

De statistische analyse van reeksen tijdstippen van gebeurtenissen: een inleiding.

N.J.D. Nagelkerke en J. Oosting

Laboratorium voor Medische Fysica, Universiteit van Amsterdam.

Samenvatting

Aan de hand van een reeks waarnemingen van tijdstippen van gebeurtenissen, worden een aantal begrippen, modellen en analysemethoden besproken die gebruikt kunnen worden bij het doorlichten van een dergelijke verzameling gegevens.

Inleiding

Het komt vaak voor dat de gegevens waarmee een (adviserend) statisticus te maken krijgt, bestaan uit waarnemingen die een functie zijn van de tijd. Deze waarnemingen kunnen een 'gewone' tijdreeks vormen zoals bijvoorbeeld waargenomen temperaturen te De Bilt, al dan niet discreet bemonsterd, maar ook kunnen deze waarnemingen het karakter hebben van een reeks tijdstippen van gebeurtenissen. Deze gebeurtenissen kunnen waargenomen bliksemschichten te De Bilt zijn, fouten die optreden bij het functioneren van een apparaat, het optreden van een aritmisch verschijnsel in een elektrocardiogram, het verrichten van een dwanghandeling door een psychiatrisch patiënt of het plegen van moorden te Amsterdam. Het analyseren van dit soort gegevens behoort echter niet tot de dagelijkse bezigheden van de meeste statistici. Dit is voor een deel toe te schrijven aan het feit dat standaard-statistische pakketten (SPSS, BMDP) veel kunnen analyseren maar het laten afweten waar het dit soort data betreft.

Een moeilijkheid bij dit soort gegevens - en bij tijdreeksen in het algemeen - is dat de vraagstelling anders is dan bij 'gewone' statistiek waar 'inference' over een populatie de hoofdmoot van de theorie vormt. Bij het bestuderen van reeksen tijdstippen ligt de nadruk op het verkennende vlak waarbij men de gegevens confronteert met een aantal modellen en modeleigenschappen en gekeken wordt in hoeverre de gegevens 'kloppen' met dergelijke modellen en waar de gegevens daarvan afwijkingen vertonen.

Een dergelijke vraagstelling is gebruikelijk bij het vak dat signaalanalyse wordt genoemd.

Daar dit tot de wat minder bekende zijpaden van de statistiek behoort, leek het ons aardig om in KM iets te vertellen over de statistische analyses van reeksen gebeurtenissen waarvan de theorie onder andere te vinden is in de monografie van Cox en Lewis (1966) over dit onderwerp. We zullen bij de bespreking gebruik maken van gegevens die bestaan uit een registratie van tijdstippen waarop bepaalde gebeurtenissen (paroxysmale activiteit) voorkomen in een elektro-encefalogram van een epilepsie-patiënt.

wel niet meer gelijkwaardig en dus heeft het zin om deze apart te bestuderen. Een belangrijk begrip in de theorie van stochastische processen is stationariteit hetgeen inhoudt dat de statistische eigenschappen van het proces niet veranderen in de tijd. Niet-stationariteit kan vele vormen aannemen. Een vorm van niet-stationariteit waarmee men vaak te maken krijgt bij de bestudering van puntprocessen is het verloop in 'intensiteit' (frequentie) in de tijd. Dit kan hinderlijk zijn indien men bijvoorbeeld geïnteresseerd is in de verdeling van de interval statistics X_i . Aan de andere kant kan het ook juist de niet-stationariteit zijn waar men in geïnteresseerd is.

Een erg bekend en zeer eenvoudig puntproces is het Poisson-proces. De eenvoud van dit proces wordt veroorzaakt door het feit dat de kans op k gebeurtenissen in een bepaald interval slechts afhangt van de lengte van dat interval, onverschillig wat er buiten dat interval gebeurt is.

Bij een Poisson-proces zijn de lengten van de intervallen onderling onafhankelijk en exponentieel verdeeld. In het algemeen heet een proces waarbij de intervallen X_i onderling onafhankelijk zijn, een renewal- (vernieuwings) proces. Bij een dergelijk proces geldt in het algemeen wel dat kennis van het tijdstip van optreden van de laatste gebeurtenis informatie geeft over de kansdichtheid van optreden van een eerstvolgende gebeurtenis.

