

SAMENVATTINGEN LEZINGEN STATISTISCHE DAG 1984

VOORSPELLEN IN VERSCHILLENDE TOEPASSINGSGEBIEDEN

door: A.C. van Eijnsbergen

Men spreekt over het *voorspellen* van een (op dat tijdstip) niet-waarneembare stochastische (multi)variabele $\underline{y}_0$ als die voorzegt wordt door een functie $\hat{\underline{y}}_0 = A(\underline{y})$ van een stochastische (multi)variabele $\underline{y}$, die wel waarneembaar is. Voorspellers kunnen onderling worden vergeleken aan de hand van de *voorspellingsfout* $\underline{y}_0 - \hat{\underline{y}}_0$.

In veel handboeken ontbreekt het voorspellen als statistische methode of de presentatie is onduidelijk of onjuist, omdat geen goed onderscheid gemaakt wordt tussen voorspeller $\hat{\underline{y}}_0$ en voorspellingsfout $\underline{y}_0 - \hat{\underline{y}}_0$.

Men kan voorspellers (partiëel) ordenen met behulp van de "mean-square error" $MSE(\hat{\underline{y}}_0) = E(\underline{y}_0 - \hat{\underline{y}}_0)(\underline{y}_0 - \hat{\underline{y}}_0)'$. In een klasse van toegelaten voorspellers kan men, indien mogelijk, de beste voorspeller (met minimale MSE) opsporen.

Als $\hat{\underline{y}}_0$ een toegelaten voorspeller is die voldoet aan de *normaal-vergelijking(en)* $E(\underline{y}_0 - \hat{\underline{y}}_0)(B(\underline{y}))' = 0$ voor alle $B(\underline{y})$ waarvoor $\hat{\underline{y}}_0 - B(\underline{y})$ ook een toegelaten voorspeller is, dan heeft $\hat{\underline{y}}_0$ minimale MSE. Het omgekeerde is ook waar als de klasse van toegelaten voorspellers een lineaire variëteit is.

Met behulp van de normaal-vergelijking kan worden afgeleid dat (bij bekendheid) de *regressiefunctie* $E(\underline{y}_0|\underline{y})$ de beste voorspeller is.

Tevens kan bewezen worden dat bij bekendheid van 1e en 2e momenten *lineaire regressie* van $\underline{y}_0$ op $\underline{y}$ de beste lineaire voorspeller levert: $\mu_0 + \Sigma_{0y} \Sigma_{yy}^{-1} (\underline{y} - \mu)$ (in een voor de hand liggende notatie).

In een aantal toepassingen heeft men bij onbekendheid van μ_0 en μ gepoogd het gebruik van schattingen voor μ_0 en μ modelmatig te motiveren. De extra eis van *zuiverheid*, $E(\underline{y}_0 - \hat{\underline{y}}_0) = 0$, wordt dan aan de lineaire voorspellers opgelegd.

Als $\mu_0 = \underline{x}_0 \beta$ en $\mu = \underline{x} \beta$ voor een onbekende β dan is de voorspellingsfout van een zuivere, lineaire voorspeller een $\begin{pmatrix} \underline{x}_0 \\ \underline{x} \end{pmatrix}$ -contrast met verwachting nul. Door alles te beperken tot zulke contrasten is existentie van μ_0 en μ niet eens noodzakelijk (vergelijk intrinsieke functies in de theorie van de geregionaliseerde variabelen).

De beste zuivere, lineaire voorspeller $A'y$ moet voldoen aan $X_0 = A'X$ (zuiverheid) en de normaal vergelijking $(\Sigma_{0y} - A'\Sigma_{yy})B = 0$ voor alle B met $B'X = 0$. Er bestaat dan een L zo, dat $\Sigma_{0y} - A'\Sigma_{yy} = LX'$. A en L moeten dus opgelost worden uit de "gerande normaal vergelijkingen"

$$(\Sigma_{0y} \ X_0) = (A' \ L) \begin{pmatrix} \Sigma_{yy} & X \\ X' & 0 \end{pmatrix}$$

Het is ook effectief om te bewijzen dat de voor de hand liggende *twee-traps oplossing*

a) μ_0 en μ via GLS-schatter van β

b) $\hat{y}_0 = \hat{\mu}_0 + \Sigma_{0y}\Sigma_{yy}^{-1}(y - \hat{\mu})$

voldoet. Deze beschouwing kan ook aangepast worden voor de formulering met contrasten.

