RAPPORTER 84/9

STATISTISKE METODER FOR ANALYSE AV SAMVARIASJON I KATEGORISKE DATA

AV
HERDIS THORÉN AMUNDSEN

STATISTISK SENTRALBYRÅ
CENTRAL BUREAU OF STATISTICS OF NORWAY

RAPPORTER FRA STATISTISK SENTRALBYRÅ 84/9

STATISTISKE METODER FOR ANALYSE AV SAMVARIASJON I KATEGORISKE DATA

AV
HERDIS THORÉN AMUNDSEN

STATISTISK SENTRALBYRÅ
OSLO — KONGSVINGER 1984

ISBN 82-537-2074-2
ISSN 0332-8422

EMNEGRUPPE
Teori og metode

ANDRE EMNEORD
Kategoriske variabler
Samvariasjon

FORORD

Denne rapporten gir en oversikt over statistiske metoder som kan brukes ved problemer der det for hver observasjonsenhet er observert to eller flere ulike typer av kjennemerke som ikke naturlig angis ved en tallverdi. F.eks. kan en type være yrke, en annen bosted, en tredje kjønn, osv. Det kan være av interesse å undersøke om data tyder på at det er en viss samvariasjon mellom disse typene.

I fremstillingen har det vært nødvendig å bruke noen matematiske symboler og en del elementære begreper fra sannsynlighetsregning og statistisk metodelære. Disse er søkt forklart i teksten og i appendiks.

Prosjektet har vært finansiert av midler fra jubileumsgaven fra Norges Bank til Statistisk Sentralbyrå.

Forfatteren ønsker på denne måte å takke en rekke kolleger for kommentarer, spesielt Birgit Eie, Petter Laake, Tore Schweder, Ib Thomsen, Tor Haldorsen, Rolf Aaberge og Per Sevaldson.

Statistisk Sentralbyrå, Oslo, 16. mai 1984

Arne Øien

I N N H O L D

TREDJE DEL. Dataanalyse.

FJERDE DEL: Avhengighetsmål.

FØRSTE DEL:

Innledning og grunnleggende forutsetninger

1. INNLEDNING

1.1. Bakgrunn.

I de fleste undersøkelsene Statistisk Sentralbyrå foretar, blir det observert flere kjennemerker for hver observasjonsenhet (telleenhet). Det kan være yrke, alder, utdanningsnivå, barnetall, husholdningens sammensetning, bosted osv., osv. Ser vi på hver kjennemerkeverdi som verdien av en variabel, så kan en god del av disse variablene kalles kategoriske, de plasserer telleenhetene i ulike kategorier (se avsnitt 1.2 om typer av variable).

Enten vi skal fremlegge resultatene av undersøkelsen i form av tabeller, eller vi skal bruke data til å foreta en analyse av et problem, så vil et hovedspørsmål være om det er samvariasjon eller ikke mellom de variable. Hvis vi kan finne ut noe om dette, kan det hjelpe oss ved valg av tabeller. Og i analyse av et problem er det jo nettopp samspill, eller mangel på samspill mellom de variable som vil være interessant. I en del tilfeller ønsker vi et eller annet slags mål for samvariasjonen.

Å finne ut om det er samspill mellom to variable, f.eks. yrke og helse, eller alder og helse, kan være forholdsvis enkelt. Med tre eller flere variable blir det vanskeligere. Hvordan blir bildet når vi ser på yrke, alder, helse, og kanskje flere variable, simultant ? Om vi ønsker å se spesielt på hvordan én variabel, f.eks. helse, varierer med de øvrige, så er dette greit hvis denne ene variable (helse) har intervall skala (jfr. 1.2), f.eks. tallet på sykedager i et år. Da kan vi kanskje bruke regresjonsanalyse. Men hvis den variable har en nominell skala, sykdomsgruppe f.eks., vil andre metoder være bedre.

Når vi skal undersøke samvariasjonen mellom flere variable uten å se på en av dem som funksjon av de andre, må vi også finne frem til egnede metoder. Metoden vil være avhengig av hva slags problem og hva slags data vi har. Vi må gjøre klart hva vi vet om

problemet på forhånd, og hva vi ønsker å finne ut. Gjelder det å bestemme hvilke variable vi bør sortere observasjonene etter når vi setter opp en tabell ? Ønsker vi å undersøke om det er samvariasjon mellom lønnsnivå og kjønn når vi samtidig kontrollerer for alder, utdannelse, yrke m.m. ?

Det finnes ulike problemstillinger i mengdevis, og det finnes statistiske metoder som kan være til hjelp i analysen av mange av dem, men kanskje ikke i alle.

Metodeutviklingen på dette feltet har skutt særlig fart de siste 10 - 15 årene. Det henger sammen med utviklingen av de elektroniske regnemaskinene, som gjør det mulig raskt å håndtere store datamengder og sortere dem etter mange kjennetegn, så vi kan få frem data for analyse av sammenhenger i mange dimensjoner.

I arbeidsprogrammet for 1980 for metodegruppen er nevnt det store arbeidet gruppen har gjort med utvikling av metoder for analyse av kategoriske data, ikke minst i tilknytning til professor Sverdrups innsats på dette område. Disse metodene er en videreutvikling og forbedring av en rekke metoder som begynner å bli kjent, men kanskje ikke godt nok. Det varer ofte en stund før nye metoder rekker frem til dem som har bruk for dem. Dels vil det alltid ta tid før nye metoder kommer med i lærebøker og oppslagsverk, og dels er det vanskelig for brukerne å holde seg a jour med alt det nye som kommer.

Hensikten med dette notatet er å gi en oversikt over en rekke av de metodene som kan tenkes å være nyttige for arbeidet i Byrået. Det er ment som en "veiviser" og gir bare en meget kortfattet omtale av hver metode, med henvisninger til hvor en kan finne en mer utførlig fremstilling, samt beregningsmuligheter m.m.

Det er så mange ulike problemstillinger som skal analyseres at det her ikke er mulig å få med metoder for mer enn de vanligste variantene. Dermed vil problemer som trenger svært spesielle metoder, ikke bli dekket. Det kan være vanskelig for en bruker å se at et problem er av denne typen. Det er derfor en god regel å ta kontakt med en trenet metode-statistiker for å diskutere problemstillingen og kanskje få råd, eventuelt videre samarbeid om metodevalget, på et tidlig tidspunkt i planleggingen av en undersøkelse.

1.2. Observasjoner, ulike typer av variable.

Det er karakteristisk for den typen observasjonsmaterialer vi skal behandle her, at de variablene (kjennemerkene, egenskapene) vi observerer, har - eller kan tilordnes - verdier som plaserer telleenhetene i ulike kategorier. Variablene kan i mange tilfeller ikke betraktes som kontinuerlige, dvs. de kan ikke anta verdier som ligger vilkårlig nær hverandre, slik som f.eks. høyde eller avstand. De variable er diskrete, dvs. at verdiene er adskilt ved intervaller på tallstreken. Noen variable har en nominell skala, dvs. at det ikke finnes noen naturlig rangordning av verdiene. For variabelen yrke f.eks., så kan vi nok nummerere yrkesgruppene, men det er mange måter å gjøre det på og det er ingen "naturlig" ordning av dem, en håndverker er ikke "større" eller "mindre" enn en fagarbeider eller en professor på yrkesskalaen.

For utdanning derimot kan vi ha en rangordning når vi definerer og ordner nivåene på en passende måte og ikke lar dem "grene ut" i forskjellige retninger fra et visst nivå av. Vi kan nummerere kategoriene (nivåene) med f.eks. 1,2,3, osv. på en ordinal skala, der vi ikke tillegger selve tallet noen egentlig mening og heller ikke differensen mellom to tall. Tallene tjener til å ordne kategoriene, og vi kan si at nivå nr. 3 kommer etter nr. 2, som igjen kommer etter nr. 1 i rangordning (eventuelt før i begge tilfeller).

En intervallskala ordner verdiene og dessuten har differenser mellom verdiene god mening. Det finnes imidlertid ikke noe "naturlig" nullpunkt, derfor har forholdstall mellom verdiene liten mening. For temperatur f.eks. vil 20° C nok være dobbelt så høy som 10° C, men ikke "dobbelt så varm", og med Fahrenheit skala blir de tilsvarene temperaturene 68° og 40°, altså et annet forholdstall.

På en forholdstallskala vil både ordning, differenser og forholdstall mellom verdier ha mening. 4 barn er dobbelt så mange som 2 barn og alder 15 år er halvparten av 30 år.

Den store mengde av statistiske metoder er utviklet for variable med forholdstalls- og kanskje intervallskala. Det var

lenge et meget dårligere utvalg for nominalskalavariable, især når det gjelder samvariasjonsproblemer, men dette endrer seg nå ganske raskt. Vi skal forsøke å formidle et inntrykk av dette.

1.3. Ordning av data, krysstabeller (hyppighetstabeller, kontingenstabeller).

Eksempler

I arbeidet med å finne frem til en egnet analysemetode gjelder det først og fremst å få problemstillingen klart frem, og dessuten å ha oversikt over data. For det siste kan det være en god hjelp å tenke seg de observerte telleenhetene ordnet i en simultan fordeling med absolutte hyppigheter ordnet(gruppert)etter de verdiene hver av de ulike variablene kan anta. Vi får en simultan hyppighetsfordeling, dvs. en tabell over tallet på telleenheter for hver av de mulige verdikombinasjonene. Disse er definert slik at de utelukker hverandre. En telleenhet vil altså høre til under én og bare én verdikombinasjon. For to variable får vi en toveistabell som tabell 1.1. nedenfor. Den har tre linjer, svarende til de tre verdiene for variabel nr. 1, arbeidstidsordning, og to kolonner, svarende til to verdier for variabel nr. 2, yrkesutdanning, og kan kalles en 3x2-tabell. Med I verdier (linjer) for variabel nr. 1 og J verdier for nr. 2 får vi en I x J-tabell. Har vi dessuten en tredje variabel med K verdier, blir det en IxJxK-tabell. Vi setter linjesummer (marginaltall) til høyre, og kolonnesummer (marginaler) nederst i tabellene. Vi bruker altså ikke de konvensjonene for dette som nå brukes i Byråets offisielle statistikk.

Tabell 1.1. Lønnstakere etter yrkesutdanning og arbeidstidsordning. Fra SØS 113.

		Yrkesutdanning		
Arbeidstid		Med	Uten	Sum
	i j =	1	2	(marginal)
Skift	1	121	82	203
Natt	2	19	6	25
Dag	3	612	480	1092
Sum (marginal)		752	568	1320

Slike hyppighetstabeller som skal brukes i analyse av samvariasjon mellom variable kalles ofte kontingenstabeller (contingency tables). Vi skal kalle dem enten krysstabeller eller komparative tabeller, se nedenfor. Vi skal gi flere eksempler senere, også med flere enn to variable. En må selvsagt bruke de vanlige "knepene" for å sette opp tabeller i flere enn to dimensjoner. For oss er det ikke akkurat det å sette opp tabellen som er det viktigste (unntatt hvis vi ønsker et rent visuelt inntrykk av den), men at vi har en systematikk som lar oss referere til tallene på en entydig måte. De problemstillingene vi skal analysere, kan være meget forskjellige.

Vi har ulike varianter av toveis- og flerveistabeller, alt etter hvordan observasjonene er fremkommet. Vi vil i dette notatet bruke navnet krysstabell når vi som i eksemplet ovenfor har ett utvalg på n telleenheter, som så er ordnet etter ulike variabelkombinasjoner, slik som tabell 1.1. Her kan n enten være et tall som er bestemt på forhånd, eller det kan være fremkommet tilfeldig, fordi tellingen er avgrenset på annen måte enn ved utvalgets størrelse. I en trafikkontroll som går over et visst tidsrom, kan totaltallet, n, på undersøkte biler være tilfeldig. For undersøkelser med frafall kan vi i mange tilfelle regne som om n var gitt på forhånd, selv om dette tallet er lavere enn det vi planla å få. I alle tilfelle kan alle andre tall enn n betraktes som tilfeldige.

Vi kan også ha toveistabeller satt sammen av enveistabeller fra flere separate utvalg. Hensikten er å foreta en sammenlikning. Det er nærliggende å kalle dem komparative tabeller. I tabell 511 i Statistisk årbok 1979 finner vi bl.a. tallet på skip i de nordiske lands handelsflåte fordelt etter alder. (I tabell 1.2 utelater vi Island.)

Tabell 1.2. Handelsskip på 100 bruttotonn og over, i 4 nordiske land pr. 31/12 1978, fordelt etter alder.

Aldersgruppe År	i	j =	Norge 1	Danmark 2	Sverige 3	Finland 4	Sum
< 5	1		501	273	121	64	959
5-9	2		434	180	114	58	786
10-14	3		302	228	81	58	669
15-19	4		143	91	36	40	310
20-24	5		125	50	27	21	223
25-29	6		67	32	21	8	128
≥ 30	7		228	101	113	87	529
Sum			1800	955	513	336	3604

I denne tabellen må vi betrakte totaltallene for hvert land som gitt, og dermed er n = 3604 gitt.

Det finnes også toveistabeller der vi må betrakte begge marginalene som gitt, men de blir svært spesielle.

Flerveistabeller kan være enten rene krysstabeller, eller en blanding av krysstabeller og komparative tabeller, eller rene komparative tabeller (ikke så ofte). Vi kommer tilbake til dette senere.

I eksemplene i dette notatet har vi lånt og tilpasset data fra diverse undersøkelser i Byrået for å bruke dem til å illustrere ulike metoder og problemstillinger. Data blir her brukt uten særlig hensyn til opprinnelig utgangspunkt, problemstilling eller analyse. Våre tall kan også avvike noe fra de opprinnelige. Vi gjør dette bl.a. for å spare plass. Vi tenker oss altså at data er "nye" for hver gang vi bruker dem til å illustrere et nytt problem.

1.4. Hovedinndeling av metodene. Forutsetninger

Det har vært vanskelig å bestemme hvordan dette notatet burde legges opp. Det er bare delvis "naturlige ordninger" av metodene, så rekkefølgen gir seg ikke uten videre av seg selv. Vi skal imidlertid ta utgangspunkt i to prinsipielt forskjellige overordnede problemstillinger for analyse av data, statistisk inferens og dataanalyse, som i det vesentlige vil bli behandlet hver for seg. Enkelte fremgangsmåter kan komme inn under begge kategorier, men med noe ulik tolking av resultatene. Vi skal forsøke å skissere de to problemstillingene.

1.4.1. Statistisk inferens

Vårt utgangspunkt er her at vi skal tolke vårt observasjonsmateriale i relasjon til et forholdsvis godt strukturert problem. Vi vet en god del om det på forhånd, enten i form av en teori, eller fordi vi av tidligere erfaring mener at det er visse forhold, f.eks. sammenhenger mellom variable, som vi tør gå ut fra som gitt. På grunnlag av denne a priori viten kan vi formulere en modell der en eller flere av de variable opptrer som stokastiske (dvs. at vi postulerer en sannsynlighetsfordeling for dem, jfr. avsnitt 2.1.). Økonometriske modeller er et godt eksempel, men

modellene behøver ikke være så omfattende eller så eksplisitte, det skal vi se etter hvert.

Vi vil så bruke data til å tallfeste, estimere, visse ukjente parametre i modellen, eller til å teste hypoteser om dem, eller kanskje til å få et grunnlag for å komme med prediksjoner om fremtidige verdier av de variable. Visse andre ønskemål kan også være aktuelle, f.eks. hvis data skal være grunnlag for å treffe beslutninger (statistisk desisjonsteori).

Poenget er altså at vi formulerer den stokastiske modell for problemet ut fra de forhåndsopplysninger vi har. Så utleder vi ved hjelp av statistisk teori en metode for testing, estimering e.l., som følger av modellen. *Først når dette er gjort ser vi på tallmaterialet*, og bruker det til å utføre testen, estimeringen e.l.

Det er altså nødvendig å tenke gjennom problemstillingen og formulere den før vi bruker statistisk teknikk, og teknikken er bestemt av problemet, ikke av de observerte tallene. For enhver statistisk metode gjelder visse forutsetninger om fordelingen av de variable, *metoden bestemmes av vårt a priori kjennskap til problemet og måten observasjonene er tatt på.*

1.4.2. Data analyse

I en del situasjoner er vårt problem lite strukturert, vi vet lite om eventuell samvariasjon på forhånd. Kanskje ønsker vi å skille interessante fra mindre interessante variable når det gjelder valg av tabeller til publisering. Eller vi vil bruke datamaterialet til å få frem om, og hvilken, samvariasjon kan være verd å undersøke nærmere, f.eks. ved en ny undersøkelse. Også i dette tilfellet kan vi kanskje postulere en sannsynlighetsfordeling for de variable, men fordi vi har liten a priori viten, vil vi ha en lite eksplisitt formulering. Vi kan ikke formulere hypoteser o.l. på samme måten som under I. Vi må nøye oss med en beskrivelse av data, uten å kunne trekke konklusjoner ut over dette.

1.4.3. Valg av problemstilling

Noen hevder at vi alltid skal "la tallene tale for seg selv", som i 1.4.2, og ikke formulere mer eksplisitte modeller (1.4.1) som vi likevel ikke kan verifisere helt ut. Det er vanskelig å se hvordan vi i så fall kan bygge opp en teori og få dypere viten og innsikt. Ved bruk av statistisk induksjon bygger vi på tidligere erfaring og kan si (med en viss, liten sannsynlighet for å ta feil): Gitt at modellen er riktig (i store trekk) så tyder data på at det er

samvariasjon tilstede (vi har forkastet nullhypotesen om ingen samvariasjon). Så kan vi gå videre ut fra dette. I data analyse må vi si: ja, disse data kan tyde på samvariasjon, men vi vet ikke om det skyldes tilfeldigheter i dette materialet eller om det ligger mer autonome sammenhenger bak. I lærebøker og artikler finner vi også målsettinger uttrykt ved: "We want to fit a model that describes the data", eller "-- which model gives the best representation of the data ?" Med dette menes vel å merke ikke at en har a priori grunn til å forsøke ulike modeller, men at en starter med en meget generell, lite spesifisert sannsynlighets-modell (som f.eks. en mettet modell som nevnt i avsnitt 2.2.), og så forsøker å redusere tallet på parametre i denne.

Men hvilken interesse har det å finne den "modellen som stemmer best med data" ? Det kan bare ha mening hvis "det ligger noe bak" som gjør at nettopp den modellen er rimelig, den kan tolkes på en fornuftig måte i det problemet vi har. Det er jo i og for seg ikke nødvendig med noen "modell" hvis vi bare vil ha frem akkurat de tallene vi har observert. Vi ønsker å tolke tallene. Vi samler jo ikke inn tallene hvis vi ikke tror det ligger "noe lovmessig" bak dem, som vi gjerne vil belyse nærmere. Dette "noe" må inn i analysen, tallene kan ikke gjøre jobben alene.

Ofte vil våre problemer og data ikke kunne henføres til bare den ene av de to problemstillingene. Vi kan trenge noe fra begge felt. Og her er det av interesse å søke å innrette seg slik at f.eks. de dataanalysemetodene vi bruker kanskje kan inngå i en analyse med statistisk inferens. Vi har muligheter for dette i multivariabel analyse i dag, ved metoder som bl.a. Erling Sverdrup, Harald Goldstein og andre har utarbeidet.

Når det gjelder å få et godt utbytte av en statistisk analyse, står vi oss på å søke å utnytte det vi har av a priori viten om problemet i modellformuleringen. Det er ikke lett, vi må ha teoretisk og praktisk innsikt på det fagområde undersøkelsen gjelder enten det nå er biologi, demografi, geografi, sosiologi, sosialøkonomi eller annet. Dessuten må vi utvikle øvelse i å formulere hensiktsmessige statistiske modeller. Samvirke mellom "fagkompetanse" og "statistikerkompetanse" er essensiell her.

1.5. Veiviser.

I kapittel 2 gir vi en kortfattet oversikt over teoretiske begreper vi vil trenge under omtalen av metodene. Begrepene gjelder spesielt annen del, om inferensmetoder, men vi har også bruk for en del av dem i tredje del, om dataanalyse, og fjerde del, om avhengighetsmål. En leser som ikke er fortrolig med de statistiske begrepene vi bruker, vil finne en litt mer utførlig omtale i appendiksene. Vi forsøker også å uttrykke verbalt det symboler og formler står for.

Annen del omfatter kapitlene 3 - 12. I 3 - 7 behandler vi stort sett tabeller der vi kan ha tall i alle ruter, dvs. alle kombinasjoner av variabelverdier kan forekomme, men gjør det ikke alltid. I 3 behandler vi gjengse metoder for toveisgrupperte data. Vi går litt mer utførlig til verks i dette kapittel enn i de to følgende. Riktignok er problemene enklere her, men det er også lettere å finne løsninger på dem, og de kan tjene som en introduksjon og et grunnlag for kapittel 4 om treveisgrupperte data og kapittel 5 om flerveisgrupperte data.

Metoder som kan brukes for situasjonene i kapitlene 3 - 5 vil en også finne i kapittel 6, om log-lineære modeller og i kapittel 7, der vi ser på én variabel som funksjon av de øvrige.

I kapittel 8 er det tatt med noen eksempler på spesielle problemstillinger, bl.a. for undersøkelser med parvis sammenlikning av data.

Kapittel 9 dreier seg om såkalt strukturelt ufullstendige modeller, der visse kombinasjoner av kjennetegn ikke kan forekomme, dvs. at vi ikke kan ha tall i alle rutene i tabellen.

I kapittel 10 forsøker vi å skissere metoder det især kan lønne seg å bruke i store datamaterialer. Jan Bjørnstad [B 1] og Harald Goldstein [1981] har foreslått nye måter å gå frem på i slike situasjoner.

Det er både likheter og forskjeller mellom mange av de metodene vi har omtalt. I kapittel 11 har vi noen kommentarer og en oversikt i forbindelse med et eksempel.

I kapittel 12 nevnes kort en del spesielle problemstillinger og metoder som vi ikke går nærmere inn på.

Tredje del, kapittel 13, omfatter metoder for data analyse. Det kommer stadig forslag til nye fremgangsmåter og dekningen av dette stoffet er ikke særlig utførlig. Vi ser på noen metoder som er forholdsvis etablerte.

I fjerde del, kapittel 14, gir vi for fullstendighets skyld en oversikt over summariske mål for samvariasjon mellom to variable. Det finnes ganske mange av dem og de har vært meget brukt i mer elementær statistisk analyse. De blir gjerne beregnet når en bruker de store statistikk-EDB-programpakkene til å skrive ut krysstabeller.

Appendiksene gir en kortfattet oversikt over noen enkle sannsynlighetsteoretiske begreper og sammenhenger.

2. GRUNNLEGGENDE FORUTSETNINGER OM DATAMATERIALET. SANNSYNLIGHETS-MODELLER

Det ligger alltid visse forutsetninger om observasjonsmaterialet og om de forhold som har generert det, til grunn for de statistiske metodene vi bruker. En del forutsetninger vil være felles for de fleste, eller for store grupper, av de metodene vi skal se på. I dette kapitlet skal vi trekke fram noen slike forutsetninger av ulik karakter.

Det er tilrådelig å ha dem i tankene ved planleggingen av data-innsamlingen, slik at valget av statistisk metode kan lettes.

Vi må faktisk ha et visst kjennskap til statistiske metoder og bruken av dem for å kunne gi hele problemet en utforming som kan føre til en brukbar statistisk analyse.

Det er lite fruktbart å beskrive statistiske metoder uten å bruke enkelte begreper fra det grunnlaget statistiske metoder bygger på, nemlig sannsynlighetsregningen. I dette notatet skal vi ikke bruke mer av teoretiske begreper enn høyst nødvendig for å formulere forutsetninger og kriterier. De fleste lesere vil for-håpentlig kunne følge framstillingen. For dem som er usikre, forsøker vi i appendiks A å gi en kortfattet beskrivelse av det nødvendige grunnlaget, samt en del formler som det blir vist til i enkelte avsnitt.

For å presisere forutsetninger og metoder er vi nødt til å ta med noen formler i teksten. I den utstrekning det er mulig, prøver vi også å gi en verbal presisering.

2.1 Hvordan er tallene fremkommet, og hva skal de brukes til? Betydningen for modell og metode.

Når vi skal tolke et observasjonsmateriale i relasjon til et gitt problem, kan arbeidsgangen skisseres slik: Vi formulerer en stokastisk modell ut fra det vi vet på forhånd om problemet (her kan både teori og empiri inngå) og dessuten om hvordan data er fremkommet. Ut fra dette forsøker vi så å finne en statistisk metode for tolkningen.

I formulering av modellen inngår det å bestemme hva vi skal betrakte som stokastiske variable og hva vi kan si om sannsynlig-hetsfordelingen for dem. Dette siste er enklere å gjøre hvis vi kan postulere at data er et *tilfeldig utvalg* i teoretisk forstand. Denne betegnelsen brukes imidlertid i ulike betydninger.

i) Den enkleste formen er et tilfeldig utvalg fra en eksisterende populasjon, dvs. et utvalg på n telleenheter (f.eks. personer eller familier) som er trukket på en slik måte at alle kombinasjoner av n telleenheter som kan dannes i populasjonen har samme sannsynlighet for å bli trukket ut. Det er viktig å ta utvalget på denne måten *hvis formålet med undersøkelsen er å trekke slutninger om visse data i populasjonen*, data som vi kunne observere hvis vi undersøkte hele populasjonen, som f.eks. antall hybelboere pr. 1/11-1980 eller antall hus bygd før 1945.

Vi vet at det er mange andre måter å ta utvalg på fra en endelig populasjon, jfr. Byråets utvalgsplan. Poenget er at de er *sannsynlighetsutvalg*, slik at vi har kontroll over sannsynlighetsfordelingen (teoretisk sett) for våre variable, og dermed har grunnlag for å utlede metodene.

ii) Betegnelsene utvalg og populasjon er blitt overført til observasjonsmaterialer av andre typer. Situasjonen er ofte at vi ønsker å trekke slutninger ut over observasjonsmaterialet. Dette kan vi gjøre når det har mening å postulere en *hypotetisk populasjon* av personer, familier e.l., med tilsvarende bakgrunn som den gruppen vi undersøker mhp. faktorer som kan spille en rolle. Her er det ikke spørsmål om gruppen er *trukket* tilfeldig, men spørsmål om den kan betraktes som et utsnitt fra en populasjon vi ønsker å generalisere til, og om variabelsettene for de forskjellige enhetene i gruppen kan antas å være stokastisk uavhengige: Vil virkningen av en behandling på hver enkelt pasient være den samme uansett hvordan virkningen er på de øvrige ? Vil graden av atferdsvansker hos ett barn være upåvirket av atferdsvanskene hos de andre barna i undersøkelsen ?

"Hypotetisk populasjon" er et hjelpebegrep. Det vi gjør, er å formulere *en sannsynlighetsmodell* for problemet, og denne modellen danner grunnlaget for utledningen av den statistiske metoden. For visse formål, f.eks. for sammenlikningen mellom land, kan vi postulere at selv dataene om Norges handelsskip i tabell 1.2 er et utvalg fra en teoretisk populasjon, med en gitt men ukjent teoretisk aldersfordeling, jfr. avsnitt 2.4.3.

Ett og samme observasjonsmateriale kan faktisk tolkes på ulike vis, avhengig av problemstillingen og hvordan observasjonene er tatt. Anta at vi har observert 10 lønnstakere med en viss bakgrunn, og funnet at 9 av dem arbeider dagskift.

Situasjon 1): Vi har ingen teori på forhånd f.eks. om at denne kategorien lønnstakere oftest arbeider dagskift, sammenliknet med andre. Vi rapporterer hyppigheten 0,9 som et "funn", som må verifiseres ved nye undersøkelser.

Situasjon 2): De 10 lønnstakerne er trukket tilfeldig fra en større gruppe med samme bakgrunn. Da kan vi estimere hyppigheten av å arbeide dagskift i denne gruppen til 0,9, eller angi et konfidensintervall for hyppigheten.

Situasjon 3): Vi har en teori om at lønnstakere med denne bakgrunn oftest arbeider dagtid, eller kanskje om at de oftere arbeider dagtid enn lønnstakere med en annen bakgrunn. Enten vi nå har trukket tilfeldig eller ikke, kan vi ut fra vårt kjennskap til situasjonen kanskje postulere stokastisk uavhengighet mellom dem mht. forekomsten av dagtidsarbeid og bakgrunnsvariable. Vi kan da estimere eller teste hypoteser om sannsynligheten for dagtidsarbeid for lønnstakere med denne bakgrunn og muligens få verifisert utsagn som teorien medfører.

Vi må imidlertid være forsiktige. Hvis 6 av disse lønnstakerne kommer fra et lite sted med én bedrift der alle arbeider dagtid, så må vi nok formulere modellen noe anderledes.

For arbeidet i Byrået er det nyttig at både data fra utvalgsundersøkelser og fra fullstendige tellinger, i visse sammenhenger f.eks. for analyseformål, ofte kan betraktes *som om* de er observasjoner av stokastiske variable med visse sannsynlighetsfordelinger, jfr. skipsaldereksemplet.

Når det gjelder utvalgsdata, vil postulering av en "superpopulasjonsmodell" være nyttig, se f.eks. Thomsens IO 76/28. Da ser vi på selve den eksisterende populasjonen som et utvalg fra en hypotetisk "superpopulasjon", og dermed blir også det utvalget vi har trukket, et utvalg fra denne superpopulasjonen. Dvs. at vi kan ha en teoretisk modell for problemet som ikke behøver å ha noe med selve trekkingen av utvalget å gjøre.

Avhengighet i data.

Vi kan ha et problem og et datamateriale der vi ikke kan postulere stokastisk uavhengighet, f.eks. om vi har visse typer av tidsrekkedata, der den verdien vi observerer på tidspunkt t kan avhenge av verdiene vi har observert på tidligere tidspunkter. Vi vil fremdeles kunne finne statistiske metoder så sant vi kan presisere hva slags avhengighet det dreier seg om. Vi skal imidlertid ikke ta opp slike problem i dette notatet.

2.2. Vanlige sannsynlighetsmodeller for krysstabeller og komparative tabeller (hyppighetstabeller).

Våre multivariable problemer og data kan gjelde høyst forskjellige forhold som krever temmelig ulike metoder. Likevel kan det lønne seg å etablere en felles terminologi og symbolbruk for datamaterialet. Vi skal innføre en del symboler i denne forbindelse.

2.2.1. En toveistabell.

La oss se på en hyppighetstabell som den vi har i tabell 1.1. De n = 1320 observasjonene er gruppert etter to variable, y (arbeidstidsordning) og x (yrkesutdanning). Vi har en toveistabell, mer spesielt en 3 x 2-tabell. Her har y tre kategorier nummerert i = 1,2 og 3, mens x har to kategorier nummerert j = 1 og 2. Tallet på observerte telleenheter (lønnstakere) som er i kategori (linje) nr. i for y og kategori (kolonne) nr. j for x vil vi kalle n_{ij} for en vilkårlig toveistabell. Tabell 2.1.a er en 3 x 2-tabell over symboler, som svarer til tabell 1.1.

I eksemplet er altså $n_{11} = 121$, $n_{21} = 19, \ldots, n_{32} = 480$.

