Arbeidsnotater

STATISTISK SENTRALBYRÅ

IO 69/17 Oslo, 19. november 1969

OM BEFOLKNINGSPROGNOSER OG DERES PRESISJON

av Tore Schweder*)

INNHOLD

*) Skrevet i gruppen for personmodeller.

1. Innledning

En prognose er en usikker forutsigelse. I dette notatet skal vi ta for oss befolkningsprognoser, og prøve å analysere deres usikkerhet.

1 A. La et fenomen på tidspunktet t være karakterisert ved størrelsen X_t, $t = 0, \pm 1, \pm 2, \ldots$. Vi skal anta at $\{X_t\}$ er en stokastisk prosess. En prognose for fenomenet på det fremtidige tidspunkt T, er en prediksjon av den stokastiske variablen X_T. Se H.T. Amundsen [1], bind 2, side 198.

Det finnes flere former for slik prediksjon. Tre av de viktigste er:

(i) Å estimere en median i fordelingen til X_T.

(ii) Å estimere forventningen til X_T.

(iii) Å estimere et "prediksjonsområde" $\mathcal{D}$ med konfidenskoeffisient β d.v.s. $P_r(X_T \in \mathcal{D}) \geq \beta$.

Vi kunne kalle prediksjoner av type (i) for medianpredikasjoner og av type (ii) for forventningspredikasjoner.

Som en ser er en prognose en karakterisering av sannsynlighetsfordelingen til en stokastisk variabel.

For å lage en god prognose må man følgelig starte med å undersøke hva man a priori kan anta om den stokastiske prosessen $\{X_t\}$. Hvis den a priori innsikt fullstendig karakteriserer den stokastiske prosessen, og hvis vi kjenner X_0 (vi lager prognosen i år 0), kunne vi regne ut forventningen ξ_T eller medianen m^T for X_T. Det er imidlertid meget skjelden at man er så heldig at den innsikt en har a priori er så omfattende. I alminnelighet har vi visse observasjoner av prosessen for eksempel $X_0, X_{-1}, \ldots, X_{-n}$. På grunnlag av disse må vi estimere sannsynlighetsstrukturen for prosessen. For å få til denne estimeringen, må vi vanligvis gjøre visse tilleggsantagelser. Disse må selvsagt være så plausible som mulig. Sammen med de antagelsene vi opprinnelig formulerte, danner disse tilleggsantagelsene prognosemodellen.

1 B. La oss med prognosens feil mene forskjellen mellom prognosen og X_T. Størrelsen på denne feilen skyldes i alminnelighet tre forskjellige feilkilder:

(i) Prognosemodellen er inadekvat.

(ii) Den sannsynlighetsstrukturen prediksjonene er laget på grunnlag av er estimert, og dermed med stor sannsynlighet litt gal.

(iii) X_T er stokastisk.

Feilkilde (i) består dels i at modellen er feilspesifisert på grunn av manglende innsikt i fenomenet eller at vi må gjøre forenklede tilleggsantagelser for at vi skal klare å bruke prognosemodellen. Dels kan feilen

komme av en såkalt feed-back effekt, d.v.s. den stokastiske prosessen $\{X_t\}$ blir påvirket av en eventuell prognose, og denne påvirkningen er ikke bygget inn i modellen.

Anta nå at prognosemodellen er adekvat. Hvis den er en parametrisk modell - hvilket er det alminnelige, må de ukjente parameterne estimeres på grunnlag av observasjonene $X_0, X_{-1},\ldots, X_{-n}$. Først når dette er gjort, er sannsynlighetsstrukturen tilstrekkelig spesifisert til at prognosen kan beregnes. Feilkilde (ii) består i at parameterne er estimert og dermed sannsynligvis forskjellige fra de riktige.

Selv om prognosemodellen er adekvat og helspesifisert vil median eller forventningsprognosen nesten sikkert avvike fra de verdier som senere blir observert. X_T er jo en stokastisk variabel, og sannsynligheten er stor for at den skal anta en verdi forskjellig fra en på forhånd oppgitt verdi. Dette er feilkilde (iii).

1 C. I dette arbeidsnotatet skal vi ta for oss befolkningsprognoser basert på den såkalte matrisemodellen. Vi skal i kapittel 2 spesifisere denne modellen til en flerdimensjonal forgreningsprosess. Videre skal vi undersøke hva den vanlige befolkningsprognosen er i denne modellen, og vi skal konstruere et tilnærmet prediksjonsområde i kapittel 3. På grunnlag av dette prediksjonsområdet skal vi så vurdere hvilken virkning feilkilde (iii) har på prognosen. Vi skal med andre ord prøve å svare på spørsmålet: Hvor presis er befolkningsprognosen hvis fødsels- og dødsratene holder seg konstante gjennom hele prognoseperioden, og hvis vi var så heldige å ha helt riktige estimater for disse ratene. På grunnlag av norske befolkningsdata for 1965 har vi i kapittel 4 numerisk analysert to prognoser, en med prognoselengde 20 år og en med 40 år, for å finne svar på spørsmålet.

Sykes [3] har studert befolkningsprognosers presisjon. Han er ikke spesielt interessert i å undersøke feilkilde (iii), men ser på denne feilkilde sammen med den feilen som er en følge av at fødsels- og dødsratene forandrer seg gjennom prognosetida. Sykes foreslår tre stokastiske prognosemodeller hvori han kan uttrykke denne usikkerheten.