Een derde manier is via de (duale) normaal vergelijkingen van het *regressie probleem met ontbrekende waarnemingen* y_0

$$\begin{pmatrix} y_0 \\ y \end{pmatrix} = \begin{pmatrix} x_0 \\ x \end{pmatrix} \beta + \begin{pmatrix} e_0 \\ e \end{pmatrix} \quad \text{met} \quad \text{VAR} \begin{pmatrix} e_0 \\ e \end{pmatrix} = \begin{pmatrix} \Sigma_{00} & \Sigma_{0y} \\ \Sigma_{y0} & \Sigma_{yy} \end{pmatrix}.$$

Dit regressie probleem is iteratief op te lossen met een soort EM-algoritme, hoewel een directe methode te prefereren is (bijvoorbeeld met covariantie-analyse).

Door "orthogonalisatie" van y op y_0 krijgt men

$$\begin{aligned} y_0 &= x_0 \beta + e_0 \\ y &= (x - \Sigma_{y0}\Sigma_{00}^{-1}x_0)\beta + \Sigma_{y0}\Sigma_{00}^{-1}y_0 + \underline{f} \end{aligned}$$

($\underline{f}$ ongecorreleerd met e_0).

De laatste vergelijking is vergelijkbaar met het "*mixed model*" (bijvoorbeeld bij toepassingen in de Zoötechniek). Door in de twee vergelijkingen y_0 op te vatten als mathematische parameter kunnen $\hat{y}_0$ en $\hat{\beta}$ bepaald worden met de zgn. mixed model normaal vergelijkingen.

In de tweede vergelijking kan y_0 ook opgevat worden als Bayesiaanse parameter, waarvan in de eerste vergelijking informatie over de apriori-verdeling gegeven is. Gebruik van deze apriori kennis levert dus de beste lineaire, zuivere voorspeller. Wordt de apriori-kennis niet gebruikt, d.w.z. wordt de regressie

alleen uitgevoerd met de tweede vergelijking, dan krijgen we de zgn. fixed modelnormaal vergelijkingen. De gevonden voorspeller blijkt dan de beste te zijn onder alle lineaire voorspellers waarvoor $E(\hat{y}_0 | y_0 = y_0) = y_0$.

De eis van deze *sterke soort zuiverheid* heeft offers gevraagd wat betreft de variantie. De beste lineaire zuivere voorspeller heeft een betere variantie, maar is onzuiver in sterke zin (vergelijk Ridge regression en schatters van Stein's type).

In factoranalyse bezit Bartlett's voorspeller van de factorscores deze sterke soort zuiverheid.

De statistische methode van het voorspellen lijkt ontwikkeld te zijn in verspreid liggende toepassingsgebieden met bijna gescheiden informatie circuits. De methode verdient meer aandacht, zelfs bij inleidende handboeken over statistiek. Hieronder volgen voor een aantal onderwerpen gerelateerd met voorspellen een beknopte verwijzing.

