PERSONNUMMERERING I NORGE: LITT ANVENDT TALLTEORI OG PSYKOLOGI

ERNST S. SELMER

»Og det skjedde i de dager at det utgikk et bud fra keiser Augustus at all verden skulle innskrives i manntall.«

Meget har endret seg siden juleevangeliet ble skrevet, og det gjelder også prinsippene for folkeregistrering. Idag fører statlige byråer omfattende statistikker over befolkningen. Avanserte statistiske metoder og moderne elektroniske regnemaskiner er uunnværlige hjelpemidler i dette arbeid.

I Norge sorterer folkeregistreringen under Statistisk Sentralbyrå, i engere kretser bare omtalt som »Byrået«. Denne hendige forkortelse skal vi også adoptere her.

For å lette den maskinelle bearbeidelse av et sentralt personregister for Norge, er det meningen å overføre det til magnetbånd. Dette vil skje i forbindelse med en total *nummerering* av hele den norske befolkning (allerede i 1964). Nummeret skal knyttes til den enkelte person, uavhengig av bopel, giftemål o. l. Registeret må selvsagt stadig ajourføres på grunnlag av meldinger om fødsler, dødsfall, flyttinger osv. Det regnes med at Byråets sentrale register etter hvert kan bli til stor nytte for en rekke andre instanser, f. eks. skattemyndigheter, trygde- og pensjonskasser, Forsvaret, Norsk Rikskringkasting o. l. I vårt fremtidige mellomværende med slike institusjoner må vi regne med at vårt tildelte personnummer kommer til å spille en viktig rolle.[1]

Prinsippet for nummereringen er enkelt og velkjent: Først kommer et sekssifret tall for *fødselsdatoen*, med to sifre for dag, to for måned og to for årstall. Deretter kommer et *individualnummer* innen hver fødselsdato. For en befolkning som Norges trenges her *tre sifre*, men det skulle også holde en stund. Hittil har det vært maksimalt ca. 250 fødsler på samme dag i Norge — naturligvis våren 1946.

Først et par bemerkninger om fødselsdatoen. Forfatteren er født 11. februar 1920, og vil altså som første del av sitt personnummer få sifrene

[1] Initiativet til personnummereringen kom ikke fra Byrået, men fra næringslivets organisasjoner. Begrunnelsen var nettopp at et fast nummer for den enkelte lønnstaker ville lette næringslivets mellomværende med skattemyndigheter, trygdeordninger o. l.

110220. Fra et registreringssynspunkt er selvsagt årstallet viktigst, dernest måneden. I den allerede innførte *svenske* personnummerering har man derfor valgt å gi opplysningene i datoen i motsatt rekkefølge, altså for eksemplet ovenfor som 200211. Imidlertid tyder svenske erfaringer på at man herved har introdusert en psykologisk betinget feilkilde; folk er nå engang vant til å oppgi dag, måned og år i denne rekkefølge.

Ved at sifrene for *århundre* sløyfes, introduserer man en viss liten mulighet for forveksling. I Oslo, hvor en lokal personnummerering med tosifret individualnummer allerede har vært i bruk en tid, opplevet man f. eks. en forveksling mellom et nyfødt barn og det ikke avsluttede dødsbo etter en millionær som var født nøyaktig 100 år tidligere. Forvekslingen fikk ingen skattemessige konsekvenser for babyen, men var jo godt avisstoff.

Med *tresifret* individualnummer er det imidlertid lettere å forebygge slike forvekslinger, idet man f. eks. kan bruke visse 100-sifre for personer som er født i forrige århundre, andre sifre for fødselsdatoer i dette århundre og atter andre 100-sifre for fødsler fra år 2000. Med under 300 fødsler pr. dag er det jo nok av individualnummer til rådighet.

Med et nisifret personnummer er det nokså stor sannsynlighet for *feil*, av to helt forskjellige typer:

1° *Punchefeil.* En rekke lokale registere føres på *hullkort* som må punches, og slike kort brukes også som hjelpemiddel ved overføring av opplysningene til magnetbånd. Ingen operatør puncher feilfritt, og gjentatt punching med sammenligning er en nokså omstendelig kontrollmetode, som man om mulig vil unngå. — I samme kategori kommer *avskriftsfeil* på steder hvor opplysningene behandles rent manuelt.

