestäm sannolikheten för minst 2 olyckor under 3 på varandra följande dagar.
- (c) Bestäm sannolikheten för ingen olycka under 4 dagar.

I följande övning antyds hur man kan förfara då säsongsvariationer föreligger.

Övning 2) Järnvägsspår som ligger utmed ett vattendrag löper risk att översvämmas. Under en tidsperiod av 15 år har man sett detta hända 18 gånger.

- (a) Skatta förväntat antal översvämningar under ett år.

Av de 18 översvämningarna skedde 1 i januari, 3 i februari, 4 i mars, 3 i september, 4 i oktober, 2 i november och 1 i december.

- (b) Fördela ut denna skattning av förväntat antal översvämningar på årets månader.

Övning 3) Belastningsprocessen i figur 1 överstiger kritiska nivån $x_T = 5$ totalt 8 gånger under de 56 dagar och 4 timmar som processen studerats. Den sammanlagda tiden processen är över nivån är 23.5 timmar. Antag att överskridandena (dvs uppåtkorsningarna) sker efter en Poissonprocess med intensiteten λ_T överskridanden per dygn.

- (a) Skatta λ_T .
- (b) Hur länge är processen över den kritiska nivån i genomsnitt?

För belastningsprocessen i figur 1 är nivån $x_T = 5$ en varningsnivå. Vi ska säga att vi har ett tillbud då lasten går över detta värde. Katastrof blir det om lasten fortsätter upp över nivån $x_K = 6.75$. Vi betecknar tillbud med T och katastrof med K . Låt $p_K = P(K|T)$.

Man kan visa att om tillbudsprocessen är en Poissonprocess med intensitet λ_T , så blir katastrofprocessen också en Poissonprocess, men nu med intensiteten $\lambda_K = p_K \lambda_T$. Detta är ju väldigt naturligt, eftersom $100p_K\%$ av tillbuden i det långa loppet utvecklas till katastrofer och katastrofer är ännu ovanligare än tillbud.

Övning 4) (Fortsättning på övning 3 ovan.) Processen i figur 1 överstiger tillbudsnivån x_T , totalt 8 gånger. Vid 2 av dessa 8 tillfällen överskrides katastrofnivån. Skatta den betingade sannolikheten $p_K = P(K|T)$ och katastrofprocessens intensitet λ_K .

Övning 5) I slutet av 70-talet skattades sannolikheten för härdsmläta i Barsebäck till $2/10^5$ per reaktordriftår i en utredning. En annan utredning kom fram till det 5 gånger större värdet $1/10^4$. Sannolikheten för ett okontrollerat utsläpp,

omfattande mer än 10% av reaktorhärden, givet härdsmlta, skattades till 0.2 resp 0.5 för de två typer av lättvattenreaktorer som används i Sverige.

Efter tillbudet i Three Mile Island (Harrisburg) 1979 förbättrades säkerheten. Fortfarande bedöms dock sannolikheten för en härdsmlta till ca $1/10^5$ per driftår men den betingade sannolikheten för ett okontrollerbart utsläpp givet härdsmlta är nu reducerad till någonstans mellan $1/10$ och $1/100$. Låt oss, för att vara konkreta, anta att den är $p = 1/50$.

Antag nu att en härdsmlta betraktas som tillbud T och en okontrollerbar sådan som katastrof K . I detta sammanhang är Poissonantagandet klart rimligt. Vad blir intensiteterna för tillbuds- respektive katastrofprocessen?

Övning 6) (Forts på övning 5 ovan.) Antag att en viss typ av reaktor har en drifttid om 30 år.

- (a) Bestäm sannolikheten för ingen härdsmlta under drifttiden.
- (b) Om ett land använder 10 sådana reaktorer i 30 år, vad blir då sannolikheten för ingen härdsmlta?
- (c) Om 10 länder använder 10 sådana reaktorer i 30 år, vad blir då sannolikheten för ingen härdsmlta?

Ett intressant sätt att tänka på Poissonprocessen är att Fru Fortuna (slumpens gudinna) utför ett slumpmässigt försök i varje (oändligt litet) tidsögonblick för att avgöra om händelsen (låt oss kalla den A) som Poissonprocessen räknar inträffar eller ej. Man resonerar då som om konvergensen i resultatet

$$\text{Bin}(n, p) \rightarrow \text{Poi}(\lambda) \text{ då } n \rightarrow \infty \text{ och } np \rightarrow \lambda$$

redan skett. Därför tänker vi oss att n är ett oändligt stort heltal och att sannolikheten $p = \lambda/n$ är infinitesimal³. Då ersätts gränsvärdesresultatet ovan av likheten $\text{Bin}(n, p) = \text{Poi}(\lambda)$.