- regressie-functie, lineaire regressie
zie Rao, Linear Statistical Inference and its Applications; Ch 4, § 9
- "gerande" normaal vergelijkingen
vergelijk lineaire regressie met lineaire restricties aan de parameters
- tweetraps procedure
zie Bulmer, The Mathematical Theory of Quantitative Genetics, Oxford University Press Oxford (1980)
- voorspellen
vergelijk "missing-data" techniek met covariantie
- Best Linear Unbiased Predictor (BLUP)
zie Henderson (1976) A simple method for computing the inverse of a numerator relationshipmatrix used in prediction of breeding values, Biometrics 32, 69 - 83
- BLUE voor Bayesiaanse parameters
zie Harville (1976), Extension of the Gauss-Markov theorem to include the estimation of random effects, Annals of Statistics 4, 384 - 395
- Kriging, universal kriging
zie Matheron G., The Theory of Regionalized variables and its Applications (1971), Les Cahiers du Centre de Morphologie Mathematique, Fontainebleau
- tijdreeksen (in econometrie)
zie Theil, Principles of Econometrics (1971) Wiley, New York.

Kriens, J., K.H.T., Tilburg, Veenstra, R.H., Nederlandse Accountants Maatschap, Amsterdam.*

Steekproefsgewijze interne controle door middel van het A.O.Q.L.-systeem.

De meeste in de accountantscontrole gebruikte steekproefsystemen zijn gericht op het keuren van populaties, dan wel op het schatten van kenmerken. Zij laten buiten beschouwing of het mogelijk is aangetroffen fouten te verbeteren. In veel gevallen kan dat ook niet of heeft het weinig zin. Daarnaast bestaan er echter situaties, waarin het mogelijk en zinvol is fouten te verbeteren. Men kan dan de controle zo inrichten dat de kwaliteit van de onderzochte populaties onder alle omstandigheden voldoet aan een minimale kwaliteitseis. In deze voordracht wordt een dergelijk systeem beschreven, namelijk het A.O.Q.L.-systeem, voluit Average Outgoing Quality Limit-systeem. Het is in het bijzonder geschikt voor toepassing bij interne controles. Men kan bijvoorbeeld denken aan een bedrijf dat van leveranciers goederen ontvangt en ook (achteraf) de bijbehorende facturen. Van belang is dan de vraag of de gefactureerde aantallen goederen ook werkelijk geleverd zijn. Correctie van eventuele fouten kan in dit geval plaatsvinden door middel van creditfacturen.

De grondgedachte van het A.O.Q.L.-systeem is als volgt. Er wordt een reeks populaties (meestal subpopulaties genoemd) achtereenvolgens gekeurd door hieruit aselechte steekproeven te nemen. De aangewezen steekproefelementen worden onderzocht. Treft men weinig fouten in de steekproef aan, dan wordt de rest van de desbetreffende subpopulatie gelaten zoals deze is, in het andere geval wordt de gehele subpopulatie gecontroleerd. Alle fouten, gevonden bij steek-

* J. Kriens, Vakgroep Wiskunde en Informatica, Katholieke Hogeschool, Hogeschoollaan 225, Tilburg, tel. 013-662351.

R.H. Veenstra, Firmant Nederlandse Accountants Maatschap, Jan van Goyenkade 11, Amsterdam, tel. 020-733012.

proefsgewijze controles, respectievelijk bij volledige controles, worden verbeterd. Doelstelling van de controle is dat er na de controle gemiddeld niet teveel fouten mogen achterblijven, bijvoorbeeld niet meer dan een vooraf vastgesteld percentage of bedrag.

Teneinde de probleemstelling scherper te kunnen omschrijven, hanteren wij de volgende symbolen. Stel hierbij dat de omvang van de subpopulaties en de steekproefomvangs steeds even groot zijn en de goedkeurgrenzen eveneens dezelfde waarde bezitten. Men kan laten zien dat deze onderstellingen niet essentieel zijn.

Grootheden en symbolen

N = omvang subpopulatie

n = omvang steekproef

k = aantal fouten in steekproef

k_0 = goedkeurgrens

$k \leq k_0$: goedkeuren

$k > k_0$: afkeuren

p = fractie fouten vóór keuring

$\bar{p}$ = gemiddelde fractie fouten na keuring.

