Arbeidsnotater

STATISTISK SENTRALBYRÅ

IO 63/2
Ref.: BB/EL, 30/10-63

Oslo, 5. desember 1963

VALG AV PERSONNUMMERSYSTEM

Av

Bjørnulf Bendiksen

Med vedlegg av Prof., dr.philos. Ernst S. Selmer

1. Bakgrunnen for Byråets arbeid med saken

Byrået har fått i oppdrag å gjennomføre en ordning med fast personløpenummer bygd på fødselsdato. Bakgrunnen for dette oppdraget er at stadig flere offentlige registre i den senere tid har tatt nummerordninger i bruk for de personer som registrene omfatter. Registrene krever da oftest at numrene påføres de oppgaver som går inn til registrene, og for oppgavegivere, f.eks. bedrifter, er det en stor ulempe at de forskjellige registre bruker ulike løpenummer for samme persom. Før en går over til å behandle selve nummeret kan det være grunn til å se litt nærmere på hvorfor en rekke registre har gått over til nummerordninger.

I praksis viser det seg som regel at registerhold blir dyrt og at det krever et stort manuelt arbeide. Dels skyldes dette at registrene må legge et stort arbeid i å få inn de opplysningene som skal inn i registeret, dels skyldes det at det er vanskelig å få mekanisert registerholdet i særlig grad, men dels skyldes det også at identifikasjonskjennetegnene svikter. Svikten i identifikasjonskjennetegn kan være av flere slag:

1. For det første kan det hende at identifikasjonene ikke i tilstrekkelig grad er skillende og entydige (f.eks. kan flere personer bære samme navn).
2. For det annet kan det hende at identifikasjonsoppgavene endrer seg over tiden (adresseendringer eller navneendringer, det siste spesielt for kvinner ved ekteskapsinngåelse).
3. For det tredje kan det hende at det på grunn av feil ikke er samsvar mellom identifikasjonsopplysningene i to registre eller i melding til registeret og det som står i selve registeret.

Personnummerets oppgave er å lette identifiseringen, og ved at det er felles for flere registre, også fremme samarbeidet mellom registrene og muliggjøre en økt mekanisering. I samsvar med det som nettopp er nevnt må personnummeret da fylle følgende krav:

1. Det må entydig kjennetegne den enkelte person slik at ikke to personer har samme personnummer.
2. Det må være fast over tiden.
3. Om mulig bør personnummeret lages slik at eventuelle oppgavefeil i personnummeret lett oppdages.

2. Rekkefølgen av sifrene for fødselsdato

I oppdraget til Byrået er det fastlagt at det nye personnummeret skal bygge på fødselsdato. Det første spørsmålet som reiser seg er da i hvilken rekkefølge en normalt skal bruke sifrene i fødselsdato. Det er to alternativer som det i Norge kan bli naturlig å velge mellom, enten fødselsdag -måned -år, eller fødselsår -måned -dag. Det første av de nevnte alternativer er det som i dag blir mest alminnelig brukt, og som derfor faller naturligst for publikum. Av denne grunn bør en følge dette alternativet. Det har vært innvendt mot alternativet at for mange arkiveringsformål vil det være naturlig å ha fødselsår først. Det vil imidlertid ikke være noe i veien for å arkivere oppgaver etter fødselsåret selv om personnummeret starter med fødselsdag.

I overgangstiden vil det av praktiske grunner forekomme at det i offentlige registre brukes rekkefølgen år - måned - dag. Således er dette tilfelle i de sivile registre for fødsler og ekteskapsinngåelse og i kirkebøkenes avdeling for døde. Så lenge en imidlertid er klar over at oppgavene her kommer inn i "omvendt rekkefølge" vil de kunne behandles riktig, selv om det i og for seg er noe mere tungvint, og at rekkefølgen av denne grunn bør endres.

3. Identifikasjonen innen fødselsdag

Det er i Norge født opp til ca. 250 personer på en og samme dag (april 1946) og fødselsdag alene er selvsagt på ingen måte tilstrekkelig til entydig identifikasjon. Innen fødselsdagen må den enkelte gis kjennetegn som entydig identifiserer vedkommende. Kjennetegnet må være fast, uforanderlig over tiden og f.eks. ikke skifte ved skifte av bopel.

Identifikasjonen innen fødselsdag kan skje på flere måter. Fellesordningen for Tariffestet Pensjon har som midlertidig ordning brukt første bokstav i for- og etternavn som identifikasjon, Oslo og Bergen ligningskontorer bruker begge en tosifret nummerering.

Forbokstavene i navnet har den fordel at de umiddelbart framgår av navnet, men systemet har to åpenbare svakheter: Identifikasjonen blir ikke entydig for så vidt som det innen de aller fleste fødselsdager vil forekomme en rekke personer som har de samme forbokstaver i navnet. Navnet kan dessuten endre seg over tiden. Dette systemet fyller altså ingen av de to første kravene vi stilte til et godt identifikasjonskjennetegn. Bokstavene fra navnet er hverken tilstrekkelig skillende eller faste.

Den nummerering som ligningskontorene i Oslo og Bergen har foretatt, er på sin side både tilstrekkelig skillende og fast så lenge vi ser på den enkelte

kommune som en enhet, men den gir ikke tilstrekkelige identifikasjoner for landsomfattende registre. Også for mange oppgavegivere, f.eks. bedrifter som har tilsatte som flytter fra en kommune til en annen, er ulempen ved skifte av personnummer fremdeles tilstede.

Det er ikke mulig å finne faste kjennetegn som fra før er knyttet til den enkelte person og som er slike at de til registreringsformål kan brukes til entydig identifikasjon sammen med fødselsdag. Det må derfor skje en systematisk tildeling av kjennetegn på samme måte som den som er foretatt i Oslo og Bergen, men denne gang på landsomfattende basis.

En naturlig løsning synes å være at en som kjennetegn bruker siffer. Om en foretar en tilsvarende tildeling av bokstaver, vil en i Norge riktignok kunne klare seg med to bokstaver istedenfor tre sifre. Umiddelbart skulle en kanskje tro at det ville bli mindre oppgavefeil om en ikke får så lange sifferserier, men i stedet en kombinasjon av bokstaver og siffer. Undersøkelser som har vært foretatt i andre land, tyder imidlertid på at dette ikke er tilfelle hvis ikke bokstavene på en logisk måte kan knyttes til det som skal kjennetegnes. I tillegg kommer det moment at bokstaver hullkortmessig er mer brysomme å ha med å gjøre enn siffer, og at bokstavene representeres på ulik måte i forskjellige maskinsystemer. Det fornuftigste vil derfor trolig være å tildele identifikasjonskjennetegnene innen fødselsdag i form av sifre. Som nevnt vil det da trenges tre sifre i tillegg til fødselsdagen for identifikasjon. Disse tre sifre er i det følgende kalt individualsifre.

Om en nøyer seg med å la fødselsåret representeres ved to sifre, kan det forekomme personer med samme personnummer men som er født i hvert sitt hundreår. Så lenge en bare holder seg til levende personer, vil det være få personer det dreier seg om, og problemet vil kunne løses relativt enkelt ved at en tildeler nye individualsifre til personer som nærmer seg hundre år. En annen løsning vil være å tildele personer født i de forskjellige århundrer ulike nummerserier. Personer født i 1800-årene kan da f.eks. tildeles individualsifre fra 100-299, mens personer født i år 1900 eller senere tildeles nummer i serien 300 og utover. Ved år 2000 vil en igjen kunne ta i bruk en ny serie med nummer. Dette siste alternativet er trolig å foretrekke, fordi en person også i lang tid etter sin død vil kunne beholde en identifikasjon som ikke deles med en annen. I visse sammenhenger kan det være en fordel.

Det har fra flere holdt vært reist spørsmål om ikke personidentifikasjonen bør lages slik at den ved hjelp av personnummeret kan skille på kjønnene. Om en lar menns nummer f.eks. ende på ulike siffer og kvinner på like siffer, vil en få et greitt skille som i liten utstrekning legger beslag på sifre.