2. Befolkningsutvikling sett som forgreningsprosess

2 A. La oss anta at en befolkning på et tidspunkt er karakterisert ved hvor mange personer det er i hver av de r disjunkte gruppene $1, 2,\ldots, r$. Antall personer i gruppe j i år t betegner vi med X_{tj}, $j = 1,\ldots, r$, og befolkningen er på dette tidspunkt karakterisert ved vektoren

$$X_t = (X_{t1},\ldots, X_{tr}).$$

Når vi i år 0 skal lage en prognose for år T består det altså i å gi en prediksjon av den stokastiske vektorvariablen X_T.

Vi skal nevne hvilke forutsetninger man må gjøre for at $\{X_t\}$ skal være en flerdimensjonal forgreningsprosess.

La

$$e_1 = (1, 0, \ldots, 0)$$
$$e_2 = (0,1,0,\ldots,0)$$
$$\vdots$$
$$e_r = (0,\ldots, 0, 1)$$

være de r r-dimensjonale enhetsvektorene. Mengden

$$E = \{e_1,\ldots, e_r\}$$

definerer vi som tilstandsrommet for en person. Hver person er dermed karakterisert ved en vektor $Y \in E$, og $Y = e_k$ når personen er i gruppe k.

Hvis det på tidspunkt t er N_t personer i befolkningen, og det i-te av disse er i tilstand Y_i, har vi

$$X_t = \sum_{i=1}^{N_t} Y_i.$$

Befolkningen utvikler seg og noen personer skifter tilstand, noen dør og det fødes nye. En person som i år t er i tilstand e_k kan i år t + 1 gi opphav til en liten slekt. For det første gir personen opphav til seg selv med ny tilstand e_a hvis personen overlever, dessuten kan personen føde nye personer for eksempel med karakteristikker e_b, e_c,..., e_h. Den delbefolkning eller slekten som personen i tilstand e_k gir opphav til året etter er karakterisert ved den stokastiske vektorvariablen

$$Z_k = e_a + \ldots + e_n = (Z_{k1},\ldots, Z_{kr}),$$

der Z_{kj} er antall personer i tilstand e_j som det omtalte individ har gitt opphav til. Det er selvsagt mulig at $Z_{kj} = 0$, $j = 1,\ldots, r$.

Når det ikke er imigrasjon i befolkningen, hvilket vi skal anta, er befolkningen i år t + 1 nettop lik summen av slektene de enkelte individene i år t gir opphav til. La oss derfor nummerere de X_{tj} individene i tilstand e_j fra 1 til X_{tj}, og la Z_j^i karakterisere den slekta det i-te individ i tilstand e_j gir opphav til i år t + 1; $i = 1,\ldots, X_{tj}$; $j = 1,\ldots, r$.

Vi kan nå skrive

$$X_{t+1} = \sum_{j=1}^{r} \sum_{i=1}^{X_{tj}} Z_j^i. \tag{2.1}$$

2 B. Hvis de stokastiske variablene Z_j^i $j = 1,..,\ r;\ i = 1,...$ tilfredsstiller antagelsene (i),..., (iv) under, er den stokastiske prosessen $\{X_t\}$ en tidshomogen flerdimensjonal forgreningsprosess.

(i) For hver $j = 1,...,\ r$ er de stokastiske variablene $Z_j^1,\ Z_j^2,...$ identisk fordelt.

(ii) De stokastiske variablene $Z_j^i;\ j = 1,...,\ r;\ i = 1,\ 2,...$ er stokastisk uavhengige.

(iii) Fordelingen til Z_j^i avhenger ikke av t; $j = 1,...,\ r;$ $i = 1,\ 2,....$

(iv) Når $X_{t+1} = \sum_{j=1}^{r} \sum_{i=1}^{X_{tj}} Z_j^i$ er Z_j^i stokastisk uavhengig av X_t $j = 1,...,\ r;\ i = 1,\ 2,....$

Slik vi har definert variablene er forutsetning (i) automatisk oppfylt. Forutsetning (ii) er ekvivalent med at personene utvikler seg, føder nye individer og dør uavhengig av hverandre. Forutsetning (iii) er ekvivalent med at det ikke er noen forandring i befolkningens utviklingstrend eller at $\{X_t\}$ er en tidshomogen stokastisk prosess. Forutsetning (iv) går ut på at personenes utvikling i løpet av år t er uavhengig av hvorledes befolkningen var ved begynnelsen av år t.

De to antagelsene (ii) og (iv) er i mange situasjoner tvilsomme, men når det gjelder menneskebefolkninger er de nokså plausible.

Å bruke forgreningsprosesser som modell for befolkningsutvikling er nokså vanlig, se Harris [2] . Fra denne boka skal vi hente de hovedresultatene vi trenger, men først må vi gjøre noen definisjoner.

2 C. Forventningsmatrisen M er definert ved

$$M = \begin{bmatrix} EZ_1 \\ \cdots \\ \cdot \\ \cdot \\ \cdot \\ \cdots \\ EZ_r \end{bmatrix}$$

Kovariansmatrisen til Z_k, σ_k som vi skal anta eksisterer, er definert ved

$$\sigma_k = E\,(Z_k - EZ_k)'(Z_k - EZ_k) \quad k = 1,...,\ r$$

La videre

$$\xi_t = EX_t$$

og

$$S_t = E\,(X_t - \xi_t)'(X_t - \xi_t)$$

være henholdsvis forventningen og kovariansmatrisen til X_t.