$\bar{p}$ wordt de Average Outgoing Quality (= A.O.Q.) genoemd en hangt af van N , n , k en p . Het verband tussen $\bar{p}$ en p wordt gegeven door de formule

$$\bar{p} = \text{kans goedkeuren} \times p \times \left(1 - \frac{n}{N}\right) + \text{kans afkeuren} \times 0 = p \left[k \leq k_0 \right] \times p \times \left(1 - \frac{n}{N}\right).$$

In figuur 1 is dit verband grafisch weergegeven.

De moeilijkheid is dat wij, ook bij gegeven N , n en k_0 , $\bar{p}$ niet kunnen bepalen omdat p onbekend is. Het is echter wel mogelijk het maximum $\bar{p}_m$ van $\bar{p}$ te berekenen, de zogenaamde Average Outgoing Quality Limit.

In de praktijk stelt men uiteraard eerst $\bar{p}_m$ vast en bepaalt daarna bij de gegeven waarde van N , de vereiste waarden van n en k_0 .

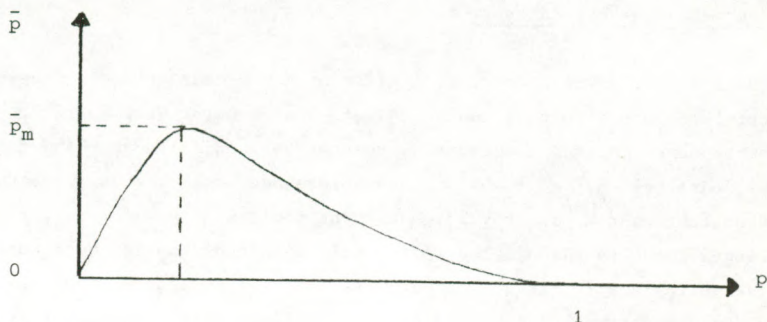


Fig. 1

Het verband tussen $\bar{p}$ en p voor gegeven waarden van N , n en k_0 . ¹⁾

Het systeem is bij Bell Telephone Laboratories ontworpen door H.F. Dodge en H.G. Romig (1). De auteurs hebben het 'vertaald' voor toepassing in de administratieve sfeer (2), waarin het onder andere toegepast wordt op de controle van inkoopfacturen bij werkmaatschappijen van het Ahold concern. Daarnaast beschikken zij over toepassingen bij onder andere controles op periodiek uitgebrachte rapporten, lonen, deelnemersbestanden van pensioenfondsen, heffingen en uitkeringen en kosten van vrachtvervoerders. Voor nadere oriëntatie wordt naar de genoemde literatuur verwezen.

Literatuur

- (1) Dodge, H.F. and Romig, H.G., Sampling Inspection Tables, John Wiley and Sons, New York, Second Edition (1959).
- (2) Veenstra, R.H. en Kriens, J., Toepassen van steekproeven in de administratie en de interne controle door middel van het A.O.Q.L.-systeem, Bedrijfskunde 54 (1982) 252-262.

1)

In deze figuur zijn langs de beide assen niet dezelfde schalen gebruikt.

Het schatten van celcyclusparameters uit de grootteverdeling van bacteriecellen in een steady-state cultuur.

Bacteriecellen zijn meestal te klein om hun groei- en deelgedrag via de lichtmicroscop direct te kunnen volgen. Daarom wordt vaak gebruik gemaakt van momentopnamen met een electronenmicroscop van een flink aantal geprepareerde (dus dode) cellen (ca. 1000) uit een zogenaamde "steady-state" cultuur. Dat is een cultuur waarin, nadat vele generaties nodig zijn geweest om initiële schommelingen kwijt te raken, elke individuele cel eigenschap een constante frequentieverdeling heeft gekregen en het totale aantal cellen exponentieel toeneemt. Bij goede experimenten kunnen statistische fluctuaties in deze frequentieverdelingen worden verwaarloosd wegens het astronomische aantal cellen en mogen de waargenomen cellen als een aselechte steekproef met teruglegging worden beschouwd.