4. Bør det utarbeides kontrollsifre ?

Fødselsdag sammen med det som ovenfor ble kalt individualsifret gir en entydig identifikasjon for den enkelte person, samtidig som vi får kjennetegn som ikke endrer seg over tiden. Dette er de to første krav vi stilte til identifikasjonskjennetegnet. Når det gjelder å sikre seg mest mulig mot å få oppgitt gale identifikasjonskjennetegn, kan en komme svært langt ved hjelp av kontrollsifre. Et kontrollsiffer dannes ved en matematisk funksjon av de øvrige sifre i personnummeret. Eksempler på slike funksjoner vil en finne i de utredninger professor Selmer har gitt, se særlig hans utredning Del 1, side 11 - 12. Hvis en i en melding får oppgitt et personnummer hvor kontrollsifferet ikke er i samsvar med de øvrige sifre, kan en med sikkerhet si at det må være feil, enten i kontrollsifferet og/eller i de øvrige sifre i personnummeret. På den annen side gir samsvar mellom personnummer og kontrollsiffer ingen garanti for at oppgavene er riktige. Det kan nemlig hende at det forekommer flere feil som under utregningen av kontrollsiffer opphever hverandre. Har en flere enn ett kontrollsiffer minker sjansen for at feil på denne måten slipper gjennom, men når det gjelder personnummeret vil det i praksis trolig bare være aktuelt å bruke enten ett eller to kontrollsifre. I professor Selmers utredning er det gjort rede for hva slags feil som ikke vil bli oppdaget under forskjellige alternativer, og det vil senere i dette notat bli gjort rede for hva det kan bety i praksis.

5. Lengden av personnummeret

Med seks sifre for spesifikasjon av fødselsdag, tre sifre til individspesifikasjon og ett eller to kontrollsifre, vil den lengden av personnummeret bli på ti eller elleve sifre. Til sammenlikning kan en nevne at det personnummeret Oslo kommune bruker i dag er på ni sifre.

En tallserie på ti eller elleve sifre virker skremmende lang om en betrakter det hele som ett tall. I alle tilfelle bør et så langt tall, blant annet for å minske tallet på oppgavefeil, splittes opp i minst to grupper. I dette spesielle tilfellet faller en slik oppsplitting så meget mer naturlig som de første seks sifre har sin selvstendige betydning. På skjemaer hvor personnummeret skal brukes, bør en derfor i alminnelighet spesifisere fødselsdag på skjemaet, og så, under en egen rubrikk, de resterende sifre. Sikkerheten vil øke om en lar to forskjellige personer fylle ut henholdsvis fødselsdag og de resterende sifre, så en ikke risikerer å få skrevet av fødselsdag og individualsifret for feil person. Hvor det gjelder et skjema som publikum selv skal fylle

ut, og som skal passere en offentlig myndighet, kan det således være praktisk å la publikum selv fylle ut fødselsdag, mens den offentlige myndighet, f.eks. folkeregisteret, fører på identifikasjonssifrene med kontrollsifferet før skjemaet sendes videre. Istedenfor et skjema som utfylt vil se omtrent slik ut:

Personnummer
12031131557

vil en altså i stedet få rubrikker av typen:

Fødselsdato	Personidentifikasjon
12/3-11	315 57

Det vil ikke være nødvendig å fylle ut fødselsdato på en bestemt måte, f.eks. vil det være helt tilfredsstillende om det skrives månedens navn istedenfor nummer (oktober istedenfor 10). Likevel bør skjemaet være slik at publikum naturlig skriver dag først og år sist. Dette kan gjøres enten ved at det skrives dag, måned, år i rubrikkhodet, eller ved at skillestrek mellom dag og måned og mellom måned og år trykkes på en slik måte i skjemaene at rekkefølgen av denne grunn blir den riktige.

6. Ett eller to kontrollsifre

Ved utarbeidingen av personnumrene står en overfor et vanskelig valg med hensyn til om en skal utarbeide personnummeret med ett eller med to kontrollsifre. Valget avhenger av hvilke formål personnummeret skal brukes til og hvilke sikkerhetsmarginer en vil kreve. Allerede ved ett kontrollsiffer tas den alt overveiende del av feilene (98-99 prosent), men ett kontrollsiffer til virker nesten like effektivt på de resterende feil som første kontrollsiffer på de opprinnelige feil. Institusjoner som har greie manuelle eller maskinelle kontrollmuligheter ved siden av de som kontrollsifferet byr på, vil muligens finne det tilstrekkelig at det brukes ett kontrollsiffer. På en annen side synes det å være mange oppgaver hvor den sikkerhet ett kontrollsiffer alene byr på er for liten. De ulemper og kostnader som feilregistreringer fører med seg vil ofte være så store at det vil være riktig å bruke to kontrollsifre og derved eliminere flest mulig feil.

Når en tar i betraktning den usikkerhet som vil rå med hensyn til den framtidige bruk av personnummeret, kan det være mye som taler for at en utarbeider personnummeret med to kontrollsifre, men slik at første kontrollsiffer eventuelt kan brukes alene. Mens det vil være en relativt billig sak å sløyfe ett kontrollsiffer, vil det være en dyr og omstendelig historie senere å gå over til to sifre, om en i første omgang utarbeider personnummeret med bare ett kontrollsiffer.

7. Oppgave- og punchefeil i personnummeret

Før vi ser nærmere på hvor godt de forskjellige kontrollsystemer "tar" feil i personnummeret, kan det være grunn til å nevne de hovedgrupper av feil som vil forekomme i et personnummer slik det er skissert ovenfor. Det kan da være greitt å skille feilene i følgende grupper:

1. Oppgavefeil i fødselsdag
2. Oppgavefeil i individual- og kontrollsiffer
3. Punchefeil.

Oppgavefeil i fødselsdato er den største feilkilden, og også den som er vanskeligst å få kontrollert ved hjelp av kontrollsifre. Hovedoppgaven må derfor være å få redusert denne feiltypen mest mulig, og en har derfor under arbeidet med personnummeret konsentrert seg vesentlig om dette. Punchefeil tas godt ved de fleste metoder for kontrollsifre, særlig fordi punchefeil som regel bare forekommer i ett siffer i et tall. Oppgavefeil i individualsifferet har vi liten oversikt over. Særlig hyppigheten vil her trolig kunne variere sterkt, avhengig av på hvilken måte sifrene er overført til punchegrunnlaget. Oftest vil det dreie seg om en avskrift, men selv i dette tilfellet kan hyppigheten av feil variere ikke lite, alt etter som avskriftgrunnlaget er. For best mulig på forhånd å kunne vurdere hvordan kontrollsifferet vil virke i praksis, har Byrået søkt å skaffe materiale som både kan vise hvor hyppig de forskjellige hovedgrupper av feil vil forekomme, og hvilke typer av feil som forekommer innen hver av hovedgruppene.

Som nevnt vil oppgavefeil i fødselsdag være den hyppigste forekomne feil. Allerede under de forberedende arbeider med personnummeret, begynte derfor Byrået å samle materiale til belysning av hvilke feiltyper en her måtte gjøre regning med. For noe over en årgang av døde, i alt ca. 35 000 personer, ble den fødselsdato som var oppgitt på dødsattesten, sammenliknet med fødselsdatoen som var oppgitt i kirkebøkenes avdeling for døde. Dette materialet lå til grunn for professor Selmers utredning Del I. Under arbeidet med personnummeret oppdaget en imidlertid at folkeregisteret i Oslo ved Folketellingen i 1960 hadde skrevet ned alle uoverensstemmelser mellom den fødselsdato på folketellingsskjemaene, og den som sto oppført i folkeregisteret. Dette materialet var oppbevart og ligger til grunn for alle senere beregninger.

Når det gjelder punchefeil, har en analysert et materiale som gjaldt punching av ca. 230 000 fødselsdatoer. I alt ble det punchet feil i 1 591 datoer.

Oppgavefeil i individualsifrene har en ikke hatt noe materiale til å bedømme hyppigheten av. Som nevnt vil det oftest dreie seg om avskriftsfeil. En undersøkelse, som ble foretatt i forbindelse med konstruksjonen av et nytt

kontonummersystem for den engelske Postsparebanken, tyder på at slike avskrifts-feil stort sett følger mønsteret for punchefeil. Selve feilhyppigheten vil imidlertid som nevnt variere sterkt alt etter grunnmaterialets karakter.