Tänk dig nu att Fru Fortuna utför n oberoende försök, i vilka hon observerar huruvida A inträffar eller ej, under tiden från 0 till 1. Nu är n oändligt. Varje tidsögonblick måste då vara $dt = 1/n$ långt och i tidsintervallet $[0, t]$ hinner hon med nt oberoende försök. Antag vidare att sannolikheten att A inträffar är $p = \lambda dt$ i varje enskilt försök. Om alla kasten är oberoende, blir antalet gånger som myntet landar med krona upp $\text{Bin}(nt, p) = \text{Poi}(\lambda t)$, ty $ntp = nt\lambda dt = \lambda t$ (obs att nt ej nödvändigtvis är ett heltal och att detta inget betyder i argumenteringen, därför att $(nt \pm 1)p = \lambda t \pm p$ är infinitesimalt nära λt , ($p = \lambda dt$ är ju infinitesimal)).⁴

Peaks Over Threshold (POT)

I övning 4 ovan skattades den betingade sannolikheten för katastrof med en metod som man brukar kalla parameterfri. Vi ska nu skatta samma sannolikhet med en parametrisk metod baserad på Paretofördelningen.

³ordet infinitesimal betyder oändligt liten

⁴Resonemanget bygger på s k icke-standard-analys, som är en pur färsk typ av matematik, där man räknar med oändligt stora och oändligt små tal som om de vore vanliga reella tal.

Vi ska antaga att maximala nivån under ett överskridande av tillbuds-nivån $x_T = 5$ följer en Paretofördelning och att maxima från olika överskridanden är oberoende. Det finns teoretiska skäl för dessa antaganden som vi ej går in på här. Täthetsfunktionen för en Paretofördelad stokastisk variabel X , som bara antar värden ≥ 5 , är

$$f(x|T) = \frac{\alpha}{5} \left(\frac{5}{x}\right)^{\alpha+1}, \quad x \geq 5$$

och medelst integration får vi

$$1 - F(x|T) = P(X > x|T) = \int_x^\infty f(y|T) dy = \left(\frac{5}{x}\right)^\alpha \quad \text{för } x \geq 5$$

(obs att $\alpha > 0$). Tanken är att skatta parametern α och sedan sannolikheten $p_K = P(X > 6.75|T)$ med hjälp av maxvärdena från de 8 överskridandena:

6.85 5.11 5.61 5.28 6.87 6.40 5.16 5.12

En sådan skattning av p_K säges vara modellbaserad (ibland parametrisk) i motsats till den vi gjorde i en övning ovan.

Vi ska skatta α med "maximum likelihood"-metoden⁵, som nämndes i anteckningarna om Bayesiansk uppdatering. Därvid bildas först den så kallade "likelihooden" eller troligheten som jag tycker är en bra svensk översättning. Troligheten är produkten av observationernas resp tätheter evaluerade i sina observationer:

$$L(\alpha; x_1, \dots, x_n) = \prod_{i=1}^n f(x_i)$$

($x_1, \dots, x_n$ betecknar här stickprovets värden; produkt blir det eftersom vi tänker oss att de olika överskridandena är oberoende av varandra). *Trolighetsskattningen* av α är det värde $\hat{\alpha}$ som maximerar troligheten:

$$\hat{\alpha} = \arg \max_{\alpha} L(\alpha; x_1, \dots, x_n)$$

Obs att $\hat{\alpha}$ beror av observationerna $x_1, \dots, x_n$ och är därför en stickprovsvariabel.

Den här typen av maximeringar är numera enkla att genomföra numeriskt eftersom vi har tillgång till mycket kraftfulla matematik-program som Matlab, Mathcad, m fl. Med Mathematica erhöj jag

$$\hat{\alpha} \approx 7.105$$

och när jag stoppar in detta värde på α i uttrycket

$$p_K = P(X > 6.75|T) = \left(\frac{5}{6.75}\right)^\alpha$$

⁵Kallas ibland ML-metoden och ibland trolighetsmetoden

får jag skattningen

$$\hat{p}_K \approx 0.119$$

Jfr med $\hat{p}_K = 0.25$ (se övning 4 ovan). Man kan och bör fundera över om våra två skattningar $\hat{p}_K = 0.25$ och $\hat{p}_K = 0.119$ av $p_K = P(K|T)$ är motstridiga eller ej.