De verdelingsfunctie van X_i wordt meestal genoteerd als $F(x)$; $R(x) = 1 - F(x)$ wordt de overlevingsfunctie genoemd en $f(x) = \frac{d}{dx}F(x)$ de kansdichtheidsfunctie. Een ander belangrijk begrip is de hazard-functie of conditionele (n.l. op $X > x$) kansdichtheid $\lambda(x) = f(x)/R(x)$. De variabele x is een tijdsvariabele die op nul gezet wordt bij het optreden van een gebeurtenis. Bij een Poisson-proces is $\lambda(x) = \lambda$ een constante.

Andere 'nuttige' intervalverdelingen zijn gamma-verdelingen, lognormale en Weibull-verdelingen en mengverdelingen van het exponentiële type.

Een voor de hand liggende methode om te kijken of een proces een renewal-proces is, is het inspecteren van de $\{\hat{\rho}_j\}$, de geschatte correlatiecoëfficiënten tussen de lengten van intervallen die j intervallen uit elkaar liggen. De correlatie tussen opeenvolgende intervallen is in deze notatie dus ρ_1 enz.

Bij de bestudering van deze $\{\rho_j\}$ is het zgn. spectrum van intervallen

$$f_+(\omega) = \frac{1}{\pi} \{1 + 2 \sum_{k=1}^{\infty} \rho_k \cos(k\omega)\} \quad (\omega > 0)$$

veelal een nuttig hulpmiddel. In het bijzonder gaat het uitvoeren van een toets of $\rho_j = 0$ voor $j \geq 1$, d.w.z. het toetsen of een proces een renewal-proces is, eenvoudig via een geschat spectrum $\hat{f}(\omega)$.

Een ander spectrum dat veel gebruikt wordt, is het spectrum van het telproces $g_+(\omega)$, gedefinieerd als

$$g_+(\omega) = \frac{m}{\pi} + \frac{m}{\pi} \int_0^{\infty} \{e^{-i\omega\tau} + e^{i\omega\tau}\} m_f(\tau) d\tau \quad (\omega > 0).$$

In deze formule is $m = 1/E(x)$ en $m_f(\tau)$ stelt de zgn. renewal-dichtheid voor,

De gegevens

De volgende 169 tijdsintervallen (in minuten) tussen opeenvolgende gebeurtenissen werden genoteerd. De gegevens zijn kolomsgewijs gerangschikt.

5.00	.05	4.07	.91	8.43	.81	2.01
1.57	8.02	2.00	1.63	.60	13.63	22.42
.95	6.01	6.01	3.24	6.81	3.21	.82
2.01	1.83	12.03	.93	3.01	1.23	.73
.04	4.01	14.09	4.03	3.41	1.02	.31
3.61	4.04	.55	.11	.66	1.61	.09
2.33	5.65	8.38	10.03	.09	3.23	1.82
1.61	20.04	3.03	2.02	.91	3.22	2.04
5.01	.37	6.03	1.71	5.03	1.62	4.01
1.01	6.02	2.01	3.23	9.23	2.01	13.01
3.21	2.64	2.80	6.02	5.21	.06	12.00
3.21	.51	7.02	.04	5.01	1.19	5.00
3.61	3.43	2.01	.02	4.00	.58	2.02
1.84	1.21	8.00	3.56	.87	4.01	6.81
1.43	2.03	4.05	3.19	10.06	2.01	16.03
1.81	15.04	2.04	4.04	.03	.41	10.01
7.04	4.04	.82	3.21	6.93	1.60	3.20
14.03	6.83	2.02	4.01	7.94	10.02	.62
4.04	3.41	6.33	2.02	3.61	2.31	.40
.70	.83	6.42	5.01	3.21	4.82	2.03
6.41	15.03	6.99	4.99	8.22	6.01	.80
6.04	1.81	6.00	6.65	4.03	3.23	5.62
6.84	.11	2.03	8.01	.06	5.62	2.01
14.04	3.01	4.41	4.03	.04	10.03	.93
						1.22