8. Enkelte alternativer for utregning av kontrollsifre

De fleste hullkortanlegg her i landet bruker punchemaskiner av IBM-fabrikat. IBM har to standardsystemer for siffer som kan brukes som kontroll på punchemaskiner, ett som bygger på en 10-modul og ett som bygger på en 11-modul. 10-modulen er den eldste og er den som Oslo ligningskontor nå bruker ved utregning av sitt kontrollsiffer. Dette kontrollutstyret er imidlertid lite hensiktsmessig når det gjelder kontroll av fødselsdato, særlig fordi det ikke "tar" ombytte av bigrammer (se professor Selmers utredning Del 1 side 12). Slike bigram-ombytter, f.eks. bytte av fødselsdag og fødselsmåned er hyppige feil i oppgaver over fødselsdag. Spesielt av denne grunn må modul 10 forkastes, og av de standardproduserte kontrollsystemer har en derfor konsentrert seg om modul 11.

a. IBM's modul 11 brukt to ganger

Professor Selmers første utredning dreier seg i det vesentlige om hvordan IBM's modul 11 vil virke om den brukes både som første og annet kontrollsiffer. Fordelen ved et slikt system vil være at en da kan bruke et standardprodusert utstyr både når en bruker personnummeret med ett og når en bruker det med to kontrollsifre. Som det framgår av utredningen har systemet imidlertid matematiske svakheter som gjør at det må forkastet.

b. IBM's modul 11 brukt som første kontrollsiffer med spesialkonstruksjon av annet siffer

Dette alternativ behandles på sidene 18 - 21 i professor Selmers utredning. Del 2. Alternativet forutsetter at punchemaskinkontrollen alltid foretas på første kontrollsiffer, mens annet kontrollsiffer brukes på en computer. Fordelen ved dette systemet er at det matematisk er vesentlig bedre enn alternativ a, og det kan brukes standardprodusert kontrollutstyr. En svakhet ved alternativet er at en ikke får kontrollert oppgavefeil eller punchefeil i annet kontrollsiffer før etter kjøring på computer. Det viser seg at de fleste feilene som oppdages ved computerkjøringen vil være slike feil.

c. Spesialkonstruksjon både av første og annet kontrollsiffer

I sin annen utredning har professor Selmer undersøkt hvor effektive kontrollsifrene kan bli om begge spesialkonstrueres. Han har satt som betingelse at samme vekttallsystem skal brukes for begge sifre. Det viser seg at han kommer

fram til et system hvor første kontrollsiffer bare slipper igjennom ca. 70 prosent av de feil som IBM's modul 11 slipper igjennom. Om en ser begge kontrollsifre under ett, må en imidlertid regne med omtrent samme sikkerhet som alternativ b. Alternativet har den ulempe at en er avhengig av spesialkonstruert utstyr på punchemaskiner, også når to sifre brukes, og da en altså ikke får større total effektivitet enn under alternativ b.

d. Spesialkonstruksjon av første kontrollsiffer, IBM's 11-modul brukt som annet kontrollsiffer

Dette alternativet er ikke utredet i detalj, men på grunnlag av det en kjenner fra de øvrige alternativer, vet en tilstrekkelig til å kunne vurdere det. Det krav vi i dette tilfellet må stille til første kontrollsiffer, er for det første at det "samarbeider godt" med annet kontrollsiffer (IBM's modul 11) og dessuten at det også er et effektivt kontrollsiffer brukt alene. En foreløpig analyse som professor Selmer har foretatt, viser at en kan komme fram til et system som når begge sifre brukes, tilfredsstiller de krav vi setter, og om første siffer brukes alene virker vesentlig bedre enn IBM's 11-modul brukt en gang. Første kontrollsiffer brukt alene virker likevel ikke fullt så effektivt som ett kontrollsiffer gjør under det alternativ som er nevnt ovenfor i punkt c.

Dette alternativet har den fordel sammenliknet med alternativ a, at det er vesentlig mere effektivt, og en får samtidig kontrollert oppgavefeil eller punchefeil i annet kontrollsiffer allerede under punchingen av materialet (som alternativ b ikke gjorde). Alternativet er, brukt med bare ett siffer, ikke så effektivt som alternativ c, men er med to siffer trolig omtrent like effektivt, og har da den fordelen at en kan bruke standardprodusert utstyr som kontroll på punchemaskinen. Brukt bare med ett kontrollsiffer må kontrollutstyret på punchemaskinen spesialbestilles, som ved alternativ c, men til gjengjeld vil en altså få et for dette formål vesentlig mer effektivt kontrollutstyr enn det standardproduserte.

9. Eksempler på effektiviteten av kontrollsifrene

Som tidligere nevnt har en, når en har studert effektiviteten av kontrollsifrene, i det vesentlige konsentrert seg om hvor effektivt sifrene virker på oppgavefeil i fødselsdato. Det materialet som her særlig har vært brukt, er det såkalte "Oslo-materialet" som besto av i alt 6 942 kort. I tillegg til analysen av oppgavefeilene er materialet også blitt prøvet på forskjellige alternativer av kontrollsifre, og tallet på feil som passerte henholdsvis ett og to kontrollsifre, stemte svært godt med det en kunne vente på

grunnlag av professor Selmers utredninger. De avvikene som forekom, var ikke større enn at de kan tilskrives tilfeldigheter i materialet. Følgende tabell viser hovedresultatene av kontrollkjøringer som ble foretatt:

Alternativ nevnt under punkt 8.	Feil i alt	Feil som passerer 1. kontroll-siffer	Feil som passerer begge kontroll-sifre
a) IBM's modul 11 brukt to ganger	6 942	133	13
b) IBM's modul 11 brukt som første kontroll-siffer med spesialkonstruksjon av annet siffer	6 942	133	1-3
c) Spesialkonstruksjon både av første og annet kontrollsiffer	6 942	93	0
d) Spesialkonstruksjon av første kontroll-siffer, IBM's modul 11 brukt som annet kontrollsiffer	6 942	(ca. 100)	(ca. 2)

Under alternativ b har en på feil som passerer to kontrollsifre satt grensene 1-3 idet en har utarbeidd forskjellige alternativer her (se alternativene i professor Selmers annen utredning) og alle alternativer lå innen disse grensene. Under punkt d er tallene ført opp i paranteser idet en ikke har foretatt aktuelle kjøringer for dette alternativet. Tallene er satt opp på grunnlag av det en for øvrig kjenner av materialet sammenholdt med det som går fram av professor Selmers utredninger.

Om en så til slutt vil forsøke å vurdere hva kontrollsifrene vil bety i praksis, må en gjøre seg opp en mening om hvor stor prosentdel av oppgavene en kan vente kommer inn med oppgavefeil eller punchefeil. Hyppigheten av oppgave-feil i fødselsdato lå både i Oslo-materialet og i den sammenlikning som ble foretatt av dødsattestene mot kirkebøkenes avdeling for døde, mellom 1 og 2 pro-sent. Under den punching av 230 000 fødselsdatoer som er nevnt ovenfor, forekom det feil i vel ½ prosent av tilfellene. Oppgavefeil i individualsifrene har en ikke noe materiale til å vurdere hyppigheten av, men en har i det følgende gått ut fra at feilene her vil ligge på ca. 1 prosent, noe som trolig er i overkant. Punchefeil i individualsifrene og kontrollsifferet har en anslått til ½ prosent. Ut fra dette kan en sette opp følgende tabell som viser hvor mange feil som vil oppstå ved 1 mill. registreringer, og hvor mange av disse feilene som vil passere henholdsvis første og annet kontrollsiffer. Som utgangspunkt har en her tatt et eksempel som ligger nærmest det som under punkt 8 er nevnt som alternativ b.

Feil ved 1 000 000 registreringer

Feiltype	Feil i alt	Feil som passerer 1. kontroll-siffer	Feil som passerer 2. kontroll-siffer
Oppgavefeil i fødselsdato	15 000	270	5
Oppgavefeil i individualsiffer	10 000	90 ?	2 ?
Punchefeil i fødselsdato	5 000	31	0,3
Punchefeil i individualsiffer + kontrollsiffer	5 000	31	0,3
Kombinasjon f.dato + annen feil	300	27	1
Individualsiffer + punchefeil	100	9	0,3
Punchefeil i begge ledd	25	2	0,07
I alt	35 425	460	≈9

Når det gjelder oppgavefeil i individualsiffer, har en gått ut fra at de feil som gjøres stort sett er i samsvar med feilene i den tidligere nevnte engelske undersøkelsen for kontonummer i Postsparebanken. Tallet for feil som passerer henholdsvis første og annet kontrollsiffer er likevel for denne gruppe usikre. Usikkerheten kan imidlertid ikke være større enn at størrelsesordenen av det totale antall feil som vil passere henholdsvis ett og to kontrollsifre er relativt pålitelige.

Som en konklusjon kan en altså si at med de forutsetninger som er gjort vil $3\frac{1}{2}$ prosent av oppgavene være feil etter at kort er punchet. Uten andre kontroller vil ca. 1 av 2 000 registreringer komme til å passere med feil, og altså bli registrert på feil person om en har ett kontrollsiffer, mens bare 1 av ca. 100 000 registreringer vil skje på feil person om en bruker to kontrollsifre.

Kontrollsifre ved personnummerering

Av Prof., dr.philos. Ernst S. Selmer

Del 1

Som svar på Deres brev av 8.4.63 (nedenfor kalt "brevet") og det 23.4. oversendte statistiske materiale (nedenfor kalt "materialet") kan jeg gi følgende opplysninger:

I samsvar med brevet har jeg gått ut fra at personnummeret skal være 9-sifret, begynnende fra venstre med dag, måned og år (2-sifret), samt et tresifret nummer. Etter disse antar jeg to kontrollsifre:

$$d_{10}\ d_1\ m_{10}\ m_1\ a_{10}\ a_1\ n_{100}\ n_{10}\ n_1\ k_1\ k_2$$

Som foreslått i brevet har jeg antatt at k_1 kan brukes separat, ved kontroll av selve personnummeret på 9 sifre. Om ønsket kan så k_2 brukes til kontroll av personnummeret og k_1.

