Avvikelsedetektion i signaler från Regina

Anders Holst[1], Jan Ekman[1], Stefan Larsen[2]

[1]SICS
[2]Bombardier Transportation AB

Sammanfattning

I Reginatågen genereras signaler om såväl s.k. "events" (meddelanden om både rutinhändelser och mer eller mindre allvarliga fel i olika enheter) som "condition" (driftsräknare för olika enheter). Det är önskvärt att övervaka båda dessa typer av signaler för att upptäcka avvikelser som kraftigt förändrad frekvens eller driftsintensitet. Sådana avvikelser skulle kunna signalera olika servicebehov, och det skulle alltså vara till användning om servicepersonal fick reda på dem i god tid innan de lett till allvarligare fel.

Vi kommer här att gå igenom en grundläggande och generellt användbar statistisk modell för detta scenario. Metoden utvärderas på autentiska data från Reginatågen.

SICS Technical Report T2006:04 ISSN 1100-3154 ISRN: SICS-T–2006/04-SE

1 Bakgrund

I Reginatågen genereras signaler om såväl s.k. "events" (meddelanden om både rutinhändelser och mer eller mindre allvarliga fel i olika enheter) som "condition" (driftsräknare för olika enheter). Det är önskvärt att övervaka båda dessa typer av signaler för att upptäcka avvikelser som kraftigt förändrad frekvens eller driftsintensitet. Sådana avvikelser skulle kunna signalera olika servicebehov, och det skulle alltså vara till användning om servicepersonal fick reda på dem i god tid innan de lett till allvarligare fel.

Det som ligger närmast till hands är att upptäcka avvikelser i ett fordon relativt alla andra, men man kan även tänka sig avvikelser i hela flottan vid en viss tidpunkt (t.ex. för att signalera extrema väderförhållanden), eller på en viss geografisk plats (t.ex. för att upptäcka balisfel eller ledningsproblem), eller vilket annan urvalskriterium som helst.

Vi kommer här att gå igenom en grundläggande och generellt användbar statistisk modell för detta scenario.

2 Händelsedata

Olika händelsemeddelanden är olika allvarliga. En del signalerar akuta fel som ovillkorligen måste åtgärdas. Dessa är de "enkla" händelserna: det behövs ingen avvikelsedetektion när man får dem, utan det är bara att boka in en verkstadstid direkt (eller åtgärda på annat sätt). Andra händelser signalerar rena rutinåtgärder, som att föraren har nycklat upp tåget, eller att någon lucka är öppen. Vid första anblicken är det kanske svårt att se hur dessa händelser skulle kunna skvallra något om servicebehov. Sen finns ett stort antal händelser som potentiellt kan signalera att något är fel, även om de också "ströhänder" slumpmässigt i normala fall. Några enstaka sådana händelser kan normalt ignoreras, utan först när samma meddelande har uppträtt ett antal gånger betraktas det som signifikant. Dessa är de svåra händelserna, där ett verktyg för att upptäcka en förändrad frekvens skulle kunna reagera innan en människa noterar att något är ovanligt. På samma sätt skulle även förändringar i frekvensen av till synes helt "harmlösa" meddelanden skvallra om att något är på gång som borde kollas upp extra.

Grundidén för avvikelsedetektionen i detta fall är att räkna hur många händelser av en viss typ som inträffar under ett visst tidsintervall, och jämföra det med hur många händelser som brukar genereras under ett lika långt intervall. Om antalet genererade händelser är signifikant annorlunda än det förväntade antalet så är detta en avvikelse. I mer detalj går det till så att vi samlar data från ett stort antal förmodat "normala" intervall, och bygger från dessa en statistisk modell över fördelningen av antalet händelser i ett intervall. Sen jämför vi ett nytt intervall med den modellen, genom att räkna ut sannolikheten att det observerade antalet händelser i det nya intervallet ska ha genererats av den statistiska modellen. Om den sannolikheten är låg, så har förmodligen det nya fallet inte genererats av denna modell över "normala" fall, utan betraktas som

avvikande.

Låt oss alltså anta att varje meddelandekod genereras slumpmässigt med en viss frekvens. (Även om vi vet att vissa meddelanden genereras mer regelbundet eller bara under specifika omständigheter så är detta en approximation som är tillräckligt bra i många fall). Det innebär att avståndet mellan två meddelanden av samma typ är *exponentialfördelat*:

$$f(d) = e^{-df} \tag{1}$$

där parametern f är medelfrekvensen för händelserna.