Tabell 2.1.a. Toveistabell over antall, n_{ij}.

y \ x	j=1	2	Marginal (sum) n_{i+}
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
3	n_{31}	n_{32}	n_{3+}
Marginal (sum) n_{+j}	n_{+1}	n_{+2}	n

Tabell 2.1.b. Toveistabell over sannsynligheter, p_{ij}.

y \ x	j=1	2	marginal p_{i+}
1	p_{11}	p_{12}	p_{1+}
2	p_{21}	p_{22}	p_{2+}
3	p_{31}	p_{32}	p_{3+}
Marginal p_{+j}	p_{+1}	p_{+2}	1

I tabellen har vi dessuten med marginal-tallene, d.v.s. summen $n_{i+} = n_{i1} + n_{i2}$ for i = 1,2 og 3, samt $n_{+j} = n_{1j} + n_{2j} + n_{3j}$ for j = 1 og 2, og totalsummen n.

La oss postulere at det ligger visse sannsynligheter "bak" tallene i tabellen, i følgende forstand: I de fleste av de problemene vi skal se på, kan vi postulere (stokastisk) uavhengighet mellom variabelkombinasjonene (i,j) fra telleenhet til telleenhet, jfr. avsnitt 2.1. Vi kan dessuten anta at alle telleenheter har samme sannsynlighet, p_{ij}, for å gi en observasjon i rute (i,j) i tabellen. Da kan vi tenke oss en tabell som 2.1.b over disse sannsynlighetene, samt marginalsannsynlighetene $p_{i+} = p_{i1} + p_{i2}$ for i = 1,2 og 3, og $p_{+j} = p_{1j} + p_{2j} + p_{3j}$ for j = 1 og 2. Videre er summen av alle sannsynlighetene i tabellen, d.v.s. $\sum_i \sum_j p_{ij} = \sum_j p_{+j} = 1$, idet vi skal ha tatt med alle de variabelkombinasjoner som kan forekomme. Disse sannsynlighetene vil i alminnelighet være ukjente, men det vi måtte vite a priori, og de hypotesene vi har, vil kunne uttrykkes som utsagn om, eller restriksjoner på, p_{ij}-verdiene. Vi kan f.eks. ha en hypotese om at alle (i,j)-kombinasjoner er like sannsynlige, da er $p_{ij} = \frac{1}{6}$ i tabell 2.1.b. Eller vi mener at de to sannsynlighetene i siste linje er like, dvs. at $p_{31} = p_{32} = 0{,}5\ p_{3+}$. Mer interessant er hypotesen om uavhengighet mellom de to variable, y og x. Vi kan også ha modeller der p_{ij}-verdiene er funksjoner av et mindre antall parametre, se f.eks. kapittel 6.

2.2.2. Strukturelle nuller

I noen problemer vil det være enkelte variabelkombinasjoner som ikke kan forekomme. Vi må da sette $p_{ij} = 0$ for hver slik kombinasjon, og vil også ha $n_{ij} = 0$. Vi sier at vi har strukturelle nuller i tabellen.

I omtalen av stokastisk uavhengighet nedenfor forutsetter vi alle $p_{ij} > 0$. Se avsnitt 2.6 og kapittel 9 om problemer med strukturelle nuller.

2.2.3. Stokastisk uavhengighet.

La oss først se på eksempel 1.1. Vi vil sammenligne fordelingen etter arbeidstidskategori for lønnstakere med og uten yrkesutdanning. Vi regner ut den relative hyppighet av å være i kategorien i = 1,2 og 3, særskilt for dem med yrkesutdanning, dvs. 121/752 = 0.161 osv., og dem uten, dvs. 82/568 = 0.144 osv., samt den marginale fordelingen, og setter disse opp i tabell 2.2.a.

Tabell 2.2.a. Fordeling etter arbeidstidsordning særskilt for lønnstakere med og uten yrkesutdanning, samt marginalt.

y	i	Betingede fordelinger x: Med j = 1	Uten 2	Marginal fordeling
Skift	1	0.161	0.144	0.154
Natt	2	0.025	0.011	0.019
Dag	3	0.814	0.845	0.827
Sum		1	1	1

Vi kan her sammenlikne den betingede hyppighetsfordelingen for dem med yrkesutdanning med den betingede fordelingen for dem uten, samt med den marginale. Vi ser at de tre fordelingene er nokså like. Det kan se ut som om arbeidstidsordning er nesten uavhengig av yrkesutdanning (bare den lille gruppen med nattskift avviker noe).

Hvis vi ser på de betingede fordelingene etter utdanningsgruppe for hver arbeidstidsordning, vil vi få et liknende bilde, "nesten-uavhengigheten" mellom de variable er gjensidig, jfr. tabell 2.2.b.

Tabell 2.2.b. Betinget fordeling etter yrkesutdanning for hver arbeidstidsordning, samt marginalt.

y	i	x: Med j = 1	Uten 2	Sum
Betingede	1	0.60	0.40	1
forde-	2	0.76	0.24	1
linger	3	0.56	0.44	1
Marginal fordeling		0.57	0.43	1

De betingede hyppighetene for gruppe i, gitt j, er $\frac{n_{ij}}{n_{+j}}$.

I den teoretiske modellen trenger vi en mer presis definisjon av uavhengighetsbegrepet. På analog måte som hyppighetene definerer vi de betingede sannsynlighetene for i, gitt j, som

$$P(\text{kategori i for y}|\text{gitt kategori j for x}) = p_{i|j} = \frac{p_{ij}}{p_{+j}} \text{ for } i = 1,2,3.$$

Her betyr P(...) sannsynligheten for innholdet i parentesen.
Vi har da at summen av alle de betingede sannsynlighetene i kolonne j er lik 1, dvs.

$$p_{1|j} + p_{2|j} + p_{3|j} = \frac{p_{1j} + p_{2j} + p_{3j}}{p_{+j}} = 1.$$

Dette gjelder særskilt for j = 1, for j = 2 og for høyere verdier av j når det er flere kolonner. Vi ser at dette er helt analogt med det vi får for betingede relative hyppigheter, som i tabell 2.2.a. Størrelsene $p_{1|j}$, $p_{2|j}$ og $p_{3|j}$ angir den betingede sannsynlighetsfordelingen for y gitt kategori j for x.

Setning 2.2.3.

Vi sier at x og y er stokastisk uavhengige når alle de betingede sannsynlighetsfordelingene for y gitt x-kategoriene er like, og lik den marginale

$$p_{i+} = p_{i|1} = p_{i|2} \ (= \text{eventuelle flere } p_{i|j}) \text{ for } i = 1,2,3 \text{ osv.}$$

Vi ser at når dette er oppfylt, dvs. når vi har

$$\frac{p_{i1}}{p_{+1}} = \frac{p_{i2}}{p_{+2}} = p_{i+}, \text{ så er } p_{i1} = p_{i+} \cdot p_{+1},$$

$$\text{og } p_{i2} = p_{i+} \cdot p_{+2} \text{ for alle i.}$$

Vi har da $p_{ij} = p_{i+} \cdot p_{+j}$ for alle (i,j) i tabellen, dvs. at sannsynligheten for kombinasjonen (i,j) er lik produktet av de to marginale sannsynlighetene i linje i og kolonne j. Dette er multiplikasjonssetningen for variable som er stokastisk uavhengige av hverandre.

Når setning 2.2.3 gjelder, så blir også de betingede sannsynlighetene for hver j-kategori for x, gitt en viss i-kategori for y, like store og lik den marginale p_{+j}:

$$P(j \text{ for } x \mid \text{gitt } i \text{ for } y) = p_{j|i} = \frac{p_{ij}}{p_{i+}} = \frac{p_{i+}p_{+j}}{p_{i+}} = p_{+j} \text{ for alle } j,$$

og for hver i.

Det er denne spesielle strukturen i p_{ij}-tabellen vi tester når vi undersøker om våre data tyder på stokastisk uavhengighet mellom de variable i toveis-tabellen.

2.2.4. Kryssproduktforholdet i en 2 x 2 -tabell

For en 2 x 2 -tabell betyr altså uavhengighet at vi har

$p_{11} = p_{1+}p_{+1}$, $p_{12} = p_{1+}p_{+2}$, $p_{21} = p_{2+}p_{+1}$, og $p_{22} = p_{2+}p_{+2}$.

Det er nok at vi vet at én av likhetene holder, da gjelder også de øvrige.

Vi ser f.eks. at likheten holder for p_{12} slik:

Vi vet at $p_{1+} = p_{11} + p_{12}$. Hvis nå $p_{11} = p_{1+}p_{+1}$ så er

$$p_{12} = p_{1+} - p_{11} = p_{1+} - p_{1+}p_{+1} = p_{1+}(1 - p_{+1}) = p_{1+}p_{+2}.$$

Dette viser videre at

$$\frac{p_{11}}{p_{12}} = \frac{p_{21}}{p_{22}} \left(= \frac{p_{+1}}{p_{+2}}\right), \tag{2.2.4a}$$

og at det såkalte kryssproduktforholdet, α, er lik 1, dvs.

$$\alpha = \frac{p_{11}p_{22}}{p_{12}p_{21}} = 1 \tag{2.2.4b}$$

ved uavhengighet. Det omvendte gjelder også: α=1 medfører uavhengighet. Det er vanlig å bruke α som et mål for graden av avhengighet i en 2x2-tabell, se avsnitt 14.4.

For vilkårlig i og j i en IxJ-tabell kan vi se på f.eks.

$$\alpha_{ij} = \frac{p_{ij}p_{IJ}}{p_{iJ}p_{Ij}} \tag{2.2.4c}$$

Ved uavhengighet finner vi

$$\alpha_{ij} = \frac{p_{i+}p_{+j}p_{I+}p_{+J}}{p_{i+}p_{+J}p_{I+}p_{+j}} = 1$$

også her. Valget av "hjørnet" (I,J) som sammenlikningsbasis er ofte brukt. Vi kan selvsagt også se på f.eks.

$$\alpha_{ij;i'j'} = \frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}}, \tag{2.2.4d}$$

for hvert valg av (i'j').

2.2.5. Treveistabeller

Vi bruker helt tilsvarende symboler og terminologi når vi har flere variable i problemet, det blir like mange fotindekser på n og p som vi har variable.

Med tre variable, y, x og z, der

y har kategoriene 1,2,...,i,..,I,
x " " 1,2,...,j,..,J,
z " " 1,2,...,k,..,K,

har vi antallene n_{ijk} og sannsynlighetene p_{ijk} for kombinasjonen (i, j, k) i tabellen.

Tabell 2.2.c gir en treveisgruppering av data for 1977-78 fra Byråets fritidsundersøkelse. Her altså $n_{111} = 52$, $n_{112} = 46$, $n_{121} = 239$, $n_{122} = 249$,... $n_{322} = 163$,...,$n_{711} = 72$, ...,$n_{722} = 39$.

Tabell 2.2.c Treveistabell der n_{ijk} er tallet på personer gruppert etter antall helgeturer (i), adgang/ikke adgang til fritidshus (j) og kjønn (k).

Tall på helgeturer i		Adgang til fritidshus: Ja, j=1 Menn k=1	Ja, j=1 Kvinner k=2	Sum over k n_{i1+}	Nei, j=2 Menn k=1	Nei, j=2 Kvinner k=2	Sum over k n_{i2+}	Sum over j og k n_{i++}
0	1	52	46	98	239	249	488	586
1 - 2	2	45	39	84	141	158	299	383
3 - 5	3	68	78	146	141	163	304	450
6 - 9	4	51	53	104	80	86	166	270
10 -14	5	58	56	114	44	53	97	211
15 -19	6	42	26	68	29	31	60	128
20 -	7	72	74	146	51	39	90	236
Sum over i		388 n_{+11}	372 n_{+12}	760 n_{+1+}	725 n_{+21}	779 n_{+22}	1504 n_{+2+}	2264 $n_{+++}=n$
Sum over i og j		$n_{++1} = 1113$			$n_{++2} = 1151$			2264

Vi får flere sett av marginaler som kan defineres etter tur som

$$n_{ij+} = \sum_{k=1}^{K} n_{ijk}, \qquad n_{i++} = \sum_{j=1}^{J} n_{ij+}, \qquad n_{+++} = \sum_{i=1}^{I} n_{i++} = n,$$

$$n_{+jk} = \sum_{i=1}^{I} n_{ijk}, \qquad n_{+j+} = \sum_{i=1}^{I} n_{ij+} = \sum_{k=1}^{K} n_{+jk}, \text{ osv.}$$

Vi har f.eks. $n_{11+} = 52 + 46 = 98$, $n_{1++} = 98 + 488 = 586$, $n_{+12} = 46 + 39 + 78 + 53 + 56 + 26 + 74 = 372$, $n_{+1+} = 388 + 372 = 760$, osv.

Tilsvarende definerer vi p_{ij+}, p_{i++}, p_{+++}, p_{+jk} osv. ved summering av p_{ijk}-verdiene. Summen av alle p_{ijk}, altså p_{+++} er lik 1.

2.2.6. Flerveistabeller med mange variable

For et vilkårlig antall variable, f.eks. m stykker, skal vi skrive tallet på telleenheter i en "rute", dvs. for variabelkombinasjonen (i,j,...,g) som

$n_{ij...g}$, der i går fra 1 til I
j " " 1 til J
---
g " " 1 til G.

Den tilsvarende sannsynligheten kalles

$$p_{ij...g}$$

De ulike marginaler fremkommer ved summering, som ovenfor. Vi setter + på fotindeksplassen for den variable vi har summert over. (I mange tekster blir 0 eller · brukt i dette øyemed.)

2.2.7. Sannsynlighetsmodeller, tellevariable

Sannsynlighetene p_{ij} i tabell 2.1.b, respektive samlingen av alle sannsynlighetene $p_{ij...g}$ i en I x J x ....x G-tabell, angir sannsynlighetene for de mulige utfallene av én observasjon. Men de betyr også at det gjelder visse sannsynligheter for de ulike, mulige verdiene på tallene n_{ij}, resp. $n_{ij...g}$. Hvert av disse kan jo i prinsippet være lik et av tallene mellom 0 og n, med en viss sannsynlighet for hver verdi. Dessuten vil enhver mulig kombinasjon av tall for hele tabellen ha en viss sannsynlighet for å forekomme. Vi må se på hvert av tallene $n_{ij...g}$ som en verdi av en stokastisk variabel. For å sondre mellom disse variablene og de kategoriske variable i tabellen skal vi kalle $n_{ij...g}$ for tellevariable (om nødvendig) og bruke $n_{ij...g}$ som symbol også for den variable, ikke bare for tallet. Hver tellevariabel kan altså anta verdiene 0,1,2,....,n, men slik at summen av alle sammen alltid er lik n.

Samlingen av sannsynlighetene for alle de mulige verdikombinasjonene av de tellevariable i et problem utgjør den simultane sannsynlighetsfordelingen for de IJ, resp. IJ....G tellevariablene (n_{ij} eller $n_{ij...g}$) vi har i tabellen.

En slik simultan sannsynlighetsfordeling kan spesifiseres (bortsett fra ukjente parametre) når vi vet, dvs. kan postulere, hvordan observasjonene er fremkommet og når $p_{ij...g}$-tabellen betraktes som gitt (den behøver ikke være kjent). Vi skal se på de vanligst forekommende tilfellene.

En multinomisk fordeling (app. B 5) får vi når observasjonene for de enkelte telleenhetene kan antas å være stokastisk uavhengige av hverandre, samme $p_{ij...g}$-tabell gjelder for hver telleenhet og totaltallet n er gitt på forhånd. Siden n er gitt, vil alltid én av de IJ....G tellevariablene være lik differensen mellom n og summen av de øvrige (IJ ...G-1) tellevariablene. Fordelingen har derfor (IJ...G-1) dimensjoner (variable) når alle $p_{ij...g} > 0$. For eksemplet 1.1 har vi en (3 x 2 - 1) = 5 dimensjonal fordeling. Selv om n ikke er gitt på forhånd, f.eks. på grunn av frafall, så vil vi ofte kunne regne som om n er gitt. Det viktige er den stokastiske uavhengigheten fra telleenhet til telleenhet, og samme $p_{ij...g}$-tabell.

Dette gjelder også når vi har Poisson-fordelte variable (app. B 7), dvs. at betingelsene ovenfor er oppfylt, bortsett fra at n er en stokastisk variabel, jfr. eksemplet med antall biler i en trafikktelling i avsnitt 1.3. Den kan vises at den betingede fordelingen for en gitt n, av tellevariablene

vil være multinomisk også i dette tilfelle.

En flerdimensjonal (generalisert) hypergeometrisk fordeling (app. B 8) kan vi ha når observasjonene er trukket tilfeldig fra en endelig populasjon (uten tilbakelegging). For store nok utvalg og populasjon vil tilnærmingen til en multinomisk fordeling være god. Hvis vår analyse dessuten gjelder mer generelle forhold og ikke det å si noe om størrelser i selve populasjonen, vil oftest en multinomisk fordeling være en bra modell her også, jfr. avsnitt 2.1 og Thomsen [1976].

En produktmultinomisk modell (app. B 6) vil vi ha hvis vi har en komparativ tabell satt sammen av flere enveis-eller krysstabeller som er stokastisk uavhengige av hverandre og hver har multinomisk fordeling, jfr. eksemplet i tabell 1.2. Her vil summen av p_{ij} - (resp. $p_{ij...g}$) verdiene være lik 1 for hver enkelt deltabell.

Vi ser altså at en multinomisk fordeling kan postuleres i mange situasjoner. Den danner derfor grunnlaget for mange av de metodene som blir brukt.

Hvis vi har strukturelle nuller i problemet, kan vi fremdeles ha multinomisk fordeling, men med så mange færre parametre ($p_{ij...g}$) som vi har nuller. Hvis $p_{32} = 0$ i tabell 2.1.b, men alle andre $p_{ij} > 0$, så vil vi ha en fordeling i 3 x 2 - 1 - 1 = 4 variable, istedenfor i 5.

I en del tilfeller vil andre modeller være aktuelle, det vil bli tatt opp etter hvert.

2.2.8. Mettet og umettet modell

Vi sier at vi har en mettet modell for de variable i en kontingenstabell hvis vi bruker alle $p_{ij...g}$-verdiene (minus en, siden summen er 1) til å beskrive sannsynlighetsfordelingen. Vi trenger da (IJ...G-1) parametre for å beskrive fordelingen (eller dette tallet minus antall strukturelle nuller). I en mettet 3 x 2-tabell trenger vi altså $3 \cdot 2 - 1 = 5$ parametre.

Har vi restriksjoner på p_{ij}-ene, f.eks. ved at flere av dem er like, eller ved uavhengighet, jfr. setning 2.2.1, vil antall parametre i alminnelighet være lavere enn (IJ...G - 1). I en 3 x 2-tabell med uavhengighet mellom X og Y trenger vi (3 - 1) + (2 - 1) = 2 + 1 = 3 parametre. Det er ofte et ønskemål å beskrive fordelingen med et lite antall parametre, det vil gjerne bety at vi har god oversikt over problemet. Vi har da i alminnelighet en umettet modell.

2.3. Statistiske metoder. Estimering, testing, prediksjon

La oss gå ut fra at vi har formulert en stokastisk modell for vårt problem, slik at vi kan se på data som observasjoner av variable med en viss sannsynlighetsfordeling. Da kan "tolkingen av data i relasjon til problemet" omformes til spørsmål om hva vi kan slutte om sannsynlighetsfordelingen ut fra data eller om hvilke restriksjoner på parametrene som er forenelige med data. Vi kan ha ulike spørsmålsstillinger, bl.a. ut fra hvor mye vi tør si på forhånd om fordelingen.

Vi viser til lærebøkene på litteraturlisten når det gjelder en systematisk innføring i statistisk metodelære. Her skal vi bare kort omtale noen av de vanligste metodene.

2.3.1. Estimering

Vi er ofte interessert i størrelsen av sannsynligheter eller andre parametre i en fordeling, og vil estimere dem ut fra data. Vi bør da vite litt om estimeringsmetoder. Dette kan vi også trenge ved hypoteseprøving. Vi skal skissere noen metoder her.

Kanskje ønsker vi å estimere p_{ij}-verdiene i en krysstabell. For en mettet modell vil det ligge nær å bruke de relative hyppighetene

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \qquad \text{for } i = 1,2 \text{ og } j = 1,2 \qquad (2.3.1a)$$

Ut fra dette må vi sette

$$\hat{p}_{1+} = \frac{n_{1+}}{n}, \quad \hat{p}_{2+} = \frac{n_{2+}}{n} \text{ osv.} \qquad (2.3.1.b)$$

Alle de estimeringsmetodene vi omtaler nedenfor vil gi oss dette resultatet, men slik er det ikke i alminnelighet. For en umettet modell, dvs. at vi har lagt restriksjoner på p-ene, f.eks. om uavhengighet, er det ikke så liketil å se direkte hvordan vi bør estimere. Vi skal derfor kort omtale noen vanlige estimeringsmetoder.

Vi bruker symboler som $\hat{\Theta}$, Θ^* o.l. for estimat (estimatorer) for en parameter Θ. En estimator er en funksjon av de observerbare variable. Et estimat får vi ved å sette inn de observerte tallene i denne.

Minste kvadraters metode. Vi uttrykker forventningen av de variable vi har observert, ved parametrene, f.eks. en forventning av n_{ij} lik np_{ij}. Som estimater, $\hat{p}_{ij}$, velger vi de verdiene som minimerer kvadratsummen av differensene mellom n_{ij} og $n\hat{p}_{ij}$, altså

$$\sum_{i=1}^{I} \sum_{j=1}^{J} (n_{ij} - n\hat{p}_{ij})^2 ,$$

for de gitte n_{ij}-verdier og med $\sum_{ij} \hat{p}_{ij} = 1$.
For en mettet modell får vi løsningen (2.3.1), altså de relative hyppighetene.

For en umettet modell vil p_{ij} være funksjoner $p_{ij}(\Theta_1,\Theta_2,\ldots)$, av et lavere antall parametre, Θ_1,Θ_2, osv. Vi må da finne de verdiene $\hat{\Theta}_1,\hat{\Theta}_2$ osv. som minimerer

$$\sum_i \sum_j (n_{ij} - np_{ij}(\hat{\Theta}_1,\hat{\Theta}_2,\ldots))^2 ,$$

for gitte n_{ij}-verdier, under betingelsen at summen av alle de estimerte $\hat{p}_{ij}$-verdiene er 1.

Sannsynlighetsmaksimeringsmetoden
(forkortet ML etter engelsk: Maximum Likelihood) er meget brukt i krysstabellestimering. For en 2 x 2 krysstabell med multinomisk fordeling er sannsynlighetsfunksjonen for de tellevariable en funksjon av n_{ij} og p_{ij} med bibetingelsene

$$p_{11} + p_{12} + p_{21} + p_{22} = 1 \quad \text{og} \quad n_{11} + n_{12} + n_{13} + n_{14} = n$$

Uttrykket for sannsynlighetsfunksjonen står i appendiks B formel (B 5.2). ML-metoden går ut på å finne de verdiene av p-ene som gjør funksjonsverdien størst mulig når vi har satt inn de observerte n_{ij}-verdiene. Uten restriksjoner på p-ene, dvs. for en mettet modell, blir løsningen igjen (2.3.1.a). Av dette følger $\hat{p}_{1+} = \frac{n_{1+}}{n}$ osv.

Hvis vi har restriksjoner, f.eks. om uavhengighet mellom arbeidstidsordning og yrkesutdanning, blir resultatet et annet. Uavhengighet betyr f.eks. at

$$p_{11} = p_{1+}\,p_{+1},\ p_{12} = p_{1+}\,p_{+2},\ p_{21} = p_{2+}\,p_{+1} \text{ og } p_{22} = p_{2+}\,p_{+2},$$

se setning 2.2.1.
Ved å sette inn dette i sannsynlighetsfunksjonen for de tellevariable og trekke sammen får vi en funksjon av n_{1+}, n_{2+}, n_{+1} og n_{+2} samt av p_{1+}, p_{2+}, p_{+1} og p_{+2}, se (B 5.3) i appendiks B, ML-metoden gir i dette tilfelle

$$\hat{p}_{1+} = \frac{n_{1+}}{n},\ \hat{p}_{2+} = \frac{n_{2+}}{n},\ \hat{p}_{+1} = \frac{n_{+1}}{n},\ \hat{p}_{+2} = \frac{n_{+2}}{n},$$

dvs. vi får her løsningene (2.3.1.b) direkte. For å oppfylle restriksjonene må vi sette estimatene for p_{ij} lik

$$p^*_{11} = \frac{n_{1+}}{n}\cdot\frac{n_{+1}}{n},\ p^*_{12} = \frac{n_{1+}}{n}\cdot\frac{n_{+2}}{n},\ p^*_{21} = \frac{n_{2+}}{n}\cdot\frac{n_{+1}}{n} \text{ og}$$

$$p^*_{22} = \frac{n_{2+}}{n}\cdot\frac{n_{+2}}{n},$$

altså forskjellig fra (2.3.1.a).

Andre restriksjoner kan gi andre resultater. I mange tilfeller, især med store flerveistabeller og mange mulige former for restriksjoner, kan vi ikke finne ML-løsningene direkte som ovenfor, men må rent numerisk iterere oss frem ved skrittvise tilnærmelser til resultatene. Det er imidlertid vist at løsningene eksisterer for en rekke aktuelle problemstillinger.

Kjikvadrat-(χ^2)-minimering er også en metode som blir en del brukt, bl.a. i forbindelse med kji-kvadrat-(χ^2-)testing, jfr. avsnitt 2.4. Vi tar utgangspunkt i en χ^2-observator, f.eks.

$$z = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-np_{ij})^2}{np_{ij}} \text{ eller } z = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-np_{ij})^2}{n_{ij}},$$

og minimerer m.h.p. de ukjente parametrene. For en mettet modell får vi i vårt eksempel resultatet (2.3.1), men for umettede modeller får vi andre resultater.

Hvilket prinsipp vi vil bruke i en bestemt situasjon, vil avhenge av situasjonen, noe som vil fremgå senere.

2.3.2. Hypoteseprøving

Slik analyse av problemer knyttet til kontingenstabeller hittil har utviklet seg, er det testing av hypoteser om diverse avhengighets- eller uavhengighetsforhold som har fått størst oppmerksomhet. La oss se på prinsippene i klassisk hypoteseprøving.

I en I x J krysstabell vil hypotesen om uavhengighet mellom de to variablene (x og y) være uttrykt ved

$$p_{ij} = p_{i+}p_{+j} \qquad \text{for } i = 1,2,\ldots,I \quad j = 1,2,\ldots,J. \qquad (2.3.2)$$

Dette er vår nullhypotese. Som alternativ til denne må vi ha en hypotese om en eller annen form for stokastisk avhengighet. Vi viser til avsnittene 3.1 og 3.2 når det gjelder ulike alternativ i 2 x 2-tabeller. Med større tabeller vil a priori spesifisering av alternativ ofte være vanskeligere. Vi har noen eksempler i avsnitt 3.3 og kapittel 4.

En vanlig situasjon er å teste (2.3.2) mot det uspesifiserte alternativet "mettet modell", og hvis det blir forkasting, så går en inn og undersøker deler av tabellen nærmere. Dette kan føre til testing av hypoteser som avhenger av resultatet av den foregående test. Vi ser litt på dette problemet under en egen overskrift nedenfor.

Vi skal søke å følge grunnprinsippene for valg av testmetode, dvs. valg av kriterium for å forkaste nullhypotesen: Vi velger en lav øvre grense, ε, for sannsynligheten for å forkaste nullhypotesen når den er riktig, altså for å begå en forkastingsfeil. Vi kaller ε forkastingsnivået, eller bare nivået, for testen. Vanlig valg av ε er $\varepsilon = 0{,}05$ eller $0{,}01$. Samtidig ønsker vi en test som gir en større (helst stor) sannsynlighet for å forkaste nullhypotesen når den er gal, dvs. når alternativet faktisk gjelder.

Vi sier at en metode har stor teststyrke, er sterk, mot et alternativ hvis den sist nevnte sannsynligheten er stor (for en gitt ε). I alminnelighet kan vi ikke oppnå stor teststyrke mot alle de mulige alternativ til en nullhypotese. Men jo bedre vi klarer å spesifisere alternativet, dvs. jo mindre klassen av alternativ er, jo større mulighet vil vi i alminnelighet ha for å finne en test med god teststyrke for de alternativ som betyr noe (er spesielt forskjellige fra nullhypotesen).

2.3.3. Prediksjon

Ikke så sjelden skal resultatet av en statistisk analyse brukes til "å spå om fremtiden", dvs. til å forutsi størrelsen på en eller

flere variable (kanskje gitt verdien av andre variable) på et fremtidig tidspunkt, jfr. befolkningsprognoser o.l.

Vi skal ikke gå særlig inn på prediksjonsmetoder her. Men vi kan nevne at vi muligens bør gå rett løs på prediksjonsproblemet uten å estimere først.

Vi vil også minne om at det ikke er sikkert at vi får en god prediksjon selv om vi har funnet et godt estimat å basere prediksjonen på. For det første hefter det usikkerhet (samplingfeil) ved estimatet og dessuten en vanligvis enda større usikkerhet ved den størrelsen som skal predikeres. Hvis denne siste har et standardavvik av størrelsesorden σ og estimatoren har $\sigma/\sqrt{n}$, så kan standardfeilen bli (i en enkel modell) $\sigma\sqrt{\frac{n+1}{n}} > \sigma$. Så vi bør ikke bli altfor skuffet om prognosen slår dårlig til. Se f.eks. kap. 15.5 i A II. (I tillegg kan det hende at selve vår prediksjon om den blir offentliggjort, har ført til endringer i folks handlemåte, og derved i forutsetningene. Dette kan medføre at prognosen helst ikke skal slå til.)

2.3.4. Testing av flere a priori oppstilte hypoteser med samme observasjonsmateriale, multiple tester

Vi ser her på situasjoner hvor vi har bestemt oss for å teste flere forskjellige hypoteser ved hjelp av samme data, uten at data medvirker ved valg av hypotesen. Hvis vi gjør dette ved bruk av flere enkelttester, en for hver nullhypotese f.eks., så må vi huske på at sannsynligheten for å begå minst én forkastingsfeil, i alminnelighet vil være større enn nivået for en enkelt test (jfr. A II, avsnitt 9.8, eller H. I, avsnitt 4).

Sett at vi har to hypoteser som begge er riktige og en test med forkastingsnivå ε (som vi kan oppnå) for hver av dem. Sannsynligheten for å forkaste den første når den er riktig, er ε. Sannsynligheten for ikke å forkaste den er da $1 - \varepsilon$. Sannsynligheten for å forkaste den andre når den er riktig er ε', der ε' kan være lik ε, eller muligens ulik ε hvis den er avhengig av den første testen. Sannsynligheten for ikke å forkaste den første men den andre, er da $(1 - \varepsilon)\varepsilon'$. Det vil si at sannsynligheten for minst en forkastningsfeil er $\varepsilon + (1 - \varepsilon)\varepsilon'$ som er større enn ε (unntatt hvis $\varepsilon' = 0$ eller $\varepsilon = 1$).

Hvis de to testobservatorene vi bruker er stokastisk uavhengige, blir dette $\varepsilon + (1 - \varepsilon)\varepsilon = 2\varepsilon - \varepsilon^2$, temmelig nær 2ε. Ved m uavhengige tester med samme ε får vi tilsvarende at m ε er en øvre skranke for sannsynligheten for å begå minst én forkastingsfeil. For å redusere denne sannsynligheten kan vi f.eks. bruke nivå $\frac{\varepsilon}{m}$ istedenfor ε for den enkelte testen.