2° *Gale opplysninger* fra publikum. At folk glemmer et tilfeldig tildelt individualnummer kan forklares, men som vi skal se er det påfallende hvor mange som angir sin egen fødselsdag galt.

For å knipe eventuelle forekommende feil kan man innføre ett eller flere *kontrollsifre* etter fødselsdatoen og individualnummeret. Her er det forfatteren kommer inn i billedet, idet jeg av Statistisk Sentralbyrå ble anmodet om å vurdere effektiviteten av forskjellige kontrollsystemer. Det har vært et omfattende arbeid, hvor både matematiske, maskintekniske og registreringsmessige problemer måtte koordineres, i nært samarbeid med byråsjef B. Bendiksen i Byrået. Nedenfor skal jeg vesentlig gi en kort fremstilling av den matematiske side av saken.[1]

[1] Nærmere detaljer finnes i min rapport »Kontrollsifre ved personnummerering«, del 1–3, som i stensillert form kan fåes ved henvendelse til Statistisk Sentralbyrå, Oslo.

Det meste av det hullkortutstyr som brukes i Norge er levert av IBM (International Business Machines), som også produserer forskjellige typer *kontrollutstyr* til punchingen. Metoden bygger på at man tar en *veiet tverrsum* av det tall som skal kontrolleres. Denne tverrsum reduseres så etter en viss modul til et ensifret tall, det egentlige kontrollsiffer, som plasseres umiddelbart til høyre for det siste siffer av tallet.

Det finnes utstyr som *beregner* kontrollsifferet samtidig med punchingen. I Norge vil imidlertid kontrollsifrene bli beregnet sentralt, på Byråets elektroniske regnemaskin, slik at det senere bare vil bli behov for utstyr som *kontrollerer* samtidig med punchingen, og varsler hvis kontrollsifferet ikke stemmer med resten av tallet.

Det kontrollsystem som vil bli tatt i bruk er det såkalte »*Modulus* 11«. Vekttallene i standard-modellen gjentas her i grupper på 6, etter følgende system:

Sifre i gitt tall:	...	x_8	x_7	x_6	x_5	x_4	x_3	x_2	x_1	x_0
Vekttall:	...	4	3	2	7	6	5	4	3	2

Av den veide tverrsum

$$t = \ldots + 4x_8 + 3x_7 + 2x_6 + 7x_5 + 6x_4 + 5x_3 + 4x_2 + 3x_1 + 2x_0$$

dannes så minste positive rest r ved divisjon med 11. Kontrollsifferet k er ikke selve denne rest, men dens 11-komplement:

$$k = 11 - r\,.$$

Uttrykt med det tallteoretiske kongruenssymbol $\equiv$ blir derfor[1]

$$k = 11 - r \equiv 11 - t \equiv -t \pmod{11}\,.$$

De mulige verdier for k er 0, 1, 2, ..., 9, 10. Siden k bare skal representere ett siffer, må vi *forkaste* $k = 10$, og altså også alle tall som leder til et slikt kontrollsiffer. Prinsippet er derfor ikke brukbart i systemer hvor det ikke foreligger noen valgfrihet; det ville f. eks. ikke være mulig å kontrollere bare fødselsdatoen på denne måte. Men nå skal det jo etter fødselsdatoen tilføyes et forholdsvis vilkårlig tresifret individualnummer. Den elektroniske regnemaskin passer da ved tildelingen på at det resulterende kontrollsiffer er lovlig. Under hver fødselsdato må derfor hvert 11-te individualnummer forkastes (men det er fortsatt rikelig å ta av).

---

[1] Som kjent betyr $a \equiv b \pmod{n}$ at differensen $a - b$ er delelig med n, altså at a og b gir samme (hoved)rest ved divisjon med n. Kongruenser kan adderes, subtraheres og multipliseres som vanlige ligninger. Når modulen n er et primtall, kan vi også fritt forkorte i en kongruens (med tall $\not\equiv 0$, altså med tall som ikke er multipla av modulen).