Jag tror inte att de är det. För det första, så är 2 överskridanden i detta sammanhang nära 1 överskridande och $1/8 = 0.125 \approx 0.119$. För det andra, antag att katastrofnivån istället för 6.75 vore 7. Då skulle den parameterfria skattningen av $p_K = P(K|T)$ bli $\hat{p}_K = 0$ (vi har ju inga observationer över nivån 7), medan den parametriska skulle bli

$$\hat{p}_K = \left(\frac{5}{7}\right)^\alpha \approx 0.092$$

Ett annat sätt att bilda sig en uppfattning om kvaliteten i den parametriska skattningen av $p_K = P(K|T)$ är att plotta data mot kvantiler (s k kvantil-plot) m a p den anpassade fördelningen som har $\alpha = 7.105$. Räkna själv ut att de åtta (teoretiska) kvantilerna svarande mot sannolikheterna

$$1/16, 3/16, 5/16, 7/16, 9/16, 11/16, 13/16, 15/16$$

är

$$5.05 \quad 5.15 \quad 5.27 \quad 5.42 \quad 5.62 \quad 5.89 \quad 6.33 \quad 7.39$$

($P(X \leq 5.05) = 1/16$, $P(X \leq 5.15) = 3/16$, etc.) Motsvarande datapunkter är

$$5.11 \quad 5.12 \quad 5.16 \quad 5.28 \quad 5.61 \quad 6.40 \quad 6.85 \quad 6.87$$

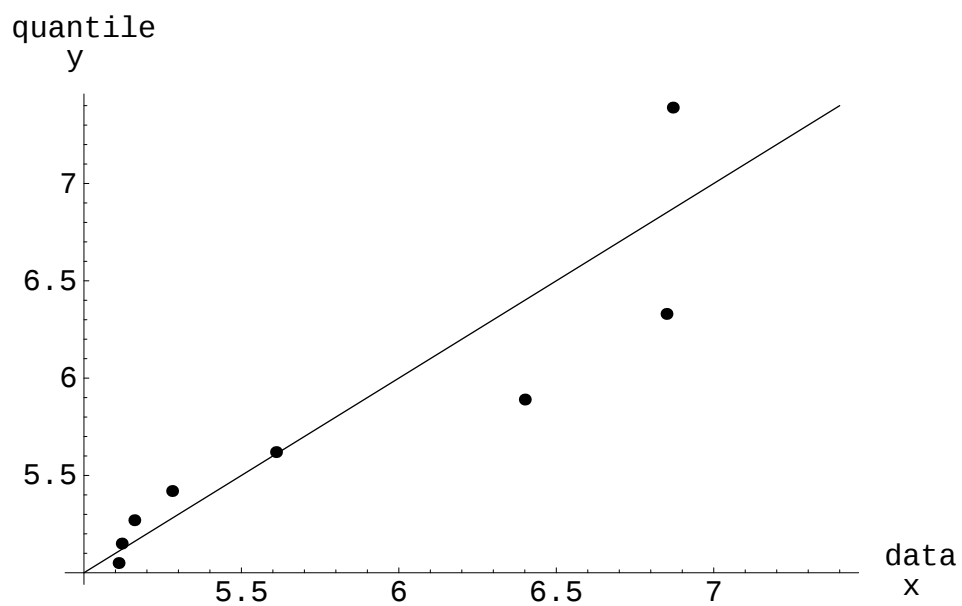
Kvantil-plotten visas i figur 3. Kurvan är inte så linjär som man kanske skulle önska sig. Fast man ska inte vänta sig att kvantil-plotten blir väldigt linjär när man endast har 8 datapunkter att anpassa till. För att få en uppfattning om olinjäriteten är vad man kan förvänta sig, simulerade jag 8 observationer av en Paretofördelning med $\alpha = 7.105$ och plottade dessa mot samma teoretiska kvantiler. Den plotten visas i figur 4. Det går inte att dra definitiva slutsatser av denna enda simulering, men nog verkar den simulerade kvantil-plotten aningen mer linjär. Eller?