Enige begrippen en modellen

Voor het beschrijven en analyseren van puntprocessen wordt gebruik gemaakt van de volgende twee manieren van voorstellen van het proces, namelijk:

- 1) de 'telproces'-voorstelling $N(t) = \#$ (gebeurtenissen vóór of op tijdstip t)
- 2) de 'interval statistics'-voorstelling $X_1, X_2, X_3, \dots$, waarbij $X_i = t_{(i)} - t_{(i-1)}$, de tijdsduur tussen de $(i-1)$ -de en i -de gebeurtenis.

Uiteraard zijn deze voorstellingen gelijkwaardig in de zin dat, indien $N(t) < k$, dan is $\sum_{i=1}^k X_i > t$.

Tweede orde-beschrijvingen (correlaties e.d.) van deze voorstellingen zijn even-

n.l. de kans om tijd τ na het optreden van een gebeurtenis nog een andere - maar niet noodzakelijk de eerstvolgende - gebeurtenis te vinden. Deze spectra kunnen (naar analogie van spectrale dichtheden van gewone tijdreeksen) worden opgevat als Fourier-getransformeerden van autocorrelatie-functies.

Het nut van deze beide spectra lijkt ons niet in de eerste plaats te liggen in eenvoudige interpreteerbaarheid, zoals bijvoorbeeld het geval is bij spectra van gewone tijdreeksen, als wel in de mogelijkheden die deze spectra bieden om te kijken (toetsen) of de gegevens door een bepaald stochastisch proces (waarvan de theoretische spectra bekend zijn) gegenereerd hadden kunnen worden.

Voor een renewal-proces is het theoretische spectrum $f_+(\omega)$ bijvoorbeeld vlak (constant) en voor een Poisson-proces geldt dit ook voor $g_+(\omega)$.

Een ander handvat waarmee men de gegevens aan kan pakken, is de variantie-tijdcurve $V(t)$ die de variantie van het aantal gebeurtenissen als functie van de lengte van een interval aangeeft. Analooq daaraan bestaat er ook een covariantie-tijdcurve.

Schatten en toetsen

Om een verzameling gegevens te confronteren met een theoretisch model waarvan een aantal grootheden (bijvoorbeeld spectra) bekend zijn, moeten we deze grootheden kunnen schatten en deze vergelijken met de theoretische waarden. Zo mogelijk moeten we toetsen of de gegevens verenigbaar zijn met het model. Enige voorzichtigheid is daarbij wel geboden. Immers, ieder model is maar een idealisering van de werkelijkheid. Aan de hand van onze gegevens zullen we illustreren wat er aan schattings- en toetsingsmogelijkheden zijn. Voor deze analyse maakten we gebruik van het (Fortran) programma SASE (auteur P.A.W. Lewis), n.l. statistical analysis of series of events, dat op een paar plaatsen in Nederland aanwezig is. Wij stellen onze versie (bron J.W. de Kwaadsteniet, VU) ter beschikking van iedere belangstellende.

De eerste analyse die het programma uitvoert, is een toets op het aanwezig zijn van een - lineaire - trend in de intensiteit. Men kan hierbij hopen dat ook andere, gecompliceerdere, niet-stationariteiten 'netto' een trend bevatten die op deze manier ontdekt kunnen worden. In onze data bleek geen trend aangetoond te kunnen worden.

Na deze toets op trend drukt het programma de waarden $ZI = t_{(i)} - t_{(i-k)}$ af voor $i = k, 2k, \dots$ en $k = 2, 6, 16, 40$. Tevens worden $\log ZI$, het middelpunt van het interval $(t_{(i-k)}, t_{(i)})$ evenals gemiddelden, standaard-afwijkingen en variatie-coëfficiënten van de individuele intervallen binnen de groepen van k intervallen berekend. Deze grootheden zijn o.a. nuttig als invoer voor een regressie-analyse met bijvoorbeeld een of andere functie van de tijd als covariabele. Verder wordt voor elke waarde van k een toets op homogeniteit van varianties verricht. Steeds

werd de hypothese van homogeniteit verworpen. De meest waarschijnlijke interpretatie - gezien ook de verdere analyses - van dit verwerpen, is dat er zwakke fluctuaties in intensiteit optreden, d.w.z. het proces is niet volledig stationair.