Om avstånden mellan händelserna är exponentialfördelade så innebär det att antalet händelser som inträffar inom en bestämd tidsperiod är *Poissonfördelat*:

$$P(n) = \frac{e^{-\lambda}\lambda^n}{n!} \tag{2}$$

där parametern λ är medelantalet händelser under en sådan tidsperiod. Förhållandet mellan parametrarna f och λ är:

$$\lambda = Tf \tag{3}$$

där T är den betraktade tidsperiodens längd.

Indata till vårt problem består av räknare över antalen inträffade händelser i ett antal intervall, samt en räknare för antalet händelser i ett specifikt intervall som vi vill avgöra hur normalt det är. Utdata är sannolikheten att det specifika antalet händelser har genererats av samma modell som de övriga. Det direkta sättet att göra detta är att skatta parametern λ från mängden av räknare, och sen använda formeln ovan för att räkna ut sannolikheten. Vi ska dock jämföra tre olika sätt att göra detta: En approximation med normalfördelning; klassisk skattning av Poissonfördelningen; samt Bayesiansk beräkning av sannolikheten.

2.1 Approximation med normalfördelning

För höga händelsefrekvenser blir det snabbt stora tal i uttrycket för Poissonfördelningen, varför den kan kännas komplicerad att beräkna. Samtidigt ser Poissonfördelningen för höga frekvenser ut ungefär ut som en normalfördelning, varför man kan vara frestad att använda denna som approximation. Samtidigt måste man komma ihåg att för låga frekvenser ser fördelningarna inte alls lika ut, och denna approximation fungerar alltså inte då. Vi tar upp normalfördelningen här för att visa hur approximationen slår, och för att ge en känsla för när den går att använda.

Normalfördelningen (d.v.s Gaussfördelningen) ser ut:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2} \tag{4}$$

där medelvärdet μ och standardavvikelsen σ kan skattas som (maximum likelihood-skattningar):

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \tag{5}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} \tag{6}$$

För att bedöma hur avvikande ett nytt fall är, så räcker det egentligen med att räkna ut hur många standardavvikelser från medlet fallet ligger, $|x_i - \mu| / \sigma$. I de två följande beräkningsmetoderna behöver man dock räkna ut själva sannolikheten för det nya fallet, varför vi kommer att göra det här också med formel (4) ovan för att kunna jämföra resultaten.

2.2 Klassisk skattning av Poissonfördelningen

Låt oss göra samma sak med Poissonfördelningen som med normalfördelningen ovan. Det är i själva verket lättare, eftersom den bara har en parameter, λ, som skattas på samma sätt som μ ovan:

$$\lambda = \frac{\sum_{i=1}^{N} x_i}{N} \tag{7}$$

(Hos Poissonfördelningen är såväl medelvärdet som variansen lika med λ.)

Nu kan sannolikheten för ett nytt fall räknas ut med hjälp av formel (2).

2.3 Bayesiansk beräkning av sannolikheten

En nackdel med den klassiska skattningen är att den inte tar hänsyn till osäkerheten i λ. Om man har många datapunkter så är precisionen förmodligen god, men om man har ganska få observationer så är skattningen väldigt känslig för slumpvariationer i data. Detta korrigeras med Bayesiansk statistik: Istället för att använda en punktskattning av λ så betraktar man denna parameter som en stokastisk variabel, och beräknar hela fördelningen över den givet mängden observerade data, $P(\lambda|D)$. För att beräkna sannolikheten för ett nytt fall x får vi integrera över alla möjliga värden av λ:

$$P(x|D) = \int_{\lambda} P(x|\lambda) P(\lambda|D) \tag{8}$$

Enligt Bayes sats så är:

$$P(\lambda|D) \propto P(D|\lambda) P(\lambda) \tag{9}$$

och om vi ansätter en uniform prior så får vi att:

$$P(\lambda|D) \propto P(D|\lambda) = \prod_{i=1}^{N} P(x_i|\lambda) = \frac{e^{-N\lambda}\lambda^{\sum_i x_i}}{\prod_i (x_i!)} \tag{10}$$