Den parameterfria skattningen av $p_K = P(K|T)$ är tydligen betydligt känsligare för valet av katastrofnivå. Om nu den skulle gå att välja. En annan poäng med POT-metoden är att med den kan vi skatta $p_K = P(K|T)$ även när vi inte har observerat några tillbud som utvecklades till katastrofer.

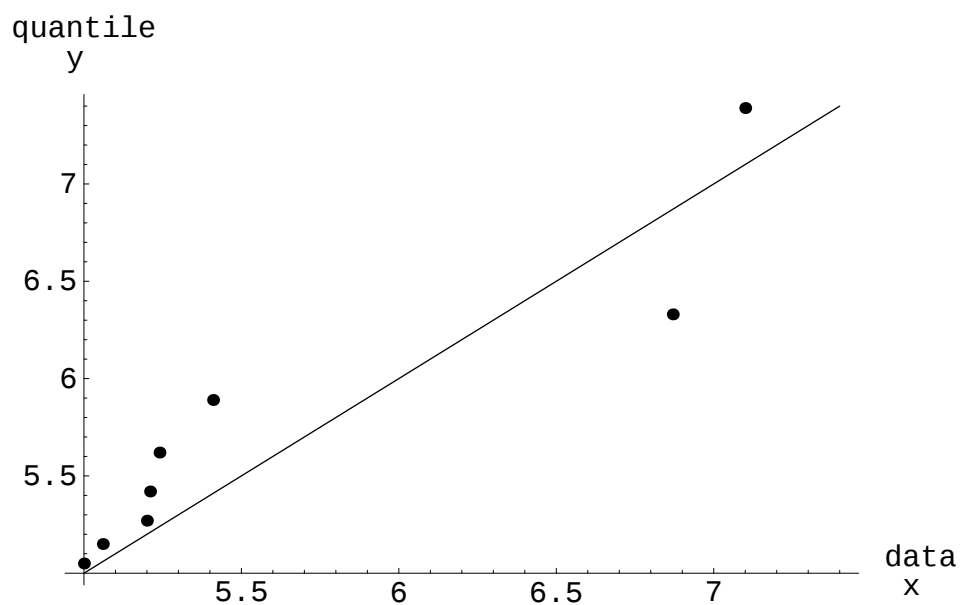
Övning 7) Använd den parametriska skattningen av $p_K = P(K|T)$ till att skatta intensiteten λ_K i katastrofprocessen (jfr övning 4).

Övning 8) Antag att du har ett stickprov $x_1, \dots, x_n$ på en stokastisk variabel X med Pareto-täthet

$$f(x) = \frac{\alpha}{5} \left(\frac{5}{x}\right)^{\alpha+1} \quad \text{för } x \geq 5$$



Figur 3: Kvantil-plot av anpassningen av överskridandedata till Paretofördelningen



Figur 4: Kvantil-plot av simulerade överskridandedata

- (a) Bestäm troligheten $L(\alpha; x_1, \dots, x_n)$.
- (b) Bestäm log-troligheten, alltså $\mathcal{L}(\alpha) = \ln L(\alpha; x_1, \dots, x_n)$.
- (c) Derivera log-troligheten map på α .
- (d) Lös ut α ur ekvationen

$$\frac{\partial \mathcal{L}(\alpha)}{\partial \alpha} = 0$$

- (e) Övertyga dig om att du har hittat det globala maximat.
- (f) Beräkna trolighetsskattningen $\hat{\alpha}$ genom att plugga in de $n = 8$ maxima i din formel.

Här följer ytterligare några övningar på trolighetsmetoden. Fundera efter varje övning på om du tycker att resultatet är rimligt eller ej.

Övning 9) Antag att du har ett stickprov $x_1, \dots, x_n$ på $X \sim E(\lambda)$. Härled trolighetsskattningen av λ .

Övning 10) Antag att du har ett stickprov $x_1, \dots, x_n$ på $X \sim \text{Geo}(p)$. Härled trolighetsskattningen av p . Hur bär man sig åt för att få ett stickprov bestående av n observationer på $\text{Geo}(p)$?

Övning 11) Antag att du har ett stickprov $x_1, \dots, x_n$ på $X \sim \text{Bin}(1, p)$. Härled trolighetsskattningen av p . (Ledning: $y = \sum_i x_i$ är en observation på $Y \sim \text{Bin}(n, p)$.)