Et eldre IBM-system (brukes nå i Oslo) anvender modulen 10, med vekttallene 1, 2, 1, 2, ... Her behøver man ikke forkaste noe kontrollsiffer, men systemet er vanskeligere å behandle matematisk, siden modulen ikke (i motsetning til 11) er et *primtall*. For vårt spesielle formål har systemet videre den ulempe at det tydeligvis gir samme veide tverrsum — altså uoppdaget feil — hvis *dag og måned ombyttes* i fødselsdatoen. Dette er nemlig en hyppig forekommende feiltype.

Selv om man velger modulen 11, kan man på bestilling få *andre vekttall* enn standardserien 7, 6, 5, 4, 3, 2. Vi vil derfor i alminnelighet betegne vekttallene med v_i, idet vi samtidig innfører følgende betegnelser for sifrene i personnummeret:

(1)

Dag		Måned		År		Nummer		
d_{10}	d_1	m_{10}	m_1	a_{10}	a_1	n_{100}	n_{10}	n_1
v_9	v_8	v_7	v_6	v_5	v_4	v_3	v_2	v_1

Så må vi se nærmere på de feil som vekttallene skal kontrollere. Som nevnt er det delvis punchefeil, delvis gale oppgaver fra publikum.

Den dominerende punchefeil er *feil i ett enkelt siffer*. Hvis det korrekte siffer kalles x, det galt punchede for x', og det tilhørende vekttall v, vil feilen forbli *uoppdaget* bare hvis

$$vx \equiv vx' \pmod{11},$$

eller

$$v(x-x') \equiv 0 .$$

(I det følgende kan »mod 11« utelates.) Da modulen er et primtall, medfører dette

$$v \equiv 0 \quad \text{eller} \quad x \equiv x' .$$

Det er klart at man alltid vil velge (det ensifrede) $v \not\equiv 0$, altså $\not\equiv 0 \pmod{11}$. For enhver modul $\geqq 10$ vil videre $x \equiv x'$ medføre $x = x'$. Feil i ett siffer vil derfor *alltid oppdages* når alle vekttall er $\not\equiv 0$. (Dette argument holder ikke når modulen er < 10. F. eks. ville et system modulus 7 ikke kunne skille mellom sifrene 0 og 7, mellom 1 og 8 eller mellom 2 og 9.)

De hyppigste punchefeil i *to sifre* er ombytting og kompensasjon. — Hvis operatøren *bytter om* to (forskjellige) sifre x_i og x_j, svarende til vekttallene v_i og v_j, vil feilen forbli uoppdaget bare hvis

$$v_i x_i + v_j x_j \equiv v_i x_j + v_j x_i ,$$

eller

$$(v_i - v_j)(x_i - x_j) \equiv 0 .$$

Da vi har antatt $x_i \not\equiv x_j$, altså $x_i - x_j \not\equiv 0$, medfører dette $v_i \equiv v_j$, altså $v_i = v_j$. En ombytting av to sifre vil altså forbli uoppdaget bare hvis de to plasser svarer til samme vekttall. For IBM's standard vekttall (grupper på 6 sifre) betyr dette en avstand på 6 plasser mellom de ombyttede sifre. En slik ombytting er imidlertid uhyre usannsynlig; de aller fleste ombyttinger skjer mellom *nabo*-posisjoner.

»*Kompensasjon*« vil si at to nabosifre begge økes eller minskes med samme beløp i forhold til de korrekte verdier, f. eks. 45 istedenfor 12. (Feilen skyldes gal håndstilling i forhold til tastaturet.) Hvis de korrekte sifre er x_i og x_{i+1}, de punchede sifre $x_i + a$ og $x_{i+1} + a$, og de tilhørende vekttall v_i og v_{i+1}, vil feilen forbli uoppdaget bare hvis

$$v_i x_i + v_{i+1} x_{i+1} \equiv v_i(x_i + a) + v_{i+1}(x_{i+1} + a) ,$$

eller

$$a(v_i + v_{i+1}) \equiv 0 .$$

Da vi antar $a \not\equiv 0$, altså $a \not\equiv 0$, medfører dette $v_i + v_{i+1} \equiv 0$, altså $v_i + v_{i+1} = 11$. Kompensasjon vil derfor alltid oppdages hvis summen av to nabovekttall aldri er 11. For IBM's standard vekttall opptrer imidlertid denne sum én gang innen gruppen, som $5 + 6$.