Når vi antar at X_0 er kjent, har vi ved elementær betinget forventning og varians på grunnlag av (2.1)

$$(2.2) \qquad \xi_t = \xi_{t-1} M = X_0 M^t$$

$$(2.3) \qquad S_t = \sum_{n=1}^{t} \{ M'^{t-n} (\sum_{i=1}^{r} \tau_i EX_{n-1\, i}) M^{t-n} \}$$

Se Harris [2] side 37.

På grunnlag av dette uttrykk for S_t, kan man vise viktige assymptotiske resultater.

Hvis prosessen $\{X_t\}$ er positivt regulær, d.v.s. det finnes et $n > 0$ slik at M^n bare har ekte positive elementer, da er den egenverdien til M som har høyest absoluttverdi, p, reell, positiv og enkel og den tilhørende høyre egenvektor $\underset{\sim}{\mu}$ og den venstre egenvektor $\underset{\sim}{\nu}$ er reelle og positive. Vi velger $\underset{\sim}{\mu}, \underset{\sim}{\nu} : \underset{\sim}{\nu}\,\underset{\sim}{\mu} = 1$.

$$(M\underset{\sim}{\mu} = p\underset{\sim}{\mu} \quad \text{og} \quad \underset{\sim}{\nu}M = p\underset{\sim}{\nu})$$

Fra Harris [2] side 44 har vi nå at hvis $p > 1$ finnes det en stokastisk variabel W slik at

$$(2.4) \qquad \frac{X_t}{p^t} \to W \cdot \underset{\sim}{\nu} \quad \text{nesten sikkert når } t \to \infty.$$

La δ være definert ved

$$\delta = (\delta_1, \ldots, \delta_r) = X_0 \sum_{j=1}^{\infty} \frac{M^{i-1}}{p^{2j-1}} = X_0 (I - p^{-2}M)^{-1}$$

da er

$$(2.5) \qquad \lim_{t \to \infty} p^{-2t} S_t = p^{-2} (\underset{\sim}{\mu}\,\underset{\sim}{\nu})' (\sum_{i=1}^{r} \delta_i \tau_i)(\underset{\sim}{\mu}\,\underset{\sim}{\nu}).$$

Videre

$$(2.6) \qquad \lim_{t \to \infty} p^{-t} \xi_t = X_0 \underset{\sim}{\mu}\,\underset{\sim}{\nu}.$$

Se Harris [2] side 45.

Dermed får vi

$$(2.7) \qquad EW = X_0 \mu$$

$$(2.8) \quad \text{og} \qquad \text{var } W = \tau^2 = p^{-2} \underset{\sim}{\mu}' (\sum_{i=1}^{r} \delta_i \sigma_i) \underset{\sim}{\mu}.$$

(2.7) følger umiddelbart av (2.4) og (2.6). Siden $p^{-2t} S_t$ konvergerer, må den konvergere mot

$$(2.9) \qquad E [W\nu - EW\nu]'[W\nu - EW\nu] = \nu' \tau \nu = \lim_{t \to \infty} p^{-2t} S_t$$

som må eksistere. Siden $(\underset{\sim}{\mu}\ \underset{\sim}{\nu})' = \underset{\sim}{\nu}'\underset{\sim}{\mu}'$ følger (2.8) ved sammenligning av (2.5) og (2.9).

2 D. Hvis det i år 0 er et stort antall personer i hver av de r gruppene, er, som vi skal vise, X_t tilnærmet multinormalfordelt for $t = 1, 2, \ldots$.

La Z_{kt}^i være den slekta som i-te person i tilstand k på tidspunkt 0 gir opphav til på tidspunkt t; $k = 1, \ldots, r$, $i = 1, \ldots, X_{0k}$

Z_{kt}^i $i = 1, 2, \ldots, X_{0k}$ er uavhengige identisk fordelte vektorvariable, og følgelig er $\sum_{i=1}^{r} Z_{kt}^i$ tilnærmet multimormalt fordelt fordi X_{0k} er forutsatt stor og Z_{kt}^i har annenordens momentmatrise. Dermed har vi at

$$X_t = \sum_{k=1}^{r} \sum_{i=1}^{X_{0k}} Z_{kt}^i$$

må være tilnærmet multinormalt fordelt med forventning ξ_t og kovariansmatrise S_t.

Følgelig vil

(2.10) $(X_t - \xi_t)' S_t^{-1} (X_t - \xi_t)$ være tilnærmet χ^2 - fordelt med r frihetsgrader.

Når t er stor, har $W\nu$ tilnærmet samme fordeling som $p^{-t} X_t$ som tilnærmet er multinormalt fordelt med kovariansmatrise $\tau^2 \nu'\nu$. Følgelig må W tilnærmet være normalt fordelt.

3. Befolkningsprognoser basert på forgreningsprosessmodellen

3 A. La oss spesifisere vår modell slik at den passer for den norske befolkning. Hvis høyeste levealder i Norge er f.eks. ω år kunne en la $r = 2\cdot\omega$ og interpretere tilstandene slik:

en kvinne som er k år gammel er i tilstand $k + 1$; $k = 0, 1, \ldots, \omega - 1$;

en mann som er k år gammel er i tilstand $\omega + k + 1$; $k = 0, 1, \ldots, \omega - 1$.