I denne forbindelse først en bemerkning om plasseringen av kontrollsifrene: I det kommersielt tilgjengelige utstyr kontrollerer et slikt siffer alle sifre til venstre for seg; det er derfor opplagt at k_2 må stå helt sist. Derimot kunne man tenke seg at k_1 sto lenger inne i tallet, f.eks. mellom a_1 og n_{100}, slik at første kontroll bare gikk på fødselsdatoen. Med det system som nedenfor foreslås, IBM Modulus 11, er dette imidlertid ikke mulig, da ikke alle sifferkombinasjoner vil gi anledning til et kontrollsiffer. Tenkes k_1 flyttet mellom n_{100} og n_{10}, vil dette av samme grunn kunne gi visse restriksjoner i valget av n_{100}, noe som er uheldig hvis n_{100} skal brukes til grovsortering av personkategorier. Og plassering av k_1 mellom n_{10} og n_1 må sies å være "søkt".

Jeg har konsentrert meg om de allerede eksisterende IBM-systemer modulus 10 og 11, for det første fordi jeg antar at nykonstruksjon av kontrollutstyr vil skape store komplikasjoner, for det annet fordi det system jeg har konsentrert meg om, modulus 11, må sies å være meget godt. Det eldre system modulus 10 (det som nå brukes av Oslo kommune) vil jeg derimot ikke anbefale. Som generelt kontrollsystem er det ikke på langt nær så godt som modulus 11, og

dessuten vil det (som påpekt i brevet) ha den store svakhet i den aktuelle forbindelse at det ikke oppdager ombytting av dato og måned o.l. Rent matematisk er også modulus 11 behageligere, da det er et rent vekttallsystem uten menteoverføring.

Modulus 11-systemet illustreres ved følgende eksempel (fra IBM-brosjyren):

Gitt tall: 9 4 3 4 5 7 8 4 2

Vekttall : 4 3 2 7 6 5 4 3 2

periode

Veiet tverrsum: 36 + 12 + 6 + 28 + 30 + 35 + 32 + 12 + 4 = 195.

Rest av denne ved divisjon med 11 = 8.

11-komplementet av resten = 11 - 8 = 3 = kontrollsiffer.

Selv-kontrollerende tall: 9 4 3 4 5 7 8 4 2 3

En stor fordel ved systemet er at modulen er primtall. Dette må være større enn 10, for at to forskjellige desimalsifre ikke skal gi samme rest (som f.eks. sifrene 1 og 8 ved modulen 7). På den annen side oppstår derved en spesiell ulempe, idet alle tall som gir kontrollsiffer 10 må forkastes. Som allerede nevnt, medfører dette at systemet ikke kan brukes for et "fastlåst" tall som f.eks. fødselsdatoen. Ved personnummereringen skal det imidlertid føyes til et nokså vilkårlig 3-sifret nummer, og her kan man da jonglere med modulen, idet gjennomsnittlig hvert 11te nummer må forkastes. Ifølge brevet skal en slik nummertildeling skje sentralt og maskinelt, hvilket da ikke vil by på problemer. Blir det derimot aktuelt med lokal nummertildeling, må man treffe spesielle forholdsregler.

Som nevnt i brevet skal jeg i det følgende vurdere styrken i et system med ett kontrollsiffer k_1 , samt den forbedring som oppstår ved å føye til et nytt kontrollsiffer k_2. Vi opererer altså med følgende vekttall:

Nummer:	d_{10}	d_1	m_{10}	m_1	a_{10}	a_1	n_{100}	n_{10}	n_1	k_1	k_2
Vekttall for k_1:	4	3	2	7	6	5	4	3	2		
" " k_2:	5	4	3	2	7	6	5	4	3	2	

(Anm.: Når det i det følgende sies at en feil "passerer" k_1, men "tas" av k_2, forutsettes det selvsagt at ved 2. kontroll er k_1 riktig punchet, slik at ikke en feil i k_1 skal kunne "dekke" en feil i selve nummeret. En slik (usannsynlig) mulighet er det nokså håpløst å gardere seg mot.)

Ved helt tilfeldige feil må man vente at ett kontrollsiffer k_1 vil la gjennomsnittlig hver 11te feil passere uoppdaget. Feilene er imidlertid gjerne

av mer systematisk natur, som nedenfor beskrevet. Man skulle også vente at bare hver 11te av de "overlevende" feil fra k_1 ville passere k_2 i gjennomsnitt, men her har modulus 11-systemet i mange tilfelle en påfallende egenskap: Annet kontrollsiffer tar ofte enten alle eller ingen av de feil som passerte første kontrollsiffer!

Vi kan skille mellom to hovedtyper av feil: Rene punche-feil (som systemet er konstruert for), og systematiske feil i fødselsdatoen. Jeg har studert virkningen av kontrollsifrene på begge typer feil, og skal nedenfor kort gjøre rede for resultatene.

1) Ett siffer galt: Oppdages alltid ved k_1. (Dette er aktuelt som feil av begge kategorier.)

2) Ombytting (transposisjon) av to sifre i vilkårlig avstand (også begge feil-typer): Oppdages alltid ved k_1 når de to sifre ikke har samme vekttall, altså avstand 6. Ved avstand 6 tas feilen hverken ved k_1 eller k_2. I siste tilfelle må ett av sifrene (p.g.a. avstanden) være hentet fra det tresifrede nummer. Det må da dreie seg om en punche-feil, men en ombytting av to sifre med så stor avstand er uhyre usannsynlig.

3) Transposisjon xyx $\leftrightarrow$ yxy, f.eks. 212 istedenfor 121. Dette er en typisk punche-feil, den oppdages alltid ved k_1.

4) "Kompensasjon", dvs. at to nabosifre begge økes med samme tall, f.eks. 34 istedenfor 12. (Denne feiltype nevnes ihvertfall som typisk i IBM-brosjyren for modulus 10; det er opplagt en punche-feil.) Oppdages av k_1 overalt unntatt i årstallet (a_{10} a_1), hvor vekttallsummen 6 + 5 = 11 nettopp gir modulen. Dette unntak vil imidlertid oppdages av k_2 (andre vekttall).

5) Tilfellene 2 og 4 er spesialtilfeller av feil på to plasser. Hvis slike feil er vilkårlige, vil gjennomsnittlig hver 11te passere k_1 uoppdaget. Det viser seg nå, som allerede nevnt, at forholdene for k_2 er ganske eiendommelige:

a) Hvis de to plasser har samme vekttall, altså avstand 6, er det klart at en uoppdaget feil ved k_1 heller ikke vil oppdages ved k_2 (også samme vekttall). Som nevnt under pkt. 2 må det her dreie seg om en punche-feil, alstå svært usannsynlig.

b) Hvis plassene er gitt ved d_1 og m_1 eller m_1 og n_{10}, vil de uoppdagede feil ved k_1 også passere k_2 uoppdaget.

c) Ved alle andre plasseringer av de to feil vil alle uoppdagede feil ved k_1 oppdages ved k_2.

Vi må se litt mer på de uoppdagede feil under pkt. b. Tilfellet m_1 og n_{10} må (p.g.a. n_{10}) skyldes punche-feil, altså usannsynlig. Derimot kan feil i enersifferet i dag og måned skyldes feil i oppgavene. I det tilsendte materiale skjuler slike feil seg under sekkeposten "Samme år, forskjell i både måned og dag", med i alt 70 tilfeller. Det bør undersøkes hvor mange av disse feil som opptrer bare i enersifrene[1].

6) Feil i fødselsdag (systematiske feil): Ifølge pkt. 1 vil feil i bare ett siffer alltid tas ved k_1. De feil i begge sifre som ikke tas ved k_1 vil ifølge pkt. 5 c alltid tas ved k_2.

I materialet er det 79 dager med feil i begge sifre. Det viser seg at 10 av disse feil ikke tas ved k_1. Dette er litt mer enn $\frac{1}{11}$ av alle feil, men avviket er ikke signifikant.

Det kan tenkes å opptre gale dager i nærheten av den riktige. Hvis 10-sifret d_{10} er det samme, vil feilen i d_1 tas ved k_1. Hvis 10-sifret er forskjellig (altså hvis den riktige og gale dag ligger like på hver side av 10, 20 eller 30), er det bare differens 5 som ikke tas allerede av k_1.