Det kan også være andre muligheter enn å velge alle enkeltnivåene like, vi må jo tenke på teststyrken ved valg av test. Her må vi se på det enkelte problem for å finne løsninger.

2.3.5. Testing etter å ha "kikket på data". Hypoteser som avhenger av utfallet av tidligere tester med de samme data

Det er mange fallgruber ved formulering av modell og hypoteser. En av dem er at vi lar data veilede oss ved oppstilling av hypoteser, uten at vi tar hensyn til dette ved valg av testmetode. Det er da stor mulighet for at vi bruker galt nivå for tester e.l., jfr. eksemplet i app. C.

En tilsvarende situasjon kommer vi i hvis vi har foretatt visse tester i et materiale, og ut fra resultatet av disse bestemmer oss for å teste andre hypoteser. Anta f.eks. at vi har brukt treveis-tabellen 2.2.c for å teste avhengighet mellom de tre variable antall helgeturer, adgang til fritidshus og kjønn, jfr. avsnitt 4.3.3. Vi får forkastet hypotesen om uavhengighet, men finner ut at vi burde teste om det er avhengighet mellom to av dem, mens den tredje kan være uavhengig av disse to. Hvordan bør vi nå gå frem for å beholde kontrollen over forkastingsnivået ?

Dette er et generelt statistisk problem, løsninger for multivariable normalt fordelte data er foreslått bl.a. av Scheffé, Tukey, Spjøtvoll m.fl. Bjørnstad [B.I] og især Sverdrup [1975 og 1978, I] har anvist generelle løsninger for multinomiske situasjoner. Vi skal gi noen eksempler i kap. 3 og 4. Poenget er her å velge en test for den opprinnelige hypotesen som gir muligheter for også å komme med utsagn om "delhypoteser" e.l. på et neste trinn.

Vi ser at hvis vi ikke finner løsninger som gir oss kontroll over forkastingsnivået, så er vi på vei over i ren dataanalyse, der "testnivå" bare brukes som et rent formelt kriterium for å skille mellom "store" og "små" forskjeller, vi kan ikke lenger si noe om forskjellene er signifikante eller ikke. Dette kan selvsagt være utveien i en del situasjoner.

2.3.6. Pretest-estimering

Vi har forøvrig det tilsvarende problem når vi estimerer visse parametre etter først å ha testet hypoteser om dem. I eksemplet foran, under estimering etter ML-metoden, vil vi vel bruke $\hat{p}_{ij} = n_{ij}/n$ som estimater hvis vi forkaster hypotesen om uavhengighet mellom de variable x og y, mens vi kanskje bruker

$$p^*_{ij} = n_{i+}\, n_{+j}/n^2$$

hvis vi ikke forkaster den. Vi bør faktisk se nærmere på en slik "pre-test-estimator", som enten er lik $\hat{p}_{ij}$ eller p^*_{ij}, alt etter utfallet av testen.

2.3.7. Testing i meget store observasjonsmaterialer. Vanskeligheten med å formulere den "riktige" nullhypotesen.

Ettersom bruken av datamaskiner har gjort det mulig å bruke meget store datamaterialer i analyseøyemed, er en for alvor blitt oppmerksom på visse ulemper ved den gjengse måten vi hittil har brukt for å formulere og teste hypoteser. Vi vil nemlig finne at når datamaterialet er tilstrekkelig stort, vil vi komme til å forkaste nesten enhver nullhypotese. Testene får stor styrke, også mot alternativ som ligger meget nær nullhypotesen, kanskje så nær at de for praktiske formål kan sies å være jevngode med denne. For små og middels store datamaterialer vil sannsynligheten for å akseptere en hypotese som ligger "nær" nullhypotesen vanligvis være så liten (nær ε) når den er riktig, at vi ikke får akseptert den, selv om den altså er riktig.

Nå vil jo vår formulering av nullhypotesen ofte være mer stringent enn vi egentlig har dekning for. Vi setter H_o: $p_1 = p_2$, mens vi kanskje mener at p_1 og p_2 er så nær hverandre at de for vårt formål kan betraktes som like.

Bjørnstad [B I] foreslår en viss måte å definere "utvidet" nullhypotese på, og tilsvarende tester. Hans forslag er en generalisering av følgende:

Istedenfor å sette : H_o: $p_1 = p_2$

kan en sette: H_o: $|p_1 - p_2| \leq c$,

der c er et lite valgt tall.

En annen sak er at vi slett ikke behøver å tro på at nullhypotesen gjelder (stringent), vi bruker den bare som et hjelpemiddel for å komme frem til en test som kan gi oss den konklusjon at p_1 og p_2 er forskjellige, og helst også om $p_1 > p_2$, f.eks. Sverdrup fremholder dette synspunktet og påpeker at våre vanlige tester kan brukes også i slike tilfelle.

Goldstein [1981] har et forslag for kontingenstabeller med mange observasjoner, som går ut på å foreta testen uten at vi behøver å vite om vi har greid å formulere "den riktige nullhypotesen", vi er bare "i nærheten av" den. Det gjelder da å bruke en test som gir fornuftig resultat også i slike situasjoner. Se kap. 10.

2.3.8. Noen statistiske termer.

I avsnittet om estimering definerte vi for bruk ved estimering: en estimator: en funksjon av de observerbare variable. F.eks. er

$$\hat{p}_{ij} = \frac{n_{ij}}{n} \text{ og } \hat{p}_{i+} = \frac{n_{i+}}{n}$$

estimatorer når vi ser på dem som funksjoner av henholdsvis n_{ij} og n_{i+}. Når vi setter inn tall fra et gitt observasjonsmateriale, f.eks.

$$\hat{p}_{ij} = \frac{480}{1320} = 0.3636 \quad \text{og} \quad \hat{p}_{i+} = \frac{1092}{1320} = 0.827\ldots,$$

så har vi funnet estimater for p_{ij} og p_{i+}.

På engelsk kalles en slik funksjon av observerbare variable ofte en "statistic". Vi har en mer generell betegnelse på norsk også, som vi bruker bl.a. ved hypoteseprøving, nemlig en

observator: en funksjon av de observerbare variable (som ovenfor).

Ved hypoteseprøving trenger vi å kjenne den sannsynlighetsfordelingen vår observator har når nullhypotesen er riktig. Vi skal iblant bruke betegnelsen

nullfordeling: fordeling av en observator når nullhypotesen gjelder (eventuelt en spesifisert versjon av denne).

Videre trenger vi et navn på de verdiene av observatoren z, som gir forkasting av nullhypotesen. Vi skal bruke betegnelsene

Nedre ε-fraktil: Den verdien z_ε som er slik at $P(z \leq z_\varepsilon) = \varepsilon$, dvs. at z_ε er den verdien av z som er slik at sannsynligheten er ε for å få $z=z_\varepsilon$ eller mindre når nullfordelingen gjelder.

Øvre ε-fraktil: Den verdien $z_{1-\varepsilon}$ som er slik at $P(z \geq z_{1-\varepsilon}) = \varepsilon$, dvs. at det er en sannsynlighet lik ε for å få $z=z_{1-\varepsilon}$ eller større når nullfordelingen gjelder.

I den standardiserte normalfordelingen er

med $\varepsilon = 0,05$	med $\varepsilon = 0,025$
$z_{0,05} = -1,64$	$z_{0,025} = -1,96$
$z_{0,95} = 1,64$	$z_{0,975} = 1,96$

A priori betingelser

Som en fellesbetegnelse på de forutsetninger vi gjør om modellen for våre variable før vi begynner å teste eller estimere, bruker vi ofte betegnelsen: A priori betingelser eller forutsetninger. Eksempel: en IxJ-tabell der alle $p_{ij} > 0$ og fordelingen er multinomisk.

2.4. Kji-kvadrat føyningstest og homogenitetstest, sannsynlighetskvotetest.

I mange problemer der vi har passelig store observasjonsmaterialer med observasjoner i alle ruter, vil vi kunne bruke en eller annen kji-kvadrat (χ^2-) test. For noen av disse vil det være nyttig å kunne vise til velkjente tester, nemlig χ^2-føyningstesten eller χ^2-homogenitetstesten. Vi skal derfor kort referere disse testene her.

2.4.1 χ^2-føyningstesten for fullt spesifisert nullhypotese i en krysstabell.

For enveis-grupperte data er denne testen beskrevet i de fleste lærebøker i statistikk, f.eks. Lillestøl, Hodges & Lehman, Hellevik, A I. Har vi to- eller flerveisgrupperte data, så er det bare å tenke seg alle kategoriene ordnet etter hverandre i én kolonne (istedenfor i krysstabellform) for å kunne bruke denne testen direkte. Vi skal imidlertid sette den opp ved hjelp av våre krysstabellsymboler (sammenlign med eksemplet nedenfor).

Vi har en I x J x ... x G-tabell og vil teste én hypotese om at alle p-ene har gitte verdier, dvs. nullhypotesen

$$H_o : p_{ij\ldots g} = p^o_{ij\ldots g} \qquad \text{for } i = 1,2,\ldots,I;\ j = 1,2,\ldots,J;\ \ldots;\ g = 1,2,\ldots,G$$

Her er alle $p^o_{ij\ldots g}$ kjente tall, med sum 1. Alternativet til nullhypotesen behøver ikke være godt spesifisert, vi kan ha to eller flere p-verdier forskjellige fra de oppgitte p^o-verdiene.

For en <u>fullstendig krysstabell,</u> der $\sum_{ij\ldots g} p_{ij\ldots g} = 1$, men ingen andre marginaler er fastlagt, kan vi bruke observatoren som er summen av alle de kvadrerte avvikene mellom "observert og forventet" antall i en rute, dividert med forventet antall, dvs.

$$z = \sum_{i=1}^{I} \sum_{j=1}^{J} \cdots \sum_{g=1}^{G} \frac{(n_{ij\ldots g} - np^o_{ij\ldots g})^2}{np^o_{ij\ldots g}} \qquad (2.4.1)$$

Under H_o vil denne være χ^2-fordelt med (I J...G - 1) <u>frihetsgrader eller df (for: degrees of freedom)</u>, med god tilnærmelse når np^o-verdiene er store nok. Vi forkaster H_o hvis vi ved innsetting av våre tall i (2.4.1) får en z-verdi større eller lik den øvre ε-fraktilen i den nevnte fordelingen.

"Store" np^o-verdier blir gjerne angitt som ≥ 5. Men i store tabeller kan en ha en del np^o-verdier helt nede i 1.

Anta at vi i eksemplet i tabell 1.1 ut fra tidligere erfaring tror at

$$p_{11} = p_{12} = 0{,}1, \quad p_{21} = p_{22} = 0{,}01 \quad \text{og } p_{31} = p_{32} = 0{,}39$$

og vil teste denne nullhypotesen. Vi finner

$$z = \sum_{i=1}^{3} \sum_{j=1}^{2} \frac{(n_{ij} - np^o_{ij})^2}{np^o_{ij}} = \frac{(121-132)^2}{132} + \frac{(82-132)^2}{132} + \frac{(19-13{,}2)^2}{13{,}2} +$$

$$+ \frac{(6-13{,}2)^2}{13{,}2} + \frac{(612-501{,}6)^2}{501{,}6} + \frac{(480-501{,}6)^2}{501{,}6} = \underline{51{,}6}$$

Nå er øvre 5-prosentfraktil i χ^2-fordelingen med $3 \cdot 2 - 1 = 5$ df lik <u>11.07,</u> så vi må forkaste H_o.

EDB-program for denne testen ser ikke ut til å finnes i SPSS-pakken. Den er imidlertid enkel å programmere.

2.4.2 Føyningstesten for komparative tabeller.

Når kontingenstabellen består av en rekke enveis- eller krysstabeller som hver for seg har et gitt antall observasjoner og sum 1 for de tilhørende p-verdiene, jfr. tabell 1.2, kan vi også lage en føyningstest. Anta f.eks. at vi har en I x J x K-tabell, som består av K toveis krysstabeller hver med I x J ruter. Da vil vi kunne regne ut en ny z-verdi som i (2.4.1) for hver av de K krysstabellene. Vi får en

$$z_k = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ijk} - n_{++k} p^o_{ijk})^2}{n_{++k} p^o_{ijk}} \qquad \text{for hver k, f.eks. for k=2}$$

Denne z_k vil være χ^2-fordelt med (I·J-1) df under H_o. Så kan vi addere alle disse z_k-ene, og får

$$z_K = \sum_{k=1}^{K} z_k = \sum_{k=1}^{K} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ijk} - n_{++k} p^o_{ijk})^2}{n_{++k} p^o_{ijk}} \quad . \quad (2.4.2)$$

Denne z_K vil være χ^2-fordelt med K(IJ - 1) = KIJ - K df under H_o. Tiltross for at uttrykkene (2.4.1) og (2.4.2) kan se svært like ut, så har z_k og z_K forskjellig fordeling fordi $\sum\sum_{ij} p^o_{ijk} = 1$ for hver av toveistabellene i (2.4.2).

Hvis tallene i tabell 1.1 var fremkommet fra ett utvalg trukket blant personer med yrkesutdanning, og et annet trukket blant personer uten, kan vi teste: $p_{11} = p_{12} = 0{,}2$, $p_{21} = p_{22} = 0{,}02$ og $p_{31} = p_{32} = 0{,}78$.

Vi finner $z_K = \frac{(121-150{,}4)^2}{150{,}4} + \frac{(19-15)^2}{15} + \frac{(612-586{,}6)^2}{586{,}6} +$

$\frac{(82-113{,}6)^2}{113{,}6} + \frac{(6-11{,}4)^2}{11{,}4} + \frac{(480-443)}{443} = 22{,}4$, som skal sammenlignes med 9,49 som er øvre 5 prosentfraktil for 2·2 = 4 df. Dette gir forkasting av nullhypotesen. Vi har her valgt de samme $p^o_{ij.}$-verdiene som i (2.4.1) for å understreke at resultatene faktisk blir ulike.

2.4.3 χ^2-homogenitetstesten for sammenligning av flere multinomiske fordelinger.

Vi vil bruke tallene i tabell 1.2 til å sammenlikne aldersfordelingen for handelsflåten i de fire nordiske landene. Nå kan

vi jo bare regne ut de relative hyppighetene for hver aldersklasse i hvert av landene for å se at det selvsagt er forskjell på de fire fordelingene. Vi er imidlertid ikke interessert i små forskjeller som kan betraktes som nokså tilfeldige. Vi vil vite om det er så store forskjeller at de tyder på en mer dyptgående strukturell ulikhet mellom landene når det gjelder f.eks. fornying av handelsflåten. Vi tenker oss altså at de observerte tallene dels bestemmes av strukturelle sammenhenger som er karakteristisk for det enkelte land, og dels av mer tilfeldige hendelser.

Da kan vi velge å betrakte data fra hvert land som et utvalg fra en "bakenforliggende" teoretisk fordeling som er karakteristisk for landet. Så undersøker vi om data tyder på at disse teoretiske fordelingene er forskjellige for de 4 landene.

Med J (4) utvalg, hvert med de samme I (7) kategoriene for den variable x, må nullhypotesen være at alle sannsynlighetene i hver linje (aldersklasse) er like store, dvs. vi har

$$H_o: \quad p_{i1} = p_{i2} = \ldots p_{iJ} \qquad \text{for } i = 1,2,..,I,$$

mot en uspesifisert klasse av alternativ, der minst to av likhetstegnene ikke gjelder. Hvis vi skulle teste en spesifisert hypotese $p_{ij} = p^o_{ij}$, der alle p^o_{ij} er kjente, ville vi bruke observatoren (2.4.2) foran. Nå bruker vi en lignende, idet vi erstatter p^o_{ij} med dens estimat under H_o, nemlig den marginale relative hyppigheten for linje nr. i, altså $\hat{p}_{ij} = \hat{p}_{i+} = n_{i+}/n$. Vi har da observatoren

$$z_h = \sum_{j=1}^{J} \sum_{i=1}^{I} \frac{(n_{ij} - n_{+j}\, n_{i+}/n)^2}{n_{+j}\, n_{i+}/n} \qquad (2.4.3)$$

Det viser seg at under H_o er denne asymptotisk χ^2-fordelt med $(J(I - 1) - (I - 1)) = (J - 1)(I - 1)$ df. I forhold til (2.4.2) "mister" vi like mange df som vi estimerer parametre, her I - 1, idet summen er gitt lik 1.

I eksempel 1.2 finner vi $z_h = 116$, som er større enn alle tabulerte fraktiler for χ^2-fordelingen med $(4 - 1)(7 - 1) = 18$ df. Vi må kunne slutte at de observerte forskjellene ikke bare kan tilskrives tilfeldigheter.

EDB-program for denne testen i toveistabeller finnes i SPSS, CROSSTABS, punkt 16.1.2.1.

2.4.4. Litt mer om χ^2-fordelte observatorer.

Prinsippet for konstruksjon av observatorer for testing av homogenitet skulle fremgår av 2.4.3. En må imidlertid våre oppmerksom på at bevisene for asymptotisk χ^2-fordeling under nullhypotesen forutsetter at estimatene $\hat{p}_{ij...g}$ er funnet enten ved ML-estimering eller ved χ^2-minimering, jfr. avsnitt 2.3. Vi kan f.eks. ikke sette inn $\hat{p}$-verdier vi har funnet på "øyemål" e.l.

Det har også vist seg at fordelingen kan avhenge av om estimeringen foregår ut fra enkeltobservasjoner eller fra de grupperte observasjonene i tabellen, se Albricht[1980]. (Å bruke enkeltobservasjoner er nærliggende hvis en estimerer et lite antall parametre, $\hat{\Theta}_1, \hat{\Theta}_2, \ldots$ som så de enkelte $\hat{p}_{ij...g}$ er funksjoner av). Vi skal regne med grupperte observasjoner i χ^2-testene.

I visse tilfeller bruker en observatorer som avviker fra dem vi har definert foran, ved at de observerte istedenfor "de beregnede" $n_{ij...g}$-verdiene er satt inn i nevnerne. Istedenfor

(2.4.3) bruker vi $z = \sum\sum_{ij} \frac{(n_{ij} - n_{+j}n_{i+}/n)^2}{n_{ij}}$ (2.4.4)

o.l.

Den asymptotiske fordelingen for (2.4.4) er den samme som for (2.4.3)

I Fienberg [1979] finnes en oversikt over bruk og egenskaper for en del χ^2-observatorer, der også problemet med "tynt besatte" ruter er tatt opp.

Vi skal foreslå alternative tester til de vanlige χ^2-testene for en del problemstillinger.

2.4.5. Yates' korreksjon

De forskjellige z-observatorene vi har nevnt, er diskrete variable, mens χ^2-fordelingen er kontinuerlig. Fordi vi approksimerer en diskret fordeling med en kontinuerlig, kan forskjellen mellom eksakt og tilnærmet fordeling bli stor når det er små tall i tabellen. Istedenfor tellerne i brøkene i (2.4.1) setter vi da inn

$$(|n_{ij..g} - np^o_{ij...g}| - 0,5)^2 \qquad (2.4.6)$$

og tilsvarende for de øvrige observatorer. Vi ser at nevnerne np^o i (2.4.1) skal være ganske små før den såkalte Yates' korreksjon betyr noe for z-verdien.

I eksemplet i 2.4.1 vil vi få en $z^{korr} = 49,9$ istedenfor den funne $z = 51,6$. Over halvdelen av forskjellen skyldes her leddene med 13,2 i nevneren.

2.4.6. Sannsynlighetskvotetest, LL-test (log likelihood ratio-test)

I problemer der de alternative hypotesene bare er delvis spesifiserte, mens vi kjenner uttrykket for den simultane sannsynlighetsfunksjonen for de observerbare variable, kan vi utlede tester ved hjelp av den såkalte sannsynlighetskvotemetoden (Likelihood ratio method), se f.eks.HTI,AII, avsnitt 9.6 og 14.3, eller Fienberg 1978, eller BFH.

Vi kan begynne med en observator, q, som er forholdstallet mellom den største verdien sannsynlighetsfunksjonen kan anta under de a priori betingelser, og den største verdien den kan anta under nullhypotesen. Vi vil forkaste nullhypotesen hvis q blir tilstrekkelig stor for våre data. For å slippe å utlede nullfordelingen for q, ser vi isteden på observatoren

$$z_L = 2 \log q,$$

der log står for naturlig logaritme. Det kan nemlig vises at denne z_L er tilnærmet χ^2-fordelt under nullhypotesen, med df lik det antall parametre som fastsettes av nullhypotesen.

For vårt eksempel i 2.4.3, vil vi finne

$$z_L = -2 \sum_{i=1}^{I} \sum_{j=1}^{J} n_{ij}(\log(n_{i+}n_{+j}/n) - \log n_{ij}),$$

som er asymptotisk χ^2-fordelt med (I-1)(J-1) df. Dette er samme antall df som testen i 2.4.3, og resultatet av de to testene vil ligge meget nær hverandre. For eksemplet i 1.2 finner vi $z_L = 113$.

Vi skal kalle slike tester der vi bruker z_L-observatorer og tilnærmet χ^2-fordeling for LL-tester (log likelihood ratio tests).

I de fleste av de tilfellene i kapitlene 3-5 der vi beskriver χ^2-tester, vil vi isteden kunne bruke sannsynlighetskvotetester hvis vi ønsker det, men vi nevner ikke dette i hvert enkelt tilfelle.

Det er især i kapittel 6 at vi får bruk for sannsynlighetskvotetestene, de brukes gjerne i forbindelse med log-lineære modeller, og i dataprogrammene for disse.

2.5 Flerdesisjonsproblemer.

Ved hypoteseprøving har vi ett av to mulige utfall av en test:
i) Vi forkaster nullhypotesen, og konkluderer med at alternativet gjelder.
ii) Vi forkaster ikke nullhypotesen, og kan ikke gi noen konklusjon.

Riktignok er vi i mange tilfeller tilbøyelige til å si at vi i tilfelle ii) vil godta nullhypotesen. Dette er fristende, fordi vi oftest ikke har noen mulighet for virkelig å få verifisert en nullhypotese. Det må vel også være tillatt å gjøre dette i en del situasjoner der vi har et visst a priori belegg for at den kan være riktig.

Men i andre situasjoner må vi la konklusjonen stå åpen, spesielt når nullhypotesen bare brukes som et hjelpemiddel, en "klareringshypotese", som vi ikke egentlig tror på, men trenger rent teknisk for å kunne "få frem" alternativet, jfr. avsnitt 2.3.7.

I situasjoner der vi vet for lite a priori til å kunne spesifisere alternativene særlig godt, vil vi kanskje ønske en test som kan gi et mer nyansert bilde.

Anta at vi vil sammenligne sannsynlighetene p_1 og p_2 for et visst kjennetegn i to utvalg: Vi kan da teste nullhypotesen om at de to sannsynlighetene er like mot det alternativ at de er forskjellige, dvs.

$$H_o: p_1 = p_2 \quad \text{mot} \quad p_1 \neq p_2 .$$

i) Hvis vi forkaster H_o, ønsker vi å kunne si

enten: $p_1 > p_2$

eller: $p_1 < p_2$

ii) Hvis vi ikke forkaster H_o vil vi vanligvis ikke kunne si noe, men det kan, som nevnt ovenfor, være situasjoner der vi har lov til å regne med at

$$p_1 = p_2.$$

En test av denne art, med nivå ε, dvs. sannsynlighet $\leq \varepsilon$ for å komme med én av påstandene under i) selv om H_o er riktig, er den følgende: vi forkaster H_o og konkluderer med $p_1 > p_2$ når verdien av den observatoren vi bruker er større eller lik den øvre ε-fraktil i den nullfordelingen vi bruker. Og vi forkaster H og konkluderer med $p_1 < p_2$ når verdien er mindre eller lik nedre ε-fraktil i nullfordelingen. (Dette svarer altså til en "vanlig" test med nivå 2ε.) Se f.eks. Sverdrup (1976) eller AA. I App. 2. Se eksempel iii) under de første testene i avsnitt 3.1.
Denne tankegangen kan generaliseres til å gjelde flere parametre og ulike typer av desisjoner. Se f.eks. avsnitt 3.3.1 og 3.3.2.
Det er også vanlig at en etter testing av en rekke parametre setter en del av dem (f.eks. q stykker) lik null og så estimerer de resterende ut fra den "reduserte" modellen, dvs. en modell der de førstnevnte q parametrene er null. Problemet er da å gå frem slik at en har kontroll over nivået for dette flerdesisjonsproblemet. Se f.eks. avsnitt 6.5 og 7.1.

2.6 "Ufullstendige tabeller". Tilfeldige og strukturelle nuller.

Så lenge asymptotiske χ^2-tester o.l. var de eneste hjelpemidler ved analyse av samvariasjonsproblemer for kategoriske variable, så ble det vanskeligheter når en hadde $n_{ij...g} = 0$ i en eller flere ruter i en tabell.

Vi må her skille mellom to situasjoner

i) $p_{ij...g} > 0$, mens $n_{ij...g} = 0$.
Det er bare slump at observasjoner mangler, gjør vi n større, kan vi til slutt få tall i alle ruter. Men det er stadig en viss sannsynlighet for å få nuller.

ii) $p_{ij...g} = 0$ for en del ruter, da er nødvendigvis også $n_{ij...g} = 0$.
Dette gjelder altså kombinasjoner (i,j,...,g) som ikke kan forekomme.

I situasjon i) har vi tilfeldige nuller. Modell og metode er i prinsippet nøyaktig de samme som når vi har tall i alle ruter. Ved bruk av de vanlige asymptotiske metodene må vi imidlertid ty til nødutveier så ikke unøyaktigheten blir altfor stor.

En utvei er som kjent å slå sammen kategorier, så ingen ruter blir tomme eller har for små tall. En fare her er at en kan få ulike testresultater, avhengig av hvordan sammenslåingen blir gjort.

En annen utvei er å sette inn et lite tall, f.eks. 0,5 e.l. istedenfor null i de tomme rutene. Også her kan resultatet bli avhengig av hva en velger, jfr. Fridstrøm [1980] og Fienberg [1979].

"Eksakte" metoder, dvs. metoder der vi kommer frem uten å måtte ty til tilnærmet beregning av sannsynlighetsnivå, vil vi kunne bruke også om vi har nuller i tabellen.

ii) Strukturelle nuller, dvs. $p_{ij\ldots g} = 0$ for en del ruter, betyr at vi har et problem av en spesiell struktur, noe vi må ta hensyn til ved valg av metode. Vi kan ikke uten videre bruke de standardmetodene som er utviklet for fullstendige tabeller.

Det er gjort en del arbeid med å formulere problemet slik at en kan bruke metoder som er analoge med de tilnærmede testene vi bruker for fullstendige tabeller. Vi kommer tilbake til dette i avsnitt 9, jfr. også avsnitt 6 og 7.

ANNEN DEL

Metoder som forutsetter a priori spesifikasjon av stokastisk modell, jfr. avsnitt 1.4 om stokastisk inferens.

De fleste av metodene her er vel kjente og finnes i elementære lærebøker. Vi skal imidlertid også ta med noen nyere metoder, foreslått av Sverdrup, Aaberge m.fl. Vi starter med toveistabeller og går så over til tre- og flerveistabeller. Vi må her huske på at konklusjoner fra f.eks. toveistabeller ikke alltid holder hvis vårt problem og materiale egentlig omfatter flere variable som bør analyseres simultant.

3. TOVEISTABELLER.

Det finnes en god del metoder for analyse av problemstillinger der tallmaterialet kan ordnes i en toveistabell. Vi starter her med 2 x 2 tabeller og går over til større tabeller etter hvert. Vi skal bruke en del gjennomgående talleksempler, dels for å sammenligne resultater, men især for å spare plass. Det er imidlertid bare selve tallene som går igjen, problemstilling og forutsetninger varierer fra punkt til punkt.

Som 2 x 2 tabeller bruker vi et sammendrag av tabell 2.2.c, samt et konstruert eksempel med små tall, men med omtrent de samme relative hyppigheter.

Tabell 3.1. Tall på personer som

drar på helgetur i	har adgang til fritidshus Ja j=1	Nei j=2	Marginal
Ja 1	662	1016	1678
Nei 2	98	488	586
Marginal	760	1504	2264

Tabell 3.2. Tall på personer som

drar på helgetur i	har adgang til fritidshus Ja j=1	Nei 2	Marginal
Ja 1	7	10	17
Nei 2	1	5	6
Marginal	8	15	23

3.1. Sammenlikning av to relative hyppigheter. Komparativ 2 x 2 tabell.

Data: Tabell 3.1 resp. 3.2

Problem (i): Vi vil undersøke om det er større tilbøyelighet til å dra på helgetur blant dem som har fritidshus enn blant dem som ikke har det. (Vi regner ikke med at det omvendte kan gjelde.)

Problem (ii): Er det forskjell på tilbøyeligheten til å dra på helgetur i de to gruppene ?

Problem (iii): Som (ii) med tilleggsspørsmål: Hvis det er forskjell, hvilken vei går den ?

Forutsetninger: Vi har trukket to tilfeldige utvalg, et på $n_{+1} = 760$ (resp. 8) personer blant personer som har adgang til fritidshus, et annet på $n_{+2} = 1504$ (resp. 15) personer blant dem som ikke har det. Vi regner med binomisk fordeling av n_{11} og n_{12}.

Vi vil teste nullhypotesen H_o: $p_{11} = p_{12}$,
dvs. ingen forskjell mellom tilbøyelighetene, mot det alternativ som interesserer oss, enten (i) eller (ii) eller (iii).

Alternativet, og dermed konklusjonen hvis vi får forkasting av H_o, er altså

$p_{11} > p_{12}$ i situasjon (i), og

$p_{11} \neq p_{12}$ i situasjon (ii).

I situasjon (iii) blir svaret ved forkasting av H_o enten $p_{11} < p_{12}$ eller $p_{11} > p_{12}$.

3.1.1. Fisher-Irwins test,

se H.I, 12.2 eller A.I, 3.6.1.

Det kan vises at vi får en god testmetode i dette tilfellet ved å regne som om n_{1+} er et gitt tall. Da har n_{11} en hypergeometrisk fordeling, gitt ved sannsynlighetsfunksjonen $f(n_{11})$ nedenfor, når nullhypotesen er riktig.

Dette leder oss til følgende testmetode i de tre ulike situasjoner:

I situasjon (i) forkastet vi H_o hvis vi observerer en n_{11} som er større eller lik øvre ε-fraktil i denne fordelingen.

I situasjon (ii) forkaster vi H_o hvis vi enten har n_{11} mindre eller lik nedre $\varepsilon/2$-fraktil eller større eller lik øvre $\varepsilon/2$-fraktil. Hvis det er vanskelig å finne slike fraktiler, kan vi velge f.eks. nedre δ-fraktil og øvre $(\varepsilon-\delta)$fraktil, slik at testens nivå blir ε. Se eksemplet nedenfor.

I situasjon (iii) konkluderer vi med $p_{11} < p_{12}$ hvis n_{11} mindre eller lik nedre ε-fraktil, og med $p_{11} > p_{12}$ hvis n_{11} større eller lik øvre ε-fraktil.

I vårt eksempel i tabell 3.2 kan vi regne ut sannsynlighetene fra de to "halene" i fordelingen; i det vi merker oss den laveste verdien n_{11} kan ha er, 2 og den største er 8 når alle marginalene er gitt som i tabellen. Vi finner for nedre hale:

$$P(n_{11} \leq 2 \mid H_o) = 0.0003 \qquad \text{mens } P(n_{11} \leq 3 \mid H_o) = 0.08$$

Fra øvre hale finner vi

$$P(n_{11} \geq 8 \mid H_o) = 0.0496 \qquad \text{og } P(n_{11} \geq 7 \mid H_o) = 0.2$$

Med nivå $\varepsilon = 0.05$ kan vi forkaste nullhypotesen hvis vi får $n_{11} = 8$ ved alternativ (i) eller (iii). Ved alternativ (ii) kan vi forkaste hvis vi får $n_{11} = 2$ eller 8. Vi har funnet $n_{11} = 7$, og kan derfor ikke forkaste nullhypotesen i noe av de tre tilfellene.