Det forekommer også andre punchefeil i to sifre, av mer tilfeldig karakter. Punchefeil i mer enn to sifre er forholdsvis sjeldne, og vanskelige å systematisere.

Vi kommer så til den annen hovedtype av feil, nemlig *gale oppgaver* fra publikum. Det viser seg, kanskje nokså overraskende, at slike feil må ventes å forekomme to-tre ganger så hyppig som punchefeil (forutsatt erfarne punche-operatører). Og her er det at titelens »anvendt psykologi« kommer inn i billedet.

Først noen ord om erfaringsmaterialet: Vi bygget opprinnelig på en sammenligning som Byrået hadde foretatt for »en årgang av døde«, ca. 35 000 personer, mellom kirkebøkenes og dødsattestenes oppgave over fødselsdato. Materialet var imidlertid hverken tilstrekkelig stort eller helt representativt hva utfyllingen av oppgavene angår (prest — lege).

Men så gjorde Byrået et funn i Oslo kommune. Som nevnt har Oslo allerede gjennomført en lokal personnummerering, i forbindelse med en folketelling i 1960. Ved denne telling skulle publikum selv fylle ut fødselsdatoen, som så ble sammenlignet med oppgavene i folkeregisteret. Man fant og noterte i Oslo ca. 8000 uoverensstemmelser, men vi fikk bare ca. 7000 av dem. De resterende tusen var personer som også hadde skrevet navnet sitt galt, og derfor var havnet i en annen skuff!

Materialet fra Oslo er så stort at det må antas å være representativt

også i landsmålestokk. De divergerende oppgaver ble punchet på kort ved Byrået og gjennomanalysert på dets elektroniske regnemaskin.

Som ved punchefeil er den dominerende feiltype at bare *ett siffer* er galt angitt. Ombytting av nabosifre er også forholdsvis hyppig forekommende. For slike feil gjelder det samme som forklart under omtalen av punchefeil.

Ved siden av ett siffer galt er den største feilpost at *begge sifre i fødselsdagen* er galt angitt (mens måned og år er riktige). Om vi holder ombytting av sifrene utenfor, gjenstår det nesten 600 feil av denne kategori. Antallet er så stort at det berettiger et nøyere studium av »psykologien« bak feilene.

Hvis den korrekte fødselsdag har sifrene $d_{10}d_1$ og den gale $d'_{10}d'_1$, følger det av (1) at feilen forblir uoppdaget hvis

$$v_9 d_{10} + v_8 d_1 \equiv v_9 d'_{10} + v_8 d'_1 ,$$

eller

$$\frac{v_9}{v_8} \equiv \frac{d'_1 - d_1}{d_{10} - d'_{10}} . \tag{2}$$

Høyresiden er uforandret ved en ombytting av merkede og umerkede sifre; ved analysen er det selvsagt ikke nødvendig å vite hvilken oppgave som var korrekt og hvilken som var gal.

Det viser seg nå at differens 1 ved 10-skifte, altså ombyttingene $09 \leftrightarrow 10$, $19 \leftrightarrow 20$ og $29 \leftrightarrow 30$, tilsammen svarer for omtrent $\frac{1}{8}$ av alle de betraktede feil i fødselsdagen. For slike ombyttinger er

$$d_{10} - d'_{10} = \pm 1, \; d'_1 - d_1 = \pm 9$$

(tegnene følger hverandre), og de vil altså ifølge (2) *ikke* bli oppdaget hvis vi velger $v_9/v_8 \equiv 9$. Denne verdi for forholdet v_9/v_8 bør derfor absolutt unngåes.