Med hjälp av detta kan vi beräkna sannolikheten för ett nytt fall som:

$$\begin{aligned} P(x|D) = \int P(x|\lambda)P(\lambda|D) &= \frac{\int_\lambda P(x|\lambda)P(D|\lambda)}{\int_\lambda P(D|\lambda)} \\ &= \frac{\int_\lambda e^{-(N+1)\lambda}\lambda^{x+\sum_i x_i}/(x!\prod_i(x_i!))}{\int_\lambda e^{-N\lambda}\lambda^{\sum_i x_i}/\prod_i(x_i!)} \\ &= \frac{1}{x!}\cdot\frac{(x+\sum_i x_i)!}{(N+1)^{x+1+\sum_i x_i}}\cdot\frac{N^{1+\sum_i x_i}}{(\sum_i x_i)!} \\ &= \frac{(x+\sum_i x_i)!}{x!(\sum_i x_i)!}\cdot\left(\frac{N}{N+1}\right)^{1+\sum_i x_i}\cdot\left(\frac{1}{N+1}\right)^{x} \end{aligned} \tag{11}$$

Eftersom samtliga uttryck ovan för att beräkna sannolikheten av ett nytt fall snabbt ger värden mycket nära noll när det nya fallet börjar bli lite ovanligt, så är det man normalt brukar bedöma den negativa log-likelihooden, dvs $-\log P(x)$. I fallet ovan, och om man använder Sterlings approximation för fakulteter, $n! \approx \sqrt{2\pi}n^{n+0.5}/e^n$, (och logaritmen av denna, $\log n! \approx (\log(n) - 1)(n+0.5) + 0.5(\log(2\pi)+1)$) så blir det man ska beräkna:

$$\begin{aligned} -\log P(x|D) \approx &\log(N+1)(x+1+\sum_i x_i) - \log(N)(1+\sum_i x_i) + \\ &+ (\log x - 1)(x+0.5) + (\log(\sum_i x_i) - 1)(\sum_i x_i + 0.5) - \\ &- (\log(x+\sum_i x_i) - 1)(x+\sum_i x_i + 0.5) + 0.5(\log(2\pi)+1) \end{aligned} \tag{12}$$

2.4 Varierande intervallängd

Ovan förutsattes att vi har data över antalet händelser inom ett stort antal lika långa intervall. En möjlig situation är att vi själva inte kan välja intervallängden, utan bara har data över antalet händelser i ett antal olika långa intervall. Lyckligtvis är detta rättframt att hantera.

Ett konkret exempel på när detta kan vara användbart i fallet med Reginor är följande: En ansats är att betrakta flottan av Reginor under den senaste veckan för att se om någon skiljer ut sig från mängden med avseende på något meddelande. Våra datapunkter kommer alltså att bli det antal av den aktuella händelsetypen som inträffat på varje Regina under den senaste veckan. Men de

flesta händelserna inträffar förmodligen i första hand på tåg i drift, varför antalet händelser beror på hur mycket varje fordon har använts. Eftersom fordonen rör sig i lite olika omlopp med lite olika utnyttjandegrad så blir de inte jämförbara. Istället för att räkna intervallet i veckodagar kanske man alltså hellre vill räkna intervallet i antal rullade kilometer (eller någon annan driftenhet). Naturligtvis kan man då med hjälp av driftsräknarna ta reda på antalet händelser de senaste 5000 kilometerna istället för den senaste veckan, men ett alternativ är att fortfarande räkna händelser den senaste veckan och därefter ta hänsyn till de olika fordonens olika intervallängd i kilometer.

Börja med att titta på förhållandet i formel (3), över relationen mellan frekvensen f i en exponentialfördelning och λ i Poissonfördelningen. Frekvensen av händelser är fortfarande konstant, men λ varierar mellan datapunkterna med intervallängderna:

$$\lambda_i = t_i f \tag{13}$$

Poissonfördelningen uttryckt i parametrarna f och t_i blir då istället:

$$P(x_i|f, t_i) = \frac{e^{-t_i f}(t_i f)^{x_i}}{x_i!} \tag{14}$$

Om man vill räkna klassiskt så skattar man nu helt enkelt parametern f som:

$$f = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} t_i} \tag{15}$$

och sätter in detta tillsammans med ett nytt par x och t i Poissonfördelningsuttrycket.

För den Bayesianska beräkningen skriver vi ner fördelningen över f givet data:

$$P(f|D) \propto P(D|f) = \prod_{i=1}^{N} P(x_i|f, t_i) = \frac{e^{-f\sum_i t_i} f^{\sum_i x_i} \prod_i t_i^{x_i}}{\prod_i (x_i!)} \tag{16}$$

vilket ger:

$$
\begin{aligned}
P(x|D) = \int P(x|f)P(f|D) = \frac{\int_f P(x|f,t)P(D|f)}{\int_f P(D|f)} \\
=& \frac{\int_f e^{-f(t+\sum_i t_i)} f^{x+\sum_i x_i} t^x \prod_i t_i^{x_i} / (x! \prod_i (x_i!))}{\int_f e^{-f\sum_i t_i} f^{\sum_i x_i} \prod_i t_i^{x_i} / \prod_i (x_i!)} \\
=& \frac{t^x}{x!} \cdot \frac{(x+\sum_i x_i)!}{(t+\sum_i t_i)^{x+1+\sum_i x_i}} \cdot \frac{(\sum_i t_i)^{1+\sum_i x_i}}{(\sum_i x_i)!} \\
=& \frac{(x+\sum_i x_i)!}{x!(\sum_i x_i)!} \cdot \left(\frac{\sum_i t_i}{t+\sum_i t_i}\right)^{1+\sum_i x_i} \cdot \left(\frac{t}{t+\sum_i t_i}\right)^x
\end{aligned} \tag{17}
$$

Efter logaritmering och Sterlings approximation får man slutligen:

$$
\begin{aligned}
-\log P(x|D) \approx& \log(t+\sum_i t_i)(x+1+\sum_i x_i) - \log(\sum_i t_i)(1+\sum_i x_i) - \log(t)x + \\
&+ (\log x - 1)(x+0.5) + (\log(\sum_i x_i) - 1)(\sum_i x_i + 0.5) - \\
&- (\log(x+\sum_i x_i) - 1)(x+\sum_i x_i + 0.5) + 0.5(\log(2\pi)+1)
\end{aligned} \tag{18}
$$

2.5 Räkneexempel

För att se hur de tre olika metoderna (för fixa intervallängder) att uppskatta sannolikheten för ett nytt fall beter sig, ska vi titta på två exempel.

Antag att vi för en viss händelsekod har observerat sex lika långa intervall med följande antal händelser i: 3, 5, 2, 3, 8, 6. Dessa utgör de data vi vill basera vår modell över normala fall på. Därefter vill vi utvärdera var och en av följande nya antal händelser för att se hur ovanliga de bedöms vara: 0, 4, 25.

Låt oss börja med normalapproximationen. Om man sätter in observationerna i formel (5) och (6) så får man skattningarna $\mu = 4.5$ och $\sigma = 2.0616$. I tabellen nedan visas i kolumnen "Normalappr." resultatet av att sätta in dessa parametervärden tillsammans med vart och ett av testfallen i formel (4).

För den klassiska beräkningen av sannolikheterna använder vi istället formel (2) med $\lambda = 4.5$. Resultaten syns i kolumnen "Klassiskt" i tabellen nedan.

Slutligen använder vi formel (11) för den Bayesianska beräkningen. Resultaten syns i kolumnen "Bayesianskt" i tabellen.

	Normalappr.	Klassiskt	Bayesianskt
0	0.01787	0.01111	0.01335
4	0.1879	0.1898	0.1750
25	$6.527 \cdot 10^{-23}$	$1.532 \cdot 10^{-11}$	$4.754 \cdot 10^{-9}$

Som syns i tabellen så ger de tre metoderna i detta exempel ungefär samma resultat för 0 och 4. Alla är också överens om att fallet 25 är klart avvikande, men de skiljer sig åt i hur mycket. Normalapproximationen ger en mycket lägre sannolikhet än de andra, vilket är typiskt: den överskattar ofta ovanligheten. Den Bayesianska beräkningen är den mest försiktiga, vilket också är typiskt. Den tar nämligen hänsyn till osäkerheten i observationsdata. Det innebär att om vi baserar vår beräkning på en mycket större mängd data med samma medelvärde så kommer den klassiska skattningen att vara oförändrad, men den Bayesianska närma sig den klassiska.

Låt oss som ett andra exempel anta att vi har en ganska ovanlig händelsekod som endast inträffar i ungefär ett intervall av femtio. Låt oss anta att vi har observerat 50 intervall och bara ett av dem med en händelse av den aktuella typen. Hur kommer då de tre varianterna att bedöma ovanligheten av ytterligare en sådan händelse? Tabellen nedan visar resultaten.

	Normalappr.	Klassiskt	Bayesianskt
1	$6.525 \cdot 10^{-11}$	0.01960	0.03769

Här är det extra tydligt att normalapproximationen överdriver ovanligheten. Anledningen är som sagt att normalfördelningen är en mycket dålig approximation till Poissonfördelningen för små antal. Återigen är den Bayesianska skattningen mer försiktig än den klassiska. Den klassiska ligger nära $1/50$ vilket vid första anblicken känns rätt i detta fall, medan den Bayesianska har vägt in att det ha varit en ren slump att endast ett fall inträffade i de 50 vi tittade på.