SPSS-programmet CROSSTABS bruker denne testen for å teste uavhengighet i en 2 x 2 tabell når $n < 21$. (Testen for uavhengighet i dette programmet svarer til testen i situasjon (ii) ovenfor, jfr. avsnitt 3.2.1).

Sannsynlighetsfunksjonen vi bruker, er

$$f(n_{11}) = \frac{\binom{n_{+1}}{n_{11}}\binom{n_{+2}}{n_{12}}}{\binom{n}{n_{1+}}} = \frac{\binom{8}{n_{11}}\binom{15}{17-n_{11}}}{\binom{23}{17}} .$$

I Biometrika Tables for Statisticians finnes tabeller for F.I-testen for $n \leq 50$. Det eksisterer også spesielle tabeller over den hypergeometriske fordeling.

Dette er hjelpemidler vi kan bruke for å regne ut de eksakte sannsynlighetene i null-fordelingene.

I eksempel 3.1 er tallene så store at vi bruker tilnærmet regning for å bestemme fraktilene fordi dette er enklere, men testene er i prinsippet de samme.

3.1.2. Tilnærmet normaltest utnytter her at den hypergeometriske fordelingen er tilnærmet normal når n er tilstrekkelig stor, og alle fire forventningene $n_{+j}p_{ij}$ (i praksis alle n_{ij}) > 5. Se henvisningene ovenfor eller H.I, avsnitt 3 om slik testing.

Vi regner ut den standardiserte differensen d mellom $\hat{p}_{11}$ og $\hat{p}_{12}$ i de to utvalg, altså

$$d = \left(\frac{n_{11}}{n_{+1}} - \frac{n_{12}}{n_{+2}}\right) \Bigg/ \sqrt{\frac{n_{1+}}{n} \cdot \frac{n_{2+}}{n} \left(\frac{1}{n_{+1}} + \frac{1}{n_{+2}}\right)}$$

og sammenlikner resultatet med fraktilene i den standardiserte normalfordelingen etter samme regler som i 3.1.1.

I eksempel 3.1 finner vi d = 10.03, mens øvre 5-prosentfraktil er 1,64, og $2\frac{1}{2}$-prosentfraktilen er 1,96. Vi får altså forkasting av H_o i alle tre situasjoner. Konklusjonene blir:

Situasjon (i): Det er større tilbøyelighet til å dra på helgetur blant dem som har adgang til fritidshus.

Situasjon (ii): det er forskjellig tilbøyelighet i de to gruppene.

Situasjon (iii): som under (i).

Bruker vi denne tilnærmede testen for eksempel 3.2 finner vi d = 1,08, som vil gi samme konklusjon som vi fant under 3.1.1, nemlig ingen forkasting av H_o.

Se 3.1.3 om korrigert d i små utvalg. I Aa.I , 4(i) har Aaberge satt opp en versjon av normaltesten med korreksjoner både i teller og nevner. Dette skal sikre større nøyaktighet ved den tilnærmede beregningen.

3.1.3. χ^2-homogenitetstesten.

Under de samme forutsetningene om $n_{+j}p_{ij}$ som i 3.1.2, kan vi bruke testen i 2.4.3, med $I = 2$ og $J = 2$. Etter sammentrekning av uttrykket for z_h finner vi at det kan skrives

$$z_h = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{n_{1+}\, n_{2+}\, n_{+1}\, n_{+2}}$$

som skal sammenliknes med øvre ε-fraktil i χ^2-fordelingen med 1 df. Vi ser etter litt regning at

$$z_h = d^2,$$

slik at i situasjon (ii) faller testene i 3.1.2 og 3.1.3 sammen. Vi har også at $(1{,}96)^2 = 3{,}84$ som er øvre 5 prosent fraktil for z_h i χ^2-fordelingen med én df.

χ^2-testen er ikke konstruert for spesiell bruk mot ensidig alternativ. I dette og lignende tilfeller kan vi likevel greie å bruke den i situasjon (i), resp. (iii). Vi kan nemlig konkludere med at $p_{11} > p_{12}$ hvis $z_h >$ øvre 10 prosent fraktil 2,71 $(= 1{,}645^2)$ og dessuten $n_{11}n_{22} - n_{12}n_{21} > 0$. (Dette blir det samme som normaltesten i 3.1.2).

I dette tilfellet kan den Yates-korrigerte kji-kvadrat observatoren skrives, jfr. avsnitt 2.4.5,

$$z_h^{korr} = \frac{n(|n_{11}n_{22} - n_{12}n_{21}| - 0{,}5n)^2}{n_{1+}\, n_{2+}\, n_{+1}\, n_{+2}}.$$

Det er denne som regnes ut av SPSS-programmet CROSSTABS når $n > 20$ (for en 2 x 2-tabell).

For eksemplet i tabell 3.2 finner vi

$$z_h = \frac{23(7\cdot 5 - 10\cdot 1)^2}{8\cdot 15\cdot 6\cdot 17} = \underline{1{,}17} \quad (= 1{,}08^2, \text{ jfr. 3.1.2}),$$

og $$z_h^{korr} = \underline{1{,}04}.$$

Vi har både 1.04 og 1.17 mindre enn fraktilen 2,71, dvs. ingen forkasting. Dette er rimelig, de tre testene brukt på data i tabell 3.2 er jo i realiteten samme test, nemlig 3.1.1. Vi har bare brukt ulike hjelpemidler ved utregningen.

3.2 En 2 x 2 krysstabell.

Data: tabell 3.1 resp. 3.2.

Forutsetninger i hele avsnitt 3.2: Vi har trukket ett utvalg på n, her 2264 (resp. 23), personer.

Multinomisk fordeling.

3.2.1 Testing av uavhengighet.

Problem: Er det avhengighet mellom det å ha adgang til fritidshus og det å dra på helgetur ?

Etter det som følger av setning 2.2.1, kan det se ut som om vi må teste: $p_{ij} = p_{i+}\, p_{+j}$ for i = 1,2 og j = 1,2. Det er imidlertid nok å teste (f.eks.)

$$H_o: p_{11} = p_{1+}\, p_{+1}, \text{ idet de øvrige } p_{ij} \text{ vil følge av denne.}$$

Alternativet er $p_{11} \neq p_{1+}\, p_{+1}$ (og tilsvarende for de øvrige p_{ij}).

I følge 2.2.1 kan uavhengigheten også uttrykkes ved at de to betingede fordelingene for i, gitt henholdsvis j = 1 og j = 2, er like, dvs. at

$$H_o: \frac{p_{11}}{p_{+1}} = \frac{p_{12}}{p_{+2}} \quad \left(\text{og } \frac{p_{21}}{p_{+1}} = \frac{p_{22}}{p_{+2}}, \text{ som følger automatisk}\right).$$

Hvis vi her ser på den betingede fordelingen for n_{11}, gitt n_{+1}, vil den være binomisk med parametre p_{11}/p_{+1} og n_{+1}. Tilsvarende er n_{12} binomisk fordelt. Ved å betrakte n_{+1} og n_{+2} som gitte tall under testingen, er vi tilbake i samme modell som i avsnitt 3.1, og kan bruke testene som er angitt der for situasjon (ii). Det kan vises teoretisk at dette er en så bra løsning som vi kan håpe å finne. I dette tilfelle er altså den såkalte χ^2-uavhengighetstesten identisk med testen i avsnitt 3.1.3.

Vi kan faktisk også teste mot mer spesielle alternativ enn bare "avhengighet", idet vi kan uttrykke situasjonene (i) og (iii) i 3.1 ved hjelp av de betingede sannsynlighetene og teste dem som i 3.1. Dette kan være mer informativt enn bare å få konklusjonen "avhengighet". Se også kap. 6.

3.2.2. Teste om p_{ij} er lik gitte tall.

Vi vil teste

$$H_o: \; p_{ij} = p^o_{ij}, \text{ der } p^o_{ij} \text{ for } i = 1,2 \text{ og } j = 1,2 \text{ er gitte tall.}$$

Når $np^o_{ij} \geq 5$ er dette et spesialtilfelle med I = J = 2 av χ^2-føynings-testen i 2.4.1. Men hva med små utvalg ? Vi kan utlede tester mot gitte alternativ, men testprosedyren vil da avhenge av alternativet. Fra Aa.I kan en utlede løsninger for visse typer av alternativ. Vi skal se på ett.

3.2.2.a. Alle utfall like sannsynlige. Vi vil teste $H_o: p_{11} = p_{12} = p_{21} = p_{22} = 0,25$, mot alternativet $p_{21} < 0,25$ og de tre resterende sannsynlighetene lik hverandre (eller iallfall større en p_{21}). Det betyr altså at det er liten sannsynlighet for ikke å dra på helgetur og ha fritidshus, men ellers er det ikke særlige forskjeller.
Vi ser at H_o innebærer både uavhengighet og at $p_{1+} = p_{2+} = p_{+1} = p_{+2} = 0.5$.

Betinget binomisk test.
Under H_o er fordelingen av n_{12} binomisk, med $p_{12} = 0,25$. Vi kommer til at vi må forkaste H_o når $n_{12} \leq$ nedre ε-fraktil i denne fordelingen.

La oss endre data i tabell 3.2 til 24 observasjoner, med fordelingen

7	10	17
2	5	7
9	15	24

Med n = 24 og p = 0,25 i den binomiske fordelingen finner vi her at

$$P(n_{12} \leq 2) = 0.0398 \qquad (\text{og } (P(n_{12} \leq 3) = 0.115).$$

Med $\varepsilon = 0.05$ vil vi altså forkaste H_o i dette tilfelle, og konkludere med at p_{12} er mindre enn de øvrige. Vi har jo nettopp $n_{12} = 2$, som skal gi forkasting.

Det finnes tabeller over den binomiske fordeling for små n, f.eks. i Biometrika tables.

For stor n kan vi regne med at n_{12} er tilnærmet normalt fordelt, og forkaste H_o når

$$d = \frac{n_{12} - 0{,}25\ n}{\sqrt{0{,}25 \cdot 0{,}75n}} \leq -1{,}64 \quad \text{for } \varepsilon = 0{,}05$$

I vårt eksempel får vi d = - 1,90. Det skulle gi forkasting, men vi bør huske på at approksimasjonen i små utvalg kanskje ikke er god. Beregner vi den korrigerte verdi, får vi

$$d^{korr} = \frac{n_{12} - 0{,}25n + 0{,}5}{\sqrt{0{,}1875\ n}} = -1{,}60$$

som ikke gir forkasting. Sammenligner vi med de binomiske sannsynlighetene ovenfor, så ser vi at normalapproksimasjonen er dårlig i dette tilfelle.

3.2.2.b. χ^2-test

Hvis vi derimot brukte χ^2-føyningstesten (vi har $np^o_{ij} = 24 \cdot 0{,}25 = 6$ i alle ruter), får vi z = 5,7, mens 5%-fraktilen for 3 df er 7,81, altså ikke forkasting.

Denne siste testen er ikke "skreddersydd" for formålet, slik den binomiske testen ovenfor er. På den annen side kan den binomiske testen være dårligere hvis vi bruker den overfor andre alternativ enn de spesifiserte.

3.2.2.c. "Nødtest" for små utvalg

Hvis vi ikke har spesifiserte alternativ, og for få observasjoner til å bruke χ^2-testen, kan vi til nød forkaste nullhypotesen når sannsynligheten under H_o for å få det observerte resultat er liten. La oss kalle denne sannsynligheten $P^o(obs)$. Den skal være så liten at den samlede sannsynligheten

$$P^o(obs) + P(\text{tabeller med sannsynlighet} \leq P^o(obs)) \leq \varepsilon.$$

Vi (eller en EDB-maskin) må altså lage en "liste" over alle de tabellene med ulike n_{ij}-verdier som er mulige å få ut med gitt n, og regne ut sannsynligheten for hver av dem når alle $p^o_{ij} = 0.25$. Så plukker vi ut dem som har små sannsynligheter for å forekomme, og regner som angitt ovenfor. Testen har nivå ε, men kan ha dårlig teststyrke mot mange alternativ, forsåvidt er den i klasse med χ^2-testen. Det finnes EDB-program for slike tester, f.eks. i Weiss (1978).

3.2.3. Sammenlikning av to sannsynligheter

Problem: Vi vil teste om sannsynligheten for at en som drar på helgetur har adgang til fritidshus er større enn sannsynligheten for at vedkommende ikke har det (uten å si noe om dem som ikke drar).

Vi må sette opp

$$H_o: \; p_{11} = p_{12} \; (= \frac{1}{2} p_{1+}) \text{ mot } p_{11} > p_{12}. \quad \text{Eller}$$

$$\frac{p_{11}}{p_{1+}} = \frac{p_{12}}{p_{1+}} = 0{,}5 \quad \text{mot} \quad \frac{p_{11}}{p_{1+}} > 0{,}5.$$

(Vi får samme test med nullhypotesen $H_o^*: p_{11} \leq p_{12}$).
Her vil den betingede fordeling for n_{11}, gitt n_{1+}, være binomisk med parametre n_{1+} og p_{11}/p_{1+}, som under H_o er 0,5. Vi forkaster H_o hvis n_{11} er større eller lik øvre ε-fraktil i denne fordelingen. I eksempel 3.2 får vi opplagt ikke forkasting, idet vi har $n_{11} = 7 < 0{,}5\, n_{1+} = 8{,}5$ (som er forventet verdi med $p = 0{,}5$). Vi kan finne de binomiske sannsynlighetene v.h. av en tabell (eller regne dem ut). I store utvalg utnytter vi at fordelingen er tilnærmet normal og forkaster H_o hvis

$$\frac{n_{11} - 0{,}5\, n_{1+}}{0{,}5\sqrt{n_{1+}}} \geq 1{,}645 \quad \text{(øvre 5 prosent fraktil i den standardiserte normalfordelingen).}$$

Kontinuitetskorreksjonen består i å trekke fra 0,5 i telleren.

Vi kan sammenligne andre p_{ij} på tilsvarende måte, f.eks. $p_{11} = p_{21}$ (med n_{+1} gitt), eller $p_{21} = p_{22}$ (med n_{2+} gitt) osv.

3.2.4. Symmetri om en diagonal

Problem: Er sannsynligheten for både å dra på helgetur og ikke ha fritidshus større enn for (både) å ha fritidshus og ikke dra? (Det finnes bedre eksempler, se f.eks. avsnitt 8.1.1 om parvise observasjoner).

$$H_o: p_{12} = p_{21} \text{ mot } p_{12} > p_{21} \quad \text{(eventuelt andre alternativ).}$$

Vi kommer frem til en test som er helt analog med den i 3.2.3. Den betingede fordelingen, gitt $(n_{12} + n_{21})$, av n_{21} er binomisk, under H_o, med parametre 0,5 og $(n_{12} + n_{21})$. Vi forkaster H_o når n_{21} er mindre enn ε-fraktilen i denne. Jfr. Aa. I, avsnitt 3.2.

Vi bruker samme testen for H_o': $p_{12} \leq p_{21}$.

I eksempel 3.2 finner vi at $P(n_{21} \leq 2) = 0{,}033$. Med $n_{21} = 1$ forkaster vi altså H_o.

Vi kan behandle hypotesen:

$$H_o: p_{11} = p_{22}$$

på helt tilsvarende måte. (Vi tør ikke trekke på eksemplet lenger, det finnes jo andre, f.eks. paneldata).

3.2.5 Fullstendig symmetri eller fullstendig parvis likhet.

En sjelden gang kan en ha et problem med

$$H_o: \; p_{11} = p_{22} \text{ og } p_{12} = p_{21} \text{ mot minst én forskjell}$$

eller,

$$H_o': \; p_{11} = p_{12} \text{ og } p_{21} = p_{22} \quad " \quad " \quad " \quad "$$

Her innebærer H_o at $p_{11} + p_{12} = p_{22} + p_{21} = 0{,}5$

og $p_{11} + p_{21} = p_{22} + p_{12} = 0{,}5$,

dvs. at $p_{1+} = p_{2+} = p_{+1} = p_{+2} = 0{,}5$.

Dette minner om eksemplet i 3.2.2, men nå har vi <u>ikke</u> nødvendigvis $p_{11} = p_{12}$.

H_o' innebærer at $p_{11} + p_{21} = p_{12} + p_{22} = 0{,}5 = p_{+1} = p_{+2}$.
Dessuten at $p_{11} + p_{22} = p_{12} + p_{21} = 0{,}5$.

En måte å teste hypotesen H_o på, er å foreta to slike symmetritester som foreslått i avsnitt 3.2.4.

For hypotesen: $p_{11} = p_{22}$ vil n_{11} være betinget binomisk fordelt, med parametre 0,5 og $(n_{11} + n_{22})$. Vi forkaster hypotesen $p_{11} = p_{22}$ når n_{11} er enten mindre enn nedre $\varepsilon/2$-fraktil eller større enn øvre $\varepsilon/2$-fraktil i denne fordelingen.

Og hypotesen: $p_{12} = p_{21}$ forkaster vi når n_{21} er enten mindre enn nedre $\varepsilon/2$-fraktil eller større enn øvre $\varepsilon/2$-fraktil i en binomisk fordeling med parametre 0,5 og $(n_{12} + n_{21})$.

Selve H_o forkaster vi hvis vi får forkasting ved én eller begge disse testene.

Sannsynligheten for <u>ikke</u> å forkaste H_o når den er riktig, må være $(1 - \varepsilon)^2$, slik at sannsynligheten for å forkaste H_o er: $1 - (1 - \varepsilon)^2 = 2\varepsilon - \varepsilon^2$. Vi bør altså bruke $\varepsilon = 0{,}025$ hvis vi ønsker en test med nivå (i underkant av) 0,05.

3.2.6. Relativ symmetri.

Problem 1: Er sannsynligheten for å dra på helgetur for en som ikke har fritidshus mindre enn sannsynligheten for ikke å dra for en som har fritidshus ? (Igjen ville kanskje et annet eksempel være bedre).

Dette minner om 3.2.4, men vi sammenlikner her de betingede sannsynlighetene. Vi setter

$$H_o : \frac{p_{12}}{p_{+2}} = \frac{p_{21}}{p_{+1}} \quad \text{mot} \quad \frac{p_{12}}{p_{+2}} < \frac{p_{21}}{p_{+1}} .$$

(Vi kan også ha H_o^1 med $\geq$).

Det viser seg her at vi ledes til en Fisher-Irwin test igjen, med n_{+1} og $n_{12} + n_{21}$ gitt.
Test 1 blir: Vi skal forkaste H_o hvis n_{21} er stor, dvs. større enn øvre ε-fraktil i den hypergeometriske fordelingen som har sannsynlighetsfunksjon

$$f_o(n_{21}) = \frac{\binom{n_{12} + n_{21}}{n_{21}}\binom{n_{11} + n_{22}}{n_{11}}}{\binom{n}{n_{+1}}}$$

Se Aa. I, 3.3., s.21-24.
For store utvalg kan vi igjen bruke en tilnærmet normaltest, med observator

$$d = \left(\frac{n_{12}}{n_{+2}} - \frac{n_{21}}{n_{+1}}\right) \Big/ \sqrt{\frac{(n_{12} + n_{21})}{n} \cdot \frac{(n_{11} + n_{22})}{n} \left(\frac{1}{n_{+2}} + \frac{1}{n_{+1}}\right)}$$

Jfr. 3.1.2. Vi må forkaste H_o når d er mindre enn nedre ε-fraktil i den standardiserte normalfordelingen. Aaberge har forøvrig satt opp asymptotiske tester med kontinuitetskorreksjoner også i nevnerne, se Aa. I,4 (iii).

Vi kunne også ha formulert et

problem 2: Er sannsynligheten for ikke å ha fritidshus for en som drar på helgetur mindre enn sannsynligheten for å ha fritidshus for en som ikke drar ?

Vi må da sette

$$H_o^2: \ \frac{p_{12}}{p_{1+}} \geq \frac{p_{21}}{p_{2+}} \quad \text{mot} \quad \frac{p_{12}}{p_{1+}} < \frac{p_{21}}{p_{2+}} .$$

Vi får en Fisher-Irwin test som ovenfor, men nå med n_{1+} gitt ved siden av $n_{12} + n_{21}$.

Test 2 blir: forkast H_o når n_{21} er stor, dvs. større enn øvre ε-fraktil funnet fra

$$f_o(n_{21}) = \frac{\binom{n_{12} + n_{21}}{n_{21}}\binom{n_{11} + n_{22}}{n_{22}}}{\binom{n}{n_{2+}}}$$

jfr. Aa. I, 3.3., s. 21-24.

Ved å sammenligne med 3.1.1 og 3.2.1 ser vi at problemet og løsningen er analoge med homogenitetstesten. Vi har bare "byttet om" linjene for j = 2 i tabellen. Og ved test 2 er det dessuten linjesummene som er gitt istedenfor kolonnesummene.

Vi kunne også formulere et problem 3, som gjør at vi må teste f.eks.

$$H_o^3: \ \frac{p_{11}}{p_{+1}} = \frac{p_{22}}{p_{+2}} \quad \text{mot} \quad \frac{p_{11}}{p_{+1}} < \frac{p_{22}}{p_{+2}}$$

Nå er $p_{11} = p_{+1} - p_{21}$ og $p_{22} = p_{+2} - p_{12}$. Setter vi dette inn i H_o^3, får vi

$$1 - \frac{p_{21}}{p_{+1}} = 1 - \frac{p_{12}}{p_{+2}}, \text{ dvs. } \frac{p_{12}}{p_{+2}} = \frac{p_{21}}{p_{+1}}, \text{ med}$$

alternativet

$$\frac{p_{12}}{p_{+2}} < \frac{p_{21}}{p_{+1}}, \quad \text{slik at problem 3 er identisk med problem 1.}$$

Det finnes flere mulige problemstillinger for 2x2 tabeller, hver av dem kan behandles spesielt.

En del av de eksemplene vi har gitt, kan også behandles ved hjelp av de log-lineære modellene og metodene i kap. 6. Dette gjelder for store utvalg.

3.3. Toveistabeller med I linjer og to kolonner, eller med to linjer og J kolonner.

Disse tabellene er spesialtilfeller av I x J-tabeller med $I \geq 2$ og/eller $J \geq 2$, som behandles i avsnitt 3.4. Metodene i 3.4 gjelder derfor også for 3.3. Imidlertid er det enkelte problemstillinger som er spesielle for tabeller med 2 linjer eller kolonner, og vi kan i blant gjøre visse forenklinger. Dette gjør at vi vil behandle slike tabeller særskilt.

Som illustrasjon i dette avsnittet skal vi bl.a. bruke 3 x 2-tabellen nr. 1.1 over utdanning og arbeidstid, idet vi tenker oss ulike måter tallene kunne være fremkommet på.

Vi viser til 2.4.1 når det gjelder å teste fullt spesifiserte nullhypoteser mot uspesifiserte alternativ i store utvalg.

3.3.1. Komparative tabeller med to utvalg.

Forutsetninger: Vi har observert J = 2 utvalg med henholdsvis n_{+1} (= 752) og n_{+2} (= 568) observasjoner.
Utvalgene er stokastisk uavhengige av hverandre.
Multinomisk fordeling innen hvert utvalg.

Problem (i): Er det forskjell mellom de to (teoretiske) fordelingene for j = 1 og for j = 2 ? I eksemplet: er arbeidstidsfordelingen ulik for personer med yrkesutdanning og for personer uten yrkesutdanning ?

Nullhypotesen er at fordelingene er like, altså $p_{i1} = p_{i2}$ for $i = 1,2,\ldots,I$ mot et uspesifisert alternativ, der minst to av likhetstegnene ikke gjelder.

For stort nok utvalg, som vi har i eksemplet, kan vi bruke χ^2-testen i avsnitt 2.4.3, med (I-1)(2-1) = I-1 df. I eksemplet har vi altså 3-1 = 2 df. Når J = 2, kan observatoren skrives

$$Z_h = \sum_{i=1}^{I} \frac{(nn_{i1}-n_{+1}n_{i+})^2}{n_{+1}n_{+2}n_{i+}} .$$

Vi finner $Z_h = 4,65$ og kan ikke forkaste nullhypotesen, siden øvre 5-prosent fraktil for 2 df er 5,99. Beregningen kan utføres med SPSS-programmet Crosstabs, jfr. avsnitt 2.4.3.

For små utvalg og/eller for problemer med mer spesifiserte alternativ, bør vi bruke andre tester. Vi kan f.eks. ha alternativ som:

problem (ii): $p_{31} > p_{32}$ (resp. $p_{i1} > p_{i2}$ for en bestemt i) uten å si noe om de øvrige p-ene. I eksemplet: er sannsynligheten for å arbeide dagskift større for personer med yrkesutdanning enn for personer uten yrkesutdanning ?

Eller vi kan ha et

multippelt problem (iii): $p_{11} < p_{12}$ og $p_{21} < p_{22}$ (da er $p_{31} > p_{32}$). I eksemplet: Er sannsynlighetene for h.h.vis skiftarbeid og nattarbeid minst for personer med yrkesutdanning ?

Mer generelt kan vi ha $p_{i1} < p_{i2}$ for $i = 1,2,\ldots,m$, der $m < I$ og der vi enten har $p_{i1} > p_{i2}$ for $i = m + 1,..,I$, eller muligens ikke har dette oppfylt for alle $i > r$. (Vi kan også ha $p_{i1} < p_{i2}$ for en del i-verdier som ikke nødvendigvis følger etter hverandre).

For problem (ii) kan vi bruke en vanlig test for sammenlikning av to relative hyppigheter. Vi har nemlig at n_{31} og n_{32} (resp. n_{i1} og n_{i2}) hver er binomisk fordelt med parametre henholdsvis (p_{31}, n_{+1}) og (p_{32}, n_{+2}), (resp. (p_{i1}, n_{+1}) og (p_{i2}, n_{+2})). Vi skal f.eks. teste

$$p_{31} = p_{32} \quad \text{mot} \quad p_{31} > p_{32}.$$

Testen blir den som er beskrevet i 3.1.1, situasjon (i), dvs. Fisher-Irwins test for små utvalg og tilnærmet normaltest for større utvalg, se 3.1.2. Vi finner i vårt eksempel at

$$\left(\frac{n_{31}}{n_{+1}} - \frac{n_{32}}{n_{+2}}\right) = -0,03$$

så vi kan ikke forkaste nullhypotesen. (Forkasting krever en tilstrekkelig stor positiv differens.)

For et multippelt problem (iii) kan vi velge mellom flere testmetoder.

(iii) a: Med a priori gitte hypoteser, jfr. avsnitt 2.5.4, kan vi utføre to, (resp. m) Fisher-Irwin tester, dvs. vi tester

$$p_{11} = p_{12} \quad \text{mot} \quad p_{11} < p_{12}$$

ved hjelp av n_{11}, n_{12}, n_{+1} og n_{+2}.
Dessuten tester vi

$$p_{21} = p_{22} \quad \text{mot} \quad p_{21} < p_{22}$$

ved hjelp av n_{21}, n_{22}, n_{+1} og n_{+2}, osv.

I hver av testene bruker vi nivå $\varepsilon/2$, resp. ε/m, for å sikre oss at sannsynligheten for å begå minst én forkastingsfeil ikke overstiger ε, jfr. avsnitt 2.3.4.

Se H.I, avsnitt 4, om et lignende problem, der alternativet er $p_{i1} \neq p_{i2}$. I eksemplet får vi ikke forkasting ved noen av de to testene.

b) Vi kan også velge først å teste nullhypotesen som angitt under problem (i).
Hvis vi får forkasting ved denne testen, går vi videre og undersøker om

$$\frac{n_{i1}/n_{+1} - n_{i2}/n_{+2}}{\sqrt{n_{i+}(n-n_{i+})(n_{+1}^{-1} + n_{+2}^{-1})}} \; n < -\sqrt{z_{1-\varepsilon,I-1}}$$

der $z_{1-\varepsilon,I-1}$ er øvre ε-fraktil i χ^2-fordelingen med (I-1)df (i eksemplet har vi 2 df). Dette gjør vi for i = 1,2,..,m. (I eksemplet for i = 1 og 2). Vi kan da konkludere med $p_{i1} < p_{i2}$ for de i-verdier der ulikheten er oppfylt.

Ved denne testen vil nivået være ε.
Se H.I, avsnitt 5, testing av kontraster. Denne testmetoden kan vi bruke selv om vi ikke på forhånd har bestemt oss for hvilke p_{i1}- og p_{i2}-verdier vi vil sammenlikne.

Mer spesifiserte situasjoner.

Aaberge har utledet metoder for testing mot alternativ som er mer spesifiserte enn dem vi har i problem (iii), idet vi også angir proporsjonalitetsfaktorer mellom forholdstallene

$$\frac{p_{11}}{p_{12}}, \frac{p_{21}}{p_{22}}, \frac{p_{31}}{p_{32}} \quad \text{osv.}$$

Hvis vi f.eks. tester nullhypotesen mot alternativet

$$\frac{p_{11}}{p_{12}} = \frac{p_{21}}{p_{22}} > \frac{p_{31}}{p_{32}},$$

så skal vi forkaste nullhypotesen hvis $y = n_{11} + n_{21}$ er større enn den øvre ε-fraktilen i nullfordelingen av y, som her er den hypergeometriske fordeling

$$g(y|n_{+1}, n_{+2}, n_{1+}, n_{2+}, n_{3+}) = \binom{n - n_{3+}}{y}\binom{n_{3+}}{n_{31}} \Big/ \binom{n}{n_{+1}}.$$

For eksemplet i tabell 1.1 bruker vi den normale approksimasjonen til denne fordelingen, og forkaster H_o når

$y = n_{11} + n_{21}$ er så stor at

$$y \geq \left(n_{+1}\frac{n - n_{3+}}{n}\right) + Z_{1-\varepsilon}\sqrt{n_{+1}\frac{n-n_{3+}}{n}\cdot\frac{n_{3+}}{n}\cdot\frac{n-n_{+1}}{n-1}}$$

der $Z_{1-\varepsilon}$ er øvre ε-fraktil i standard normalfordelingen.
I eksemplet har vi $y = 121 + 19 = \underline{130}$. Høyresiden blir, med $\varepsilon = 0{,}05$,

$$752\,\frac{228}{1320} + 1.64\sqrt{752\,\frac{228}{1320}\cdot\frac{1092}{1320}\cdot\frac{568}{1319}} = \underline{141}.$$

Vi kan altså heller ikke forkaste denne nullhypotesen her.

Aaberges tester kan tilpasses en rekke andre situasjoner der vi a priori kan spesifisere andre proporsjoner mellom forholdstallene. Vi kommer tilbake til enkelte slike i avsnitt 3.4.

3.3.2. Komparative tabeller med J > 2 binomiske utvalg.

Det finnes mange forskjellige testmetoder for de ulike situasjonene i dette tilfellet. Vi skal bruke data fra tabell 3.3 a og b nedenfor til å illustrere tester i dette avsnittet.

Tabell 3.3.a. Personer som ikke disponerer fritidshus, fordelt etter alder og om de var på helgetur eller ikke i observasjonsperioden.