Övning 12) Antag att du har ett stickprov $x_1, \dots, x_n$ på $X \sim N(\mu, \sigma^2)$. Härled trolighetsskattningen av (μ, σ^2) .

Det finns andra metoder än trolighetsmetoden som man kan använda till att härleda skattningar. En sådan är momentmetoden (som också skulle kunna kallas den naturliga metoden). Hur den fungerar för Paretofördelningen ska vi gå igenom nu.

Paretofördelningens väntevärde är

$$\begin{aligned} \mu &= \int_u^\infty x \frac{\alpha}{u} \left(\frac{u}{x}\right)^{\alpha+1} dx \\ &= \dots = \frac{\alpha}{\alpha-1} u \end{aligned}$$

Obs att här måste $\alpha > 1$. Annars är inte integralen konvergent. Inverterar vi uttrycket kan vi uttrycka α , som vi söker, i μ , som skattas av $\bar{x}$, och u , som är känt. Vi får att

$$\alpha = \frac{\mu}{\mu - u}$$

Momentmetodens skattning av α är därför

$$\hat{\alpha} = \frac{\bar{x}}{\bar{x} - u}$$

I exemplet med lastprocessen gäller $u = 5$ och $\bar{x} = 5.8$, vilket ger momenskattningen $\hat{\alpha} = 7.25$. Jfr med trolighetsskattningen $\hat{\alpha} = 7.105$.

Vilket av dessa två värden som bäst approximerar det sanna värdet är svårt att säga. En sätt att ta reda på detta är medelst simulering. Man gör då många upprepade simuleringar av $n = 8$ observationer på en Paretofördelning med $u = 5$ och, t ex, $\alpha = 7.2$. I varje simulering beräknas trolighets- och moment-skattningen av α . Sedan jämföres resultaten.

Övning 13) Antag att du har ett stickprov $x_1, \dots, x_n$ på X . Härled moment-skattningen av parametern (parametrarna) i X 's fördelning, om X är

- (a) $E(\lambda)$
- (b) $\text{Geo}(p)$
- (c) $\text{Bin}(1, p)$
- (d) $N(\mu, \sigma)$

I konstruktionssammanhang är man ofta intresserad av så kallade *återkomstnivåer*. Man definierar t ex 100-dygnsnivån x_{100} som den nivå som i genomsnitt över-skrides en gång på 100 dygn. Vi ska nu visa har man kan skatta x_{100} för last-processen vi studerar. Obs först att $x_{100} > 5$, ty nivån 5 överstigs i snitt 1 gång på $1/0.142 \approx 7$ dygn.

Beteckna med K' händelsen att lastprocessen överstiger nivån x_{100} . Vårt krav på x_{100} innebär att $0.142P(K'|T) = 1/100$, vilket ger $P(K'|T) \approx 0.0702$ och ur $P(K'|T) = (5/x_{100})^\alpha$ får vi att

$$x_{100} = 5 \cdot P(K'|T)^{-1/\alpha} \approx 7.27$$

Övning 14) Bestäm x_{50} och x_{200} för lastprocessen vi studerar.

Övning 15) Låt Y_t beteckna en viss fonds slutkurs dag t . Antag att de procentuella kursförändringarna $X_{t+1} = 100(Y_{t+1} - Y_t)/Y_t$ är oberoende stokastiska variabler. Under 100 på varandra följande börsdagar har kursen vänt nedåt med mer än 2% 8 gånger.

(a) Skatta $p_{2\%} = P(X_t < -2)$.

De 8 kursförändringar $< -2\%$ var

$$-2.31 \quad -2.10 \quad -2.29 \quad -2.38 \quad -2.44 \quad -2.54 \quad -2.64 \quad -2.33$$

Man vill skatta sannolikheten för en kursminskning om minst 4% trots att inga sådana kraftiga nedgångar observerades och gör därför Pareto-ansatsen

$$P(-X_t > x | -X_t > 2) = \left(\frac{2}{x}\right)^\alpha, \quad x \geq 2$$

(b) Skatta α och därefter $p_{4\%} = P(X_t < -4)$.