3 Driftsdata

Den andra typen av signaler från Reginorna som vi vill kunna hantera är driftsdata, d.v.s räknare över driftstiden (eller sträckan) hos olika system. Om någon av dessa räknare plötsligt börjar öka i en större takt än den brukar, till exempel kompressorgångtiden, så kan detta vara ett tecken på servicebehov.

Först måste man bestämma sig för vilken statistisk modell som passar bäst för driftsräknarna. Man kan föreställa sig att under ett fixt tidsintervall så har en räknare en viss medelökning med en viss standardavvikelse. Om vi antar att denna ökning är normalfördelad (vilket inte är strikt sant eftersom ökningen inte kan bli negativ, men borde vara en tillräckligt god approximation i detta fall) så får man om man lägger ihop ökningen under flera sådana intervall en slumpvandring (dvs en Wienerprocess, eller Brownsk rörelse) med drift åt ena hållet. Egenskapen hos en sådan slumpvandring är att om ökningen under ett intervall av valfri längd är normalfördelad med medelvärde μ och standardavvikelse σ så är ökningen under t sådana intervall normalfördelad med medelvärde $t\mu$ och standardavvikelse $\sqrt{t}\sigma$.

Precis som för händelsedata så räknas driftsräknarna upp under drift av fordonen, varför intervallen återigen kanske inte bör räknas i dagar utan i t.ex. antal driftskilometer. Därför bör man även här kunna ta hänsyn till olika inter-

vallängd för olika fordon. Varje datapunkt består alltså av ett par av värden: ökningen a_i under intervallängden t_i.

3.1 Klassisk skattning av fördelningen

Ökningen a under det givna intervallet t är alltså normalfördelad:

$$P(a|\mu,\sigma,t) = \frac{1}{\sqrt{2\pi t}\sigma} e^{-(a-t\mu)^2/2t\sigma^2} \tag{19}$$

För att hitta maximum-likelihood-skattningen av parametrarna μ och σ multiplicerar man ihop sannolikhetsfördelningarna för alla datapunkter, deriverar resultatet med avseende på parametrarna, och sätter detta till noll. Då får man (räkningarna utelämnade) skattningarna:

$$\mu = \frac{\sum_{i=1}^{N} a_i}{\sum_{i=1}^{N} t_i} \tag{20}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (a_i - t_i\mu)^2/t_i}{N}} \tag{21}$$

Genom att sätta in dessa skattningar i fördelningen ovan tillsammans med ett nytt a och t så får man veta hur sannolikt detta nya fall är enligt modellen.

3.2 Bayesiansk beräkning av sannolikheten

För att kunna göra den Bayesianska beräkningen måste vi först ta fram en fördelning över parametrarna μ och σ. Om vi återigen antar uniforma apriorifördelningar för parametrarna så får vi:

$$P(\mu,\sigma|D) \propto P(D|\mu,\sigma) = \prod_{i=1}^{N} P(a_i|\mu,\sigma,t_i) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \frac{1}{\sqrt{\prod_i t_i}} e^{-\sum_i (a_i - t_i\mu)^2/2t_i\sigma^2} \tag{22}$$

vilket ger sannolikheten för ett nytt a och t som:

$$
\begin{aligned}
P(a|D,t) = \int_{\mu,\sigma} P(a|\mu,\sigma,t)P(\mu,\sigma|D) &= \frac{\int_{\mu,\sigma} P(a|\mu,\sigma,t)\prod_{i=1}^{N} P(a_i|\mu,\sigma,t_i)}{\int_{\mu,\sigma}\prod_{i=1}^{N} P(a_i|\mu,\sigma,t_i)} \\
&= \frac{\int_{\mu,\sigma}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{N+1}\frac{1}{\sqrt{t\prod_i t_i}}e^{-(a-t\mu)^2/2t\sigma^2-\sum_i(a_i-t_i\mu)^2/2t_i\sigma^2}}{\int_{\mu,\sigma}\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{N}\frac{1}{\sqrt{\prod_i t_i}}e^{-\sum_i(a_i-t_i\mu)^2/2t_i\sigma^2}} \\
&= \frac{\frac{1}{2}\frac{1}{\sqrt{2\pi}^N}\frac{1}{\sqrt{t\prod_i t_i}}\frac{1}{\sqrt{t+\sum_i t_i}}\Gamma(\frac{N-1}{2})/\left(\frac{a^2}{2t}+\sum_i\frac{a_i^2}{2t_i}-\frac{\left(a+\sum_i a_i\right)^2}{2t+2\sum_i t_i}\right)^{\frac{N-1}{2}}}{\frac{1}{2}\frac{1}{\sqrt{2\pi}^{N-1}}\frac{1}{\sqrt{\prod_i t_i}}\frac{1}{\sqrt{\sum_i t_i}}\Gamma(\frac{N-2}{2})/\left(\sum_i\frac{a_i^2}{2t_i}-\frac{\left(\sum_i a_i\right)^2}{2\sum_i t_i}\right)^{\frac{N-2}{2}}} \\
&= \frac{1}{\sqrt{\pi t}}\frac{\sqrt{\sum_i t_i}}{\sqrt{t+\sum_i t_i}}\frac{\Gamma(\frac{N-1}{2})}{\Gamma(\frac{N-2}{2})}\frac{\left(\sum_i\frac{a_i^2}{t_i}-\frac{\left(\sum_i a_i\right)^2}{\sum_i t_i}\right)^{\frac{N-2}{2}}}{\left(\frac{a^2}{t}+\sum_i\frac{a_i^2}{t_i}-\frac{\left(a+\sum_i a_i\right)^2}{t+\sum_i t_i}\right)^{\frac{N-1}{2}}}
\end{aligned}
\tag{23}
$$