		Aldersgruppe				
På helgetur	i	15 - 24 j = 1	25 - 34 2	35 - 54 3	55 - 74 4	Sum
Nei	1	51	74	143	220	488
Ja	2	219	261	291	245	1016
Sum		270	335	434	465	1504

Tabell 3.3.b. Konstruerte tall for et lite utvalg fordelt som i tabell 3.3.a.

		Aldersgruppe nr.			
På helgetur	i	j = 1	2	3	Sum
Nei	1	1	2	2	5
Ja	2	4	4	2	10
Sum		5	6	4	15

Forutsetninger: Vi har observert J uavhengige utvalg, med henholdsvis $n_{+1}, n_{+2}, \ldots, n_{+J}$ observasjoner. Hver av variablene $n_{11}, n_{12}, \ldots, n_{1J}$ er binomisk fordelt, med parametre h.h.vis (p_{11}, n_{+1}), (p_{12}, n_{+2}), osv. (Vi forutsetter altså her at tallene i tabell 3.3.a er fremkommet på en annen måte enn de faktisk gjorde i fritidsundersøkelsen.)

Problem (i). Vi vil teste om det er forskjell mellom p-ene i de J utvalgene, dvs.: er det ulik sannsynlighet for å dra på helgetur i de fire aldersgruppene ?

Som nullhypotese setter vi

$$H_o: p_{11} = p_{12} = \ldots = p_{1J}, \quad \text{mot alternativet } p_{1j_1} \neq p_{1j_2} \text{ for to eller flere utvalgspar.}$$

Test a. For store nok utvalg kan vi bruke testen 2.4.3, med I = 2 og J = J. Vår testobservator kan i dette tilfelle skrives

$$Z_h = \sum_{j=1}^{J} \frac{(n\, n_{1j} - n_{1+}n_{+j})^2}{n_{1+}\, n_{2+}\, n_{+j}} .$$

Under H_o er observatoren χ^2-fordelt med (J - 1)df. Vi forkaster H_o når den observerte Z_h er større eller lik øvre ε-fraktil i denne fordelingen.

I eksemplet i 3.3.a finner vi $Z_h = 86$. Vi har 3 df, med øvre 1-prosent-fraktil 11.34.

Vi forkaster altså nullhypotesen og konkluderer med at det er forskjell på sannsynlighetene.

Det blir påstått at χ^2-testen gir god approksimasjon i 2 x J tabeller selv om forventet antall i rutene i tabellene er så lavt som 1. Jfr. Lewontin & Felsenstein (1965).

I eksempel 3.3.b. finner vi

$$Z_h = \frac{(15\cdot 1 - 5\cdot 5)^2}{5\cdot 10\cdot 5} + \frac{(15\cdot 2 - 5\cdot 6)^2}{5\cdot 10\cdot 6} + \frac{(15\cdot 2 - 5\cdot 4)^2}{5\cdot 10\cdot 4} = 0.9.$$

Selv om vi her p.g.a. de små utvalgene kan være tilbøyelig til å velge en noe større ε enn vanlig, f.eks. 0,10, så får vi ikke forkasting. Den øvre 10-prosentfraktilen for to df er 4,61.

Nærmere analyse av tabellen etter forkasting av H_o.

Vi ønsker ofte å undersøke om det er spesielle (i,j)-kombinasjoner som skiller seg ut fra de andre. Vi kan da gå videre, hvis vi har fått forkasting med χ^2-testen ovenfor, og sammenlikne to aldersgrupper f.eks. nr. 1 og 2, dvs. vi vil teste

$$p_{21} = p_{22} \quad \text{mot} \quad p_{21} > p_{22}$$

Vi estimerer da variansen v for $\hat{p}_{21} - \hat{p}_{22}$, ved

$$\hat{v} = (1/n_{+1} + 1/n_{+2}) \frac{n_{21} + n_{22}}{n_{+1} + n_{+2}} \cdot \frac{n_{11} + n_{12}}{n_{+1} + n_{+2}},$$

og forkaster hypotesen $p_{21} = p_{22}$ hvis

$$(\hat{p}_{21} - \hat{p}_{22})/\sqrt{\hat{v}} \geq \sqrt{z_{1-\varepsilon,J-1}},$$

hvor $z_{1-\varepsilon,J-1}$ er øvre ε-fraktil i χ^2-fordelingen med (J-1) df. I vårt første eksempel ovenfor altså 11,34.

Vi finner der $\hat{p}_{21} = 0{,}811$, $\hat{p}_{22} = 0{,}779$, $\hat{v} = 0{,}001096$ og $0{,}032/0{,}0331 = 0{,}97$ som er mindre enn $\sqrt{11{,}34} = 3{,}37$. Vi får altså ikke forkasting. Hvis vi på forhånd hadde bestemt oss for å teste nettopp disse to p-ene, uten å se på resultatet av andre tester, ville vi brukt fraktilen 1,64 i normalfordelingen.

Ved å bruke $z_{1-\varepsilon,J-1}$ isteden, sikrer vi oss at sannsynligheten vil være høyst ε (tilnærmet, for stor n) for å forkaste nullhypotesen når den er riktig. Dette vil også gjelde om vi foretar flere sammenlikninger av p-er på tilsvarende måte. Ja, vi kan undersøke hvor mange såkalte kontraster mellom p-ene som vi ønsker, dvs. lineærkombinasjoner

$$\sum_j c_j p_{ij}, \text{ der } \sum_j c_j = 0.$$

Vi viser til Sverdrup (1975) og Haldorsen IO 77/41.

b. For små utvalg kan vi, om vi er redde for at χ^2-testen ikke er god nok, bruke Fisher-Irwin testen (3.1) og sammenlikne p-ene parvis. Vi må da bruke nivå ε/m på hver enkelt test, der m er tallet på tester vi utfører.

I eksempel 3.3.b. kan vi ikke forkaste noen av de tre hypotesene $p_{11} = p_{12}$, $p_{11} = p_{13}$ eller $p_{12} = p_{13}$ på noe rimelig nivå med denne testen.

Problem (ii). Vi vil teste om det er en viss ordning av p_{ij}-verdiene etter størrelse. Vi kan f.eks. ha

$$H_o: p_{11} = p_{12} = \ldots = p_{1J} \quad \text{mot} \quad p_{11} \leq p_{12} \leq p_{13} \leq \ldots \leq p_{1J},$$

med minst én ulikhet. Vi utelukker altså a priori at $p_{1r} > p_{1j}$ for noen $r < j$.

Her er det også flere mulige testmetoder.

a. Vi kan sammenlikne utvalg nr. 1 og utvalg nr. J ved å teste

$$H_o' : p_{11} = p_{1J} \quad \text{mot} \quad p_{11} < p_{1J},$$

ved hjelp av n_{11}, n_{+1} og n_{1J}, n_{+J}. Her kan vi igjen bruke Fisher-Irwin testen eller den tilnærmede normaltesten som i 3.2. Hvis vi forkaster H_o' så har vi samtidig forkastet H_o. Teststyrken ved denne testen kan være for dårlig, bl.a. fordi vi ikke utnytter alle observasjonene.

For data i 3.3.a får vi z = 7,7 og dermed forkasting ved normaltesten (3.1.2). For data i 3.3.b. får vi ikke forkasting.

b. Vi kan slå sammen hele observasjonsmaterialet i to grupper, en bestående av de j første utvalgene og en av de J-j siste. Så utfører vi den tilsvarende testen som under a, men nå ved hjelp av

$$(n_{11} + n_{12} + \ldots + n_{1j}),\ (n_{+1} + n_{+2} + \ldots + n_{+j}) \text{ og}$$

$$(n_{1\,j+1} + n_{1\,j+2} + \ldots + n_{1J}),\ (n_{+j+1} + n_{+j+2} + \ldots + n_{+J}).$$

Også her vil forkasting av nullhypotesen innebære forkasting av H_o.

c. FIA-testen fra Amundsen og Ljøgodt (1979 og 1981). Det er en feil i den første artikkelen som korrigeres i den andre og som gjør testen mer tungvint å bruke enn opprinnelig antatt. For forholdsvis få utvalg, f.eks. J = 3 eller 4, med få observasjoner, egner den seg bra. For eksemplet i 3.3.b sammenlikner vi først utvalg 1 og 2 ved hjelp av Fisher-Irwins test. Vi innfører betegnelsene $v_1 = n_{11}$ og $v_2 = n_{11} + n_{12}$. Så beregner vi sannsynlighetene under H_o for å få v-verdier lik eller mindre enn de observerte , gitt h.h.vis $n_{11} + n_{12}$ og n_{1+}. Vi finner

$$P(v_1 \leq n_{11}) = P(v_1 = 0) + P(v_1 = 1) = 0,576.$$

fra den hypergeometriske nullfordelingen for v_1 som her er

$$g_o(v_1|v_2) = \frac{\binom{5}{v_1}\binom{6}{3-v_1}}{\binom{11}{3}} .$$

Deretter slår vi sammen de to første utvalgene og sammenlikner resultatet med det siste, dvs. vi finner

$$P(v_2 \leq n_{11} + n_{12}) = P(v_2 \leq 3) = 0.407$$

fra

$$g_o(v_2|v_3) = \frac{\binom{11}{v_2}\binom{4}{5-v_2}}{\binom{15}{5}} .$$

Vi ser at sannsynligheten for begge resultatene, som her er lik produktet av de to sannsynlighetene, blir

$$0,576 \cdot 0,407 = 0,234.$$

Vi kan ikke forkaste nullhypotesen. Hadde vi isteden fått resultatet 0,023, ville vi forkastet H_o når $\varepsilon = 0,05$.

d. En trinnvis test er foreslått av Bjørnstad II. Alle testene nedenfor er Fisher-Irwin (eller tilnærmet normal-)tester.

1. trinn. Vi begynner som i a, med å teste de to p-ene som står lengst fra hverandre, vi kan si de har avstand J-1, altså

$$p_{11} = p_{1j} \quad \text{mot} \quad p_{11} < p_{1j} \quad \text{med nivå } \varepsilon.$$

Hvis vi ikke kan forkaste H_o, stopper vi med dette.

Hvis vi kan forkaste H_o og dermed si at $p_{11} < p_{1j}$ går vi til

2. trinn. Vi tester nå de p-ene som har avstand J-2, nemlig

(i) $p_{11} = p_{1(J-1)}$ mot $p_{11} < p_{1(J-1)}$ med nivå ε

samt

(ii) $p_{12} = p_{1J}$ mot $p_{12} < p_{1J}$ med nivå ε

Hvis vi ikke får forkasting, stopper vi. Hvis vi får forkasting ved minst en av testene, fortsetter vi til 3. trinn.

3. trinn. Vi tester de p-ene som har avstand J-3, dvs.

$$p_{1i} = p_{1(J-3+i)} \text{ mot } p_{1i} < p_{1(J-3+i)} \text{ med nivå } \frac{\varepsilon}{2},$$

for i = 1,2,3 hvis begge trinn ovenfor ga forkasting.
Hvis (1) ikke ga forkasting, sløyfer vi testene med i = 1 og 2.
Hvis (2) ikke ga forkasting, sløyfer vi testene med i = 2 og 3.

Vi tester altså ikke p-er mot hverandre når de ligger mellom to som ikke gir forkasting. Slik fortsetter vi, med testing av p_{1i} mot $p_{1(J-4+i)}$ på nivå $\frac{\varepsilon}{3}$, mot $p_{1(J-5+i)}$ på nivå $\frac{\varepsilon}{4}$, osv. så lenge prosessen ikke stopper opp fordi vi ikke får forkasting. Hvis vi får minst en forkasting på alle trinn, blir den siste testen

(J - 1). trinn. Test $p_{1i} = p_{1i+1}$ mot $p_{1i} < p_{1i+1}$ og med nivå $\varepsilon/k-2$, for en eller flere i-verdier, bestemt av hvilke forkastinger vi har fått underveis.

Valget av de ulike nivåene skal sikre at sannsynligheten for minst en feilaktig forkasting av nullhypotesen ikke skal overstige ε.
I vårt eksempel vil vi ikke få forkasting i 1. trinn og dermed må vi stoppe.

e. En test som passer for et forholdsvis glatt forløp av p_{1j}-rekken og som egner seg best for store utvalg, kan vi få ved å sette

$$p_{1j} = \alpha + \beta t_j \text{ (+ eventuelle ledd med } t_j^2, t_j^3 \text{ e.l.)},$$

og teste $\beta = 0$ mot $\beta > 0$, ved vanlige regresjonsmetoder. Her må vi kunne velge noen rimelige tall $t_1 < t_2 < t_3 < \ldots < t_J$ som kan gi et visst summarisk bilde av p_{1j}-rekken. Se f.eks. Sverdrup II.

3.3.3. Krysstabeller med I linjer og to kolonner, eller med to linjer og J kolonner. Testing av uavhengighet

Forutsetninger: Vi har observert ett utvalg med n observasjoner. Multinomisk fordeling med gitt n.

Problem: Er det avhengighet mellom x og y, dvs. mellom linjegruppering og kolonnegruppering ? I eksemplet i tab. 1.1 er spørsmålet altså: er det avhengighet mellom arbeidstidsordning og yrkesutdanning ?

Nullhypotesen er "uavhengighet". Uavhengighet betyr i følge avsnitt 2.3 at $p_{ij} = p_{i+}p_{+j}$ for alle (i,j)-kombinasjoner. Videre betyr det at de betingede fordelingene for gitt x, dvs. respektive for j = 1 og j = 2, er like, dvs. at

$$\frac{p_{i1}}{p_{+1}} = \frac{p_{i2}}{p_{+2}} \; (= p_{i+}) \quad \text{for } i = 1,2,\ldots,I.$$

Dette leder oss til samme testmetode som i 3.3.1, problem (i), dvs. at vi for store utvalg bruker χ^2-testen i 2.4.3 med (I-1)df som om n_{1+} og n_{2+} er gitte tall, se f.eks. A II 14.2.2.

Tilsvarende har vi med to linjer og J kolonner at

$$\frac{p_{1j}}{p_{1+}} = \frac{p_{2j}}{p_{2+}} \; (= p_{+j}) \quad \text{for } j = 1,2,\ldots,J$$

og vi bruker χ^2-testen i 2.4.3 med (J-1)df, på samme måte som i avsnitt 3.3.2.

Ut fra tallene i tabell 3.3.a vil vi altså konkludere med at det er avhengighet mellom alder og det å dra på helgetur (for personer som ikke disponerer fritidshus).

Nærmere analyse av tabellen etter forkasting av H_o kan vi foreta på samme måte som i 3.3.2, idet vi nå bruker betingede tester, gitt $n_{+1}, n_{+2}, \ldots n_{+j}$.

En annen metode er foreslått av Lancaster og Irwin, se f.eks. Lancaster (1969), Darroch (1974). Den går ut på å sette opp (J-1) uavhengige χ^2-fordelte observatorer, hver med 1 df, som addererer seg opp til vår Z_h, som jo har (J-1)df. Hver av addendene dannes ut fra en 2 x 2-tabell, og de ulike 2 x 2-tabellene velges etter et bestemt mønster som bestemmes a priori, f.eks. slik:

$$\begin{array}{|cc} n_{11} & n_{12} \\ n_{21} & n_{22} \end{array} \qquad \text{med } z_1 = \frac{n^2(n_{22}n_{11}-n_{12}n_{21})^2}{n_{1+}n_{2+}n_{+1}n_{+2}(n_{+1}+n_{+2})},$$

$$\begin{array}{|cc} (n_{11}+n_{12})n_{13} \\ (n_{21}+n_{22})n_{23} \end{array} \qquad \text{med } z_2 = \frac{n^2\, n_{23}(n_{11}+n_{12}) - n_{13}(n_{21}+n_{22})^2}{n_{1+}\, n_{2+}(n_{+1}+n_{+2})n_{+3}\,(n_{+1}+n_{+2}+n_{+3})},$$

osv., med addisjon av en ny kolonne ad gangen til den foregående første kolonne, og til slutt

$$\begin{array}{|c} \sum_{j=1}^{J-1} n_{1j} \quad n_{1J} \\ \sum_{j=1}^{J-1} n_{2j} \quad n_{2J} \end{array} \qquad \text{med } z_{J-1} = \frac{n^2\left(n_{2J}\sum_{1}^{J-1} n_{1j} - n_{1J}\sum_{1}^{J-1} n_{2j}\right)^2}{n_{1+}n_{2+}\left(\sum_{1}^{J-1} n_{+j}\right) n_{+J}\, n}$$

Ved å ordne kolonnene på passende måte på forhånd kan vi få frem spesielle grupper vi kan være interessert i.

Hver av z_j-verdiene sammenlignes så med øvre ε-fraktil i χ^2-fordelingen med 1 df, $z_{1-\varepsilon,1}$ og vi påstår signifikant forskjell for tabell nr. j når $z_j \geq z_{1-\varepsilon,1}$. I vårt eksempel 3.3.a har vi

51	74	
219	261	
270	335	605

$$\text{med } z_1 = \frac{1504^2(261\cdot 51-74\cdot 219)^2}{488\cdot 1016\cdot 270\cdot 335\cdot 605} = 0{,}70$$

125	143	
480	291	
605	434	1039

$$\text{med } z_2 = \qquad\qquad = 17{,}4$$

268	220	488
771	245	1016
1039	465	1504

med $z_3 = \dfrac{1504(245\cdot 268 - 220\cdot 771)^2}{488\cdot 1016\cdot 1039\cdot 465} = 67{,}9$

Vi har $z_1 + z_2 + z_3 = 86 = z_h$ i punkt 3.3.2. Både z_2 og z_3 er større enn øvre 5 prosentfraktilen 3,84 i χ^2-fordelingen med 1 df, dvs. det er signifikante forskjeller mellom nest eldste og de to yngste gruppene sett under ett og mellom den eldste og de tre øvrige sett under ett.
(NB: Vi bør bruke nivå $\varepsilon/3$.)

For <u>små utvalg</u> og/eller for problemer med mer spesifiserte alternativ bør vi som vanlig prøve å finne bedre tester.

Noen av metodene i 3.3.1 og i 3.3.2 kan brukes også her, nå som betingede tester.

Vi kan ha problemer som kan uttrykkes som problem (ii) eller (iii) i 3.3.1, men nå med <u>betingede sannsynligheter.</u> Vi kan f.eks. ha:

<u>problem (ii):</u> $\dfrac{p_{i1}}{p_{+1}} > \dfrac{p_{i2}}{p_{+2}}$ for en gitt i,

eller

<u>problem (iii):</u> $\dfrac{p_{i1}}{p_{+1}} < \dfrac{p_{i2}}{p_{+2}}$ for i = 1,2,...,m der m < I.

Vi kan foreta parvise sammenlikninger i (2 x J)-tabellen, slik som i 3.3.2.b, nå med de betingede sannsynligheter, og vi kan teste mot alternativet $\dfrac{p_{11}}{p_{+1}} \leq \dfrac{p_{12}}{p_{+2}} \leq \ldots \leq p_{1J}$, slik som angitt for 3.3.2, problem (ii).

Det er også mulighet for å teste mot visse typer av spesifiserte alternativ ved Aaberges metoder eller ved å formulere modeller som hos Sverdrup.

3.4 Toveistabeller med I linjer og J kolonner

Vi skal nå se på tabeller med $I \geq 2$ og $J \geq 2$. Vi forutsetter stadig multinomisk fordeling.

Som eksempel skal vi bruke (tilnærmede) tall fra fritidsundersøkelsen fordelt etter kommunetype for bosted.

Tabell 3.4.a. Personer som ikke disponerer fritidshus, fordelt etter type av bostedskommune og tallet på helgeturer i observasjonsperioden.

Helgeturer		Bostedskommunetype				
		Landbruk, fiske m.m.	Bland.landbr., industri m.m.	Særlig sentrale bl.tj.yting/ industri	Øvrige bl. tj.yting/ industri	Sum
Gruppe	antall	j = 1	2	3	4	
1	0	122	125	137	104	488
2	1-2	71	100	78	52	301
3	3-5	64	83	87	67	301
4	6-9	32	52	57	29	170
5	10-14	15	24	39	14	92
6	15-19	7	18	23	12	60
7	20 og over	10	17	37	28	92
	Sum	321	419	458	306	1504

Vi minner om at metoder for analyse av problemer ved hjelp av data i en slik tabell også finnes i kapitlene 6 og 7.

3.4.1. Spesifiserte sannsynligheter under null-hypotesen, lite spesifiserte under alternativet. Føyningstester

Vi ønsker her å teste nullhypotesen

$$H_o: p_{ij} = p_{ij}^o \quad \text{for} \quad \begin{matrix} i = 1,2,\ldots,I \\ j = 1,2,\ldots,J, \end{matrix}$$

hvor p_{ij} er kjente tall.

For krysstabeller har vi χ^2-testen i avsnitt 2.4.1, med (IJ-1) frihetsgrader, jfr. det gjennomregnede eksempel der. Ifølge Lewontin & Felsenstein [1965] er approksimasjonen til χ^2-fordelingen god selv om forventet antall observasjoner i de enkelte ruter er så lavt som 1.

For komparative tabeller, der kolonnesummene (eller linjesummene) er gitt, har vi testen i 2.4.2, med (I-1)J df. (respektive I(J-1) df.).

For tabeller med små tall og/eller mer spesifiserte alternativ kan det utvikles spesielle tester.

"Nødtest". For tabeller med meget små tall og kanskje tilfeldige nuller, samt lite spesifiserte alternativ, kan vi bruke en test analog med "nødtesten" i avsnitt 3.2.2. Vi regner ut sannsynligheten P_o (obs.) under H_o for å få akkurat de tallene i tabellen som vi har fått. Dessuten regner vi ut sannsynlighetene for så mange andre mulige tabeller (eventuelt med de gitte marginaler), at vi er sikre på å ha fått med alle som har mindre sannsynlighet for å forekomme enn den vi faktisk har fått. Vi forkaster nullhypotesen hvis summen av alle disse små sannsynlighetene (mindre eller lik P^o (obs.)) er mindre eller lik vårt valgte nivå ε.

3.4.2. Sammenlikning av fordelingene for kolonner (resp. linjer) i komparative tabeller. Homogenitetstester

Forutsetninger: Komparative tabeller med J utvalg som har h.h.vis $n_{+1}, n_{+2}, \ldots, n_{+J}$ observasjoner. Multinomisk fordeling innen hvert utvalg og uavhengighet mellom utvalgene.

Problem (i): Er det forskjell mellom de J (teoretiske) fordelingene ? I eksemplet i tabell 3.4.a: Er fordelingen m.h.t. antall turer forskjellig for personer bosatt i de 4 ulike kommunetypene ?

En homogenitetstest for store utvalg for dette problemet er gitt i avsnitt 2.4.3.
For eksemplet finner vi

$$z_h = \sum_{j=1}^{4} \sum_{i=1}^{7} \frac{(n_{ij} - n_{+j}\, n_{i+}/n)^2}{n_{+j}\, n_{i+}/n} = 41{,}3.$$

Nå er øvre 5%-fraktil i χ^2-fordelingen med (7-1)(4-1) = 18 df. lik 28,87. Vi kan derfor konkludere med at fordelingen på antall turer er forskjellig.

Nærmere analyse av tabellen etter forkasting av nullhypotesen

Hvis vi ønsker det, kan vi gå videre med sammenlikning av spesielle sannsynligheter i tabellen og det uten å miste kontrollen over sannsynlighetsnivået for den multiple testen som vi dermed bruker, jfr. avsnitt 3.3.2.

Anta at vi (gjerne etter å ha sett på tallene i tabellen) vil undersøke om sannsynligheten for å dra på 20 turer eller mer er større for personer i kommunetype 3 enn i type 1. Dvs. vi tester

$$p_{73} \leq p_{71} \quad \text{mot} \quad p_{73} > p_{71}.$$

Vi kan bruke observatoren

$$d = \left(\frac{n_{73}}{n_{+3}} - \frac{n_{71}}{n_{+1}} \right) \sqrt{\frac{n_{73}+n_{71}}{n_{+3}+n_{+1}} \left(1 - \frac{n_{73}+n_{71}}{n_{+3}+n_{+1}} \right) \left(\frac{1}{n_{+3}} + \frac{1}{n_{+1}} \right)}^{-1}$$

og må ha

$$d \geq \sqrt{28,87} = 5.37$$

for å kunne påstå $p_{73} > p_{71}$.
Vi finner d = 2,86 og kan ikke påstå noe.

Vi bruker her kvadratroten av χ^2-fraktilen fra homogenitetstesten fordi vi dermed sikrer oss kontroll over nivået. Vi kan foreta så mange sammenligninger av denne typen vi bare vil. Sannsynligheten vil høyst være 5% for å gjøre en (eller flere) forkastingsfeil. Se Sverdrup [1977] avsnitt 2, eller H. I.

Hvis vi på forhånd, uten å se på resten av problemet, hadde bestemt oss for å sammenlikne disse to sannsynlighetene, ville vi brukt fraktilen 1,64 jfr. avsnitt 3.2.1 i normalfordelingen (og her fått forkasting).

For små utvalg og/eller mer spesifiserte alternativ, er det i en del situasjoner mulig å utvikle spesielle tester. Dette er bl.a. gjort for problemer der det foreligger en viss ordning av kategoriene i tabellen. Vi kan f.eks. ha at kategoriene i forspalten er ordnet i stigende rekkefølge, slik som i tabell 3.4.a, og så vil vi teste om vi har en viss ordning av p_{ij}-ene etter kommunetype, f.eks. at sannsynlighetene for å dra på få turer avtar fra type 1 til 4, mens sannsynlighetene for å dra på mange turer øker fra type 1 til 4

$$p_{11} \geq p_{12} \geq p_{13} \geq p_{14}$$

$$p_{21} \geq p_{22} \geq p_{23} \geq p_{24}$$

$$p_{i1} \geq p_{i2} \geq p_{i3} \geq p_{i4}$$

$$p_{i+1,1} \leq p_{i+1,2} \leq p_{i+1,3} \leq p_{i+1,4}$$

$$p_{71} \leq p_{72} \leq p_{73} \leq p_{74}$$

Vi kan da f.eks. sette

$$p_{ij} = \alpha_i + \beta_i t_j, \quad \text{der} \sum_{i=1}^{7} \alpha_i = 1 \text{ og } \sum_{i=1}^{7} \beta_i = 0$$

mens t_j er valgte tall, f.eks. $t_1 = 1$, $t_2 = 2$, $t_3 = 3$, $t_4 = 4$.
Vi behøver ikke tro at denne modellen er "riktig", vi bruker den for å undersøke

om det er en viss trend i p-verdiene. Vi tester om $p_{ij} = \alpha_i$ (dvs. den samme for alle grupper) ved å teste om

$$\beta_i = 0 \quad \text{mot} \quad \beta_i \neq 0 \quad \text{for } i = 1,2,\ldots,J,$$

som i et vanlig regresjonsanalyseproblem. Vi kan også undersøke om vi kan konkludere med $\beta_i < 0$ eller $\beta_i > 0$.

3.4.3. Uavhengighet i krysstabeller med I linjer og J kolonner

Forutsetninger: Vi har observert ett utvalg med n observasjoner. Multinomisk fordeling med gitt n.

Problem (i): Er det avhengighet mellom linjegruppering og kolonnegruppering ? I eksemplet i tabell 3.4.a: er det avhengighet mellom type av bostedskommune og tallet på helgeturer i løpet av et år ?

Nullhypotesen kan formuleres som at alle de betingede sannsynlighetene i en linje er like (for alle linjene), dvs.

$$H_o: \quad \frac{p_{i1}}{p_{+1}} = \frac{p_{i2}}{p_{+2}} = \ldots = \frac{p_{iJ}}{p_{+J}}, \quad \text{for } i \text{ er } 1,2,\ldots,I.$$

Også her ledes vi til å bruke en betinget test, gitt $n_{+1}, n_{+2},\ldots,n_{+J}$. For store utvalg kan vi derfor bruke χ^2-testen i 2.4.3.

Vi regner ut z_h akkurat som vi gjorde i avsnitt 3.4.2, og forkaster H_o hvis vi har z_h større enn øvre ε-fraktil i χ^2-fordelingen med $(I - 1)(J - 1)$ frihetsgrader. For eksemplet i tabell 3.4.a finner vi $z_h = 41,3$ og må forkaste hypotesen om uavhengighet, idet øvre 5-prosent-fraktil i χ^2-fordelingen med 18 df er 28,87.

Vi kan også bruke observatoren (2.4.5) på akkurat samme måte.

Nærmere analyse av tabellen etter forkasting av uavhengighetshypotesen kan vi foreta på tilsvarende måte som i avsnitt 3.4.2, nå med betingede tester.

3.4.4. Regresjon for å teste monotonitet

Hvis vi kan kvantifisere kategoriene begge veier i tabellen, f.eks. ved å gi kategoriene for antall dager "verdiene" y = 0,1,5,4 7,5, 12,17 og 25, og kommunetypene verdiene t = 1,2,3,4, så kan vi estimere regresjonen

$$y = \alpha + \beta t$$

og teste om $\beta = 0$ mot $\beta > 0$ (f.eks.). Her må vi igjen huske at regresjonen bare gir et røfft uttrykk for den eventuelle samvariasjonen mellom de variable. Poenget er å få frem en stigende/avtagende tendens, men ikke å bruke regresjonen som "modell" for sammenhengen i andre forbindelser.

En Lancaster-Irwin test, som beskrevet i avsnitt 3.3.3, kan gi en del informasjon.
Vi bør også være oppmerksom på at de vanlige variansformlene neppe gjelder her, fordi det er vanskelig å tenke seg samme varians for alle y-ene.

3.5. Toveistabeller med gitte marginalsannsynligheter. Iterativ skalering.

Vi skal kort nevne en problemstilling som forekommer sjelden, men som nevnes i noen lærebøker, nemlig at marginalsannsynlighetene $p_{1+}, p_{2+}, \ldots, p_{I+}$ og $p_{+1}, p_{+2}, \ldots p_{+J}$ i en I x J-tabell er kjente tall, og vi ønsker å estimere sannsynlighetene p_{ij} i tabellen, eller teste hypotesen om f.eks. uavhengighet.

Her kan p_{ij} estimeres ved en skrittvis beregningsmåte som kalles iterativ skaleringsprosedyre (ISP) eller iterativ proporsjonal føyning IPFP. Se Plackett [1974], avsnitt 3.4 og Fienberg [1970].

Den samme beregningsmåten kan utnyttes ved uavhengighetstesting i problemer der ML-estimatene for p_{ij}-verdiene ikke kan angis eksplisitt. Se f.eks. BFH, avsnittene 3.5 og 3.6.

4. TREVEISTABELLER.

I avsnitt 2.2.5 og tabell 2.2.c (tallet på personer gruppert etter antall helgeturer, adgang fritidshus og etter kjønn) er det gitt notasjon for og eksempel på en treveistabell. Se også tabellene i avsnitt 4.2. nedenfor.

Både formulering av problemstillingen og selve analysen vil ofte være mer komplisert ved treveisgrupperte data enn ved toveistabeller, kanskje med slike unntak som nevnes i avsnitt 4.1.