Celgrootte is bij bacteriecellen nog een van de best meetbare eigenschappen en bovendien wetenschappelijk interessant wegens de waarschijnlijk belangrijke rol bij het delingsmechanisme. Er is daarom veel belangstelling voor wiskundige en statistische methoden waarmee men uit metingen aan verschillende cellen uit een steady-state cultuur met onbekende leeftijd op het moment van waarneming, de normale groei- en delingscyclus van een enkele cel kan reconstrueren.

Een vaak bestudeerde bacteriecel is Escherichia coli. Deze is staafvormig met ongeveer halfbolvormige uiteinden en een diameter die vrijwel niet verandert tijdens de celcyclus. De cellengte wordt daarom gevoeglijk als maat voor de celwandoppervlakte genomen. Als voorbereiding op de celdeling ontstaat een duidelijk zichtbare insnoering, vrij nauwkeurig in het midden, die steeds dieper wordt totdat tenslotte op het moment van celdeling de zogenaamde "moedercel", met lengte M , vervangen wordt door twee "babycellen", met lengten B_1 en B_2 ($B_1 + B_2 = M$; aselechte toewijzing van indices 1 en 2 aan de twee babies). Per definitie is een babycel een cel met leeftijd 0 en een moedercel een cel met leeftijd gelijk aan zijn leeftijd bij deling. De verdelingsdichtheid van cellengte M bij deling, $\phi(x)$, en van cellengte B bij geboorte, $\psi(x)$, zijn dus voorwaardelijke kansdichtheden van de lengte X van een aselekt getrokken cel, waarvan we de kansdichtheid met $\lambda(x)$ aangeven. We kunnen dus wel λ schatten uit onze steekproef van cellengten, maar niet ϕ of ψ . In de praktijk behelpt men zich door van de waargenomen diep ingesnoerde cellen beide toekomstige dochtercellen te meten. De verhouding B/M van de lengte van een willekeurige baby tot die van zijn moeder blijkt daarbij nagenoeg onafhankelijk te zijn van M . Deze

verhouding, gedefiniëerd op moedercellen, waarbij men aselekt één der twee babies kiest, wordt met P aangeduid. Zijn kansdichtheid, die ook weer een voorwaardelijke is, wordt meestal met $K(p)$ aangegeven. Als alle cellen zuiver door-midden delen, is P steeds $1/2$ en geldt blijikbaar

$$(1) \quad \psi(x) = 2\phi(2x).$$

Bij onzuivere deling en onafhankelijkheid van P en M (wat dikwijls wordt aangenomen) geldt de relatie

$$(2) \quad \psi(x) = \int_0^1 \phi(x/p)K(p)/p \, dp.$$

De kansdichtheid $\lambda(x)$ hangt behalve van $\phi(x)$ en $\psi(x)$ ook nog af van $v(x)$, de gemiddelde lengtegroeisnelheid van alle aanwezige cellen van lengte x , volgens de formule van Collins en Richmond (1962),

$$(3) \quad \lambda(x)v(x)/\mu = \int_0^x [2\psi(y) - \phi(y) - \lambda(y)] \, dy.$$

Hierin is μ de specifieke populatiegroeiconstante: het aantal cellen neemt per tijdseenheid toe met een fractie μ . Voor μ , ϕ en ψ geldt de relatie

$$(4) \quad \int_0^\infty e^{\mu t(y)} [2\psi(y) - \phi(y)] \, dy = 0,$$

waarbij $t(x)$ de tijd is die een cel nodig heeft om met de gemiddelde groeisnelheid v te groeien van lengte a (willekeurig maar gelijk voor alle cellen) naar lengte x , dus

$$(5) \quad t(x) = \int_a^x dy/v(y).$$

Uit (3) kunnen we λ oplossen. Er geldt

$$(6) \quad \lambda(x) = \frac{\mu}{v(x)} e^{-\mu t(x)} \int_0^x e^{\mu t(y)} [2\psi(y) - \phi(y)] \, dy.$$

Dus λ wordt bepaald door het trio (v, ϕ, ψ) of, bij onafhankelijkheid van P en M in moedercellen, door (v, ϕ, K) .