Når all reproduksjon i befolkningen er knyttet til kvinnen, vil sannsynlighetsfordelingen til Z_k være gitt ved

$$\Pr(Z_k = e_{k+1}) = \begin{cases} p_{k-1}^K & k = 1, 2, \ldots, a, b + 1, \ldots, \omega - 1. \\ p_{k-\omega-1}^M & k = \omega + 1, \ldots, 2\omega - 1. \end{cases}$$

$$\Pr\left(Z_k = e_{k+1} + n\,e_1 + m\,e_{\omega+1}\right) = p_{k-1}^{n,m} \qquad \begin{array}{l} k = a+1,\ldots, b; \\ n = 0,\ldots, 5; \\ m + n = 0,\ldots, 5. \end{array}$$

$$\Pr\left(Z_k = n\,e_1 + m\,e_{\omega+1}\right) = q_{k-1}^{n,m} \qquad \begin{array}{l} k = a+1,\ldots, b; \\ n = 0,\ldots, 5; \\ m + n = 1,\ldots, 5. \end{array}$$

Her betyr p_j^K sannsynligheten for at en j-årig kvinne skal bli $j + 1$ år. p_j^M er den tilsvarende sannsynligheten for en j-årig mann. $p_j^{n,m}$ er sannsynligheten for at en kvinne som er i reproduktiv alder d.v.s. $j \in \{a,\ldots, b-1\}$, føder n pikebarn og m guttebarn i løpet av året og at hun selv blir $j + 1$ år. $q_j^{n,m}$ er sannsynligheten for at en j-årig kvinne skal føde n piker og m gutter i løpet av året men selv dø før året er omme.

La

$$p_j^K = \sum_{n+m=0}^{5} p_j^{n,m} \;;$$

$$r_j = \sum_{n=1}^{5} n \sum_{m=0}^{5-n} \left(p_j^{n,m} + q_j^{n,m}\right)$$

$$s_j = \sum_{m=1}^{5} m \sum_{n=0}^{5-m} \left(p_j^{n,m} + q_j^{n,m}\right)$$

være henholdsvis sannsynligheten for at en j-årig kvinne skal bli $j + 1$ år; $j = 0,\ldots, \omega - 1$, forventet antall piker og forventet antall gutter en j-årig kvinne føder i løpet av er år; $j = a,\ldots, b$.

Med denne notasjonen kan vi skrive M opp slik

$$M = \begin{bmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{bmatrix}$$

der $\omega \times \omega$ delmatrisene M_{11}, M_{12}, M_{22} er gitt ved

$$M_{11} = \begin{bmatrix}
0 & p_0^K & 0 & 0 & \cdots & \cdots & \cdots & \cdots & & & 0 \\
0 & 0 & p_1^K & 0 & \cdots & \cdots & \cdots & \cdots & & & 0 \\
\vdots & & \ddots & & & & & & & & \\
0 & 0 & \cdots & \cdots & 0 & p_{a-1}^K & 0 & 0 & \cdots & \cdots & 0 \\
r_a & 0 & \cdots & \cdots & 0 & 0 & p_a^K & 0 & \cdots & \cdots & 0 \\
\vdots & & & & & & & \ddots & & & \\
r_b & 0 & & & & & & & p_b^K & & 0 \\
0 & & & & & & & & & p_{b+1}^K & 0 \\
\vdots & & & & & & & & & & \\
0 & & & & & & & & & & p_{\omega-1}^K \\
0 & & & & & & & & & & 0
\end{bmatrix}$$

$$M_{12} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & \vdots \\ S_a & 0 & & 0 \\ \vdots & \vdots & & \vdots \\ \vdots & \vdots & & \vdots \\ S_b & 0 & & 0 \\ 0 & 0 & & \vdots \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix} \qquad M_{22} = \begin{bmatrix} 0 & p_0^M & 0 & 0 & \cdots & 0 \\ 0 & 0 & p_1^N & 0 & \cdots & 0 \\ \vdots & & \ddots & & & \vdots \\ & & & \ddots & & \\ 0 & 0 & 0 & 0 & \cdots & p_{\omega-1}^M \\ 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

Når initialbefolkningen er X_0 har vi av (2.2)

$$EX_t = \xi_t = X_0 M^t =$$

$$= X_0 \begin{bmatrix} M_{11}^t & \sum_{k=1}^{t} M_{11}^{t-k} M_{12} M_{22}^{k-1} \\ 0 & M_{22}^t \end{bmatrix}$$

Hvis en ønsker å gi en forventningsprognose for år T går dette ut på å beregne $\hat{\xi}_T$, og gi den som prognosevektoren. Når $\hat{M}$ er estimator for forventningsmatrisen M, er $\hat{\xi}_T$ gitt ved $\hat{\xi}_T = X_0 (\hat{M})^T$.

Siden X_T tilnærmet er multinormal fordelt, vil $\hat{\xi}_T$ også være en tilnærmet medianprognose.

