Statistikgruppen
Handelshøjskolen i København

MPAS
Tue Tjur
September 2003

Notat 1:

Statistiske metoder i forbindelse med projekt 1.

Det vigtigste af de spørgsmål, som projekt 1 lægger op til, er:

Hvor mange og hvilke kunder er faldet fra i perioden?

Kriteriet for at en kunde er faldet fra må naturligvis være, at tiden fra sidste transaktion til dag 183 er lang i forhold til den tid der normalt går mellem kundens transaktioner. Vi kan omsætte dette vage kriterium til et eksakt mål for kundens "grad af frafald" på (mindst) to måder.

1. Poissonprocessen.. En *homogen Poissonproces* (se f.eks. noter i sandsynlighedsregning på 2. år, opgave 6.2.2) kan beskrives som en følge af stokastiske tidspunkter

$$0 < T_1 < T_2 < T_3 < \dots$$

hvis fordeling er beskrevet ved at ventetiderne

$$V_1 = T_1\,,\ V_2 = T_2 - T_1\,,\ V_3 = T_3 - T_2 \dots$$

er stokastisk uafhængige, eksponentialfordelte med samme skalaparameter $\beta = 1/\lambda$, hvor parameteren λ i denne sammenhæng kaldes *intensiteten*. Størrelserne $T_1, T_2, \dots$ fortolkes i denne forbindelse som de tidspunkter hvor en bestemt type af hændelse (punktering, taxaankomst, ...) indtræffer.

Alternativt kan Poissonprocessen beskrives ved at antallet af hændelser $N[s,t]$ i tidsintervallet $[s,t]$ er Poissonfordelt med parameter $\lambda(t-s)$, og antal hændelser i disjunkte tidintervaller er stokastisk uafhængige. Heraf følger, at intensiteten λ kan fortolkes som det forventede antal hændelser pr. tidsenhed.

Intuitivt er Poissonprocessen måske bedst beskrevet ved følgende egenskab: Den betingede sandsynlighed for at en hændelse indtræffer i et lille tidsinterval $[t, t+h]$, givet alt hvad der er sket før tid t, er uafhængig af hvad der er sket før tid t og approksimativt (for h lille) lig med λh.

Til beskrivelse af kundetransaktioner for en bestemt kunde er Poissonprocessen ikke helt rimelig, fordi man må formode, at jo længere tid der

er gået siden kunden har købt noget, jo mere sandsynligt er det at der snart sker et køb. Men som en tilnærmelse kan den måske bruges, og den giver i hvert fald følgende mulighed for kvantificering af den sikkerhed, hvormed vi kan sige at en kunde er faldet fra.

Antag at en kunde har haft transaktioner til tidspunkter

$$0 < t_1 < t_2 < \cdots < t_k \leq 183\,.$$

Idet tidsskalaen opfattes som kontinuert, og idet vi ignorerer alle problemer vedrørende weekender, jul osv., antages at denne følge af tidspunkter stammer fra en Poissonproces med (ukendt) intensitet λ. Vi observerer imidlertid kun den del af processen som ligger efter et bestemt tidspunkt a (typisk $a = 1$) og før et bestemt tidspunkt b (typisk $b = 183$). Den fælles skalaparameter for de $k-1$ observerede ventetider

$$v_2 = t_2 - t_1, \ldots, v_k = t_k - t_{k-1}$$

estimeres (ved simpel ML–estimation i modellen for uafhængige identisk fordelte eksponentialfordelte variable med ukendt skalaparameter) ved

$$\hat{\beta} = \frac{v_2 + \cdots + v_k}{k-1} = \frac{t_k - t_1}{k-1}$$

svarende til at $\lambda = 1/\beta$ naturligt estimeres ved

$$\hat{\lambda} = \frac{k-1}{t_k - t_1} = \frac{\text{antal hændelser}}{\text{tidsintervallets længde}}\,.$$