Gammafunktionen $\Gamma(n)$ kan man om man vill undvika genom att återigen använda Sterlings approximation:

$$
\frac{\Gamma(\frac{N-1}{2})}{\Gamma(\frac{N-2}{2})} \approx \frac{\left(\frac{N-3}{2}\right)^{\frac{N-2}{2}}e^{\frac{N-4}{2}}}{\left(\frac{N-4}{2}\right)^{\frac{N-3}{2}}e^{\frac{N-3}{2}}} = \sqrt{\frac{N-3}{2}}\left(\frac{N-3}{N-4}\right)^{\frac{N-3}{2}}e^{-1/2} \tag{24}
$$

Om man är ute efter den negativa log-likelihooden så blir alltså det man ska beräkna (inklusive Sterlingapproximationen):

$$
\begin{aligned}
-\log P(x|D) \approx &\left(\frac{N-1}{2}\right)\log\left(\frac{a^2}{t}+\sum_i\frac{a_i^2}{t_i}-\frac{\left(a+\sum_i a_i\right)^2}{t+\sum_i t_i}\right) - \\
&-\left(\frac{N-2}{2}\right)\log\left(\sum_i\frac{a_i^2}{t_i}-\frac{\left(\sum_i a_i\right)^2}{\sum_i t_i}\right)+ \\
&+\left(\frac{N-3}{2}\right)\log\left(\frac{N-4}{N-3}\right)+ \\
&+\frac{1}{2}\left(1-\log\left(\frac{N-3}{2}\right)+\log(\pi t)+\log(t+\sum_i t_i)-\log(\sum_i t_i)\right)
\end{aligned}
\tag{25}
$$

3.3 Räkneexempel

För att illustrera dessa formler ska vi använda ett litet utdrag ur autentiska driftsdata från en Regina. Tabellen nedan visar huvudkompressorns gångtid under ett antal intervall vars längder är givna både i antal dagar och antal körda kilometer.

Dagar	Kilometer	Driftstid
8	10	168
7	29	15
21	144	62
28	215	76
14	118	40
11	42	24
11	74	44
7	46	21
3	18	8
1	8	3
2	11	5

Den första driftstiden är mycket högre än alla andra, utan att intervallet är ovanligt långt, varför denna siffra är uppenbart konstig (förmodligen en artefakt i datamängden snarare än en verklig gångtid, men så ser data ut). Låt oss för exemplets skull ta undan detta första samt det sista fallet (som inte är särskilt konstigt) som testfall, och träna modellerna på de övriga. Vi ska också jämföra att använda dagar respektive kilometer som intervallängd.

För den klassiska beräkningen får vi enligt formel (20) och (21) att $\mu = 2.845$ och $\sigma = 1.632$ när intervallängden är mätt i dagar, och $\mu = 0.4222$ och $\sigma = 0.7688$ när intervallängden är mätt i kilometer. Genom att sätta in dessa parametrar, samt de värden vi vill testa, i formel (19) så får vi resultaten i tabellen nedan.

För den Bayesianska skattningen använder vi istället formel (23) och sätter in värdena från tabellen ovan direkt.