3 B. Når X_{0i} er store, kan vi lage tilnærmet prediksjonsområde for X_T. I følge (2.10) har vi nemlig

$$\Pr\left((X_T - \xi_T) S_T^{-1} (X_T - \xi_T)' \le Z_{r,\beta} \right) \approx \beta$$

når $Z_{r,\beta}$ er β fraktilen i χ^2-fordelingen med r frihetsgrader. Følgelig er

$$(3.1) \qquad \mathcal{P}_T = \{ x \in \mathbb{Z}_+^r \mid (x - \hat{\xi}_T) S_T^{-1} (x - \hat{\xi}_T)' \le Z_{r,\beta} \}$$

et tilnærmet β prediksjonsområde for X_T. ($\mathbb{Z}_+^r$ er mengden av r dimensjonale vektorer med heltallige ikkenegative elementer.)

3 C. Man kan også utnytte (2.4) og og det faktum at W tilnærmet er normalt fordelt til å angi et assymptotisk prediksjonsintervall. Vår prosess er jo opplagt positivt regulær og $p > 1$.

Når z er $\frac{1}{2}(1 + \beta)$ fraktilen i $N(0,1)$ fordelingen, har vi

$$\Pr\left(|W - X_0 \underset{\sim}{\mu}| < \tau z \right) \approx \beta$$

Følgelig

$$\Pr\left(\lim_{t \to \infty} X_t p^{-t} \in \{ w \cdot \underset{\sim}{\nu} \mid |w - X_0 \underset{\sim}{\mu}| < \tau z \} \right) \approx \beta$$

Dette betyr at prediksjonselipsoiden $\mathcal{P}_t$ degenererer assymptotisk til et intervall $\mathcal{P}'_t$ langs vektoren $\underset{\sim}{\nu}$. Når T er stor er altså $\mathcal{P}_T$ en "tynn" elipsoide om hovedaksen

$$(3.2) \qquad \mathcal{P}'_T = \{p^T \; w \cdot \underset{\sim}{\nu} \mid |w - X_0 \underset{\sim}{\mu}| < \tau z\}.$$

For stor T, vil altså X_T ha fordeling konsentrert om dette linjestykket. $\mathcal{P}'_T$ er selvsagt ikke selv et prediksjonsområde, fordi

$$P_r(X_T \in \mathcal{P}'_T) = 0.$$

Allikevel er $\mathcal{P}'_T$ nyttig som mål for spredningen i fordelingen til X_T, og vi har valgt å kalle $\mathcal{P}'_T$ et tilnærmet assymptotisk β prediksjonsområde.

3 D. Det assymptotiske prediksjonsområdet $\mathcal{P}'_T$ er mye lettere å presentere enn $\mathcal{P}_T$, som jo er en elipsoide i de r dimensjonale rom, (i kapittel 4 har vi $r = 45$).

Konvergensen i (2,4) er avhengig av størrelsen på p. Når p er stor, f.eks. $p = 1.03$, vil det ikke være så stor forskjell på $\mathcal{P}'_T$ og $\mathcal{P}_T$ for moderat store T, f.eks. $T = 30$. Likeledes for små p og store T. I slike tilfeller kan en nøye seg med å studere $\mathcal{P}'_T$.

Når det er vesentlig forskjell på $\mathcal{P}'_T$ og $\mathcal{P}_T$ må en selvsagt ta for seg $\mathcal{P}_T$. En måte å få inntrykk av størrelsen på denne r dimensjonale elipsoiden, er å legge et aksesystem gjennom sentrum i den, og så beregne halvaksenes lengder.

Siden $\mathcal{P}_T$ assymptotisk konvergerer mot et intervall langs vektoren $\underset{\sim}{\nu}$, vil den lengste aksen i $\mathcal{P}_T$, for moderate T, tilnærmet ha samme retning som $\underset{\sim}{\nu}$. Det er interesant å undersøke "lengden" til $\mathcal{P}_T$ som vi definerer som den største avstand mellom to punkter i $\mathcal{P}_T$. For å finne et tilnærmet uttrykk for denne lengden, kan vi la en av de omtalte aksene ha samme retning som $\underset{\sim}{\nu}$. Et inntrykk av elipsoidens "tykkelse" om hovedaksen, kan vi få ved å la de andre r-1 aksene være lineært uavhengige og normale på den første (som har retning $\underset{\sim}{\nu}$).

La oss innføre vektorene

$$x_1 = (\Sigma \nu_i^2)^{-\frac{1}{2}} \underset{\sim}{\nu}$$

$$x_2 = (\nu_1^2 + \nu_r^2)^{-\frac{1}{2}} (\nu_{45}, 0, \ldots, 0, -\nu_1)$$

$$x_3 = (\nu_2^2 + \nu_r^2)^{-\frac{1}{2}} (0, \nu_{45}, 0, \ldots, 0, -\nu_2)$$

$$\vdots$$

$$x_r = (\nu_{r-1}^2 + \nu_r^2)^{-\frac{1}{2}} (0, \ldots, 0, \nu_r, -\nu_{r-1}).$$

Disse vektorene har alle lengde 1 og de har de egenskapene vi krevde av aksesystemet.

Et punkt med koordinater $c_1, \ldots, c_r$ i det nye aksesystemet, er representert ved vektoren

$$\xi_T + \sum_{i=1}^{r} c_i x_i .$$

La a_i være lengden av den halvaksen som har samme retning som x_i. På grunnlag av (3.1) er a_i gitt ved

$$(\xi_T + a_i x_i - \xi_T) S_T^{-1} (\xi_T + a_i x_i - \xi_T)' = Z_{r,\beta} ,$$

som gir

$$a_i = (x_i S_T^{-1} x_i')^{-\frac{1}{2}} \sqrt{Z_{r,\beta}} ; \quad i = 1, 2, \ldots, r.$$

3 E. Når x og y er to "befolkningsvektorer", angir $\sum_{i=1}^{r} |x_i - y_i|$ hva vi kan kalle absoluttforskjellen mellom totalbefolkningen svarende til x og y.