Hvis skalaparameteren β var kendt, ville vi naturligt teste hypotesen $b = 183$ (dvs. kunden er ikke faldet fra) på følgende måde. Ventetiden $v_{k+1} = t_{k+1} - t_k$ fra sidst observerede transaktion til den næste er i princippet uobserverbar, men vi har observeret at den er $\geq b - t_k$. Sandsynligheden for at dette indtræffer under antagelse af at $b = 183$ er

$$P\left(\frac{V_{k+1}}{\beta} \geq \frac{183 - t_k}{\beta}\right) = \exp\left(-\frac{183 - t_k}{\beta}\right)$$

fordi $\frac{V_{k+1}}{\beta}$ er normeret eksponentialfordelt. Hvis denne sandsynlighed (P–værdi) er lille, kan vi med rimelig sikkerhed sige at kunden må være faldet fra. Da vi imidlertid ikke kender β forsøger vi at erstatte β med $\hat{\beta}$, og får så i stedet følgende argument: Vi har observeret at

$$\frac{V_{k+1}}{\hat{\beta}} = \frac{V_{k+1}}{(V_2 + \cdots + V_k)/(k-1)}$$

$$\geq \frac{183 - t_k}{(V_2 + \cdots + V_k)/(k-1)} = \frac{183 - t_k}{(t_k - t_1)/(k-1)}\,.$$

Så hvis størrelsen $\frac{183-t_k}{(t_k-t_1)/(k-1)}$ ligger ekstremt langt ude i fordelingen af $\frac{V_{k+1}}{(V_2+\cdots+V_k)/(k-1)}$, kan vi konkludere, at kunden nok er faldet fra. Men denne fordeling ses let (v.h.a. Γ–fordelingens foldningsegenskab, χ^2–fordelingens sammenhæng med Γ–fordelingen og F–fordelingens definition) at være en F–fordeling med $(2, 2(k-1))$ frihedsgrader. Så et kriterium (der kan fortolkes som en P–værdi ved test af hypotesen om "intet frafald") er størrelsen

$$P\left(F(2,2(k-1)) \geq \frac{183-t_k}{(t_k-t_1)/(k-1)}\right)$$

(hvor vi, lidt sjusket, har brugt $F(2,2(k-1))$ som betegnelse for en stokastisk variabel med den pågældende F–fordeling).

2. Den rektangulære fordeling. En anden meget simpel tankegang, som kan bruges til "kvantificering af frafaldsgrad", er følgende. Antag at tidspunkterne $t_1, t_2, \ldots, t_k$ er helt tilfældigt spredt ud over kundens aktive periode $[a, b]$, i den forstand at de lige så godt kunne være genereret som tilfældige tal fra en ligefordeling på intervallet $[a, b]$, og derefter ordnet efter størrelse. For en kunde der har været aktiv i hele perioden skulle tidspunkterne altså ligne tilfældige tal fra en ligefordeling på [0,183]. Hvis kunden derimod er faldet fra, vil tidspunktet t_k for sidste transaktion typisk ligge langt til venstre i fordelingen af den største observation $R_{(k)}$ blandt k ligefordelte observationer $R_1, \ldots, R_k$ på intervallet [0,183]. Nu gælder der

$$P\left(R_{(k)} \leq t_k\right) = P\left(R_1, \ldots, R_k \leq t_k\right)$$

$$= P\left(R_1 \leq t_k\right) \ldots P\left(R_k \leq t_k\right) = \left(\frac{t_k}{183}\right)^k .$$

Hvis denne størrelse er meget lille, må vi formode at kunden er faldet fra.

En lidt mere raffineret version af dette argument, som også tager højde for at kunden ikke nødvendigvis har været med fra begyndelsen, ser ud som følger. Det vi ved om kundens starttidspunkt a kan sammenfattes i tidspunktet t_1. Vi erstatter derfor starttidspunktet 0 med tidspunktet t_1 for første transaktion, og betragter således en ligefordeling på intervallet $[t_1, 183]$ i stedet for $[0, 183]$. Tilbage er der så kun $k-1$ transaktioner der sker til tidspunkterne $t_2, \ldots, t_k$. Sandsynligheden for at sidste transaktion sker til det tidspunkt hvor den faktisk skete, eller endnu tidligere, bliver så

$$P\left(R_{(k-1)} \leq t_k\right) = \left(\frac{t_k - t_1}{183 - t_1}\right)^{k-1} .$$

Denne størrelse kan fortolkes som en slags P–værdi ved test af hypotesen "kunden er ikke faldet fra".

Vi har her udnyttet følgende simple egenskab ved den rektangulære fordeling: Hvis $R_1, \ldots, R_k$ er uafhængige, rektangulært fordelte på (for nemheds skyld) enhedsintervallet [0,1], og $R_{(1)}, \ldots, R_{(k)}$ som sædvanligt betegner de ordnede variable, så er den betingede fordeling, givet $R_{(1)}$, af $R_{(2)}, \ldots, R_{(k)}$, karakteriseret ved at $R_{(2)}, \ldots, R_{(k)}$ kan opfattes som et ordnet sæt af $k-1$ uafhængige variable fra en ligefordeling på intervallet $[R_{(1)}, 1]$.