Sondringen mellom krysstabeller og komparative tabeller gir her flere mulige varianter enn for toveistabeller. I en ren krysstabell er bare n gitt a priori, alle de tre settene med marginaler er stokastiske. Så kan vi ha en komparativ tabell som består av to eller flere toveistabeller. Her er da antall observasjoner i hver krysstabell gitt, f.eks. tallene n_{+1+} og n_{+2+} i tabell 4.2 hvis vi har tatt ett utvalg for j=1, dvs. folk med adgang fritidshus, og et annet utvalg for j=2. Endelig kan vi ha en komparativ tabell med flere enveistabeller som så er ordnet etter de ulike par av kjennetegn for de to andre variable, f.eks. for j og k. I tabell 4.2 kan vi ha 4 enveistabeller etter helgetur/ikke helgetur, hvis vi har tatt 4 separate utvalg på h.h.v. n_{+11}, n_{+12}, n_{+21} og n_{+22} personer fra grupper med de 4 ulike kombinasjoner av kjønn og adgang fritidshus.

I dette kapitlet skal vi bare se på noen forholdsvis enkle problemstillinger, og ellers vise til kapitlene 6 og 7. Der blir det innført parametre som beskriver p_{ijk} på måter som forenkler formuleringen og testingen av en rekke hypoteser.

4.1. Fullt spesifisert nullhypotese

Det er angitt i avsnitt 2.4.1 hvordan vi bruker χ^2 føyningstesten til å teste hypoteser av formen

$$p^o_{ijk} = \frac{1}{IJK} \text{ for alle kombinasjoner } (i,j,k).$$

Med få observasjoner og uspesifiserte alternativ kan det isteden lages "nødtester" analoge med 3.2.2.c.

Mot spesifiserte alternativ vil vi i enkelte tilfeller kunne konstruere spesielle tester, men da bør alternativet ha en forholdsvis enkel struktur.

4.2. En 2x2x2-tabell

Den enkleste formen for en treveistabell er den som har bare to kategorier for hver variabel. Vi har da 2x2x2 = 8 ruter med observerte antall n_{ijk} for i=1,2, j=1,2 og k=1,2. Den kan settes opp på ulike måter, vi har her valgt varianten i tabell 4.2.a og b.

Tabell 4.2.a. Notasjon i en 2x2x2-tabell.

	j=1		j=2		Sum over k		Sum over
i	k=1	k=2	k=1	k=2	j=1	j=2	j og k
1	n_{111}	n_{112}	n_{121}	n_{122}	n_{11+}	n_{12+}	n_{1++}
2	n_{211}	n_{212}	n_{221}	n_{222}	n_{21+}	n_{22+}	n_{2++}
Sum over i	n_{+11}	n_{+12}	n_{+21}	n_{+22}	n_{+1+}	n_{+2+}	n_{+++}
Sum over i og k	n_{+1+}		n_{+2+}				n_{+++}
Sum over i og j	n_{++1}	n_{++2}			n_{+++}		$n=n_{+++}$

Tabell 4.2.b. Antall personer, n_{ijk}, gruppert etter om de har vært på helgetur eller ikke (i), om de har adgang fritidshus eller ikke (j) og kjønn (k). Sammendrag fra tabell 2.2.c.

Vært på helgetur i	Adgang fritidshus Ja, j=1		Nei, j=2		Sum over k		Sum over
	Menn, k=1	Kv., k=2	Menn, k=1	Kv., k=2	j=1	j=2	j og k
Ja, i=1	336	326	486	530	662	1016	1678
Nei, i=2	52	46	239	249	98	488	586
Sum over i	388	372	725	779	760	1506	2264
Sum over i og k	760		1504				2264
Sum over i og j	n_{++1}=1113, n_{++2}=1151				2264		2264

4.2.1 Komparativ tabell med sammenlikning av 2x2 enveistabeller

Forutsetninger: Vi har 2x2=4 uavhengige utvalg, der f.eks. n_{+11}, n_{+12}, n_{+21} og n_{+22} er gitte tall.

Binomisk fordeling av n_{ijk} innen hvert utvalg.

Problem (i): Er det forskjell mellom sannsynlighetene p_{ijk} (for å dra på helgetur) i de fire utvalgene, d.v.s. for de fire forskjellige kombinasjoner av kjønn og adgang til fritidshus?

Vi ser at nullhypotesen

$$p_{111} = p_{112} = p_{121} = p_{122}$$

kan testes på samme måte som ved sammenlikning av binomiske utvalg i avsnitt 3.3.2, svarende til J=4 der, og med antall df = 3.

Med symbolene i tabell 4.2.a kan χ^2-observatoren i homogenitetstesten 3.3.2.a skrives

$$z_h = \sum_{k=1}^{2} \sum_{j=1}^{2} \frac{(nn_{ijk} - n_{1++}n_{+jk})^2}{n_{1++}n_{2++}n_{+jk}} .$$

I vårt eksempel i 4.2.b får vi $z_h = 117,3$, som er større enn 5%-fraktilen 9,81 (og andre tabulerte fraktiler på lavere nivå), slik at vi forkaster nullhypotesen.

Nærmere analyse av enkeltproblemer etter forkasting av H_o kan vi foreta på tilsvarende måte som i avsnitt 3.3.2. Vi kan ønske å undersøke om menn har større sannsynlighet for å dra på helgetur enn kvinner, særskilt for grupper med fritidshus og grupper uten, dvs. vi tester de to hypotesene

$$p_{111} = p_{112} \quad \text{mot} \quad p_{111} > p_{112}$$

og

$$p_{121} = p_{122} \quad \text{mot} \quad p_{121} > p_{122}$$

hver for seg.

Vi forkaster den første hypotesen hvis

$$(\hat{p}_{111} - \hat{p}_{112}) / \sqrt{(1/n_{+11} + 1/n_{+12}) n_{11+} n_{21+} / n_{+1+}^2} \geq \sqrt{z_{1-\varepsilon,3}}$$

der $\hat{p}_{11k} = n_{11k}/n_{+1k}$. Vi finner verdien 0,41, altså ingen signifikant forskjell.

Den andre tester vi på tilsvarende måte ved å skifte ut j=1 med j=2 i testobservatoren.

På tilsvarende måte kan vi også undersøke om det er forskjell på menn som har fritidshus og menn som ikke har det, og analogt for kvinner

Ved å bruke $\sqrt{z_{1-\varepsilon,3}}$ istedenfor fraktilen i normalfordelingen, sikrer vi oss at nivået for den simultane testen er ε (tilnærmet), uansett hvor mange sammenlikninger av denne typen vi foretar, jfr. Sverdrup [1975] eller Haldorsen I, avsnitt 5.

For enkelte andre problemer vil vi kunne gå frem på tilsvarende måte som i avsnitt 3.3.2.

4.2.2. Komparativ tabell med sammenlikning av to toveis krysstabeller

Forutsetninger: Vi har 2 uavhengige utvalg, der f.eks. n_{++1} og n_{++2} er gitte tall, d.v.s. vi har tatt ett utvalg av menn og ett av kvinner. Tellevariablene n_{ij1} er multinomisk fordelt, det samme er n_{ij2}.

Problem: Er det forskjell mellom sannsynlighetene p_{ij1} og p_{ij2}, d.v.s. er det forskjell på de ulike kombinasjoner av turgåing og adgang fritidshus mellom menn og kvinner?

Vi har nullhypotesen

$$p_{111} = p_{112},\ p_{121} = p_{122},\ p_{211} = p_{212} \text{ og } p_{221} = p_{222}$$

Her følger forøvrig den siste av de tre andre fordi vi har

$$p_{111} + p_{121} + p_{211} + p_{221} = 1 \text{ og } p_{112} + p_{122} + p_{212} + p_{122} = 1.$$

Hvis vi har stort nok datamateriale og ingen bestemte alternativ, kan vi igjen bruke homogenitetstesten 2.4.3, jfr. 3.3.1. Uttrykket for z_h kan her skrives

$$z_h = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(nn_{ij1} - n_{++1}n_{ij+})^2}{n_{++1}n_{++2}n_{ij+}},$$

og vi har 3 df.

Vi finner $z_a = 2$ i vårt eksempel og kan ikke påstå at det er forskjell på menn og kvinner i dette tilfelle.

Andre problemstillinger kan vi undersøke på tilsvarende måte som i avsnitt 3.3.1.

Se også avsnitt 4.3.2.b. Der har vi en test som kan brukes i små utvalg hvis vi a priori kan anta at

$$p_{111}/p_{211} = p_{112}/p_{212}$$

(I vårt eksempel vil dette bety at det ikke er samvirkning mellom kjønn og adgang fritidshus.)

4.2.3. En 2x2x2 krysstabell

Forutsetninger: Vi har ett utvalg, med n gitt. Multinomisk fordeling av n_{ijk}, d.v.s. av 7 tellevariable, idet summen av de 8 variable er lik n.

Problem (i): Er det avhengighet mellom de tre kategoriske variablene? I eksemplet: er det avhengighet mellom tilbøyelighet til å dra på helgetur, kjønn og det å ha adgang til fritidshus?

Vi har her nullhypotesen

$$H_o: p_{ijk} = p_{i++}p_{+j+}p_{++k} \quad \text{for } i=1,2,\ j=1,2,,\ k=1,2.$$

Med tilstrekkelig antall observasjoner og uten bestemte alternativ, kan vi bruke en χ^2-uavhengighetstest helt analogt med hva vi gjorde for toveistabeller (se 3.2.1 eller 3.4.3), d.v.s. vi ledes til samme testobservator som i en homogenitetstest. Nullhypotesen innebærer jo at de betingede sannsynlighetene i hver linje er like, f.eks. at

$$p_{1/11} = p_{1/12} = p_{1/21} = p_{1/22} = p_{1++}, \text{ osv.}$$

Vi kan her skrive testobservatoren på formen

$$z_h = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \frac{(n^2 n_{ijk} - n_{i++}n_{+j+}n_{++k})^2}{n^2 n_{i++}n_{+j+}n_{++k}}$$

og vi forkaster nullhypotesen hvis vi finner at z_h er større enn øvre ε-fraktil i χ^2-fordelingen med 4 df. I vårt eksempel finner vi z_h=100,9, altså forkasting.

Nærmere analyse

Hvis vi ikke får forkastet nullhypotesen, kan vi ikke påstå at det er noen avhengighet mellom de tre variable. Men får vi forkasting, kan vi som i avsnitt 3.3. eller 4.2.1, ønske å se nærmere på hvordan avhengigheten arter seg.

4.2.3.a Første ordens avhengighet

Kanskje er det bare to av de variable som er avhengige, mens den tredje er uavhengig av disse to? D.v.s., kanskje det er avhengighet mellom det å dra på helgetur og det å ha adgang til fritidhus, men det

er ikke avhengighet mellom kjønn og kombinasjonene av de to første.

I så fall må vi ha for alle (ij)-kombinasjoner at

$$H_{o1}: p_{ijk} = p_{++k}p_{ij+}.$$

Tilsvarende kan vi ha andre hypoteser, f.eks. at helgetur er uavhengig av kombinasjonen kjønn/fritidshus, d.v.s.

$$H_{o2}: p_{ijk} = p_{i++}p_{+jk}$$

eller at fritidshus er uavhengig av helgetur/kjønn, d.v.s.

$$H_{o3}: p_{ijk} = p_{+j+}p_{i+k}.$$

For å teste H_{o1} kan vi bruke observatoren

$$z_h^{o1} = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \frac{(n_{ijk} - n_{++k}n_{ij+}/n)^2}{n_{++k}n_{+jk}/n} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(nn_{ij1} - n_{++1}n_{ij+})^2}{n_{++1}n_{++2}n_{ij+}}$$

Hvis vi tester hypotesen H_{o1} uten først å ha testet uavhengighetshypotesen H_o ovenfor, så kan vi forkaste H_{o1} når z_h^{o1} er mindre enn øvre ε-fraktil i χ^2-fordelingen med 3 df, altså 7,81. Vi ser at det gir samme resultat som i 4.2.2. Vi finner $z_h=2$ og kan altså ikke si at det er avhengighet mellom kjønn og kombinasjonen turgåing/fritidshus. Men har vi først testet og forkastet H_o, bør vi bruke samme fraktil som i testingene av denne, altså 4 df. Dermed sikrer vi oss at nivået for kombinasjonen av de to testene blir ε. Dette kan vi også gjøre om vi foretar flere tester, som av H_{o2} eller H_{o3}.

4.2.3.b Annen ordens avhengighet

Et annet problem når vi har avhengighet mellom de tre variable, er om avhengigheten varierer mellom de ulike kombinasjoner av variablene. Er f.eks. avhengigheten mellom turgåing og adgang fritidshus den samme for de to kjønn? Hvis ikke, vil vi gjerne vite hvor avhengigheten er sterkest.

Med kryssproduktforholdet som avhengighetsmål, jfr. avsnitt 2.2.4, kan vi da undersøke om avhengighetsforholdet er det samme for de to kjønn eller ikke. Vi tester

$$H_{ob}: \frac{p_{111}p_{221}}{p_{211}p_{121}} = \frac{p_{112}p_{222}}{p_{212}p_{122}}$$

mot f.eks. alternativt

$$H_{1b}: \frac{p_{111}p_{221}}{p_{211}p_{121}} < \frac{p_{112}p_{222}}{p_{212}p_{122}}$$

d.v.s. om avhengighetsmålet er større for kjønn 2, d.v.s. kvinner, enn for menn.

Vi ser at hvis hypotesen H_{o1} i 4.2.3.a gjelder, så vil også H_{ob} gjelde, men det omvendte behøver ikke være oppfylt. Vi kan være interessert i hypotesene H_{ob}/H_{1b} også når det er avhengighet mellom alle tre faktorer.

Denne problemstillingen er omtalt i avsnitt 4.6 hos Everitt, men testmetoden er ikke skrevet ut; det vises til avsnittet om log-lineære modeller. Problemet er også tatt opp i Sverdrup I, avsnitt 8. Følger vi hans metode i vårt eksempel, finner vi nedenstående test. Forøvrig viser vi til kapittel 6 om log-lineære modeller.

Vi kan forkaste hypotesen H_{ob} mot alternativet H_{1b} hvis vi finner at følgende gjelder:

$$f(n) = \sum_{i,j,k}(-1)^{i+j+k} \log n_{ijk} \leq \sqrt{z_{1-\varepsilon} \cdot \sum_{i,j,k} (1/n_{ijk})},$$

der log er naturlig logaritme.

Her er $z_{1-\varepsilon}$ den øvre ε-fraktil i χ^2-fordelingen med 4 frihetsgrader hvis vi først har testet H_o foran. Har vi derimot først forkastet en nullhypotese som innebærer at H_{ob} er riktig, men som ikke er fullt så streng som H_o, bruker vi en fraktil svarende til antall df ved denne testen.

Har vi ikke testet noen annen hypotese først, så skal vi velge $\sqrt{z_{1-\varepsilon}}$ som øvre ε-fraktil i den standariserte normalfordelingen. I alle

I vårt eksempel finner vi

$$f(n) = 0{,}0467$$

$$z_{095,4} = 9{,}49$$

Vi kan ikke påstå at det er noen forskjell for de to kjønn ut fra de estimerte avhengighetsmålene som blir

3,18 og 3,33.

Dette virker jo rimelig, i og med at resultatet av testen i 4.2.3.a kan tyde på uavhengighet.

4.3. Treveistabeller med IxJxK ruter

Problemstillingene for tabeller med fler enn 2 kategorier er i og for seg de samme som i 2x2x2-tabeller, men det er rom for flere varianter av analysene. Vi bruker tallene i tabell 4.3.a i eksemplene nedenfor.

Tabell 4.3.a Antall personer, n_{ijk}, gruppert etter antall helgeturer(i), adgang fritidshus eller ikke (j), og alder (k).

Helgeturer		Adgang fritidshus Ja, j=1				j=2						
Antall	kategori i	Aldersgrupper 15-24, 25-34, 35-54, 55-74 k=1	2	3	4	1	2	3	4	Sum over k n_{i1+}	n_{i2+}	Sum over j og k n_{i++}
0	1	14	18	27	39	51	74	143	220	98	488	586
1-5	2	51	34	80	65	111	144	181	167	230	603	833
6-9	3	18	22	35	29	38	47	48	33	104	166	270
10-14	4	30	22	36	26	36	28	13	20	114	97	211
15 og fl.	5	42	31	92	49	34	42	49	25	214	150	364
Sum	n_{+jk}	155	127	270	208	270	335	434	465	760	504	2264
Sum over i og k,	n_{+j+}	760				1504						2264
Sum over i og j,	n_{++k}	425	462	704	673					2264		2264 = n

4.3.1. Komparativ treveistabell med sammenlikning av JxK enveistabeller

Forutsetninger: Vi har JK uavhengige utvalg, der $n_{+11}, n_{+12}, \ldots, n_{+1K}$, $n_{+21}, \ldots, n_{+JK}$ er gitte tall.

Innen hvert av utvalgene, d.v.s. for hver kombinasjon av j og k, er n_{ijk} multinomisk fordelt.

Problem: Er det forskjell mellom sannsynlighetene p_{ijk} (for å dra på et antall helgeturer svarende til kategori nr. i) i de JK utvalgene, d.v.s. mellom de ulike kombinasjoner av adgang fritidshus og aldersgruppe?

I store utvalg kan vi her teste nullhypotesen

$$H_o: p_{i11} = p_{i12} = \quad = p_{i21} = p_{i22} = p_{i2k} = \quad = p_{iJK} \quad \text{for } i=1,2,\ldots,I,$$

på samme måte som i avsnitt 2.4.3 og 3.4.2. Forskjellen er at vi skal sammenlikne JK utvalg istedenfor J utvalg. Homogenitetsobservatoren kan her skrives

$$z_h = \sum_{j=1}^{J} \sum_{k=1}^{K} \sum_{i=1}^{I} \frac{(n_{ijk} - n_{+jk}n_{i++}/n)^2}{n_{+jk}n_{i++}/n}$$

Vi må forkaste H_o hvis vi finner en z_h-verdi større enn øvre ε-fraktil i χ^2-fordelingen med (JK-1)(I-1)df.

Med tallene i tabell 4.3.a finner vi z_h=376,5. Fraktilen for (8-1)(5-1) = 28 df er 41,34 på 5%-nivået, altså må vi slutte av sannsynlighetsfordelingen for antall helgeturer er ulik i de forskjellige grupper.

En nærmere analyse hvis vi forkaster nullhypotesen, kan vi foreta etter tilsvarende retningslinjer som i avsnitt 3.4.2 eller 4.2.1.

En spesiell situasjon kan være at vi har ordnede kategorier for to eller alle tre variable og vi ville undersøke om f.eks. p_{ijk} avtar (ev. øker) med stigende j og/eller k for i=1,2... til en viss i-verdi, mens p_{ijk}-endringen med j/k går i motsatt retning for høyere i-verdi. Vi kan f.eks. ha fallende sannsynlighet for mange turer når alderen øker, og høyere nivå med adgang fritidshus enn uten. For å uttrykke dette, kan vi velge en enkel modell som ikke behøver å være riktig, men som kan være en tilnærmelse til den sanne modellen, som f.eks. å sette

$$p_{ijk} = \alpha_i + \beta_i d_j + \gamma_i t_k.$$

her er $d_1=1$ og $d_2=0$ (d.v.s. ikke adgang fritidshus), mens vi velger t_k stigende med alderen. Vi kan f.eks. sette $t_1=20$, $t_2=30$, $t_4=45$ og $t_5=65$. Ved vanlige regresjonsmetoder kan vi så regne ut regresjonen for n_{ijk} m.h.p. d_j og t_k, og teste hypotesene

$$\gamma_i = 0 \quad \text{mot} \quad \gamma_i < 0 \quad \text{for høye i-verdier}$$

og

$$\gamma_i = 0 \quad \text{mot} \quad \gamma_i > 0 \quad \text{for lave i-verdier.}$$

Og vi kan teste

$$\beta_i = 0 \quad \text{mot} \quad \beta_i > 0 \text{ for i-verdier større enn 1.}$$

Vi kan vel også forsøke å se på antall helgeturer som en variabel y_i med stigende verdier, og her sette $y_1=0$, $y_2=3$, $y_3=7{,}5$, $y_4=12$ og $y_5=20$ (f.eks.) og simpelthen finne koeffisientene i regresjonen

$$y_{ijk} = a + b_1 d_j + b_2 t_k + b_3 t_k^2.$$

Vi har tatt med t_k^2 fordi det er tenkelig at forventet antall turer stiger med alderen i yngre aldersklasser, men avtar i eldre aldersklasser, og dette kan vi få frem hvis b_2 er positiv og b_3 negativ. Vi kan så teste koeffisientene til d_j, t_k og t_k^2 på vanlig måte.

Når vi foretar flere tester, f.eks. m stykker, så bør vi bruke et lavt sannsynlighetsnivå, f.eks. ε/m, for hver enkelt test, jfr. avsnitt 2.3.4.

4.3.2. Komparativ tabell med sammenlikning av K toveis krysstabeller

Forutsetninger: Vi har K uavhengige utvalg, der n_{++k} er gitte tall for k=1,2,...,K. Tellevariablene n_{ijk} er multinomisk fordelt for hver verdi av k, d.v.s. innen hver aldersgruppe i eksemplet.

Problem: Er det forskjell mellom sannsynlighetene p_{ij1}, p_{ij2},...,p_{ijk}? I eksemplet: har de ulike kombinasjoner av antall helgeturer og adgang fritidshus forskjellig sannsynlighet i de ulike aldersgruppene?

Vår nullhypotese kan formuleres

$$H_o: p_{ij1} = p_{ij2} = \quad = p_{ijK} \qquad \text{for alle kombinasjoner av i og j.}$$

Med stort nok datamateriale kan vi bruke homogenitetstesten 2.4.3, jfr. avsnitt 4.2.2. Uttrykket for z_h blir nå

$$z_h = \sum_{k=1}^{K} \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(n_{ijk} - n_{++k}\, n_{ij+}/n)^2}{n_{++k}\, n_{ij+}/n}$$

Vi forkaster H_o hvis den oberverte verdi av z_h blir større enn øvre ε-fraktil i χ^2-fordelingen med $K(I-1)(J-1)$df. I vårt eksempel finner vi z_h=162, mens 5%-fraktilen i χ^2-fordelingen med $4(5-1)(2-1) = 16$df er 26,30. Vi kan altså forkaste nullhypotesen.

Nærmere analyse av enkelte problemer etter at vi har forkastet H_o, kan vi foreta etter tilsvarende retningslinjer som i avsnitt 3.

Mange problemer vil være enklere å håndtere ved hjelp av metodene i avsnitt 6 (eventuelt avsnitt 7), også fordi vi der har ferdige EDB-programmer.

4.3.2.b. <u>En spesiell sammenlikning av K 2x2-tabeller.</u>

I det spesielle tilfellet at $I = 2$ og $J = 2$, har Sverdrup, III, utarbeidet en test for å sammenligne sannsynlighetene p_{11k} med p_{12k} når vi a priori kan anta at

$$p_{11k}/p_{21k} = \theta\, p_{12k}/p_{22k} \qquad \text{for } k = 1,2,\ldots,K.$$

Dette betyr at det ikke er noen "samvirkning" mellom kategorisk variabel nr. 2 og nr. 3.

Vi kan tenke oss et eksempel analogt med det i tabell 4.3.a, men der den første variable har de to kategoriene helgetur/ ikke helgetur. Vi har altså tatt $K = 4$ uavhengige utvalg, ett fra hver aldersgruppe, slik at n_{++k} er gitte tall og mener (for å illustrere denne problemstillingen !) at det ikke er noen samvariasjon mellom adgang fritidshus og alder.

Vi vil undersøke om tilbøyeligheten til å dra på helgetur er større hos personer <u>med</u> fritidshus enn hos dem uten. Vi tester da nullhypotesen

$$H_o : \theta = 1 \text{ (eller } \theta \leq 1) \text{ mot alternativet } \theta > 1.$$

Vi vil forkaste nullhypotesen hvis n_{11+} er stor, dvs. større enn øvre ε-fraktil i fordelingen av n_{11+} under nullhypotesen. Denne fordelingen kan finnes ved å gå ut fra nullfordelingene av de enkelte n_{11k} når n_{+1k} betraktes som gitt (slik som i avsnitt 3.1.1) for $k = 1,2,\ldots,K$. Sverdrup angir en fremgangsmåte for beregningene.

4.3.3. Treveis krysstabeller. Testing av uavhengighet.

Det vanligste spørsmålet en ser på i forbindelse med treveis-tabeller med I x J x K ruter, er om det er avhengighet mellom de tre kategoriske variablene. Hvis en finner at dette må være tilfelle, går en ofte videre for å undersøke arten av avhengighet, analogt med f.eks. avsnitt 3.

Forutsetninger: Vi har ett utvalg med n som et gitt tall. Telle-variablene n_{ijk} er multinomisk fordelt.

Problem: Er det avhengighet mellom de tre kategoriske variablene? I eksemplet i tabell 4.3a.: Er det avhengighet mellom antall helgeturer, adgang fritidshus og alder?

Nullhypotesen kan uttrykkes:

$$H_0: \; p_{ijk} = p_{i++}\, p_{+j+}\, p_{++k} \quad \text{for alle kombinasjoner av } i, j \text{ og } k.$$

Igjen kan vi bruke en χ^2-uavhengighetstest, analog med testen i 4.2.3, forutsatt at vi har tilstrekkelig mange observasjoner. Undersøkelser tyder på at vi kan ha helt ned i 1 observasjon (men ikke null) i en del ruter, uten at tilnærmelsen til χ^2-fordelingen med IJK-I-J-K+2 df. blir dårlig.

(Tallet på frihetsgrader fremkommer som differensen mellom antall df. i en I x J x K-tabell ved testing av en fullt spesifisert null-hypotese (p^o_{ijk} kjent) dvs. (IJK-1), og tallet på estimerte parametre i marginalene, nemlig (I-1) + (J-1) + (K-1).) Vår observator er

$$z_u = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(n_{ijk} - n_{i++} n_{+j+} n_{++k} / n^2\right)^2}{n_{i++} n_{+j+} n_{++k} / n^2}$$

I eksemplet finner vi $z_u = 408{,}7$, mens øvre ε-frakil i χ^2-fordelingen med 40-5-2-4+2 = 31 df. er 45,0. Vi kan altså anta at det er en viss avhengighet mellom de tre variablene.

Nærmere analyse.

Som i avsnitt 4.2.3 kan vi ønske å gå videre med analyse av avhengigheten.

4.3.3a, jfr. 4.2.3a.

Vi kan ha hypoteser som H_{01}, H_{02} eller H_{03}, om at det er avhengighet mellom to av de variable, mens dette paret kan være uavhengig av den tredje. Vi kan foreta testen av H_{01} ved hjelp av en observator som z_h^{01} i 4.2.3a, bare at det nå er IJK addender i summen istedenfor 2x2x2.

4.3.3b, jfr. 4.2.3b.

Annen ordens uavhengighet kan vi også undersøke. Er f.eks. avhengigheten mellom turgåing og alder den samme enten man har fritidshus eller ikke? Og hvis ikke, hvor er den da sterkest?

På tilsvarende måte som i 4.2.3b kan vi teste nullhypotesen

$$H_{0b}: \frac{p_{i1k} p_{I1k}}{p_{I1k} p_{i1k}} = \frac{p_{iJk} p_{IJK}}{p_{IJk} p_{iJK}}$$

mot det alternativ vi finner relevant enten H_{1b}, der likhetstegnet er erstattet med >, eller H_{2b}, der tegnet er <, eller H_{3b} der tegnet er ≠.

I vårt eksempel er J = 2. Vi kan foreta testingen for en eller flere verdier av i.

Problemstillingen, men ingen testmetode, diskuteres i avsnitt 4.6 hos Everitt, det vises til log-lineære modeller (jfr. vårt avsnitt 6.5).

Ifølge Sverdrup I , avsnitt 8, kan vi bruke følgende test. Vi forkaster H_{0b} mot H_{1b} hvis vi finner at

$$f(n)=\log \frac{n_{i1k}n_{I1K}}{n_{I1k}n_{i1K}} - \log \frac{n_{iJk}n_{IJK}}{n_{IJk}n_{iJK}} \geqq \hat{\sigma}\sqrt{z_{1-\varepsilon}},$$

der $$\hat{\sigma}^2 = \frac{1}{n_{i1k}} + \frac{1}{n_{I1K}} + \frac{1}{n_{I1k}} + \frac{1}{n_{i1K}} + \frac{1}{n_{iJk}} + \frac{1}{n_{IJK}} + \frac{1}{n_{IJk}} + \frac{1}{n_{iJK}}$$

Hvis vi bruker fraktilen i χ^2-fordelingen med antall frihetsgrader som ved testing av H_0 i 4.3.3, altså IJK-I-J-K+2, så kan vi foreta testen for så mange kombinasjoner av i og k som vi ønsker, og fremdeles ha et forkastingsnivå tilnærmet lik ε.

Ønsker vi bare å foreta testen for noen få (ijk)-kombinasjoner, kan vi vanligvis få større teststyrke ved å velge $z_{1-\varepsilon}$ annerledes. Vi kan f.eks. velge $\sqrt{z_{1-\varepsilon}}$ lik fraktilen i normalfordelingen med nivå $\frac{\varepsilon}{m}$ hvis vi har m tester i alt.

Ved testing mot H_{2b} vil vi forkaste H_{ob} når

$$f(n) < -\hat{\sigma}\sqrt{z_{1-\varepsilon}}$$

Tilsvarende test mot H_{3b} blir:

$$|f(n)| \geqq \hat{\sigma}\sqrt{z_{1-\varepsilon}},$$

og konkluderer med at > gjelder hvis

$$f(n) \geqq \hat{\sigma}\sqrt{z_{1-\varepsilon}}$$

og med at < gjelder hvis

$$f(n) \leqq -\hat{\sigma}\sqrt{z_{1-\varepsilon}}.$$

4.3.4. Ordnede kategorier ?

Se også avsnitt 6.9.

I eksemplet i tabell 4.3.a har vi naturlig ordnede kategorier for to variable: antall helgeturer og aldersgrupper. Hvis vi har en hypotese om at antall helgeturer (y) øker (eller avtar) med alderen (x), samtidig som antallet også avhenger av adgang fritidshus, kan vi til nød sette opp en regresjonslikning av formen

$$y = \alpha + \beta_1 d + \beta_2 x, \quad \text{der vi f.eks. velger}$$

y-verdiene lik 0, 3, 7.5, 12, 20 og d = 1 for adgang fritidshus og d = 0 uten, samt x = 20, 30, 45, 65. Vi kan bruke et vanlig regresjonsprogram for å estimere koeffisientene. Vi kan vel til nød også teste koeffisientene på vanlig måte, selv om testene bare kan betraktes som tilnærmede fordi forutsetningen om konstant varians for alle y er tvilsom her. I andre tilsvarende problem er det ikke sikkert at denne innvendingen behøver gjelde. I alle tilfelle må vi huske at regresjonslikningen bare er en røff tilnærmelse til sammenhengen mellom de variable.

5. FIRE- OG FLEREVEISTABELLER.

De fleste problemer vi vil analysere når vi har fire eller flere kategoriske variable simultant, vil være lettere å formulere med den notasjon vi skal innføre i kapittel 6, eventuelt i kapittel 7.

Enkelte enkle problemer kan imidlertid formuleres og løses på tilsvarende måte som i to- og treveistabeller. Vi skal kort omtale noen slike, selv om det vel ikke er så ofte de dukker opp.

5.1. Fullt spesifisert nullhypotese

Vi har en nullhypotese som kan skrives

$$p_{ijklg} = p^{o}_{ijklg} \quad \text{for} \quad \begin{array}{ll} i = 1,2,\ldots,I, & j = 1,2,\ldots,J \\ k = 1,2,\ldots,K, & l = 1,2,\ldots,L \text{ og} \\ g = 1,2,\ldots,G, & \end{array}$$

i en femveistabell, der alle p^{o}_{ijklg} er kjente tall.