La

$$V = \sup_{y \in \mathcal{P}_T} \sum |y_i - \xi_{ti}|$$

V kan vi kalle største totalavstand fra prognosen.

Siden $\mathcal{P}_T$ har sin lengste akse tilnærmet i retning $\underset{\sim}{\nu}$, og siden $\underset{\sim}{\nu}$ er en positiv vektor, har vi

$$V \approx a_1 \sum_{i=1}^{r} x_{1i} = a_1 \left(\sum_{i=1}^{r} \nu_i^2 \right)^{-\frac{1}{2}} \sum_{i=1}^{r} \nu_i$$

Siden $\mathcal{P}_T$ har konfidenskoeffisient β, vil absoluttforskjellen mellom prognosfisert og realisert totalbefolkning, med sannsynlighet β være høyst V. Vi tenker her på forventningsprognosen ξ_T.

3 F. Ett summarisk mål for presisjonen til prognosen ξ_T kan en få ved å se på forholdet mellom V og forventet totalbefolkning.

$$P = V / \sum_{i=1}^{r} \xi_{Ti} .$$

Når konfidensgraden er β, vil en med sannsynlighet β høyst gjette relativt $P \cdot 100$ % galt på totalbefolkningen.

For det assymptotiske prediksjonsområdet $\mathcal{P}_T'$, får vi av (2,6) og (3,2)

$$P = \frac{T}{x_0} \frac{Z}{\underset{\sim}{\mu}}$$

som er uavhengig av T.

Den relative feilen vil altså for fast prediksjonsnivå for store T være tilnærmet konstant. Det er med andre ord ikke hva vi her har kalt det rent stokastiske element som gjør langtidsprognosene så usikre. Vi har vist at hvis sannsynlighetsstrukturen holdes konstant vil prediksjonsområdenes relative størrelse være assymptotisk konstant.

Det må altså være skift i sannsynlighetsstrukturen som er årsak til langtidsprognosenes usikkerhet. Fødsels- og dødsintensitetene forandrer seg med tida, og det er fordi det er vanskelig å forutse hvordan denne utviklingen blir at langtidsprognosene blir usikre.

4. Resultater basert på norske befolkningsdata for 1965

4 A. I den praktiske undersøkelsen har vi begrenset oss til den kvinnelige befolkning i alder 0 til 44 år. Dette regnes for å dekke den perioden da kvinnen er reproduktiv, slik at vi har fått med all (kvinnelig) reproduksjon i modellen. Grunnen til at vi har begrenset oss, er den at formelen for S_T, (2.3), er komplisert å regne ut. For å få regnet ut S_T på Byråets IBM maskin uten å bruke enormt mye tid, kunne ikke dimensjonen på matrisene være for store.

4 B. På grunnlag av norske befolkningsdata for 1965 er det enkelt å konstruere M.

Siden sannsynligheten for at en kvinne skal føde 2 eller flere piker i løpet av et år er negligibel, har vi

$$r_j \approx \sum_{m=0}^{4} (q_j^{1,m} + p_j^{1,m}) = \Pr(\text{en j-årig kvinne føder en pike i løpet av året})$$

og

$$\text{var } Z_{j1} \approx r_j (1 - r_j). \tag{4.1}$$

Vi har antatt at begivenhetene "den j,årige kvinnen lever minst ett år til" og "hun får et pikebarn i løpet av året", er stokastisk uavhengige.

Dette gir

$$\sigma_j = \begin{bmatrix} r_j(1-r_j) & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & & 0 & 0 & 0 & & 0 \\ \vdots & & & & & & & \\ 0 & 0 & & 0 & 0 & 0 & & 0 \\ 0 & 0 & & 0 & p_j^K(1-p_j^K) & 0 & & 0 \\ 0 & 0 & & 0 & 0 & 0 & & 0 \\ \vdots & & & & & & & \\ 0 & 0 & & 0 & 0 & 0 & & 0 \end{bmatrix}$$

σ_j består bare av nuller intatt diagonalelementene i linje 1 og linje j+2; $j = 0, 1, \dots, 44$. Av 3 A. følger det at $r_j = 0$ $j = 0, \dots, a$, og vi hat antatt $a = 15$. Videre har vi satt $b = 43$ og selvsagt $p_{44} = 0$.

Med antagelsen om uavhengighet og tilnærmelsen (4.1) er det blitt mulig å få et tilnærmet uttrykk for σ_j. Antagelsen om uavhengighet er nok ikke riktig, det er grunn til å tro at det er en meget liten positiv kovarians mellom $Z_{j,1}$ og $Z_{j,j+1}$. Videre er det klart at

$$\text{var } Z_{j1} > r_j (1 - r_j)$$

men det er altså gode grunner til å tro at forskjellen ikke er så stor fordi sannsynligheten for multiple fødsler er så liten.