Det er også andre, om ikke så markerte psykologiske »fallgruber« blant de forekommende feil i fødselsdagen. Den beste oversikt får man ved å lage en tabell over antall uoppdagede feil for de forskjellige verdier av forholdet v_9/v_8 (mod 11). Som ventet er forholdet 9 desidert det ugunstigste (123 feil ialt), mens det på den annen side viser seg at forholdet $v_9/v_8 \equiv 2$ er det beste valg (38 feil, hvorav ingen markerte fallgruber). Dette gjelder i hvert fall i Oslo-materialet, men avstanden til nest beste valg (48 feil) er såpass stor at man kanskje kan vente samme tendens i landsmålestokk.

På samme måte kan man behandle de nokså vanlige feil med begge sifre gale i måneden eller i årstallet. Når det gjelder feil i måned, domine-

res disse av ombyttingen september ↔ oktober, altså 09 ↔ 10. Her kommer nok også fonetikken inn i billedet; det er lett å høre feil på »niende« og »tiende«. Også ved årstall er differens 1 ved 10-skifte en alminnelig feil.

I forbindelse med årstallet dukket det opp en psykologisk kuriositet. Av (1) følger at det er verdien av forholdet v_5/v_4 (mod 11) som er avgjørende. Av visse grunner kunne man vente at $v_5/v_4 \equiv 2$ ville være et meget godt valg, men dette slo ikke til i Oslo-materialet. En nærmere undersøkelse ga forklaringen: Dette forhold tar ikke ombyttingen $00 \leftrightarrow 19$, som viste seg å være ganske vanlig. Grunnen er at »år nittenhundre« blir til årstallet 19.

Ved å ta hensyn til slike psykologiske tendenser i Oslo-materialet, kan man *»skreddersy«* et sett vekttall som gir færre uoppdagede feil enn IBM's standard vekttall. Uansett hvilket system man bruker, viser det seg imidlertid at antall uoppdagede feil ved bare ett kontrollsiffer vil bli nokså stort, anslagsvis i nærheten av 0,5 promille av samtlige registreringer. For å øke sikkerheten er det derfor bestemt at man i Norge skal bruke *to kontrollsifre*; derved regner man med å komme ned i omtrent én uoppdaget feil (punche- eller oppgavefeil) pr. 100 000 registreringer.

Av mange grunner er det naturlig å bruke *samme modul* 11 ved begge kontrollsifre, men vekttallene kan godt være forskjellige. Vi tenker oss derfor (1) utvidet med et nytt sett vekttall w_i:

$$
\begin{array}{cccccccccccc}
 & d_{10} & d_1 & m_{10} & m_1 & a_{10} & a_1 & n_{100} & n_{10} & n_1 & k_v & k_w \\
(3) & v_9 & v_8 & v_7 & v_6 & v_5 & v_4 & v_3 & v_2 & v_1 & & \\
 & w_9 & w_8 & w_7 & w_6 & w_5 & w_4 & w_3 & w_2 & w_1 & w_0 &
\end{array}
$$

Ved hjelp av vekttallet w_0 kan annet kontrollsiffer k_w også kontrollere første kontrollsiffer k_v.

Vi skal også se litt på den matematiske behandling av to sett vekttall, og begynner med *feil på to vilkårlige plasser* i og j. Med de tidligere betegnelser vil slike feil forbli uoppdaget hvis

$$v_i x_i + v_j x_j \equiv v_i x_i' + v_j x_j' ,$$

og tilsvarende for w, altså

$$
\begin{aligned}
v_i(x_i - x_i') + v_j(x_j - x_j') &\equiv 0 \\
w_i(x_i - x_i') + w_j(x_j - x_j') &\equiv 0 .
\end{aligned}
$$

Da vi antar at $x_i - x_i' \text{ og } x_j - x_j' \not\equiv 0$, medfører disse homogene kongruenser at

$$\begin{vmatrix} v_i & v_j \\ w_i & w_j \end{vmatrix} \equiv 0, \text{ eller } \frac{v_i}{w_i} \equiv \frac{v_j}{w_j} .$$

Feil på to vilkårlige plasser vil derfor alltid tas med to kontrollsifre om alle forhold v_i/w_i er inkongruente (mod 11). Hvis derimot $v_i/w_i \equiv v_j/w_j$ for et visst valg av i og j, vil samtlige feil i x_i og x_j som passerte uoppdaget ved første kontrollsiffer *også* passere uoppdaget ved annet siffer.