Et spesialtilfelle av dette er at vi har samme sannsynlighet i alle ruter, d.v.s. at vi i en ren krysstabell har

$$p^{o}_{ijklg} = \frac{1}{IJKLG} .$$

Hvis vi også har et tilstrekkelig stort datamateriale og ikke spesifiserte alternativ, så kan vi bruke χ^2-tester som i avsnitt 2.4.

5.1.1 For en ren krysstabell har vi testen i 2.4.1. Hvis vi f.eks. har en 2x2x2x2x2 tabell, blir det $2^5-1 = 31$ df. for denne testen.

5.1.2. For en komparativ tabell bygger vi opp testen som i 2.4.2. Først lager vi en z-observator for hver av de uavhengige krysstabellene vi har, og deretter summerer vi disse til én observator, som da har antall df som er summen av df-tallet for de enkelte tabeller.

Anta f.eks. at vi har krysstabeller med IJK ruter, en for hver kombinasjon lg av de siste to variable i en femveistabell. Da danner vi

$$z_{lg} = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(n_{ijklg} - n_{+++lg} p^{o}_{ijklg})^2}{n p^{o}_{ijklg}}$$

for hver lg-kombinasjon. Hver av disse har (IJK-1)df. Den totale z-verdi blir

$$z_{total} = \sum_{l=1}^{L} \sum_{g=1}^{G} z_{lg}$$

som har LG(IJK-1)df.

For en 2x2x2x2x2-tabell blir det altså 4(8-1) = 28 df.

5.1.3. "Nødtester" for utvalg med få observasjoner og uten spesifiserte alternativ, kan konstrueres som i 3.3. Vi setter inn p^o-verdien i den multinomiske (eller produktmultinomiske) fordelingen som gjelder for vårt problem. Så regner vi ut sannsynligheten P_o under H_o for det resultatet vi har observert. Deretter leter vi (d.v.s. EDB-maskinen) oss frem til alle kombinasjoner av data med de gitte summer som hver har sannsynlighet lik eller mindre enn P_o. Er summen av alle disse små sannsynlighetene, inkludert P_o, mindre eller lik ε, så forkaster vi H_o.

Det kan vel være tvilsomt om det lønner seg å foreta slike beregninger når en har mange variable.

5.1.4. Bestemte alternativ til H_o bør her som ellers få oss til å lete etter bedre testmetoder.

5.2. Avhengighet/uavhengighet

Som i to- og treveistabeller kan vi ønske å teste ulike typer av uavhengighet mellom de variable. Dette blir behandlet mer systematisk i kapittel 6 (og tildels i 7), vi ser på noen spesielle tilfeller her.

5.2.1. Uavhengighet mellom alle de variable i en krysstabell. Analogt med i 3. 2.1 og 4.3.3 kan vi uttrykke nullhypotesen ved:

$$p_{ij\ldots g} = p_{i+\ldots+}p_{+j+\ldots+}\cdots p_{++\ldots+g} \qquad \text{for alle kombinasjoner av } ij\ldots g.$$

For store datamaterialer og uten bestemte alternativ kan vi igjen bruke en χ^2-test, nemlig en generalisert versjon av homogenitetstesten i 2.4.3. Den svarer til at vi estimerer alle de marginale p-ene ved de tilsvarende relative hyppighetene, altså:

$$\hat{p}_{i+\ldots+} = \frac{n_{i+\ldots+}}{n},\ \hat{p}_{+j+\ldots+} = \frac{n_{+j+\ldots+}}{n},\ \ldots\ \hat{p}_{++\ldots+g} = \frac{n_{++\ldots+g}}{n}$$

Så regner vi ut

$$z_h = \sum_{i=1}^{I}\sum_{j=1}^{J}\cdots\sum_{g=1}^{G} \frac{(n_{ij\ldots g} - n\hat{p}_{i+\ldots+}\hat{p}_{+j+\ldots+}\cdots\hat{p}_{++\ldots+g})^2}{n\hat{p}_{i+\ldots+}\hat{p}_{+j+\ldots+}\cdots\hat{p}_{++\ldots+g}}.$$

Når nullhypotesen gjelder vil denne observatoren være χ^2-fordelt med IJ....G-1-(I-1)-(J-1)....(G-1) = IJ....G-I-J-...-G+(m-1)df. (m er tallet på variable). For en 2x2x2x2x2-tabell blir det 2^5-2x5+4=26df. Vi kan altså forkaste hypotesen om uavhengighet hvis vi finner en z større enn øvre ε-fraktil i χ^2-fordelingen med dette antall df.

Hvis vi ikke får forkasting av nullhypotesen, vil vi i alminnelighet ikke ha noen sjanse for å finne ut noen interessante sammenhenger mellom de variable ut fra deler av tabellen.

Hvis vi forkaster nullhypotesen, kan det hende at vi ønsker en nærmere analyse, og det kan være ulike ønskemål og fremgangsmåter.

6. LOG-LINEÆRE MODELLER

Det å studere avhengighet/uavhengighet mellom variable i flerdimensjonale tabeller kan i visse tilfeller gjøres enklere hvis vi uttrykker sannsynlighetene $p_{ij...g}$ ved andre parametre. En måte å gjøre dette på, som er blitt meget populær, er å innføre de såkalte log-lineære modeller. Utgangspunktet er en multiplikativ modell som vi kan sette opp ved hjelp av multiplikasjonssetningen i sannsynlighetsregningen.

Multiplikative modeller kan uttrykkes som lineære modeller i logaritmene til de sannsynlighetene som inngår. En lineær modell er vanligvis enklere å arbeide med. Goodman's log-lineære modell er den som har vunnet størst utbredelse. Vi skal bare gi en kort omtale av modellen og bruken av den. Forøvrig viser vi til lærebøker og annen litteratur, f.eks.:

Tor Haldorsen: Om log-lineær analyse av flerveistabeller. IO 77/46. Heretter kalt H II.

S.E. Fienberg: The analysis of cross-classified categorical data. MIT Press 1978. Ny utgave 1980.

B.S. Everitt: The analysis of contingency tables. Ch. 5.

Bishop, Fienberg and Holland: Discrete multivariate analysis. Heretter kalt BFH.

L.A. Goodman: (ed. J. Magidsen): Analyzing qualitative/categorical data.

De tre første er lettest tilgjengelige. Den fjerde er mest omfattende. Den siste inneholder et utvalg av artikler som Leo Goodman og andre har skrevet om log-lineære modeller.

Det er utviklet EDB-programmer for log-lineær analyse, bl.a. ECTA, som finnes i Byrået og programmet P4F i BMDP-pakken (1979). Dessuten har TROLL et mindre program, TEP, D0070M. SAS-pakken har også et program.

Alle $p_{ij...g} > 0$.

I hele dette kapitlet forutsetter vi at $p_{ij...g} > 0$ for alle kombinasjoner av (i,j,...g) der i = 1,2,...I, j=1,2,...J ... og g = 1,2,...,G. Se kapittel 9 om problemer der noen $p_{ij...g} = 0$.

6.1. Log-lineære parametre i en toveistabell

Anta at vi som f.eks. i avsnitt 3.4, har en toveistabell, der i angir gruppe for antall helgeturer og j angir type av bostedskommune. Vi kan da uttrykke p_{ij} som et produkt av en generell faktor λ, som er felles for alle p-ene, en spesiell faktor λ_i for p-er i gruppe i, en spesiell faktor λ_j for p-er i gruppe (kommunetype) j og en faktor λ_{ij} som er spesiell for akkurat kombinasjonen (i,j). Vi setter altså

$$p_{ij} = \lambda\lambda_i\lambda_j\lambda_{ij}, \quad \text{for alle i og j.}$$

En hypotese om at det ikke er noe samspill mellom kommunetype og antall helgeturer kan vi her uttrykke ved at $\lambda_{ij} = 1$. En hypotese om at kommunetype ikke betyr noe for antall helgeturer, må tilsi at $\lambda_j = 1$ og $\lambda_{ij} = 1$ osv.

For at λ-ene skal kunne uttrykke sannsynligheter må vi ha visse betingelser oppfylt. Dette kan gjøres som angitt nedenfor. Vi innfører konstanter $\mu = \log \lambda$, $\mu_i^1 = \log \lambda_i$, $\mu_j^2 = \log \lambda_j$ og $\mu_{ij}^{12} = \log \lambda_{ij}$, der log står for naturlig logaritme med grunntall e = 2,71828...

Vi har altså

$$e^{\mu} = \lambda, \; e^{\mu_i^1} = \lambda_i, \; e^{\mu_j^2} = \lambda_j \text{ og } e^{\mu_{ij}^{12}} = \lambda_{ij}.$$

Så uttrykker vi den naturlige logaritmen til sannsynlighetene p_{ij} som en sum av parametre (konstanter). idet vi finner

$$\log p_{ij} = \mu+\mu_i^1+\mu_j^2+\mu_{ij}^{12} \qquad \begin{matrix} \text{for } i = 1,2,\ldots,I \\ j = 1,2,\ldots,J. \end{matrix} \qquad (6.1.1)$$

Her er altså μ et konstantledd som inngår i $\log p_{ij}$ for alle i og j. Resten av parametrene kan variere med i og/eller j. Toppindeksen 1 og fotindeksen i angir at μ_i^1 refererer seg til variabel nr. 1 og til i-te verdi av denne. Tilsvarende refererer μ_j^2 seg til variabel nr. 2 og til j-te verdi av denne. Den fjerde parameteren, μ_{ij}^{12}, refererer seg til kombinasjonen av i for variabel nr. 1 og j for nr. 2. Ved å ta antilogaritmen av (6.1.1) finner vi

$$p_{ij} = e^{\mu + \mu_i^1 + \mu_j^2 + \mu_{ij}^{12}} = e^{\mu} e^{\mu_i^1} e^{\mu_j^2} e^{\mu_{ij}} = \lambda\lambda_i^1\lambda_j^2\lambda_{ij}^{12}$$

som foran.

Alternativ uttrykksmåte. I en del lærebøker og dataprogrammer arbeider en med forventningsverdiene

$$m_{ij} = np_{ij},$$

istedenfor med p_{ij} direkte. Nå er

$$\log m_{ij} = \log n + \log p_{ij},$$

slik at det log-lineære uttrykket som svarer til (6.1.1) blir

$$\log m_{ij} = \mu_m + \mu_i^1 + \mu_j^2 + \mu_{ij}^{12},$$

der

$$\mu_m = \log n + \mu.$$

Vi har her et konstantledd som er større enn i (6.1.1), men alle de øvrige parametrene er de samme og har samme tolkning som vi skal finne nedenfor.

Analogi med variansanalyse. Den som er fortrolig med de vanlige modeller i variansanalyse, vil se analogien med disse i valget av parametre i (6.1.1), og dessuten kravene nedenfor i (6.1.2).

Antall parametre vi har innført i (6.1.1) er

$$1 + I \text{ (med toppindeks 1)} + J \text{ (med toppindeks 2)} + IJ = IJ + I + J + 1.$$

Dette er for mange til å gi en en-entydig korrespondanse mellom μ-ene og de IJ p-ene vi har. Vi innfører derfor noen restriksjoner som gjør at en del μ-er kan uttrykkes ved de øvrige. Vi krever at summer av μ-er med samme toppindeks skal være lik null, d.v.s.

$$\sum_{i=1}^{I} \mu_i^1 = 0, \quad \sum_{j=1}^{J} \mu_j^2 = 0, \quad \sum_{i=1}^{I} \mu_{ij}^{12} = 0 \quad \text{for hver } j \tag{6.1.2}$$

og

$$\sum_{j=1}^{J} \mu_{ij}^{12} = 0 \text{ for hver } i.$$

I alt kan da 1+1+I+J-1 = I+J+1 μ-er uttrykkes ved de øvrige, slik at vi får IJ μ-er som "selvstendige" parametre, altså det samme antall som vi har av p_{ij}-er.

Dessuten må vi huske kravet om at summen av alle p_{ij}-ene er lik 1, altså at vi må ha

$$\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij} = 1, \text{ d.v.s. } \sum_{i=1}^{I} \sum_{j=1}^{J} e^{\mu+\mu_i^1+\mu_i^2+\mu_{ij}^{12}} = 1. \tag{6.1.3}$$

Alt dette fører til at vi har (IJ-1) "selvstendige" μ_{ij} akkurat som vi har (IJ-1) p_{ij}-er.

For å få en tolkning av hva de ulike μ-ene står for, kan vi foreta noen summeringer av uttrykkene i (6.1.1). Tar vi summen over alle i og j, finner vi, når vi tar hensyn til (6.1.2), at alle summer av μ_i^1-er, μ_j^2-er og μ_{ij}^{12}-er blir null, så vi har

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \log p_{ij} = IJ\,\mu.$$

Altså kan vi tolke

$$\mu = \frac{1}{IJ} \sum_i \sum_j \log p_{ij} \tag{6.1.4}$$

som et uttrykk for "gjennomsnittsnivået" av alle log p_{ij}-ene.

Videre får vi ved å holde indeks i fast og summere (6.1.1) over j, at

$$\sum_{j=1}^{J} \log p_{ij} = J(\mu+\mu_i^1) \quad \text{for} \quad i = 1,2,\ldots,I.$$

Skrevet på en annen måte gir dette en tolkning av

$$\mu_i^1 = \frac{1}{J} \sum_j \log p_{ij} - \mu, \tag{6.1.5}$$

som en "egenvirkning" eller effekt av i-te nivå for variabel nr. 1 (gjennomsnittlig over j) utover det som er med i totalgjennomsnittet.

Tilsvarende finner vi

$$\mu_j^2 = \frac{1}{I} \sum_i \log p_{ij} - \mu$$

som en gjennomsnittlig egenvirkning eller effekt av j-te nivå for variabel nr. 2.

La oss nå se på logaritmen til kryssproduktforholdet

$$\alpha_{ij} = \frac{p_{ij}}{p_{iJ}} \cdot \frac{p_{IJ}}{p_{Ij}}$$

fra avsnitt 2.2.2, og uttrykke dette ved μ-ene. Vi finner

$$\begin{aligned}
\log \alpha_{ij} &= \log p_{ij} + \log p_{IJ} - \log p_{iJ} - \log p_{Ij} = \\
&= \mu + \mu_i^1 + \mu_j^2 + \mu_{ij}^{12} + \mu + \mu_I^1 + \mu_J^2 + \mu_{IJ}^{12} + \\
&\quad - \mu - \mu_i^1 - \mu_J^2 - \mu_{iJ}^{12} - \mu - \mu_I^1 - \mu_j^2 - \mu_{Ij}^{12} = \\
&= \mu_{ij}^{12} + \mu_{IJ}^{12} - \mu_{iJ}^{12} - \mu_{Ij}^{12}.
\end{aligned} \tag{6.1.6}$$

Ser vi spesielt på en 2x2-tabell, der vi har I=J=2, så vil vi p.g.a. kravet (6.1.2) ha at

$$\mu_{11}^{12} = \mu_{22}^{12} = -\mu_{12}^{12} = -\mu_{21}^{12} \tag{6.1.7}$$

Dette medfører at

$$\log \alpha = 4\mu_{11}^{12} \tag{6.1.8}$$

Ved stokastisk uavhengighet er α=1, d.v.s. log α=0 og dermed $\mu_{11}^{12} = 0$.

Omvendt, hvis $\mu_{11}^{12} = 0$ så må vi ha $\alpha=1$, d.v.s. uavhengighet. Vi ser at μ_{11}^{12} er et "mål" for den stokastiske avhengigheten (eller samspillet) mellom de variable i en 2x2-tabell, og at uavhengighet uttrykkes ved $\mu_{11}^{12} = 0$.

En tilsvarende tolkning av μ_{ij}^{12}-leddene kommer vi frem til ved (6.1.6). Vi kan tolke μ_{ij}^{12}-leddene som uttrykk for samspill mellom variabel nr. 1 og nr. 2 for ulike (i,j)-kombinasjoner. $\mu_{ij}^{12} = 0$ for alle kombinasjoner betyr uavhengighet mellom de variable, og omvendt.

La oss se på et enkelt eksempel på sammenhengen mellom p-er og μ-er i en 2x2-tabell. P.g.a. (6.1.2) er $\mu_2^1 = -\mu_1^1$ og $\mu_2^2 = -\mu_1^2$. Dessuten gjelder altså (6.1.7). Anta at vi har p_{ij} som i tabell 6.1.a.

Tabell 6.1.a. Sannsynligheter p_{ij}

Variabel nr. 1	Variabel nr. 2		Marginal
i	j = 1	2	
1	0,25	0,15	0,4
2	0,4	0,2	0,6
Marginal	0,65	0,35	1

Ifølge (6.1.1) og (6.1.2) for I=2 og J=2, har vi fire likninger mellom log p_{ij} og μ-ene, nemlig

$$\log p_{11} = -1,386 = \mu + \mu_1^1 + \mu_1^2 + \mu_{11}^{12}$$

$$\log p_{12} = -1,897 = \mu + \mu_1^1 - \mu_1^2 - \mu_{11}^{12}$$

$$\log p_{21} = -0,916 = \mu - \mu_1^1 + \mu_1^2 - \mu_{11}^{12}$$

$$\log p_{22} = -1,609 = \mu - \mu_1^1 - \mu_1^2 + \mu_{11}^{12}$$

Løser vi disse m.h.p. μ-ene, finner vi

$$\mu=-1,452,\ \mu_1^1=-0,190,\ \mu_1^2= 0,301,\ \mu_{11}^{12}=-0,045.$$

Ved innsetting i likningene ser vi at dette stemmer, bortsett fra avrundingsfeil. Vi kan "tolke" egen- og samvirkningen ved å si: variabel nr. 1 har en positiv virkning når vi går fra i=1 til i=2, vi har jo $\mu_2^1 = 0{,}19$. For variabel nr. 2 er det omvendt, negativ virkning av å gå fra j=1 til j=2. I tillegg er det en svak negativ virkning av samvariasjonen for i=1 og j=1.

Nå vil vi jo vanligvis ikke ha gitt verken p-er eller μ-er (muligens bortsett fra hypotetiske verdier) i et problem vi skal analysere. Men med et datamateriale kan vi estimere μ-ene, og så estimere p-ene ut fra μ-estimatene ved hjelp av likningssystemet (6.1.1), jfr. avsnitt 6.4.

La oss ta et eksempel til, med p_{ij} som i tabell 6.1.b.

Tabell 6.1.b. Sannsynligheter p_{ij}

Variabel nr. 1	Variabel nr. 2 j=1	2	Marg p_{i+}
1	0,26	0,14	0,4
2	0,39	0,21	0,6
Marg p_{+j}	0,65	0,35	1

Her finner vi: $\mu = -1{,}454$, $\mu_1^1 = -0{,}203$, $\mu_1^2 = 0{,}310$ og $\mu_{11}^{12} = 0$. Ifølge (6.1.8) betyr $\mu_{11}^{12} = 0$ at det er stokastisk uavhengighet mellom de to variable. Og vi ser at dette stemmer, vi har

$$p_{ij} = p_{i+}p_{+j} \qquad \text{for } i = 1,2 \text{ og } j = 1,2.$$

Også her virker variabel nr. 1 "positivt" når vi går fra i=1 til i=2, mens variabel nr. 2 virker "negativt" ved overgang fra j=1 til j=2. I denne tabell er hver av disse virkningene den samme på begge nivåer av den annen variabel, det er intet samspill.

6.2. Log-lineære parametre i en flerveis krysstabell

For en flerveis-tabell innfører vi flere sett med parametre og restriksjoner analogt med (6.1.1 og .2), slik at vi har like mange "selvstendige" μ-parametre som vi har sannsynligheter $p_{ij...g}$ i modellen, antallet er IJ...G-1. Vi setter i en m-veis tabell:

$$\log p_{ij...g} = \mu+\mu_i^1+\mu_j^2+...+\mu_g^m+\mu_{ij}^{12}+...+\mu_{ig}^{1m}+...+\mu_{jg}^{2m}+...+\mu_{qg}^{(m-1)m}+ \\ + \mu_{ijk}^{123}+...+\mu_{ijg}^{12m}+...+\mu_{jkg}^{23m}+...+\mu_{ijk...q}^{123...m-1}+\mu_{ijk...g}^{123...m} \qquad (6.2.1)$$

med restriksjonene

$$\sum_{i=1}^{I}\mu_i^1 = \sum_{j=1}^{J}\mu_j^2 =...= \sum_{g=1}^{G}\mu_g^m = 0, \quad \sum_{i=1}^{I}\mu_{ij}^{12} = \sum_{j=1}^{J}\mu_{ij}^{12} = 0,...\sum_{i=1}^{I}\mu_{ig}^{1m} = 0,$$

$$\sum_{g=1}^{G}\mu_{ig}^{1m} = 0,... \quad \sum_{g=1}^{G}\mu_{qg}^{(m-1)m}=0,... \quad \sum_{j=1}^{J}\mu_{jkg}^{23m} =...=0,... \quad \sum_{i=1}^{I}\mu_{ijk...g}^{123...m} = 0 \text{ osv.}$$

Alle disse kravene skal gjelde for hver verdi, resp. kombinasjon av verdier, av variable som det ikke summeres over, jfr. (6.1.2).

Dette ser formidabelt ut, men la oss se hvordan vi kan tolke de ulike μ-symbolene:

μ: et uttrykk for gjennomsnittlig log p-nivå. Hvis alle andre μ-er er null, så er <u>alle</u> $\log p_{ij...g} = \mu$, d.v.s. vi har en modell hvor alle sannsynligheter er like.

$\mu_i^1, \mu_j^2, \mu_g^m$: uttrykk for gjennomsnittlig "egenvirkning" av h.h.v. første variabel på nivå i, annen variabel på nivå j, o.s.v.

$\mu_{ij}^{12},..,\mu_{ik}^{13},..,\mu_{qg}^{(m-1)m}$: som "første ordens samspill" mellom de variable <u>parvis</u>.

$\mu_{ijk}^{123},..,\mu_{iqg}^{1(m-1)m}$ osv.: som "annen ordens samspill" mellom de variable tre og tre.

$\mu_{ijk...qg}^{123...(m-1)m}$: som samspill mellom <u>alle</u> de m kategoriske variable i tabellen simultant.

Hvis denne siste er null for alle kombinasjoner ijk...qg, så er det uavhengighet mellom alle de m variable simultant, men det kan fremdeles være samspill mellom alle eller noen undergrupper av (m-1) variable, med mindre de respektive $\mu_{ijk...g}^{123...m}$ (for m-1 variable) også er null.

Hvis alle samspill-leddene er null, d.v.s. at vi har

$$\log p_{ij...g} = \mu+\mu_i^1+\mu_j^2+...+\mu_g^m \quad \text{for alle ij...g-kombinasjoner,}$$

så er det uavhengighet mellom alle de variable. Jfr. avsnitt 2b i H II. Vi har at $p_{ij...g}$ er lik produktet av de marginale sannsynlighetene $p_{i+...+}$, $p_{+j+...+}$ osv. til $p_{++...+g}$ for hver (ij...g)-kombinasjon.

Vi kan altså uttrykke hypoteser om uavhengighet (manglende samspill) mellom visse av de variable ved å sette de tilsvarende μ-ledd lik null. Og vi kan teste hypoteser om samspill ved å teste om visse μ-ledd er null, jfr. avsnitt 6.5 og eksemplet i avsnitt 11.1.3. Vi må da være klar over at samspill mellom f.eks. variablene 1 og 2 ikke bare uttrykkes ved μ_{ij}^{12}, men også ved $\mu_{ijk}^{123},,,\mu_{ijg}^{12m}...\mu_{ijk...g}^{123...m}$. Derfor vil modellene i avsnitt 6.3 være enklere å behandle.

6.3. Hierarkiske modeller

Testingen i en log-lineær modell blir enklere hvis uavhengigheten er slik at manglende samspill mellom f.eks. to variable impliserer manglende samspill av høyere orden der disse to variablene inngår. Hvis f.eks. $\mu_{jk}^{23}=0$ for alle (j,k)-kombinasjoner, kreves det at

$$\mu_{ijk}^{123} = 0,..,\mu_{jkg}^{23m} = 0, \ \mu_{ijkq}^{123(m-1)} = 0, \text{ o.s.v.}$$

opp til

$$\mu_{ijk...qg}^{123...(m-1)m} = 0.$$

Vi setter ikke et lavere ordens samspill lik null hvis et høyere ordens samspill som omfatter de samme variable, ikke er lik null. Kombinasjonen av variabel nr. 2 og 3 vil altså ikke forekomme i den

reduserte modellen vi kan sette opp istedenfor den mettede modellen (6.2.1). Slike modeller kalles hierarkiske.

Med tre variable vil en mettet modell bli skrevet

$$\log p_{ijk} = \mu + \mu_i^1 + \mu_j^2 + \mu_k^3 + \mu_{ij}^{12} + \mu_{ik}^{13} + \mu_{jk}^{23} + \mu_{ijk}^{123}. \qquad (6.3.1)$$

En hierarkisk modell der $\mu_{ik}^{13} = 0$, vil også ha $\mu_{ijk}^{123} = 0$, så den kan skrives

$$\log p_{ijk} = \mu + \mu_i^1 + \mu_j^2 + \mu_k^3 + \mu_{ij}^{12} + \mu_{jk}^{23}. \qquad (6.3.2)$$

Se også H II , avsnitt 2e.

6.4. Estimering av parametrene i en log-lineær modell

Estimering av μ-ene (og dermed av p-ene) i en log-lineær modell gjør vi ved sannsynlighetsmaksimeringsmetoden (ML-metoden). Fordelingen av de tellevariable, $n_{ij...g}$, forutsetter vi da multinomisk, jfr. avsnitt 2.2.0.

Estimering i modellen (6.3.1) med data i 2x2x2-tabellen 4.2.b gir f.eks. estimert log np for at en person drar på helgetur, har adgang fritidshus og er mann, lik

$$\log \hat{np}_{111} = 5{,}36+0{,}66-0{,}51+0{,}003+0{,}29-0{,}017+0{,}035-0{,}006 = 5.82$$

dvs. $\hat{np} = 336$, som er det observerte tall, idet vi har en mettet modell. Noen desimaler er sløyfet, så vi får litt usikkerhet i siste siffer.

Her er altså $\hat{\mu} = 5{,}36$ gjennomsnittet av log $\hat{np}_{ijk}$-verdiene for hele tabellen. Dette svarer til $e^{\hat{\mu}} = 212$ som er det geometriske gjennomsnittet av alle $\hat{np}_{ijk}$-ene (dvs. åttende rot av produktet av dem).

Så har vi "egenvirkningsleddene". Her er $\hat{\mu}_1^1 = 0{,}66$ og $\hat{\mu}_2^1 = 0{,}66$, som viser at estimert forventet antall som drar på helgetur, $\hat{np}_{1jk}$, gjennomsnittlig sett ligger over totalgjennomsnittet, mens $\hat{np}_{2jk}$ (som ikke drar) ligger under. Vi vet ikke om dette gjelder for alle (j,k)-kombinasjonene før vi også har tatt med de øvrige leddene i uttrykket. (I dette eksemplet er $\hat{np}_{1jk} > \hat{np}_{2jk}$ for alle (j,k)-kombinasjonene, noe vi også kan se direkte av tabell 4.2.b). Videre er $\hat{\mu}_1^2 = -0{,}51$ og

$\hat{\mu}_2^2 = 0{,}51$. Dette antyder lavere estimert sannsynlighet for å ha adgang til fritidshus enn for ikke å ha det ut fra dette materialet.

Så er $\hat{\mu}_1^3 = 0{,}003$ og $\hat{\mu}_2^3 = -0{,}003$. Dette synes å tyde på en svak overvekt av menn i forhold til kvinner i tabellen, men vi ser direkte at dette ikke stemmer. Her kommer det tydelig frem at "egenvirkningen" må tolkes sammen med de øvrige leddene. ($\hat{\mu}_k^3$-verdiene er for øvrig ikke signifikant forskjellige fra null, men dette berører ikke tolkningen foreløpig.)

Vi ser så på første ordens samspillene eller "tofaktorkombinasjonene". Vi har $\hat{\mu}_{11}^{12} = 0{,}29$, $\hat{\mu}_{12}^{12} = -0{,}29$, $\hat{\mu}_{21}^{12} = -0{,}29$ og $\hat{\mu}_{22}^{12} = 0{,}29$. Tolket uten hensyn til de øvrige leddene så viser dette at de estimerte sannsynligheter med kombinasjonen "helgetur, adgang fritidshus" og kombinasjonen "ikke-helgetur, ikke-fritidshus" får et tillegg i forhold til "helgetur, ikke-fritidshus" og til "ikke-helgetur, adgang fritidshus". Det er samspill mellom det å dra på helgetur og å ha adgang fritidshus (som vi måtte vente!). Vi må imidlertid huske at også dette er gjennomsnittstall og at $\hat{\mu}_{ijk}^{123}$ kan bety noe her, selv om den ikke gjør det i dette eksemplet.

Videre er $\hat{\mu}_{11}^{13} = -0{,}017$, $\hat{\mu}_{12}^{13} = 0{,}017$, $\hat{\mu}_{21}^{13} = 0{,}017$ og $\hat{\mu}_{22}^{13} = -0{,}017$. Dette antyder et lite fradrag i estimert sannsynlighet for å dra på tur og være mann, samt for ikke å dra og være kvinne, mens de to andre kombinasjonene har et lite tillegg. Igjen må vi være forsiktige med tolkingen.

Så er $\hat{\mu}_{11}^{23} = 0{,}035$, $\hat{\mu}_{12}^{23} = -0{,}035$, $\hat{\mu}_{21}^{23} = -0{,}035$ og $\hat{\mu}_{22}^{23} = 0{,}035$, dvs. et lite tillegg for "adgang fritidshus, mann" og for "ikke-fritidshus, kvinne", med tilsvarende fradrag for de to andre kombinasjonene.

Til slutt har vi samspillet mellom alle tre faktorer: $\hat{\mu}_{111}^{123} = -0{,}006$, $\hat{\mu}_{112}^{123} = \hat{\mu}_{121}^{123} = \hat{\mu}_{211}^{123} = 0{,}006$, $\hat{\mu}_{122}^{123} = \hat{\mu}_{212}^{123} = \hat{\mu}_{221}^{123} = -0{,}006$ og $\hat{\mu}_{222}^{123} = 0{,}006$. Dette antyder et ubetydelig fradrag for "helgetur, fritidshus, mann" og tillegg f.eks. for "ikke-helgetur, ikke-fritidshus, kvinne".

Ingen av de tre siste settene med koeffisienter er signifikante, sett enkeltvis. Nedenfor har vi estimert en "avkortet" modell, med $\hat{\mu}_{ik}^{13} = \hat{\mu}_{jk}^{23} = \hat{\mu}_{ijk}^{123} = 0$. Som angitt i avsnitt 6.5 tester vi denne og finner $z_L = 1{,}99$ med 3df, dvs. ikke forkasting av hypotesen om "avkortet modell". Med den valgte avrunding er for øvrig de enkelte koeffisientestimatene uforandret, bortsett fra $\hat{\mu}_k^3$ som har skiftet fortegn.