7) Feil i måned (systematisk): Det er her så få muligheter at det er lett å regne seg gjennom alle, og det viser seg at ombyttingene mai $\leftrightarrow$ oktober, juni $\leftrightarrow$ november og juli $\leftrightarrow$ desember er de eneste som ikke tas allerede ved k_1 (men de tas ifølge pkt. 5 c alle ved k_2).

I materialet er det 97 feil i (bare) måned, hvorav 20 med begge sifre forskjellige. Av typen ovenfor er det tre feil, alle med ombytting av mai og oktober.

8) Feil i årstall (systematisk): Som vanlig vil feil i bare ett siffer tas ved k_1. I materialet er det 33 gale årstall med feil i begge sifre, hvorav omkring halvparten er i nærheten av det riktige, men altså på hver side av et multiplum av 10. Slike feil vil som regel tas bare ved k_1 såfremt differensen ikke er 11 (altså meget stor).

Dette gjelder imidlertid ikke ved århundreskifte, idet det isåfall er differens 1 som ikke tas ved k_1, altså feilen 99 $\leftrightarrow$ 00. Denne feil tas først ved k_2.

9) Ombytting dag $\leftrightarrow$ måned, dag $\leftrightarrow$ år eller måned $\leftrightarrow$ år: Dette er en meget viktig (systematisk) feiltype, og faktisk en viktig grunn til at modulus 10-systemet måtte forkastes. I materialet er det 22 ombyttinger av dag og måned, men derimot ingen av dag og år. Mulige ombyttinger av måned og år er ikke oppgitt.

---

1) Byråets merknad: Opptellingen viser 12 feil.

For ombytting av "bigrammer" (grupper på to sifre) gjelder igjen spesielle forhold for annet kontrollsiffer: Av de ombyttinger som ikke tas av k_1, vil enten alle eller ingen tas av k_2, avhengig av bigrammenes plassering i forhold til kontrollsifrene. Det viser seg videre at ombyttinger dag ↔ måned og måned ↔ år alle tas av k_2, men ved ombyttinger dag ↔ år gir k_2 ingen forbedring i forhold til k_1.

Vi må se litt mer på de enkelte tilfelle. Der hvor måneden inngår, er antall muligheter så lavt at de er lette å angi.

a) Ombytting dag ↔ måned: Det er bare tre kombinasjoner som ikke tas ved k_1:

10.05 ↔ 05.10, 11.06 ↔ 06.11 og 12.07 ↔ 07.12.

Som nevnt vil alle disse tas av k_2.

b) Ombytting måned ↔ år: Det er bare en kombinasjon som ikke tas av k_1:

10.09 ↔ 09.10,

og denne vil tas av k_2.

c) Ombytting dag ↔ år: Her vil altså de samme feil passere k_1 og k_2. Det viser seg at disse feil er kjennetegnet ved at bigrammene for dag og år har samme tverrsum. Når det ene bigram skal være en av datoene fra 01 til 31, finner man lett at det er 32 kombinasjoner som overhodet ikke oppdages.

Selv om som nevnt en ombytting av dag og år ikke forekommer i materialet, må dette store antall kombinasjoner sies å utgjøre en svakhet ved hele systemet.

Bruk av bare k_2: Det kan tenkes at man bruker to kontrollsifre, men at bare det ene brukes ved den eksterne bearbeidelse (punching) av materialet, mens annet siffer kun brukes ved den sentrale behandling på data-maskin. Hvis det er første siffer k_1 som brukes eksternt, er det allerede redegjort ovenfor for hvilke feil som vil oppdages og hvilke som vil passere. Hvis derimot siste siffer k_2 brukes eksternt, vil det være litt andre feil som ikke oppdages. For ordens skyld skal jeg også gjøre rede for hva som skjer i et slikt tilfelle:

Punktene 1, 2 og 3 er uforandret. I pkt. 4 er det "kompensasjon" for sifrene a_1 og n_{100} som ikke tas ved k_2 alene.

Punkt 5 er uforandret ved ombytting av k_1 og k_2.

Punkt 6: Jeg har ikke tellet opp hvor mange av feilene i materialet som ikke tas av k_2 alene (resultatet ville i alle fall bli nokså tilfeldig). Hvis den riktige og den gale dato ligger på hver side av et multiplum av 10, er det bare differens 6 som ikke tas av k_2 alene.

Punkt 7: Følgende ombyttinger:

juli ↔ oktober, august ↔ november og september ↔ desember tas ikke av k_2 alene (men alle ved k_1 etterpå). Det er 3 slike ombyttinger i materialet.

Punkt 8: Feil på hver side av et multiplum av 10 vil ikke tas ved k_2 alene hvis differensen er 7, unntatt ved århundreskifte, hvor det er differens 4 som ikke tas.

Punkt 9: Både under 9 a (dag $\leftrightarrow$ måned) og 9 b (måned $\leftrightarrow$ år) vil følgende ombyttinger ikke tas ved k_2 alene:

10.01 $\leftrightarrow$ 01.10, 11.02 $\leftrightarrow$ 02.11 og 12.03 $\leftrightarrow$ 03.12

Punkt 9 c (dag $\leftrightarrow$ år) er uforandret ved ombytting av k_1 og k_2.

Konklusjon: Som nevnt vil jeg foreslå kontroll ved modulus 11-systemet. For første kontrollsiffers vedkommende er de av IBM valgte vekttall like gode som noe annet valg, og det er ingen grunn til å foreslå modifikasjoner hvis man bare vil ha ett kontrollsiffer.

Hvis man derimot bestemmer seg for to kontrollsifre, for å oppnå større sikkerhet, viser det seg alstå at det er visse "svake ledd i kjeden", nemlig punktene 5 b og 9 c ovenfor. Disse svakheter skyldes det spesielle valg av vekttall. Jeg har sett litt nærmere på dette, og det ser ut som om man ved andre valg av vekttall kan eliminere svakhetene, uten å ødelegge noen av systemets mange gode sider. Et nærmere studium av dette vil imidlertid ta endel tid, og jeg vil ikke gå i detaljer før jeg har fått undersøkt om en modifikasjon av vekttallene kan gjennomføres uten kompliserte tekniske komplikasjoner i utstyret. Jeg har allerede rettet en forespørsel til IBM om disse problemer.

Kontrollsifre ved personnummerering

Av

Prof., dr.philos. Ernst S. Selmer

Del 2.

Den foreliggende rapport er skrevet i umiddelbar tilknytning til rapporten av 11.7., nedenfor kalt Del 1. Etter at denne var skrevet, er det kommet til endel nye opplysninger om sannsynlige punchefeil, om mulige modifikasjoner av IBM-systemet, samt spesielt et materiale på nesten 7 000 mennesker som hadde oppgitt gal fødselsdato ved personnummereringen i Oslo kommune. Dette "Oslo-materialet" er i løpet av sommeren blitt grundig analysert av Byrået, etter de feilkategorier som var spesifisert i Del 1.

Punchefeil. Smlgn. Deres brev av 17.8.63. Begge de prinsipper jeg foreslår nedenfor vil ved annet kontrollsiffer ta alle feil i to siffer. Med et system for første kontrollsiffer som enten er et umodifisert IBM modulus 11, eller et modifisert system som ihvertfall ikke er svakere, skulle derved sannsynligheten for uoppdagede punchefeil være forsvinnende liten. Et fortsatt studium av slike feil har neppe noen hensikt, ihvertfall ikke for det foreliggende formål.

Modifikasjoner. Fra IBM er det kommet opplysninger om allerede foretatte modifikasjoner av modulus 11-systemet. Dels dreier det seg om andre vekttall, dels også om andre moduler.

Som forklart i Del 1, er modulen 11 sannsynligvis den beste for formålet, og det er ingen grunn til å reflektere på systemer med modul 7, 9 eller 13. Modulo 11 har IBM også brukt i etsystem hvor vekttallene er suksessive potenser av 2. Heller ikke dette har etter min mening noen særlig interesse, av følgende grunner:

Som eneste kontrollsiffer har systemet ingen særlige fortrinn framfor det konvensjonelle, hverken når det gjelder punchefeil (bortsett fra at kompensasjon oppdages på samtlige plasser) eller systematiske feil i datoen. Skal man gå til en modifikasjon av vekttallene, bør det nye system (smlgn. ett av mine forslag nedenfor) være slik at de samme vekttall kan brukes også for annet kontrollsiffer. Og på dette punkt er potensene av 2 helt ubrukelige, da de har nøyaktig samme virkning for k_1 og k_2 (vekttallene for k_2 er jo bare proporsjonale med - nemlig det dobbelte av - vekttallene for k_1).

Forslag til annet kontrollsiffer. I det følgende vil jeg framlegge to forskjellige forslag. Valget må foretas av Byrået, men jeg skal her knytte noen kommentarer og vurderinger til de to forslag:

1) Første kontrollsiffer (k_1) beholdes umodifisert, og det innføres et "elektronisk" annet kontrollsiffer (k_2), med nye vekttall. Dette siffer brukes bare ved den sentrale, maskinelle bearbeidelse av informasjonen.