De volgende 'analyse' die het programma verricht, is een listing en een plot van de order statistics $\{X_{(i)}\}$ d.w.z. de op volgorde van grootte gezette interval statistics $\{X_i\}$. Tevens worden de empirische overlevingsfunctie en de logaritme daarvan afgedrukt. Eveneens worden de verwachte order statistics van een Poisson-proces, de zgn. exponentiële scores, geprint. Deze grootheden zijn nuttig om afwijkingen van een exponentiële intervalverdeling te bestuderen. Ook worden de schattingen van de eerste vier (centrale) momenten van de intervalverdeling gegeven.

Voor onze data pasten de empirische order statistics, met uitzondering van de zeer kleine waarden, uitstekend bij het verwachte (onder de Poisson-hypothese) patroon. De verhouding tussen gemiddelde (4.21) en variantie (16.05) van de intervallengten versterkte de gedachten in de richting van een Poisson-proces. Om te onderzoeken of de hypothese dat we met een Poisson-proces te maken hebben, verworpen kan worden, toetst het programma of de n gebeurtenissen de observatieduur T 'random' in stukken hakt. Hiervoor zijn verschillende toetsen ontwikkeld, gebaseerd op afstandsmaten tussen de empirische en verwachte verdelingsfunctie. Deze toetsen kunnen zowel worden toegepast op de oorspronkelijke waarnemingen X_i als op de grootheden

$$W_{(i)} = \frac{1}{T} \{X_{(1)} + X_{(2)} + \dots + X_{(i-1)} + (n+2-i)X_{(i)}\},$$

de zgn. modificatie volgens Durbin.

Geen van deze toetsen was in staat de Poisson-hypothese te verwerpen. Een andere toets evenwel, een toets die gevoelig is tegen het alternatief van renewal-processen met intervalverdelingen van het gamma-type (met index > 1), werd verworpen. Het verwerpen gebeurde echter aan de 'verkeerde kant': de waarde was groter dan verwacht i.p.v. kleiner, wat het geval zou zijn geweest bij gamma-verdelingen.

Aangezien de toets gebaseerd is op de grootheid

$$\log \bar{x} - \frac{1}{n} \sum \log x_i,$$

versterkte dit ons vermoeden dat er schommelingen in intensiteit optraden. Een volgende serie grootheden die het programma bekijkt, is de $\hat{\rho}_j$, de correlaties tussen de lengtes van intervallen, j -intervallen uit elkaar. Geen van de $\hat{\rho}_j$'s ($j = 1, 2, \dots, 83$) kwam - absoluut gezien - boven de 0.18, zodat we hier ook zonder formele toets wel durfden stellen dat we de renewal-hypothese niet konden verwerpen.

Een formele toets kan verricht worden met behulp van het intervalspectrum. Onder de renewal-hypothese geldt dat de 'ongesmoothde' schatters van het spectrum voor

waarden $\omega_k = 2\pi k/n$ ($k = 1, 2, \dots, \frac{n}{2}$) bij benadering realisaties zijn van onafhankelijke, exponentieel verdeelde grootheden met de verwachting $\frac{1}{\pi}$. Afwijkingen kunnen getoetst worden door de trendtoets en de toetsen die onderzoeken of waarnemingen een interval 'random' (homogeen) in stukken hakken, toe te passen op het spectrum. Onze data gedroegen zich - zoals te verwachten was - netjes en doorstonen de toets.