	Klassiskt	Bayesianskt
Första fallet, dagar	$9.394 \cdot 10^{-217}$	$6.279 \cdot 10^{-10}$
Första fallet, km	$6.332 \cdot 10^{-987}$	$2.253 \cdot 10^{-12}$
Sista fallet, dagar	0.1653	0.1402
Sista fallet, km	0.1549	0.1310

Alla varianterna är överens om att det första fallet är konstigt medan det sista är helt normalt. Skillnaderna för det sista fallet är liten mellan metoderna, men för det första fallet ger den klassiska skattningen en extremt låg sannolikhet. Någon entydig skillnad av att räkna i dagar eller kilometer syns tyvärr inte i detta exempel, men man kunde förvänta sig mer pålitligare värden av med kilometer eftersom det avspeglar den verkliga användningen bättre än antal

Fordon	Händelse		NegLogLikelihood
3029	8408	HVAC kommunikationsfel	9316.7
9024	5002	Tågvärme: säkringsfel	8900.6
9004	8325	Fel hastighet axel 4	5605.0
1002	8330	Stor hastighetsskillnad axel 3	3213.6
9046	4041	Strömriktare: låg kylvattennivå	3185.0
3044	8322	Fel hastighet axel 1	3046.3
9052	8329	Stor hastighetsskillnad axel 2	3024.3
9049	8328	Stor hastighetsskillnad axel 1	2834.8
1054	3147	Kolskeneövervakning urkopplad	2371.1
1053	3147	Kolskeneövervakning urkopplad	2304.7

Tabell 1: De 10 händelsekoderna med mest avvikande frekvens.

dagar.

4 Tillämpningsexempel

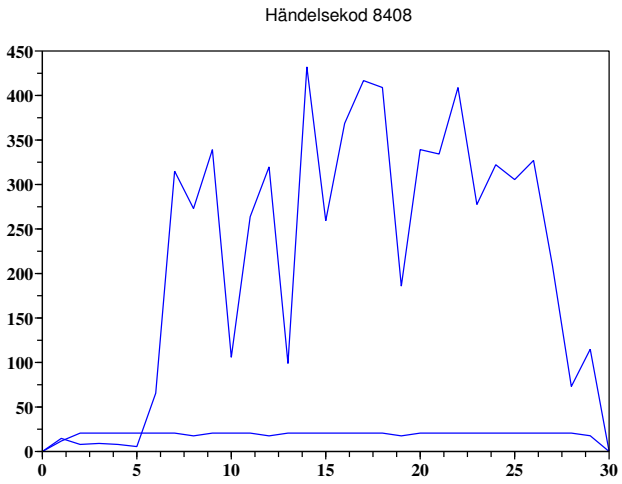
För att visa hur formlerna kan tillämpas i ett verkligt fall, använder vi återigen autentiska data från Reginatågen, både händelsedata och driftsdata. Data kommer från 57 Reginatåg (totalt 123 vagnar) och rör perioden 2003-10-01 - 2004-04-19. Antalet olika händelsekoder som hanteras på Reginorna är 1013 stycken, varav 695 stycken förekommit under den aktuella tidsperioden. Antalet driftsräknare är ursprungligen 32 stycken, varav 19 är relevanta. (De övriga är antingen oanvända eller i stort identiska med någon av de 19). Ett Reginatåg kan ha två eller tre vagnar. Det finns tre olika typer av vagnar, med lite olika profil vad gäller genererade händelser och uppdaterade driftsräknare. En två-vagnars-Regina består av en A-vagn och en B-vagn, medan en tre-vagnars-Regina dessutom har en T0-vagn i mitten.

Beroende på hur man använder formlerna kan man nu antingen ta reda på till exempel om något fordon skiljer sig från de andra sett över hela perioden, eller om ett fordon skiljer sig i tiden från hur det brukar vara, eller om ett fordon en viss vecka skiljer sig från alla fordon under alla veckor.

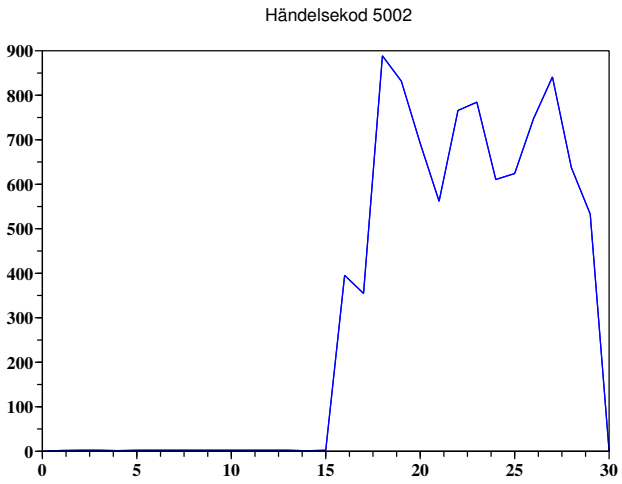
4.1 Avvikelser i händelsedata

För var och en av de 695 förekommande händelsekoderna har varje Reginavagn jämförts med alla andra vagnar av samma typ, under tidsperioden som helhet. De koder och vagnar som uppvisade störst avvikelse är listade i tabell 1.