2) Det innføres nye vekttall, tilpasset slik at de kan brukes for begge kontrollsifre.

Totalt sett (ved bruk av både k_1 og k_2) gir begge systemer like gode resultater. Det er fortrinn og mangler ved begge systemer:

Forslag 1 har selvsagt den store fordel at man kan bruke standard IBM-utstyr. På den annen side viser jo opplysningene fra IBM at det er fullt mulig å få modifiserte vekttall. Ulempen ved dette vil delvis avhenge av omkostningene, som vel igjen er sterkt avhengig av antall modifiserte enheter som kreves. Leveringstid for de første modifiserte enheter og for senere suppleringer må også tas i betraktning.

Forslag 2 har sin styrke i at det gir bedre resultater for k_1 alene, idet systemet er "skreddersydd" etter de psykologiske tendenser i Oslo-materialet (som må antas å være representativt for landet som helhet). Ennvidere gir det muligheter for ekstern kjøring av k_2, hvis dette skulle bli påkrevet. Endelig er det ett moment som også er nevnt i Deres brev av 9.8.: Om noen år kan man kanskje få punchemaskiner som verifiserer begge kontrollsifre simultant. Det er vel rimelig å anta at noe slikt er lettere å realisere når k_1 og k_2 har de samme vekttall (bare forskjøvet en plass i forhold til hverandre).

Nedenfor følger en detaljert beskrivelse av de to forslag.

Forslag 1: Elektronisk k_2. Det dreier seg her om nye vekttall bare for k_2, etter en eller annen modul. Som modul kan fortsatt velges 11, eller en modul ≤ 10. Det siste har den fordel at man ikke mister noen flere

. .

personnummer innen samme fødselsdato, mens ny bruk av modul 11 medfører at $\frac{1}{11}$ av de nummer som kunne brukes for k_1 nå må forkastes for k_2. Imidlertid har man rikelig med nummer til disposisjon, og ett helt avgjørende moment taler for fortsatt bruk av modulen 11: Hvis man f.eks. brukte modul 10 for k_2, måtte man regne med at mange flere feil passerte k_2 uoppdaget. Det dreier seg ikke om fulle 10 % av de overlevende feil etter k_1, idet disse ofte er av en så systematisk natur at man kunne "skreddersy" de nye vekttall modulo 10. Men på visse punkter er det helt umulig å gardere seg på denne måten, f.eks. når det gjelder å ta alle feilkombinasjoner på to vilkårlige plasser. Dette er enkelt nok med modul 11 for k_2, men ved modul 10 for k_2 må man nok regne med, uansett hvordan vekttallene tilpasses, at gjennomsnittlig 10 % av de overlevende slike feil etter k_1 ikke kan tas. Av denne grunn har jeg holdt meg til modul 11 også for k_2, selv om det betyr en ytterligere reduksjon i antall brukbare nummere under hver dato.

Vi opererer nå med følgende to sett vekttall

	d_{10}	d_1	m_{10}	m_1	a_{10}	a_1	n_{100}	n_{10}	n_1	k_1	k_2
For k_1:	$v_9=4$	$v_8=3$	$v_7=2$	$v_6=7$	$v_5=6$	$v_4=5$	$v_3=4$	$v_2=3$	$v_1=2$		
For k_2:	w_9	w_8	w_7	w_6	w_5	w_4	w_3	w_2	w_1	w_0	

I Del 1 er det redegjort for hvilke feil som tas av k_1 (v-ene) alene. Systemet er så effektivt at vi bare behøver å stille følgende krav til w-ene:

a) Vilkårlige feil på to plasser skal alltid tas ved bruk av begge kontrollsifre. Dette vil være oppfylt hvis alle forhold

$$(1) \qquad \frac{w_i}{v_i}, \quad i = 1, 2, \ldots, 9$$

er inkongruente modulo 11, altså hvis

$$\frac{w_i}{v_i} \not\equiv \frac{w_j}{v_j} \text{ eller } w_i v_j - w_j v_i \not\equiv 0 \pmod{11}$$

for alle $i \neq j$. ($\not\equiv 0$ betyr "ikke delelig" med modulen 11).

På denne måte oppnår vi at alle feiltyper 1 - 8 i Del 1 tas ved kombinert bruk av k_1 og k_2.

b) Når det gjelder den siste feiltype i Del 1 (nr. 9), ombytting av bigrammene for dag/måned/år, vil vi forlange at alle slike ombyttinger tas ved kombinert bruk av k_1 og k_2. Derimot har det liten hensikt å forsøke å gardere seg mot bigram-ombytting på andre plasser. Dette er ingen systematisk feil, og nevnes aldri blant aktuelle punchefeil.

Skal f.eks. ombytting av dag og måned tas ved k_1 og k_2, kreves det at

$$(2) \qquad \begin{vmatrix} v_9 - v_7 & v_8 - v_6 \\ w_9 - w_7 & w_8 - w_6 \end{vmatrix} \not\equiv 0 \pmod{11},$$

og tilsvarende inkongruenser for dag/år og måned/år.

Det viser seg nå å være et meget stort antall mulige valg av kontrollsifrene w, når vi bare krever betingelsene (1) og (2) oppfylt. Det er uråd å angi alle muligheter; isteden har jeg konsentrert meg om å finne noen sekvenser som er "pene" å se på (og derfor lette å huske).

Først en alminnelig bemerkning: En kontrollsiffer-serie tar nøyaktig de samme feil om den erstattes med proporsjonale tall, altså om man multipliserer hele sekvensen med en fast faktor (og om nødvendig reduserer de framkomne vekttall modulo 11). Om to proporsjonale serier vil vi si at de ikke er vesentlig forskjellige.

Det viser seg at det er mulig, på fire vesentlig forskjellige måter, å lage serier w som bare er bygget opp av to forskjellige siffer. Det er en pussig tilfeldighet at disse serier alle kan skrives med 5 ettall og 4 tretall:

w_9	w_8	w_7	w_6	w_5	w_4	w_3	w_2	w_1
3	3	3	1	1	3	1	1	1
3	1	1	1	1	3	1	3	3
3	3	1	1	1	3	1	1	3
1	1	3	1	1	3	3	3	1

Når det tilslutt gjelder kontrollsifferet w_0 (som kontrollerer k_1), er det selvsagt bare nødvendig å gardere seg mot punchefeil, av følgende typer:

Kompensasjon (mot n_1).

Transposisjon xyx↔yxy (mot n_{10} og n_1).

Ombytting med ett av de nærmeste sifre, f.eks. n_{100}, n_{10} og n_1.

De nødvendige matematiske betingelser for kompensasjon og transposisjon er gitt under forslag 2 nedenfor; for ombytting kreves at vekttallene er forskjellige på de berørte plasser. Den "peneste" løsning får vi for 1. tilfelle ovenfor, hvor $w_0=3$ kontrollerer ombyttingen mot n_{100}, n_{10} og n_1 (som alle har samme vekttall 1), samtidig som betingelsene for kompensasjon og transposisjon er oppfylt. Dette gir da kontrollsifrene

	d_{10}	d_1	m_{10}	m_1	a_{10}	a_1	n_{100}	n_{10}	n_1	k_1	k_2
For k_1:	4	3	2	7	6	5	4	3	2		
For k_2:	3	3	3	1	1	3	1	1	1	3	

En annen mulighet for spesielt enkle serier av w er å velge de seks siste sifre w_6, w_5, ..., w_1 alle <u>like</u>, f.eks. = 1. Det viser seg at det i alt er 20 slike muligheter. To "pene" serier er f.eks.

w_9	w_8	w_7	w_6	w_5	w_4	w_3	w_2	w_1
9	3	3	1	1	1	1	1	1
7	4	2	1	1	1	1	1	1

Hvis vi for den første velger w_0=3, får vi sekvensene

	d_{10}	d_1	m_{10}	m_1	a_{10}	a_1	n_{100}	n_{10}	n_1	k_1	k_2
For k_1:	4	3	2	7	6	5	4	3	2		
For k_2:	9	3	3	1	1	1	1	1	1	3	

Selv om f.eks. Oslo-materialet er beheftet med en rekke "psykologiske", systematiske feil, er det antagelig meget vanskelig å forutsi om noen serie for w er bedre enn en annen. Det er også liten hjelp i å prøvekjøre flere kontrollserier for sammenligningens skyld, idet antall feil som passerer både k_1 og k_2 i alle fall blir så lavt at eventuelle forskjeller mellom seriene neppe er signifikante.