En naturlig måte å unngå disse unøyaktighetene er selvsagt å estimere var Z_{j1} og cov $(Z_{j,1}, Z_{j,j+1})$. Disse estimatene kan en få ved å utnytte personregisteret. Hvis en ønsker mer pålitelige resultater enn de vi gir her, må en altså estimere σ_j $j = 0, \dots, 44$.

Et Fortran program er skrevet som regner ut S_t på grunnlag av gitte fødsels- og dødsrater. Byråets IBM 360/40 datamaskin bruker noe under en $\frac{1}{2}$ time på å regne ut S_{40}.

Egenvektorene $\underset{\sim}{\mu}$ og $\underset{\sim}{\nu}$ og den tilhørende egenverdien p er det nokså enkelt å regne ut fordi M har så enkel struktur. Det er skrevet et Algol program som regner ut disse størrelsene. I tabell 1 er p, $\underset{\sim}{\nu}$, EW og var W ført opp.

Som nevnt i 3.D. er prediksjonselipsoiden $\mathcal{P}_T$ nokså vanskelig å presentere. Vi har fulgt oppskriften i dette avsnittet, og regnet ut halvaksenes lengder a_i $i = 1, \dots, 45$ på grunnlag av den beregnede kovariansmatrise S_T. Resultatet er gitt i tabell 2 for $T = 20$ og $T = 40$ år og konfidensgrad $\beta = 0.75, 0.90, 0.95, 0.99$.

I tabell 3 er $\mathcal{P}_T$ angitt ved lengden a_1, d.v.s. halve lengden av $\mathcal{P}_T$ (langs $\underset{\sim}{\nu}$) for $T = 40, 50, 75$ og 100 år og konfidensgrad $\beta = 0.70, 0.95, 0.99$.

Legg merke til at for samme T og β er a_1 for $\mathcal{U}_T$ vesentlig mindre enn a_1 for $\mathcal{U}'_T$. Dette er ikke urimelig, for fordelingen til X_T har ennå langt fra degenerert. Det viser de relativt store verdiene på a_i $i = 2,\ldots, 45$ i forhold til a_1 for $\mathcal{U}_T$.

Den siste tabellen, 4, oppsummerer resultatene ved at den gir verdiene for P og V for de forskjellige prognoselengdene T og konfidensgradene β for $\mathcal{U}_T$ og $\mathcal{U}'_T$.

Vi vil som konklusjon anføre at verdiene for P er bemerkelsesverdig lave. P gir jo et slags uttrykk for hvor stor relativ prognosefeil som skyldes det rent stokastiske element, og for de data vi har behandlet viser altså denne feilen seg å være meget liten.

TABELL 1. Den største egenverdien = p = 1.01138

Den tilhørende venstre egenvektoren $\underset{\sim}{\nu}$ =

[1.00	.99	.97	.96	.95	.94	.93	.92	.91	.90
.89	.88	.87	.86	.85	.84	.83	.82	.81	.80
.79	.78	.77	.76	.75	.74	.74	.73	.72	.71
.70	.69	.69	.68	.67	.66	.65	.65	.64	.63
.62	.61	.61	.60	.59]					

EW = 33131.5 var $W' = \tau^2$ = 111870.3

TABELL 2. ψ_T angitt ved a_i i = 1,..., 45; for T = 20 og forskjellige β.

i	β = 0.75	β = 0.90	β = 0.95	β = 0.99
1	1 484	1 576	1 644	1 747
2	1 093	1 161	1 211	1 287
3	1 146	1 217	1 269	1 349
4	1 148	1 219	1 272	1 351
.	1 150	1 222	1 274	1 354
.	1 152	1 224	1 277	1 357
.	1 154	1 226	1 279	1 359
	1 156	1 228	1 281	1 361
	1 158	1 230	1 283	1 363
	1 160	1 232	1 285	1 365
	1 161	1 234	1 287	1 367
	1 163	1 235	1 289	1 369
	1 164	1 237	1 290	1 371
	1 166	1 238	1 291	1 372
	1 167	1 239	1 293	1 374
	1 168	1 240	1 294	1 375
	1 168	1 241	1 294	1 375
	1 169	1 242	1 295	1 376
	1 169	1 242	1 295	1 376
	1 170	1 244	1 297	1 378
	1 094	1 162	1 212	1 288
	1 091	1 160	1 209	1 285
	1 088	1 156	1 206	1 281
	1 089	1 157	1 207	1 282
	1 089	1 157	1 206	1 282
	1 091	1 159	1 209	1 285
	1 094	1 163	1 212	1 288
	1 094	1 162	1 212	1 288
	1 099	1 168	1 218	1 294
	1 096	1 164	1 214	1 290
	1 097	1 165	1 215	1 291
	1 099	1 167	1 217	1 294
	1 097	1 165	1 216	1 292
	1 089	1 158	1 207	1 283
	1 098	1 167	1 217	1 293
	1 096	1 164	1 214	1 290
	1 107	1 176	1 227	1 304
	1 115	1 185	1 236	1 313
	1 125	1 196	1 247	1 325
	1 107	1 177	1 227	1 304
	1 098	1 167	1 217	1 293
	1 074	1 141	1 190	1 264
	1 058	1 124	1 173	1 246
	1 025	1 089	1 136	1 207
45	1 036	1 101	1 148	1 220