Het is belangrijk om op te merken dat λ bij zeer verschillende verdelingen van de groeisnelheid van cellen met dezelfde lengte x toch dezelfde μ hoort, als de gemiddelde snelheid $v(x)$ maar dezelfde is. Dus ook bij gespecificeerde hypothetische ϕ en ψ en gegeven μ (nauwkeurig meetbaar) is niet elke hypothese over de groeisnelheidsverdeling te toetsen op grond van een steekproef uit λ .

Het is betrekkelijk eenvoudig uit (3) formules af te leiden voor de momenten van λ (Voorn, 1983). Als we aannemen dat $v(x)$ evenredig is met x , in welk geval we spreken van exponentiële lengtegroei, en dat P en M onafhankelijk zijn in moeders, blijken die momenten evenredig te zijn met die van dezelfde orde van ϕ . Nu is exponentiële lengtegroei een vaak gepostuleerde groeiwijze voor *E.coli* cellen, dus dat komt goed uit. De evenredigheidsconstanten zijn eenvoudige functies van de dichtheid K en daardoor voldoende nauwkeurig te schatten uit de metingen aan diep ingesnoerde cellen (zij het niet geheel zuiver). We kunnen dus, wat dikwijls gewenst wordt, gemiddelde en variatiecoëfficiënt (cv) van ϕ schatten en wel eenvoudig op grond van die van de steekproef en zonder aannemen te hoeven maken over het type verdeling van ϕ . Bij zuivere deling ($P = 1/2$ met kans 1) geldt dan bijvoorbeeld zeer eenvoudig dat het gemiddelde van ϕ gelijk is aan dat van λ gedeeld door $\ln 2$.

We kunnen dit schattingsprobleem uitbreiden tot gemiddelde en cv van ook andere celgebeurtenissen dan alleen deling, bijvoorbeeld het begin van de insnoeringsperiode. Bij exponentiële lengtegroei blijkt ook ook in dat geval de oplossing eenvoudig te zijn. Bijvoorbeeld, als α de fractie geconstricteerde cellen in de populatie is, dan is de gemiddelde cellengte bij overgang naar de insnoeringsperiode gelijk aan die bij celdeling gedeeld door $1 + \alpha$. Ook de cv is heel simpel te schatten, maar dan hebben we wel naast gemiddelde en cv van de hele steekproef ook die van de ingesnoerde cellen nodig.

Deze eenvoudige schattingsmethoden hebben echter hun prijs. Op de eerste plaats is er een verlies aan efficiëncy. Bij vergelijking met de de methode der grootste aannemelijkheid met voor de praktijk representatieve hypothetische verdelingen voor cellengte bij deling en overgang naar de insnoeringsperiode blijken de schatters voor de gemiddelden een relatieve efficiency van rond de 80 procent te hebben en die voor de cv 's iets onder de 50 procent. Op de tweede plaats blijkt de robuustheid tegen afwijkingen van de veronderstelde exponentiële lengtegroei nogal matig te zijn. Bij vergelijking met de methode der grootste aannemelijkheid is die voor de cv schattingen zelfs vrij slecht, terwijl laatstgenoemde methode tamelijk ongevoelig blijkt voor de vereiste specificatie van het type lengteverdeling bij de celgebeurtenissen. Wie over een goed computerprogramma beschikt om de meest aannemelijke schatters uit te rekenen, doet er dus goed aan dat te gebruiken. De eenvoudige schatters kunnen daarbij dan dienen als startwaarden voor het iteratieve rekenproces.

Literatuur:

- Collins, J.F. and M.H. Richmond. 1962. Rate of growth of *Bacillus cereus* between divisions. J. Gen. Microbiol. 28: 15-33.
- Voorn, W.J. 1983. Statistics of cell size in the steady-state.

Proefschrift, Universiteit van Amsterdam.