<u>Forslag 2: Nye vekttall.</u> Det dreier seg her om samme vekttall-serie for k_1 og k_2, altså

	d_{10}	d_1	m_{10}	m_1	a_{10}	a_1	n_{100}	n_{10}	n_1	k_1	k_2
For k_1:	v_9	v_8	v_7	v_6	v_5	v_4	v_3	v_2	v_1		
For k_2:	v_{10}	v_9	v_8	v_7	v_6	v_5	v_4	v_3	v_2	v_1	

Det gjelder å konstruere serien slik at den, ihvertfall for vårt formål, er <u>bedre</u> enn vekttallene på IBM's standardutstyr når det gjelder bruk av <u>bare</u> k_1. For <u>både</u> k_1 og k_2 vil vi stille samme krav som under forslag 1 ovenfor.

Vi skal se systematisk på feiltypene, og bruker samme nummerering som i Del 1:

1) Ett siffer galt oppdages alltid såfremt alle vekttall $\neq 0$.
2) Ombytting av to sifre oppdages alltid ved k_1 såfremt de

tilhørende vekttall er forskjellige. For plasser i forholdsvis stor avstand fra hverandre er ikke dette så viktig, såfremt alle feil på to vilkårlige plasser tas ved hjelp av k_2.

3) Transposisjon xyx $\leftrightarrow$ yxy tas alltid ved k_1 såfremt

$$(3) \qquad v_{i+2} - v_{i+1} + v_i \not\equiv 0 \pmod{11}, \quad i = 1, 2, \cdots, 7.$$

4) Kompensasjon tas alltid ved k_1 såfremt

$$(4) \qquad v_{i+1} + v_i \not\equiv 0 \pmod{11}, \quad i = 1, 2, \cdots, 8.$$

Hverken i (3) eller (4) er det nødvendig å trekke inn kontrollsiffer v_{10}, som bare opptrer for k_2.

5) Vilkårlige feil på to plasser. Betingelse (1) under forslag 1 antar nå formen

$$(5) \qquad \frac{v_{j+1}}{v_j} \not\equiv \frac{v_{i+1}}{v_i} \pmod{11}, \quad 1 \leq i < j \leq 9.$$

Da dette forutsetter bruk også av k_2, kommer her v_{10} inn i bildet.

6) Feil i fødselsdag tas alltid ved k_1 om bare ett siffer er galt eller om sifrene er ombyttet. Det er imidlertid et meget stort antall andre feil (573 i Oslo-materialet). Selv om disse iflg. pkt. 5 alltid tas ved tilføyelse av k_2, ville det være en fordel om vekttallene var konstruert slik at flest mulig ble tatt allerede ved k_1. Hvilke av disse feil som tas avhenger av vekttallene v_9 og v_8, eller rettere sagt av deres forhold modulo 11. En nærmere analyse av Oslo-materialet gir følgende tabell:

Forhold $v_9 : v_8 \pmod{11}$:	2	3	4	5	6	7	8	9
Feil som ikke tas ved k_1 :	38	52	51	61	55	48	88	123

Forhold 1 ($v_9 = v_8$) og - 1 (i strid med betingelse (4)) er ikke tatt med. IBM's standard kontrollsifre svarer til forholdet $4 : 3 \equiv 5 \pmod{11}$, med 61 feil som ikke tas av k_1 alene.

Antallene i tabellen varierer så meget at utslagene må sies å være signifikante. Det største antall, 123, inkluderer en rekke psykologiske "fallgruber", hvorav ombyttingene 09 $\leftrightarrow$ 10 og 29 $\leftrightarrow$ 30 svarer for 35 og 30 feil henholdsvis! (men bare 8 feil for 19 $\leftrightarrow$ 20). På den annen side gir forholdet

$$(6) \qquad \frac{v_9}{v_8} \equiv 2 \pmod{11}$$

bare 38 feil. Det later ikke til å være noen spesielle fallgruber blant disse; de største utslag i Oslo-materialet skyldes ombyttingene 05 ↔ 13, 15 ↔ 23 og 12 ↔ 20, med henholdsvis 5,5 og 6 feil.

I landsmålestokk kan det selvsagt vise seg at forskjellen mellom det beste og de "nestbeste" forhold ikke var signifikant, men det er ihvertfall mest sannsynlig at vi får en viss overensstemmelse, slik at (6) er det beste valg vi etter omstendighetene kan foreta. Tilsvarende bemerkninger gjelder for flere av de senere betingelser.

7) Feil i måned (begge sifre) avhenger av forholdet $v_7 : v_6$. I Oslo-materialet finner vi

Forhold $v_7 : v_6$ (mod 11):	2	3	4	5	6	7	8	9
Feil som ikke tas ved k_1:	12	14	15	6	10	26	42	70

Forskjellene er opplagt signifikante. Det største antall feil (=70) skyldes utelukkende ombyttingen september ↔ oktober, altså 09 ↔ 10. Selv om det ikke er så markerte forskjeller mellom de minste antall, er det nærliggende å velge

$$(7) \qquad \frac{v_7}{v_6} \equiv 5 \pmod{11}.$$

Dette er samme forhold som ved IBM's standard kontrollsifre; de aktuelle ombyttinger som ikke tas ved k_1 er gitt under pkt. 7 i Del 1.

8) Feil i årstall (begge sifre) avhenger av forholdet $v_5 : v_4$. I Oslo-materialet finner vi

Forhold $v_5 : v_4$ (mod 11) :	2	3	4	5	6	7	8	9	(10
Feil som ikke tas ved k_1 :	19	9	17	11	14	14	26	76	14)

Forskjellene er igjen signifikante. Det største antall feil (=76) skyldes i alt vesentlig differens 1 ved 10-skifte som ikke er århundreskifte (altså 09 ↔ 10, 19 ↔ 20 osv.). IBM's standardtall svarer til forholdet $10 \equiv -1 \pmod{11}$, som ikke er aktuelt p.g.a. (4). Det er naturlig å velge

$$(8) \qquad \frac{v_5}{v_4} \equiv 3 \pmod{11},$$

særlig da det nestbeste forhold 5 er i strid med betingelsene (5) og (7) (alle forhold $\frac{v_{i+1}}{v_i}$ skal være forskjellige). Det valgte forhold svarer

til differens 7 (eller 18) ved vanlig 10-skifte, og til differens 4 (eller 15) ved århundreskifte.

Som et apropos kan jeg nevne en psykologisk kuriositet i Oslo-materialet. Man kunne kanskje vente å få enda mindre antall feil ved forholdet $v_5 : v_4 \equiv 2$, som gir de <u>større</u> differenser 8 (eller 19) ved vanlig 10-skifte, og 6 (eller 17) ved århundreskifte. Når antallet i virkeligheten blir større, skyldes det delvis ombyttingen $00 \leftrightarrow 19$, idet "år nittenhundre" blir til årstallet 19!

9) Ombyttinger dag/måned/år. Som i forslag 1 vil vi forlange at <u>alle</u> slike tas ved <u>kombinert</u> bruk av k_1 og k_2. Betingelse (2) antar nå formen

$$(9) \qquad \begin{vmatrix} v_9 - v_7 & v_8 - v_6 \\ v_{10} - v_8 & v_9 - v_7 \end{vmatrix} \not\equiv 0 \pmod{11},$$

og tilsvarende inkongruenser for dag/år og måned/år.

Ombyttinger dag/år og måned/år forekommer svært sjelden, mens derimot ombyttinger dag/måned er sterkt representert i Oslo-materialet. Selv om vi holder utenfor de tilfeller hvor dag og måned har felles 1-er eller felles 10-er (altså en enkel ombytting av de to øvrige siffer), blir det tilbake 74 tilfelle av ombytting uten felles siffer. Det er nærliggende å undersøke hvorledes disse forholder seg overfor bare k_1.

Det viser seg at dette avhenger av forholdet

$$f \equiv \frac{v_9 - v_7}{v_8 - v_6} \pmod{11}.$$

I Oslo-materialet finner vi følgende antall ombyttinger dag/måned som ikke tas av k_1 alene:

f:	1	2	3	4	5	6	7	8	9	10
Antall:	14	4	7	6	10	6	10	7	4	6

Forskjellene er ikke særlig store, men ihvertfall det største antall (=14) er antagelig signifikant, da det domineres av den nærliggende ombytting $01.10 \leftrightarrow 10.01$.

Det viser seg at hvis de tidligere oppstilte betingelser (3) - (9) skal oppfylles, er det <u>ikke mulig å velge f optimalt.</u> I forslaget nedenfor har jeg måttet bruke <u>f = 7</u>, svarende til 10 ombyttinger. IBM's standardtall gir f = 5, også med 10 ombyttinger som ikke tas.

De ombyttinger dag/måned som for f = 7 ikke tas av k_1 er gitt ved

07.10 ↔ 10.07, 08.11 ↔ 11.08 og 09.12 ↔ 12.09.