Til slutt skal vi se på en av de store psykologiske fallgruber, nemlig *ombytting av fødselsdag og måned* (mens ombyttinger dag ↔ år og måned ↔ år ikke forekommer så hyppig). Av (3) ser vi at en slik feil vil passere uoppdaget hvis

$$v_9 d_{10} + v_8 d_1 + v_7 m_{10} + v_6 m_1 \equiv v_9 m_{10} + v_8 m_1 + v_7 d_{10} + v_6 d_1 ,$$

og tilsvarende for w, altså

$$\begin{aligned} (v_9 - v_7)(d_{10} - m_{10}) + (v_8 - v_6)(d_1 - m_1) &\equiv 0 \\ (w_9 - w_7)(d_{10} - m_{10}) + (w_8 - w_6)(d_1 - m_1) &\equiv 0 . \end{aligned}$$

Som ovenfor finner vi derfor at alle ombyttinger av dag og måned vil tas med to kontrollsifre hvis

$$\begin{vmatrix} v_9 - v_7 & v_8 - v_6 \\ w_9 - w_7 & w_8 - w_6 \end{vmatrix} \not\equiv 0 .$$

Til slutt skal vi kort beskrive den kombinasjon av kontrollsifre som ble valgt, etter omfattende undersøkelser og vurderinger.

Man kunne tenke seg å bruke IBM's standard vekttall for begge kontrollsifre, eller samme »skreddersydde« vekttal for begge. Det er også mulig å bruke ett standard og ett skreddersydd kontrollsiffer, i den ene eller annen rekkefølge, og det var en av de siste løsninger som ble valgt.

Første kontrollsiffer k_v vil bli skreddersydd for å motvirke punchefeil og psykologisk betingede oppgavefeil. *Annet* kontrollsiffer k_w blir derimot av standard IBM type. Dette kontrollsiffer skal da brukes »i felten«, ved hjelp av umodifisert kontrollutstyr, og samtidig får man også en kontroll av sifferet k_v. De aller fleste feil vil tas ved denne kontroll.

Kontrollsifferet k_v brukes bare hver gang opplysningene kommer inn til Byrået for ajourføring av det sentrale personregister (på magnetbånd). Disse opplysninger må da likevel passere den elektroniske regnemaskin, og for denne er det like lett å kontrollere et skreddersydd som et standard vekttallsystem. Praktisk talt alle feil som passerte den første kontroll vil da bli oppdaget.

Det kan tenkes at visse institusjoner for internt bruk helst vil nøye seg med ett kontrollsiffer, altså utelate k_w. En slik institusjon må da spesialbestille utstyr til bruk for det skreddersydde kontrollsiffer k_v, men får til gjengjeld et for formålet bedre system enn standard-vekttallene gir.

For å tilfredsstille nysgjerrige lesere skal jeg angi de skreddersydde vekttall v_i:

v_9	v_8	v_7	v_6	v_5	v_4	v_3	v_2	v_1
3	7	6	1	8	9	4	5	2

Det er klart at et slikt system er bestemt (mod 11) bare på en proporsjonalitetsfaktor nær.

Hver norsk borger skal altså i nær fremtid tildeles et 11-sifret kjenningsnummer. Det kan virke langt, men nummeret vil av praktiske grunner bli delt i to deler, en sekssifret for fødselsdatoen og en femsifret for individualnummer og kontrollsifre. Sin egen fødselsdato bør man jo kunne huske (selv om Oslo-materialet viser at det kan by på visse problemer), og resten av nummeret er tross alt ikke lenger enn et gjennomsnittlig norsk telefonnummer.