På grund af en simpel sammenhæng mellem den betingede fordeling af tidspunkternes placering i Poissonprocessen, givet antallet af hændelser i et bestemt interval, og den rektangulære fordeling, gælder følgende

Resultat. *"P–værdierne"*

$$P\left(F(2, 2(k-1)) \geq \frac{183 - t_k}{(t_k - t_1)/(k-1)}\right)$$

(baseret på Poissonprocessen) og

$$P\left(R_{(k-1)} \leq t_k\right) = \left(\frac{t_k - t_1}{183 - t_1}\right)^{k-1}$$

(baseret på den rektangulære fordeling) er sammenfaldende.

Resultatet følger i øvrigt også direkte af sammenhængen mellem F–fordelingen og B–fordelingen, når det udnyttes at fordelingen af den største observation $R_{(k)}$ netop følger en B–fordeling (Ssr. side 116–117 og side 126–127).

Det var jo dejligt — der er ikke grund til at forsøge med begge metoder, de giver nøjagtigt samme resultat.

Eksempel. Kunde nr. 1 har transaktioner til tidspunkter 17, 43, 71, 120 og 150. P–værdien baseret på Poissonprocessen bliver

$$P\left(F(2, 2(5-1)) \leq \frac{183 - 150}{(150 - 17)/(5-1)}\right) = P\left(F(2, 8) \leq 0.9925\right)$$

som (f.eks. v.h.a. WinT) udregnes til 0.4121. P–værdien baseret på den rektangulære fordeling bliver ligeledes

$$\left(\frac{150 - 17}{183 - 17}\right)^{5-1} = 0.4121\,.$$

Der er således ikke noget der tyder på, at kunde nr. 1 er faldet fra.

3. Test for frafald, estimation af antal frafald og operationel brug af frafaldskriterier.

Hvis vi for hver kunde beregner den således udledte "P–værdi" skulle vi vente, under forudsætning af at frafald overhovedet ikke forekommer, at den samlede fordeling af disse størrelser ligner en ligefordeling på enhedsintervallet. Hvis derimod histogrammet viser en en tydelig overvægt af små værdier, må vi konkludere at frafald faktisk forekommer.

Dette test er baseret på Poissonprocessens (muligvis ikke helt realistiske) antagelser. Desuden ignorerer det problemerne med weekender og jul, samt den approksimation der ligger i at Poissonprocessen opererer med kontinuert tid. En mere kvalificeret vurdering kunne baseres på simulationer. For eksempel kunne man i et manipuleret datasæt, hvor kundernes transaktionsantal er de samme som i det oprindelige datamateriale, men hvor transaktionstidspunkterne er genererede som uafhængige variable med samme fordeling som den marginale fordeling af samtlige transaktionstidspunkter, gentage udregningen af ovennævnte P–værdier, og se på hvordan de så fordeler sig. Eventuelt kunne man gøre det 1000 eller 10000 gange og danne gennemsnit, og derved få en fordeling der er mere korrekt som sammenligningsgrundlag end den rektangulære.

Ud fra billedet af hvor meget fordelingen af P–værdierne har forskudt sig i forhold til den "frafalds–rensede" fordeling må det være muligt at estimere (eller, i det mindste, give en nedre grænse for) antallet af frafald. Det må vi vende tilbage til senere.

Som en sidste krølle på disse overvejelser kunne man forestille sig at udføre konsekvensberegninger, der helt konkret undersøger konsekvenserne af en strategi baseret på frafaldskriterier af denne art. Det kriterium vi har opstillet for frafald til tid 183 kan naturligvis bruges på et hvilket som helst tidspunkt (undtagen lige i starten). Man kan derfor spørge: Hvad ville der være sket, hvis JohnsonDiversey (for eksempel) en gang om ugen havde foretaget en eller anden form for markedsføringsinitiativ overfor de kunder, der udviste en P–værdi på højst (for eksempel) 0.01? Hvor mange af disse initiativer ville have været overflødige, i den forstand at de var rettet mod kunder der alligevel senere ville vende tilbage? Kan man, ved passende justering af parametrene "en gang om ugen" og "højst 0.01" opnå noget (i passende forstand, som skal defineres) mere effektivt?