Övning 16) (Forts på ovanstående uppgift.) För hela datamängden i ovanstående uppgift erhöles medelvärdet $\bar{x} = 0.157$ och standardavvikelsen $s = 1.334$. Antag att X_t -variablerna är oberoende normalvariabler. Beräkna ur stickprovsvariablerna $\bar{x}$ och s skattningar av

(a) $p_{2\%} = P(X_t < -2)$,

(b) $p_{4\%} = P(X_t < -4)$.

Data till ovanstående två övningar simulerades från $N(0.2, 1.2)$ -fördelningen. Vi kan därför jämföra resultaten från ovanstående två övningar med de sanna värdena $p_{2\%} = P(X_t < -2) = \Phi(-1.833) \approx 0.0334$ och $p_{4\%} = P(X_t < -4) = \Phi(-3.5) \approx 0.000233$.

Övning 17) I övn 15 ansattes Paretomodellen $P(-X > x | -X > 2) = (x/2)^{1-\alpha}$, $x > 2$ och vi såg att $\alpha \approx 5.84$, samt att $P(-X > 2) \approx 0.08$. Bestäm approximativt

(a) x_{50} , alltså den negativa kursförändring som i genomsnitt överskrider 1 gång på 50 dagar,

(b) x_{250} .

Sammanfattande synpunkter

Vi har infört Poissonprocessen för att kunna modellera när i tiden ovanliga händelser inträffar. Vill man dessutom modellera styrkan av de ovanliga händelserna är Paretoantagandet vi gjorde inte oävet. Beteckna tillbudsnivån med u . Låt X beteckna totala nivån på ett tillbud, vilket som helst. Då gäller, analogt med ovan, att tätheten för X är

$$f(x|T) = \frac{\alpha}{u} \left(\frac{u}{x}\right)^{\alpha+1} \quad \text{för } x \geq u$$

och, medelst integration, fås

$$1 - F(x|T) = P(X > x|T) = \left(\frac{u}{x}\right)^{\alpha} \quad \text{för } x \geq u$$

Övning 18) Om $\alpha = 8$ och tillbudsnivån är $u = 10$, vad blir $p = P(X > v|T)$ om katastrofnivån v är (a) 12 resp (b) 13?

Övning 19) Antag att $\alpha = 5$ och att tillbudsnivån är $u = 15$. Hur stor är katastrofnivån v om i genomsnitt 1 tillbud av 3 resulterar i katastrof?

Genom att argumentera som i övning 8 inses att trolighetsskattningen av α är

$$\hat{\alpha} = \frac{n}{\sum_i \ln(x_i/u)}$$

Vi tänker oss här att vi har sett n tillbud med respektive nivåer $x_1, \dots, x_n$.

Övning 20) Trolighetsskatta α om det är känt att $u = 10$ och man har överskridandena

(a) 10.24, 20.04, 11.69, 10.41, 13.63, 11.06.

(b) 14.60, 10.56, 10.11, 10.67, 11.92, 12.17.

(c) Fundera själv på hur värdena i datamängden påverkar skattningen av α .

Paretofördelningen har en generalisering, som brukar definieras i termer av själva överskridandet (excess på engelska) $Y = X - u$. Den generaliserade Paretofördelningen uppfyller

$$P[Y > y] = \left(1 + \gamma \frac{y}{\sigma}\right)^{-1/\gamma} \quad \text{för } y > 0$$

Här är σ är en skalparameter och γ är en parameter som definierar formen på fördelningen. Jfr med den vanliga Paretofördelningen

$$P[X > x] = \left(\frac{u}{x}\right)^\alpha \quad \text{för } x > u$$

som bara har en parameter α . Vi har alltså att $Y + u = X$. Vi kan skriva om den vanliga Paretofördelningen m h a detta samband. Då fås

$$P[Y > y] = \left(1 + \frac{y}{u}\right)^{-\alpha} \quad \text{för } y > 0$$

Den vanliga Paretofördelningen är alltså även en generaliserad sådan. Sambandet mellan parameteruppsättningarna u, α och σ, γ , fås ur ekvationerna $-\alpha = -1/\gamma$ och $1/u = \gamma/\sigma$ till

$$\begin{aligned}\gamma &= \frac{1}{\alpha} \\ \sigma &= \frac{u}{\alpha}\end{aligned}$$

Observera att det finns flera orsaker⁶ till att många statistiker tycker att det är bättre att anpassa "peaks over threshold"-data, som den datamängd vi tittat på under denna föreläsning, till den generaliserade Parameterfördelningen. Skälen till att jag ovan anpassat data till den vanliga Paretofördelningen är mestadels pedagogiska.