Een volgende functie die geschat wordt, is de variantie-tijdcurve $V(t)$. Van belang voor de epilepsie-gegevens is alleen dat voor een renewal-proces geldt dat $V(t)/E(N(t))$ asymptotisch gelijk is aan $C^2(x)$, waarbij $C(x)$ de variatie-coëfficiënt van de lengte van een interval is. De geschatte waarden voor deze grootheid (voor verschillende waarden van t) lagen steeds tussen 0.9 en 1.1, zodat de afwijking van een Poisson-proces in dit opzicht niet groot kan zijn.

Tot slot geeft het programma nog schattingen van de covariatie-tijdcurve, de renewal-dichtheid en het spectrum van het telproces $g_+(\omega)$. Vooral de laatste twee grootheden zijn van belang voor controle of we met een Poisson-proces te maken hebben: beide grootheden - mits uiteraard enigszins gesmoothed - zijn dan vlak (d.w.z. constant), hetgeen redelijk het geval bleek te zijn. Voor het spectrum $g_+(\omega)$ is het mogelijk om een (asymptotische) toets uit te voeren op het Poisson-zijn van het proces. Deze toets gaf geen verwerping van de Poisson-hypothese, al moeten we opmerken dat het onderscheidingsvermogen van die toets tegen specifieke alternatieven vermoedelijk laag is, gezien het zeer algemene karakter van de toets.

Na al deze analyses te hebben bestudeerd, mogen we wel concluderen dat het proces redelijk beschreven kan worden als een Poisson-proces. De toetsen die de Poisson-hypothese verwierpen, n.l. de toets op homogeniteit van varianties en de toets van Moran, wijzen er echter op dat er geringe fluctuaties in de intensiteit optreden.

Om na te gaan in hoeverre deze resultaten van dag tot dag verschillen, hebben we dezelfde analyse toegepast op een registratie van dezelfde patiënt op een andere dag. De resultaten waren opvallend gelijk. Zo was bijvoorbeeld de gemiddelde intervallengte in de eerste registratie 4.21 minuten en in de tweede registratie 4.64 minuten. Als belangrijkste verschil vonden we dat de toetsen op homogeniteit van varianties en Moran's toets niet meer significant waren.

Tot slot hebben we nog registraties van drie andere patiënten bekeken. Wat bleek: twee van deze drie patiënten bleken zeer sterke niet-stationariteiten te vertonen die zich onder andere in de vorm van significante trends uitten. Om te onderzoeken of het gedrag 'plaatselijk' enigszins Poisson-achtig was, berekenden we het gemiddelde van de geschatte variatie-coëfficiënten van de lengtes van intervallen, berekend per groepje van twee opeenvolgende intervallen. Voor een Poisson-proces zijn deze schatters homogeen verdeeld op $(0, \sqrt{2})$ en hebben dus

verwachting $\frac{1}{2\sqrt{2}}$ (er is een negatieve bias!). De gevonden waarden waren 0.65, 0.50 en 0.66. Bij de twee registraties van de eerste patiënt waren de waarden 0.64 en 0.69.

Bij de patiënt met de registratie met de waarde 0.50 als gemiddelde variatie-coëfficiënt leek de oorzaak van de kleine waarde te liggen in het ondervertegenwoordigd zijn, ten opzichte van een exponentiële verdeling, van korte intervallen. Misschien is het zo dat bij die patiënt het optreden van één gebeurtenis het optreden van een volgende gebeurtenis blokkeert (?). Deze verklaring houdt het (plaatselijk) Poisson-karakter van het optreden van paroxysmale activiteit in wezen gehandhaafd. Het prettige van een Poisson-proces is dat er een eenvoudige 'verklaring' voor het mechanisme dat het genereert, te vinden is; namelijk dat paroxysmale activiteit de reactie is van een abnormaal gedeelte van de hersenen op het optreden van bepaalde constellaties in normale achtergrondruis. Meer dan een verklaring voor de vuist weg is dit echter niet. Een meer gefundeerde hypothese valt echter op grond van een paar simpele statistische analyses niet te geven.

Referenties

- Cox, D.R. en Lewis, P.A.W. (1966) The statistical analysis of series of events. Chapman and Hall.
- Lewis, P.A.W. (1966) A computer program for the statistical analysis of series of events. IBM Systems Journal, Vol. 5, No. 4.