TABELL 2. ψ_T angitt ved a_i for T = 40 (fortsatt)

i	β = 0.75	β = 0.90	β = 0.95	β = 0.99
1	2 184	2 320	2 419	2 571
2	1 288	1 369	1 428	1 517
3	1 357	1 441	1 503	1 597
4	1 359	1 443	1 505	1 600
5	1 360	1 445	1 507	1 602
.	1 362	1 447	1 509	1 604
.	1 364	1 449	1 511	1 606
.	1 365	1 450	1 513	1 607
	1 367	1 452	1 514	1 609
	1 368	1 453	1 516	1 611
	1 369	1 455	1 517	1 612
	1 371	1 456	1 519	1 614
	1 372	1 458	1 520	1 616
	1 374	1 459	1 522	1 617
	1 375	1 461	1 524	1 619
	1 377	1 463	1 526	1 621
	1 378	1 465	1 527	1 623
	1 380	1 467	1 530	1 625
	1 382	1 469	1 532	1 628
	1 385	1 471	1 534	1 630
	1 387	1 473	1 537	1 633
	1 389	1 476	1 539	1 635
	1 391	1 478	1 541	1 638
	1 393	1 480	1 543	1 640
	1 394	1 481	1 545	1 642
	1 396	1 483	1 547	1 644
	1 398	1 485	1 549	1 646
	1 399	1 487	1 550	1 647
	1 400	1 488	1 552	1 649
	1 402	1 489	1 553	1 650
	1 402	1 490	1 554	1 651
	1 403	1 490	1 554	1 652
	1 403	1 491	1 555	1 652
	1 403	1 490	1 554	1 652
	1 402	1 490	1 554	1 651
	1 391	1 477	1 541	1 627
	1 399	1 487	1 551	1 648
	1 397	1 484	1 548	1 645
	1 394	1 481	1 545	1 642
	1 395	1 482	1 545	1 642
	1 237	1 315	1 371	1 457
	1 231	1 308	1 364	1 449
	1 224	1 300	1 356	1 441
	1 223	1 300	1 356	1 440
45	1 221	1 298	1 353	1 438

TABELL 3. $\hat{z}_T$ angitt ved a_1 ($= \sqrt{\sum_{i=1}^{45} \nu_i^2}\ Z\ \tau\ p^T$) og p^T for forskjellige T og β.

T	p^T	a_1		
		β = 0.90	β = 0.95	β = 0.99
40	1.573	4 584	5 464	7 182
50	1.716	5 139	6 116	8 040
75	2.337	6 816	8 120	10 669
100	3.101	9 042	10 775	14 161

TABELL 4. Absolutte og relative variasjonsbredder V og P for $\hat{z}_T$ og $\hat{p}$ for forskjellige T og β.

	For $\hat{z}_T$				For $\hat{p}$			
	β = 0.90		β = 0.99		β = 0.90		β = 0,99	
T	V	P	V	P	V	P	V	P
20	10 433	0.007	11 565	0.008	30 346	0.017	47 545	0.026
40	15 358	0.008	17 020	0.009	34 000	0.017	53 225	0.026
100					59 858	0.017	93 746	0.026

5. Sammendrag

I dette arbeidsnotatet har vi sett en prognose som en prediksjon av en stokastisk variabel. En befolkningsprognose er spesielt en prediksjon av den stokastiske (vektor) variable X_T som representerer befolkningens størrelse og sammensetning på det framtidige tidspunkt T. Det spørsmålet vi har stilt er om befolkningsprognoser er usikre fordi X_T er en stokastisk variabel med stor spredning.

For å svare på spørsmålet har vi antatt at fødsels- og dødssannsynlighetene holdt seg konstant i hele prognose-tida og at de nettop var de fødsels- og dødsratene som ble beregnet på grunnlag av norske befolkningsdata for 1965. I den modellen for befolkningsutvikling som vi har antatt, har det vært mulig å beregne kovariansmatrisen til X_T og dermed et tilnærmet prediksjonsområde $\mathcal{U}_T$. $\mathcal{U}_T$ blir en elipsoide i utfallsrommet til X_T som med fastsatt sannsynlighet β (= 0.99 f.eks.) inneholder X_T: $P_r(X_T \in \mathcal{U}_T) \geq \beta$. Midtpunktet i $\mathcal{U}_T$ er nettop den vanlige befolkningsprognosen ξ_T.

Som et summarisk mål for presisjonen til prognosen ξ_T har vi regnet ut

$$P = \frac{\text{Den største forskjell mellom } \xi_T \text{ og en annen befolkningsvektor i } \mathcal{U}_T}{\text{totalbefolkningen svarende til } \xi_T}$$

Av tabell 4 ser vi at når $\beta = 0.99$, $T = 20$ er $P = 0.008$.

Den konklusjonen som må trekkes på grunnlag av utregningene er at befolkningsprognosenes usikkerhet skyldes i alt vesentlig andre faktorer enn det rent stokastiske element. Hvis en nemlig kjenner fødsels- og dødssannsynlighetene, og de holder seg konstant i hele prognosetida vil prognosens relative feil med sannsynlighet $\beta = 0.99$ høyst være 8 promille når $T = 20$ år.

Referanser

[1] H.T. Amundsen: Innføring i teoretisk statistikk; Universitetsforlaget.

[2] T.E. Harris: The Theory of Branching Processes; Springer-Verlag, Berlin, 1963.

[3] Z.M. Sykes: Some Stockastic versions of the Matrix Model for Population Dynamics; Journal of American Statistical Association. March 1969.