Observera också att vi inte studerat hur man bedömer osäkerheten i skattningarna. Till att göra också detta räcker inte en föreläsning.

Reklam

Det finns en E-kurs kallad Stationära stokastiska processer som den som vill lära sig mer om stokastiska processer i kontinuerlig tid, med fördel kan ta. Inom TM-programmet finns en kurs i tidsserieanalys i vilken man bl a tittar på s k ARMA-modeller för stokastiska processer i diskret tid.

Facit till övningarna

- 1) (a) 0.175 st/vecka (b) ca 0.0027 (c) ca 0.90
- 2) (a) $\hat{\lambda} = 18/15 = 1.2$ st/år (b) $\hat{\lambda}_1 \approx 0.067$, $\hat{\lambda}_2 = 0.200$, $\hat{\lambda}_3 \approx 0.267$, $\hat{\lambda}_4 = \hat{\lambda}_5 = \hat{\lambda}_6 = \hat{\lambda}_7 = \hat{\lambda}_8 = 0$, $\hat{\lambda}_9 = 0.200$, $\hat{\lambda}_{10} \approx 0.267$, $\hat{\lambda}_{11} \approx 0.133$, $\hat{\lambda}_{12} \approx 0.067$ st/månad

⁶som vi inte går in på här

- 3) (a) $\hat{\lambda}_T \approx 0.142$ överskridanden per dygn (b) $\bar{x} \approx 2.88$ timmar
- 4) $\hat{p}_K = 0.25$, $\hat{\lambda}_K \approx 0.0356$ per dygn
- 5) $\lambda_T \approx 1/10^5$, $\lambda_K \approx 2/10^7$ impulser per driftsår
- 6) (a) $e^{-30/10^5} \approx 0.9997$ (b) $e^{-300/10^5} \approx 0.997$ (c) $e^{-3000/10^5} \approx 0.97$ (Obs att i alla tre fallen fungerar approximationen $e^{-x} \approx 1 - x$ bra.)
- 7) $\hat{\lambda}_K = 0.0169$
- 8) (d) $\alpha = n / \sum_i \ln(x_i/5)$ (f) $\hat{\alpha} \approx 7.105$
- 9) $\hat{\lambda} = 1/\bar{x}$
- 10) $\hat{p} = 1/\bar{x}$
- 11) $\hat{p} = \bar{x}$
- 12) $\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = (n-1)s^2/n$
- 13) (a) $\hat{\lambda} = 1/\bar{x}$ (b) $\hat{p} = 1/\bar{x}$ (c) $\hat{p} = \bar{x}$ (d) $\hat{\mu} = \bar{x}$, $\hat{\sigma}^2 = (n-1)s^2/n \Rightarrow \hat{\sigma} = \sqrt{(n-1)s^2/n}$
- 14) 6.59 resp 8.01
- 15) (a) $p_{2\%} = 0.08$ (b) med trolighetsskattningen $\hat{\alpha} \approx 5.837$ blir $\hat{p}_{4\%} \approx 0.00140$, med momentskattningen $\hat{\alpha} \approx 6.281$ blir $\hat{p}_{4\%} \approx 0.00103$
- 16) (a) ≈ 0.0529 (b) ≈ 0.000916
- 17) (a) $x_{50} \approx 2.54$ (b) $x_{250} \approx 3.34$
- 18) (a) $p \approx 0.23$ (b) $p \approx 0.12$
- 19) $v \approx 18.7$
- 20) (a) $\hat{\alpha} \approx 4.526$ (b) $\hat{\alpha} \approx 6.812$

Tack till

Joachim Johansson för flera vettiga synpunkter på en tidig version.

Referenser

- [1] Coles, Stuart: An Introduction to Statistical Modelling of Extreme Values. Springer 2001.
- [2] Devore, Jay & Nicholas Farnum: Applied Statistics for Engineers and Scientists. Duxbury Press, 1999.
- [3] Göran Grimvall, Per Jacobsson & Torbjörn Thedéen: Risker i tekniska system. Studentlitteratur. Lund 2003.
- [4] A. Ramachandra Rao & Khaled H. Hamed: Flood Frequency Analysis. CRC press, 2000.