Nu kan man titta närmare på dessa genom att jämföra en vecka i taget av den utpekade vagnen med alla andra vagnar av samma typ (och under alla veckor i tidsperioden). Då får man en tidsserie över hur den utpekade händelsekoden har varierat i "ovanlighet" över tiden. Sådana diagram för de mest avvikande händelsekoderna visas i bild 1 och 2.



Figur 1: Avvikelse per vecka av händelsekod 8408 hos fordon 3029 (hög avvikelse) och fordon 3028 (låg avvikelse).



Figur 2: Avvikelse per vecka av händelsekod 5002 hos fordon 9024.

Fordon	Ackumulator	NegLogLikelihood
1054	Aux. compressor	1057.9
1029	In traffic	902.8
1046	In traffic	720.8
1054	In traffic	337.6
1054	Door 3 open	288.2
1046	Aux. compressor	285.0
1053	In traffic	256.4
1029	Main compressor	250.5
1055	Door 3 open	242.3
9054	Door 4 open	227.3

Tabell 2: De 10 mest avvikande driftsräknarna

4.2 Avvikelser i driftsdata

Precis som för händelsedata har för var och en av de 19 driftsräknarna varje Reginavagn jämförts med alla andra vagnar av samma typ, under tidsperioden som helhet. De största avvikelserna är listade i tabell 2 och utvecklingen av den största avvikelsen i tiden visas i figur 3.

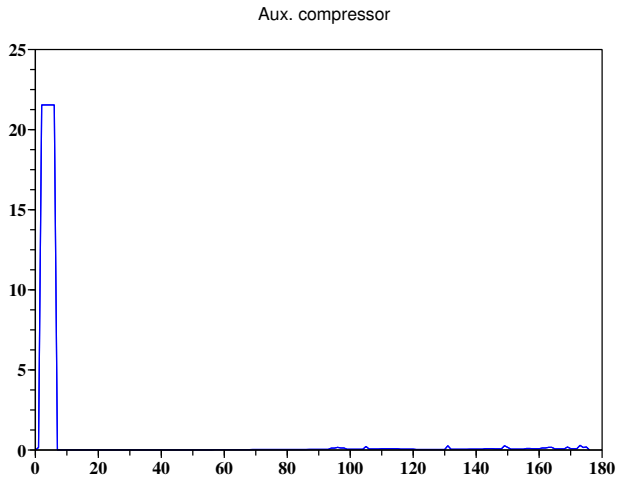
4.3 Slutsatser

Flera av de detekterade avvikelserna ovan kvarstod på fordonen under ett stort antal veckor innan de slutligen försvann, förmodligen på grund av någon serviceåtgärd. Samtidigt ser man att dessa avvikelser framstår tydligt redan den första veckan, och skulle i vissa fall förmodligen kunna detekteras redan efter ett par dagar. Användandet av denna typ av avvikelsedetektion skulle alltså kunna ge stöd till verkstaden om vad som behöver åtgärdas så att felet hittas långt tidigare än det annars skulle gjorts, och därmed förhoppningsvis gör felet enklare att åtgärda, minskar slitaget på kringliggande delar, och framför allt ökar tiden som tåget har full funktionalitet.

5 Avslutning

Ovan har vi givit de formler som behövs för att implementera avvikelsedetektion av signalerna från Reginatåg. Tillämpningsexemplet visar att de direkt kan komma till stor nytta i underhållet av Reginatåg.

Samtidigt är den behandlade problemställningen, att upptäcka avvikelser i diskreta händelseströmmar eller i räknarvärden av olika slag, långt ifrån unik för denna specifika domän. Ett exempel på det förstnämnda kan vara analys av logmeddelanden i olika sammanhang, och på det senare att detektera fel i komplexa reglersystem där ett begynnande fel i en komponent maskeras av att kringliggande komponenter kompenserar genom att arbeta hårdare. Därför



Figur 3: Avvikelsen över tiden av driftsräknare "Aux. compressor" för vagn 1054.

anser vi att de formler som tagits fram här har mycket stor användningspotential i flera olika domäner.