Ingen av disse gir spesielt store antall feil.

10) Vi har nå betraktet alle feil-kategorier fra Del 1. I Oslo-materialet forekommer imidlertid visse andre feil-typer i temmelig stort antall, nemlig a) feil i enersifrene i dag og måned, dag og år eller måned og år, samt b) feil i tre sifre i dag og måned.

a) Feil i enersifrene avhenger av følgende forhold modulo 11:

$$\text{Feil i } d_1 \text{ og } m_1\text{: } \frac{v_8}{v_6}$$

$$\text{" " } d_1 \text{ og } a_1\text{: } \frac{v_8}{v_4}$$

$$\text{" " } m_1 \text{ og } a_1\text{: } \frac{v_6}{v_4}$$

En opptelling av feil som passerer k_1 for de forskjellige verdier av forholdene gir ikke særlig markerte forskjeller, med unntak av store antall for forholdene + 1 og - 1. Forhold + 1 (like vekttall) lar nemlig ombyttinger passere, dette har vi for så vidt allerede utelukket ved pkt. 2 ovenfor. Det er mer påfallende med de store antall for forhold - 1, som ikke tar "kompensasjon" på de angjeldende plasser. Det overveiende antall feil skyldes her at enersifrene begge er angitt en enhet for store eller en enhet for små. Vi må derfor absolutt sette som ny betingelse at

$$\frac{v_8}{v_6}, \frac{v_8}{v_4} \text{ og } \frac{v_6}{v_4} \not\equiv -1 \pmod{11}. \tag{10}$$

Når det på den annen side gjelder små antall feil, er det i grunnen bare ett markert utslag: Vi får bare 1 uoppdaget feil ved k_1 i enersifrene d_1 og m_1 hvis vi velger

$$\frac{v_8}{v_6} \equiv 7 \pmod{11}. \tag{11}$$

Ved IBM's standardtall er forholdet ≡ 2, som gir hele 12 uoppdagede feil i enersifrene.

b) Feil i tre sifre i dag og måned er selvsagt håpløst å systematisere ved vilkårlige forhold mellom vekttallene v_9, v_8, v_7 og v_6. Nå har vi jo imidlertid ved (6) og (7) fastlagt $v_9 : v_8$ og $v_7 : v_6$, og alle

fire vekttall er dermed effektivt fastlagt ved å angi enda ett forhold, f.eks. $v_8 : v_6$. Jeg har gjennomført dette for Oslo-materialet, nærmest for å få bekreftet antagelsen om at man ikke vil få signifikante utslag ved feil i tre siffer. Antall uoppdagede feil viste seg da også å variere meget lite med forholdet $v_8 : v_6$. Det eneste utslag, som sikkert skyldes en tilfeldighet, var at antallet var desidert lavest (= 4) nettopp for forholdet (11). Det tilsvarende antall ved IBM-tallene er 9.

Så skal kabalen gå opp! Det viser det seg da også at den gjør, idet jeg har funnet en vekttallserie som oppfyller samtlige betingelser (3) - (11) ovenfor. Som nevnt under pkt. 9 var det derimot ikke mulig å kombinere disse betingelser med det optimale valg av forholdet f for ombytting av dag og måned.

Den funne serie er gitt ved

v_{10}	v_9	v_8	v_7	v_6	v_5	v_4	v_3	v_2	v_1
8	3	7	5	1	2	8	9	5	3

Når forholdene $\frac{v_9}{v_8}$, $\frac{v_8}{v_7}$, ..., $\frac{v_2}{v_1}$ alle skal være forskjellige (pkt. 5), og ingen av dem $\equiv +1$ (pkt. 2) eller -1 (pkt. 4), er det lett å se at vi <u>må</u> få gjentagelsen $v_1 = v_9$. Videre må $\frac{v_{10}}{v_9}$ være forskjellig fra de øvrige forhold, d.v.s. $\not\equiv \pm 1$, altså $v_{10} \not\equiv \pm 3$; jeg har valgt $-3 \equiv 8 \pmod{11}$. Gjentagelsen fra v_4 spiller ingen rolle, da ombyttingene mellom to plasser er kontrollert allerede ved k_1; for så vidt kunne vi også ha valgt $v_{10} = 3$ uten noen konsekvenser.

Den uunngåelige gjentagelse $v_9 = v_1$ betyr at en ombytting av d_{10} og n_1 <u>først vil tas ved k_2</u>. P.g.a. n_1 er dette en <u>punche- eller avskrifts-feil</u>, og derfor uhyre usannsynlig. I den foreslåtte serie var det ikke mulig å unngå enda en slik gjentagelse, nemlig $v_7 = v_2$, svarende til den usannsynlige ombytting av m_{10} og n_{10}, i innbyrdes avstand 5.

I Oslo-materialet er det 133 feil som ikke tas ved IBM's standard kontrollserie for k_1. Med forslaget ovenfor vil dette antall reduseres til ca 90 feil. Selv om man ikke kan vente en like stor prosentvis reduksjon på landsbasis, må man regne med at effekten av første kontrollsiffer er vesentlig forbedret ved det nye forslag. En ytterligere forbedring er at den nye serie tar <u>kompensasjon</u> (punchefeil) på alle plasser, mens IBM-serien svikter for plassene a_{10} og a_1 (pkt. 4 i Del 1).

Kontrollsifre ved personnummerering

Av

Prof., dr.philos. Ernst S. Selmer

Del 3.

Etter at Del 1 og 2 var skrevet, er det nå bestemt at man skal bruke to kontrollsifre, hvorav det første skal være "skreddersydd" for de psykologiske tendenser i Oslo-materialet, og det annet skal være et standard IBM kontrolltall, som beskrevet i Del 1. Den nærmere begrunnelse for dette valg er gitt i Byråets notat av 30.10.1963, "Valg av personnummersystem", alternativ 8 d s. 8.

Dette betyr at vekttallene nå vil se slik ut:

	d_{10}	d_1	m_{10}	m_1	a_{10}	a_1	n_{100}	n_{10}	n_1	k_1	k_2
For k_1:	v_9	v_8	v_7	v_6	v_5	v_4	v_3	v_2	v_1		
For k_2:	$w_9=5$	$w_8=4$	$w_7=3$	$w_6=2$	$w_5=7$	$w_4=6$	$w_3=5$	$w_2=4$	$w_1=3$	$w_0=2$	

Alle nødvendige betingelser for vekttallene v_i er gitt i Del 2. Det viste seg nå faktisk å være vanskeligere enn i det tidligere forslag 2 å få kabalen til å gå opp. Etter mange forsøk er jeg blitt stående ved følgende forslag til 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

Alle vekttall er forskjellige. Nedenfor skal kort redegjøres for i hvilken utstrekning betingelsene i Del 2 er oppfylt.

Følgende betingelser er tilfredsstillet:

(1), (2), (3), (4), (6), (10) og (11).

(Etter det som er opplyst om punchefeil i Byråets notat av 22.10.1963, er f.ø. betydningen av betingelse (3) svært liten.)

Betingelsene (5) og (9) er ikke aktuelle (de erstattes av (1) og (2) henholdsvis). I følgende tilfeller har det ikke vært mulig for meg å få optimale forhold mellom vekttall:

Betingelse (7): I forslaget ovenfor er $v_7/v_6 \equiv 6$, svarende til 10 uoppdagede feil ved k_1, mot optimalt 6 feil.

Betingelse (8): Denne er i strid med (1), da $w_5/w_4 \equiv 3$. Det var heller ikke mulig å bruke nest beste verdi; i forslaget ovenfor er $v_5/v_4 \equiv 7$, svarende til 14 uoppdagede feil (på tredje plass), mot optimalt 9.

Ombytting dag $\leftrightarrow$ måned: Forholdet f (Del 2, s. 8) antar for vekttallene ovenfor verdien $f \equiv 5$, svarende til 10 uoppdagede feil. Dette er samme (ikke optimale) antall som for vekttallene i forslag 2.

I forhold til forslag 2 økes altså antall uoppdagede feil med 9 p.g.a. betingelsene (7) og (8). På den annen side viser det seg at det nye forslag - antagelig ved en tilfeldighet - gir noe bedre resultater når det gjelder feil i bare enersifrene i dag og år eller i måned og år. Alt i alt er det derfor meget liten forskjell mellom det nye forslag og forslag 2 når de anvendes på Oslo-materialet. Forbedringen i forhold til IBM's standard vekttall er så markert at man sikkert kan regne med samme tendens også på landsbasis.

Vanligvis er det selvsagt bare IBM-vekttallene w_i som skal brukes alene. Virkningen av disse vekttall er beskrevet i Del 1, 3. 5-6 ("Bruk av bare k_2").