For å se på "virkningen" av faktoren fritidshus ut fra tallene ovenfor, kan vi sammenlikne $n\hat{p}_{\cdot ijk}$-verdiene parvis:

Med adgang fritidshus har vi:

$$\log \hat{np}_{111} = 5{,}82,\ \log \hat{np}_{112} = 5{,}79,\ \log \hat{np}_{211} = 3{,}96,\ \log \hat{np}_{212} = 3{,}84$$

Uten fritidshus:

$$\log \hat{np}_{121} = 6{,}20,\ \log \hat{np}_{122} = 6{,}28,\ \log \hat{np}_{221} = 5{,}48,\ \log \hat{np}_{212} = 5{,}52.$$

Innen hver kombinasjon av turgåing og kjønn er det altså større forventet antall uten fritidshus enn med, men forskjellen varierer fra 0,38 til 1,68 for log $\hat{np}$-verdiene, dvs. med en faktor fra 1,46 til 5,37 for $\hat{p}$-ene. Dette svarer til gjennomsnittet 0,51·2 = 1,02 vi så ovenfor.

Sammenlikner vi menn og kvinner, så er forventet antall menn ubetydelig større enn antall kvinner for både turgåere og ikke turgåere som har fritidshus. Det er omvendt for dem som ikke har fritidshus. Til slutt ser vi at forventet antall turgåere er større enn ikke-turgåere for alle kombinasjoner av fritidshus og kjønn (vi har log $\hat{np}_{111}$ - log $\hat{np}_{211}$ = 1,86,...,log $\hat{np}_{122}$ - log $\hat{np}_{222}$ = 0,66. Også her varierer forskjellene. Gjennomsnittet er 0,66·2 = 1,32 (ca.).

I dette eksemplet kunne vi ha diskutert tallene direkte ut fra tabell 4.2.5, men fullt så enkelt er det jo ikke bestandig.

Ved ML-estimering er det en-entydig sammenheng mellom $\hat{p}$-ene og $\hat{\mu}$-ene, og vi får de samme estimatene enten vi 1) først estimerer p-ene ved ML-metoden og så regner ut $\hat{\mu}$-ene ut fra ligningssystemet (6.4.1), som svarer til (6.2.1) for parametrene, eller om vi 2) først estimerer μ-ene og så regner ut $\hat{p}$-ene.

$$\log \hat{p}^{123\ldots m}_{ijk\ldots g} = \hat{\mu} + \hat{\mu}^{1}_{i} + \hat{\mu}^{2}_{j} + \ldots + \hat{\mu}^{m}_{g} + \hat{\mu}^{12}_{ij} + \ldots + \hat{\mu}^{123\ldots m}_{ijk\ldots g} \qquad (6.4.1)$$

Et eksempel på estimerte μ-er i en mettet modell er gitt i H II, tabell 3.2. Se også vårt avsnitt 11.1.3.

Setter vi visse μ-ledd lik null, så er det enkleste først å estimere de øvrige ved ML-metoden, og så finne $\hat{p}$-ene ved (6.4.1) hvis vi ønsker dem. ML-likningene for $\hat{\mu}$-ene (eller $\hat{p}$-ene) kan ikke alltid løses eksplisitt i slike tilfelle. EDB-programmene vil likevel gi løsningene når data og modell tilfredsstiller de gitte krav. Estimering av modellen

$$\log np_{ijk} = \mu_m + \mu_i^1 + \mu_j^2 + \mu_k^3 + \mu_{ij}^{12}, \text{ gir}$$

$$\log \hat{np}_{111} = 5{,}36+0{,}66-0{,}51-0{,}02+0{,}29 = 5{,}78$$

med $\hat{np}_{111} = 324$ for data i tabell 4.2.b. Vi får $\hat{np}_{112} = 336$, likevel ser gjennomsnittsforskjellen mellom menn og kvinner mer "riktig" ut her.

Programmet kan også gi estimater for tilnærmede standardavvik på de enkelte koeffisienter, og forholdstall mellom estimert koeffisient og standardavvik. Hvis en da antar at fordelingen av en slik standardisert koeffisient er tilnærmet normal, kan en teste om den f.eks. er signifikant forskjellig fra null ved å sammenligne med en passende ε-fraktil i den standardiserte normalfordelingen. En viss forsiktighet må vi vise ved slik "testing", jevnfør de neste avsnitt.

6.5. Testing i log-lineære modeller

Den vanlige testobservatoren som brukes ved log-lineære modeller, er LL-testen (log likelihood ratio test), som er omtalt i avsnitt 2.4.6.

Hvis vi f.eks. vil teste samspillet mellom variabel nr. 1 og 3 i (6.3.1), så kan vi sette opp nullhypotesen

$$H_o: \mu_{ik}^{13} = 0 \quad \text{og} \quad \mu_{ijk}^{123} = 0 \quad \text{for alle kombinasjoner av i,j,k.}$$

Under H_o gjelder altså (6.3.2). Vi estimerer μ-ene i dette uttrykket ved ML-metoden og deretter p-ene, og så regner vi ut verdien av observatoren

$$z_L = 2 \sum_{i,j,k} n_{ijk} (\log n_{ijk} - \log \hat{p}_{ijk})$$

Når nullhypotesen er riktig, så skal z_L være asymptotisk χ^2-fordelt med antall df lik antall "selvstendige" μ-parametre som settes lik null under nullhypotesen. Ved å sette $\mu_{ik}^{13} = 0$ for alle i og k, har vi (I-1)(K-1) df, og $\mu_{ijk}^{123} = 0$ gir (I-1)(J-1)(K-1) df, altså ialt

$$(I-1)(K-1) + (I-1)(J-1)(K-1) = IJK - IJ - JK + J$$

df i dette tilfelle.

Betrakter vi tabell 4.2.b som en krysstabell, med n fast, så vil nullhypotesen $\mu_{ik}^{13} = 0$ og $\mu_{ijk}^{123} = 0$ innebære at det ikke er samspill mellom det å dra på helgetur og kjønn. ECTA-programmet gir oss her $z_L = 0{,}35$ med 2 df. Nullhypotesen kan ikke forkastes. Det viser seg videre at forskjellen mellom menn og kvinner på alle nivåer i tabellen er så liten at oppdeling etter kjønn kan sløyfes, jfr. $\hat{\mu}_k^3$-estimatene i avsnitt 6.4.

For eksemplet i tabell 4.3.a, med personer fordelt etter antall helgeturer, adgang fritidshus og alder, kan vi bruke den mettede modellen 6.3.1.

Testing av hypotesen

$$H_o: \mu_{ijk}^{123} = 0 \quad \text{for alle de 40-1=39 kombinasjonene av i, j og k,}$$

gir som resultat at $z_L = 21,4$. LL-testen har her (5-1)(4-1)(2-1) = 12 df, med øvre 5 prosent fraktil lik 21,03. Vi må altså forkaste nullhypotesen, og regne med at det er samspill, i betydningen $\mu_{ijk}^{123} \neq 0$.

Tabell 6.5.a. Estimerte koeffisienter i den mettede log-lineære modellen for data i tabell 4.3.a.

	i =	1	2	3	4	5
	resp. j =	1	2			
	resp. k =	1	2	3	4	
$(\log n+\hat{\mu})$	3.780					
$\hat{\mu}_i^1$		0,120	0,728	-0,293	-0,533	-0,022
$\hat{\mu}_j^2$		-0,247	0,247			
$\hat{\mu}_k^3$		-0,165	-0,142	0,215	0,091	
$\hat{\mu}_{i1}^{12} = -\hat{\mu}_{i2}^{12}$		-0,507	-0,248	0,007	0,351	0,398
$\hat{\mu}_{i1}^{13}$		-0,427	-0,015	-0,037	0,426	0,053
$\hat{\mu}_{i2}^{13}$		-0,144	-0,109	0,143	0,126	-0,016
$\hat{\mu}_{i3}^{13}$		0,025	0,072	0,024	-0,362	0,242
$\hat{\mu}_{i4}^{13}$		0.545	0,052	-0,130	-0,190	-0,278
$\hat{\mu}_{1k}^{23} = -\hat{\mu}_{2k}^{23}$		-0,027	-0,164	0,131	0,060	
$\hat{\mu}_{i11}^{123} = -\hat{\mu}_{i21}^{123}$		0,148	0,136	- 0,099	- 0,167	- 0,018
$\hat{\mu}_{i12}^{123} = -\hat{\mu}_{i22}^{123}$		0,221	-0,058	0,030	-0,059	-0,135
$\hat{\mu}_{i13}^{123} = -\hat{\mu}_{i23}^{123}$		-0,203	-0,043	-0,047	0,262	0,032
$\hat{\mu}_{i14}^{123} = -\hat{\mu}_{i24}^{123}$		-0,166	-0,035	0,116	-0,036	0,122

Ser vi på de enkelte $\hat{\mu}^{123}$-verdiene, jfr. tabell 6.5.a, så kan vi imidlertid ikke fastslå at de er signifikante. Standardavvikene viser seg å variere mellom 0,07 og 0,1, og ingen av de standardiserte koeffisientene er større i tallverdi enn 2,377. Vi må sammenlikne med en fraktil som tar hensyn til at vi tester 12 koeffisienter, f.eks. fra normalfordelingen,

$$z_{1-\frac{\varepsilon}{2}} = 2{,}86 \text{ der } \varepsilon = \frac{0.05}{12} = 0.00417$$

Det er altså kombinasjonen av alle verdiene som gir signifikant utslag i testingen av H_o ovenfor.

Den mettede modellen gir oss nøyaktig de samme estimatene $\hat{p}_{ijk}$ som den vanlige multinomiske modellen ved ML-estimering. Det som kan være av en viss interesse, er å se på de enkelte koeffisientene i forhold til hverandre, og på fortegnene for dem som vi gjorde i avsnitt 6.4. Se også H II, avsnitt 5.a. Vi skal ta saken opp igjen i et noe enklere eksempel i avsnitt 11.1.3.

I de nevnte dataprogrammene kan vi nå også få skrevet ut faktorene

$$e^{\hat{\mu}}, \; e^{\hat{\mu}^1_i} \quad \text{o.s.v.},$$

slik at vi kan danne estimatene $\hat{p}_{ijk}$ (eller $\hat{np}_{ijk}$) direkte ved multiplikasjon av de relevante faktorene, istedenfor å gå veien om log $\hat{p}_{ijk}$ (eller log $\hat{np}_{ijk}$).

Estimering og testing i dette kapitlet er foretatt med P3F-programmet, der alle tallene i de to opprinnelige tabellene er tillagt 0,5 under analysen. Dette er en "korreksjon" som brukes når det er små tall i tabellene, men overflødig med så store tall som vi har. Forskjellen på beregningsresultatene med og uten tillegget 0,5 er imidlertid så liten at vi ikke har funnet det nødvendig å foreta nye kjøringer.

Det finnes kanskje ikke så mange eksempler i litteraturen på testing av eksplisitt formulerte nullhypoteser. Derimot er det meget vanlig at modell og program brukes slik som skissert nedenfor.

For små utvalg har Cox & Plackett [1980] visse undersøkelser og forslag.

Reduksjon av antall parametre i modellen

I mange tilfeller har vi ikke så meget a priori informasjon at vi kan sette opp meningsfylte hypoteser. Isteden ønsker vi kanskje å komme frem til en modell som er så enkel som mulig, vi vil "kaste ut" alle overflødige parametre, d.v.s. slike som er null eller nær null, fra (6.2.1). Samtidig ønsker vi ikke å kaste ut parametre som bør være med. Hvis vi tør forutsette en hierarkisk modell, kan vi gå skrittvis frem: Først tester vi $\mu_{ij\ldots g}^{12\ldots m} = 0$. Hvis nullhypotesen ikke forkastes, antar vi at det er forsvarlig å sette $\mu_{ij\ldots g}^{12\ldots m} = 0$ i arbeidet videre. Så tester vi en eller flere av $\mu_{ij\ldots q}^{12\ldots(m-1)}$, $\mu_{ij\ldots g}^{2\ldots m}$ o.s.v. på neste trinn, i den reduserte modell, og setter dem lik null som vi ikke får forkastet. For de kombinasjonene av variable dette gjelder går vi så videre og tester på trinnet som omfatter m-2 variable i den ytterligere reduserte modell. Vi fortsetter slik inntil en hypotese blir forkastet.

Ved denne fremgangsmåten kan vi i en viss forstand ha kontroll over sannsynlighetsnivået. Det er nemlig slik at om vi bruker nivå ε på hver enkelt test, og vi utfører m slike skrittvise tester, så vil sannsynligheten for minst en feilaktig forkasting ikke overstige $m\varepsilon$ (jfr. avsnitt 2.3.4). Vi bør altså velge nivået for den enkelte test svært lavt, og lavere jo flere tester vi akter å utføre.

Fremgangsmåter for mer eller mindre systematisk reduksjon av parametertallet finnes i lærebøker og artikler. I mange tilfeller er det uklart om det er kontroll over sannsynlighetsnivået for testene simultant. Jfr. avsnitt 6.6.

6.6. Log-lineære modeller og datanalyse

Meget av den bruk som gjøres av log-lineære modeller må betegnes som det vi her kaller datanalyse, jfr. avsnitt 1.4.2 og kapittel 13. Dette gjelder især når en ikke har a priori hypoteser om uavhengighet mellom visse variable, men isteden bruker mer eller mindre trinnvis testing til å redusere antall parametre (μ-er) i modellen mest mulig, og uten å kontrollere nivået for de enkelte testene i forhold til hverandre.

Slik bruk av testing i log-lineære modeller og dataprogrammer hører derfor i prinsippet hjemme i del III av dette notatet.

6.7. Tilfeldige nuller i de observerte krysstabellene

I hele dette kapitlet forutsetter vi at alle $p_{ij...g} > 0$. Likevel kan det hende at det mangler observasjoner i en eller flere ruter, både inne i tabellen og i enkelte marginaler. Vi kan altså ha $n_{ij...g} = 0$ og kanskje også $n_{ij+} = 0$ for enkelte (i,j,...)-kombinasjoner. Likevel kan vi i mange tilfelle bruke log-lineære modeller med tilhørende tester. Hvordan vi da må gå frem, vil bl.a. avhenge av hvordan nullene forekommer i tabellen. Vi kan ikke ta dette opp her, men viser til Fienberg [1978], ch. 8.1.

Aggregerte tabeller. En velkjent måte å komme utenom problemet med nuller (eller meget små tall) i en del ruter på, er å slå sammen data til en mindre tabell. Dette kan gjøres ved å utelate variable, f.eks. slå sammen en fireveis-tabell til en treveis-tabell. Eller en slår sammen ulike verdier av en variabel så en får f.eks. I-2 linjer istedenfor I linjer i tabellen.

Begge disse fremgangsmåtene kan føre til feilslutninger hvis man ikke er tilstrekkelig oppmerksom. Vi viser til avsnitt 7 i H II og ch. 2.4.1 og 2.5.3 i BFH.

6.8. Én variabel som funksjon av de øvrige i en log-lineær modell

Den log-lineære modellen uttrykker de simultane sannsynlighetene $p_{ij...g}$ som funksjoner av egenvirkninger og samspill mellom de variable. I mange tilfeller kan vi ønske å uttrykke én bestemt av de variable som funksjon av de øvrige variable i en eller annen forstand. I eksempel 4.2.b kan vi f.eks. ønske å se på det å dra på helgetur som en "funksjon" av adgang fritidshus og kjønn. Mest nærliggende er det å se på logaritmen til den betingede sannsynligheten for å dra på helgetur, gitt de ulike kombinasjonene av adgang fritidshus og kjønn. Det viser seg at i den log-lineære modellen får en frem et enkelt uttrykk ved heller å se på "log odds" som nedenfor. For 2x2x2-modellen har vi de betingede sannsynligheter for verdien 1 for den første variable, gitt henholsvis j og k for de to siste:

$$p(1|j,k) = \frac{p_{1jk}}{p_{+jk}} \quad \text{og} \quad p(2|j,k) = \frac{p_{2jk}}{p_{+jk}} = 1 - p(1|j,k).$$

Vi innfører "odds" $\Omega_{1|jk}$ som er forholdet mellom de to sannsynlighetene,

$$\Omega_{1|jk} = \frac{p(1|j,k)}{p(2|j,k)} = \frac{p_{1jk}}{p_{2jk}} \tag{6.8.1}$$

Vi setter inn fra (6.3.1) og husker at $\mu_2^1 = -\mu_1^1$ osv. som i (6.1.2) og (6.1.7) og dessuten $\mu_{2jk} = -\mu_{1jk}$. Det gir uttrykket for "log odds" eller <u>logit</u> $p(1|jk)$.

$$\log \Omega_{1|jk} = 2\mu_1^1 + 2\mu_{1j}^{12} + 2\mu_{1k}^{13} + 2\mu_{1jk}^{123} \tag{6.8.2}$$

For å forenkle uttrykket kan vi innføre nye parametre:

$$\gamma = 2\mu_1^1, \ \gamma_1 = 2\mu_{11}^{12}, \ \gamma_2 = 2\mu_{11}^{13} \quad \text{og} \quad 2\mu_{111}^{123} = \gamma_3. \tag{6.8.3}$$

Videre innfører vi en variabel z_1 som er lik 1 når j=1 og lik -1 når j=2, samt en variabel z_2 som er lik 1 når k=1 og lik -1 når k=2. Da kan vi skrive:

$$\log \Omega_{1|jk} = \gamma + \gamma_1 z_{1j} + \gamma_2 z_{2k} + \gamma_3 z_{1j} z_{2k},$$

dvs., vi har fått uttrykt log odds som en "lineær regresjonslikning" i z_1, z_2 og $(z_1 z_2)$, jfr. f.eks.(7.1.1). I alminnelighet vil vi <u>ikke</u> estimere koeffisientene γ, γ_1 osv. ved minste kvadraters metode, dvs. et regresjonsprogram. Vi bruker ML-estimering som omtalt i avsnitt 6.4, dvs. vi kan estimere γ-ene ved å sette inn estimatene for μ-ene i (6.8.3). Estimering og testing kan også her gjøres ved nyere versjoner av ECTA-programmet, mens P4F-programmet ikke regner ut $\hat{\gamma}$-ene.

I eksemplet for tabell 4.2.b finner vi estimert log odds for å dra på helgetur, gitt de forskjellige kombinasjonene av adgang fritidshus og kjønn, når vi bruker den mettede modellen:

$$\log \hat{\Omega}_{1|jk} = 1,318 + 0,586z_{1j} - 0,034z_{2k} - 0,012z_{1j}z_{2k}$$

Dette gir

$$\log \hat{\Omega}_{1|11} = 1,858,\ \log \hat{\Omega}_{1|21} = 0,778,\ \log \hat{\Omega}_{1|12} = 1,95,\ \log \hat{\Omega}_{1|22} = 0,754$$

som gir

$$\hat{\Omega}_{1|11} = 6,41,\ \hat{\Omega}_{1|21} = 2,18,\ \hat{\Omega}_{1|12} = 7,03,\ \hat{\Omega}_{1|22} = 2,13.$$

Dette stemmer bra med de tallene vi får direkte fra tabellen, især hvis vi legger til 0,5 i hver rute.

Vi kan tolke de estimerte koeffisientene i log odds-likningen slik: Det er en svært liten, og ikke signifikant forskjell mellom menn og kvinner når det gjelder å dra på helgetur. Med adgang fritidshus er tilbøyeligheten til å dra på helgetur en god del større (ca. 3 ganger så stor) enn når det ikke er adgang fritidshus. Sløyfer vi oppdelingen etter kjønn i tabellen, finner vi

$$\log n\, \hat{p}_{11} = 1,316 + 0,584z_{1j},$$

som gir praktisk talt samme resultat som ovenfor, nemlig

$$\hat{\Omega}_{1|1} = 6,69,\ \hat{\Omega}_{1|2} = 2,08.$$

Ved løsning av likningen

$$\hat{\Omega}_{1|1} = \frac{\hat{p}(1|1)}{\hat{p}(2|1)} = \frac{\hat{p}(1|1)}{1-\hat{p}(1|1)}$$

finner vi

$$\hat{p}(1|1) = \frac{\hat{\Omega}_{1|1}}{1+\hat{\Omega}_{1|1}} = 0,87$$

og tilsvarende

$$\hat{p}(1|2) = \frac{\hat{\Omega}_{1|2}}{1+\hat{\Omega}_{1|2}} = 0,675,$$

som stemmer med tallene

$$\frac{n_{11+}}{n_{+1+}} = \frac{662}{760} = 0,87 \quad \text{og} \quad \frac{n_{12+}}{n_{+2+}} = \frac{1016}{1506} = 0,675$$

i tabell 4.2.b.

Vi får altså ikke andre estimater for p-ene (vi har også her en mettet modell), men koeffisientene 0,584 og -0,584 gir oss en idé om "betydningen" av å ha fritidshus for det å dra på helgetur (iallfall når vi har vent oss til å tenke i log odds).

Det er imidlertid i større tabeller med flere variable og kategorier at en tolking av koeffisientene virkelig har interesse.

7. EN KATEGORISK VARIABEL BETRAKTET SOM EN FUNKSJON AV DE ØVRIGE KATEGORISKE VARIABLE.

Hittil har vi stort sett, og bortsett fra avsnittene om ordnede kategorier og avsnitt 6.8, omtalt metoder for å analysere samvariasjonen mellom kategoriske variable uten eksplisitt å se på hvordan én av de variable varierer med de øvrige.

Ikke så sjelden har vi imidlertid en variabel som vi er primært interessert i (f.eks. turgåing), og vi ønsker å finne ut hvordan samvariasjonen er mellom denne variable på den ene siden og de øvrige (f.eks. adgang fritidshus, kjønn, alder osv.) på den andre siden.

Vi skal kalle den variable vi er primært interessert i for

y, den har kategorier $1,2,\ldots,i,\ldots,I$,

jfr. tabell 2.1.

De øvrige variable skal vi kalle x_1, x_2 osv. til x_m, slik at

x_1 har kategorier $1,2,\ldots,j,\ldots,J$,
x_2 har kategorier $1,2,\ldots,k,\ldots,K$,
---
x_m har kategorier $1,2,\ldots,g,\ldots,G$.

Vi tenker oss altså her data i en (m+1)-veistabell, mens vi foran har hatt en m-veistabell. Vi skal først se på tilfeller der hver av de variable bare omfatter to forskjellige kategorier, de er dikotome variable. Det kan ofte være greit å sette de to verdiene til 1 (telleenheten har et bestemt kjennetegn, f.eks. fritidshus) og 0 (telleenheten har ikke kjennetegnet, dvs. fritidshus). Vi skal kalle slike variable for **binære**.

Siden mange som er vant til å bruke regresjonsanalyse i tilsvarende problemer med kvantitative variable, ønsker å bruke dette også for kvalitative variable, skal vi se litt på en slik analysemetode først. Vi viser til A II avsnitt 14.5, samt til A [1974] og A [1976,2] for en mer utførlig behandling.

7.1. Lineær regresjon for binære variable (binær regresjon)

Vi skal bruke data i tabell 4.2.b som eksempel. De variable og deres verdier er definert ved

$y = 0$ har ikke vært på helgetur
$= 1$ har vært på helgetur

$x_1 = 0$ har ikke adgang fritidshus
$= 1$ har adgang fritidshus

$x_2 = 0$ kvinne
$= 1$ mann

Vi setter opp tallene på nytt i tabell 7.1.a. De er her ordnet på en litt annen måte enn i tabell 4.2.b for å svare til vanlig ordning av de nye variabelverdiene.

Tabell 7.1.a Antall personer, $n_{yx_1x_2}$, gruppert etter om de har vært på helgetur, y, adgang fritidshus, x_1, og kjønn, x_2.

y	$x_1 = 0$		$x_1 = 1$		Sum
	$x_2 = 0$	1	$x_2 = 0$	1	
0	249	239	46	52	586
1	530	486	326	336	1678
Sum	779	725	372	388	2264

Et første forsøk på å sette opp en regresjonslikning for y med henblikk på x_1 og x_2 ville kanskje se slik ut:

$$y = b_o + b_1x_1 + b_2x_2 + \text{tilfeldig restledd},$$

der koeffisientene b_o, b_1 og b_2 er ukjente konstanter som vi ønsker å estimere ut fra data. Dette betyr at vi har

$y = b_o$ + noe tilfeldig når $x_1 = x_2 = 0$,
$y = b_o + b_1$ + noe tilfeldig når $x_1 = 1$ og $x_2 = 0$,
$y = b_o + b_2$ + noe tilfeldig når $x_1 = 0$ og $x_2 = 1$, og
$y = b_o + b_1 + b_2$ + noe tilfeldig når $x_1 = x_2 = 1$.

Her er "virkningen" av x_1 og x_2 på y additiv. Vi får altså ikke frem at det kan være en "samvirkning" av x_1 og x_2 på y, slik at vi burde ha en mulighet for et fradrag eller et tillegg i den siste likningen når både $x_1 = 1$ og $x_2 = 1$.

Dette siste får vi til hvis vi velger følgende regresjonslikning for y med hensyn på x_1 og x_2:

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \text{tilf. restledd} \qquad (7.1.1)$$

Vi forutsetter nå at det tilfeldige restleddet er ukorrelert med x_1 og x_2 og har forventning null. Da betyr (7.1.1) at den betingede sannsynligheten for $y = 1$ gitt x_1 og x_2, som også er forventningen av y for gitt x_1 og x_2, er lik

$$P(y=1|x_1,x_2) = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \qquad (7.1.2)$$

Setter vi inn de mulige kombinasjonene av x_1 og x_2, så har vi altså de fire betingede sannsynlighetene

$$\begin{aligned}
P(y=1|0,0) &= \beta_o && \text{for } x_1 = 0,\ x_2 = 0 \\
P(y=1|1,0) &= \beta_o + \beta_1 && \text{for } x_1 = 1,\ x_2 = 0 \\
P(y=1|0,1) &= \beta_o + \beta_2 && \text{for } x_1 = 0,\ x_2 = 1 \\
P(y=1|1,1) &= \beta_o + \beta_1 + \beta_2 + \beta_3 && \text{for } x_1 = 1,\ x_2 = 1
\end{aligned} \qquad (7.1.3)$$

Ved hjelp av et vanlig regresjonsprogram kan vi estimere de fire koeffisientene i (7.1.1). Vi må imidlertid være oppmerksom på at den betingede variansen for y, gitt x_1 og x_2, ikke er konstant (i alminnelighet), men er lik

$$\text{var}(y|x_1,x_2) = \frac{P(y=1|x_1x_2)(1-P(y=1|x_1x_2))}{n_{+x_1x_2}} \qquad (7.1.4)$$

De vanlige variansformlene i et regresjonsprogram vil derfor ikke gjelde. (De kan gjelde med en viss tilnærmelse når det er liten forskjell mellom alle P-verdiene).

I eksemplet 7.1.a finner vi den estimerte regresjonen

$$\hat{P}(y|x_1,x_2) = 0{,}680 + 0{,}196x_1 - 0{,}010x_2 + 0\cdot x_1x_2.$$

Ser vi bort fra sampling feilene, kan vi tolke dette slik: Det er ingen samvirkning mellom kjønn og adgang fritidshus når det gjelder helgeturer. En kvinne uten adgang til fritidshus har en estimert sannsynlighet 0,68 for å dra på helgetur. Med fritidshus er estimatet 0,196 høyere, dvs. 0,876. En mann uten adgang fritidshus har en estimert tursannsynlighet 0,01 lavere enn en kvinne, enten han har adgang fritidshus eller ei.

Regresjonsestimatorene $\hat{\beta}_j$ er forventningsrette. Variansene må vi regne ut ved hjelp av (7.1.3) og (7.1.4). For å teste signifikansen av koeffisientene, må vi bruke de spesielle formler for variansen på hver av dem. Vi kan regne med tilnærmet normal fordeling av

$$\frac{\hat{\beta}_j}{\hat{\sigma}_j}, \quad \text{der } \hat{\sigma}_j^2 = \text{estimert varians } \hat{\beta}_j,$$

når observasjonsmaterialet er stort nok, som her. For $\hat{\beta}_2$ vil variansen kunne estimeres til

$$\hat{\sigma}_2^2 = \frac{0{,}68\cdot 0{,}32}{779} + \frac{0{,}67\cdot 0{,}33}{725} = 0{,}000584265,$$

altså

$$\hat{\sigma}_2 = 0{,}024$$

Estimert standardavvik er her større enn den estimerte koeffisienten, dvs. at denne ikke er signifikant forskjellig fra 0. For $\hat{\beta}_1$ har vi variansestimatet

$$\hat{\sigma}_1^2 = \frac{0{,}68\cdot 0{,}32}{779} + \frac{0{,}876\cdot 0{,}124}{372} = 0{,}0005713,$$

$$\hat{\sigma}_1 = 0{,}024$$

dvs. $\hat{\beta}_1$ er signifikant forskjellig fra null. Det er adgang fritidshus

som betyr noe for tilbøyeligheten til å dra på helgetur, mens kjønn ikke betyr noe her. Dette er samme resultat som vi fant i avsnitt 6.8.

Tar vi regresjonen for y med henblikk på x_1 alene, får vi

$$y = 0{,}675 + 0{,}195x_1$$

Dette gir

$$\hat{p}_{1|1} = 0{,}87 \quad \text{og} \quad \hat{p}_{1|0} = 0{,}675.$$

Her er det ingen forskjell på estimatene i 6.8 og 7.1, fordi vi begge steder har en mettet modell, og da blir estimatene identiske.

Forskjellen på tolkningen av den log-lineære og den lineære modellen er at vi i den første får et additivt tillegg i log odds, mens vi i den lineære får et additivt tillegg i $\hat{p}_{1|j}$ direkte: Uten fritidshus er estimert sannsynlighet for å dra på helgetur lik 0,675. Med fritidshus er den 0,195 høyere.

For et problem med ialt m+1 variable blir den fulle (mettede) regresjonslikningen

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \beta_{12} x_1 x_2 + \ldots + \beta_{1m} x_1 x_m + \ldots + \beta_{23} x_2 x_3 + \ldots$$

$$+ \beta_{(m-1)m}\, x_{m-1} x_m + \beta_{123} x_1 x_2 x_3 + \ldots + \beta_{(m-2)(m-1)m}\, x_{m-2} x_{m-1} x_m + \ldots +$$

$$+ \ldots + \beta_{12\ldots m}\, x_1 x_2 \ldots x_m + \text{tilf. avvik} \qquad (7.1.5)$$

Foruten det tilfeldige leddet har vi like mange additive ledd i regresjonslikningen som vi har betingede sannsynligheter for de ulike kombinasjoner vi kan danne av x-enes null- og én-verdier, nemlig $s = 2^m$. (Vi har gitt β-ene fotindekser som svarer til de variabelkombinasjoner de hører til. I et konkret problem vil vi gi β-ene fortløpende fotindekser fra 0 til (s-1).)

Hvis det er strukturelle nuller i vårt problem, dvs. at det finnes kombinasjoner av x-ene som ikke kan forekomme, så sløyfer vi de tilsvarende leddene i regresjonen. Dvs. vi setter β-ene for disse leddene lik null, og får igjen en regresjon der vi har like mange β-er som p-verdier forskjellige fra null. Se A [1974] avsnitt 2.2 (siste del) eller [1976,2] avsnitt 3.

Estimering og testing

Koeffisientene i (7.1.5) kan vi estimere ved et vanlig regresjonsprogram.

Variansen på den enkelte koeffisient, $\hat{\beta}_j$, kan vi estimere ved å finne ut hvordan $\hat{\beta}_j$ kan uttrykkes som en lineærkombinasjon av $\hat{P}(1|x_1,x_2,\ldots x_m)$-verdier, og deretter estimere variansen ut fra dette, jfr. regneeksemplet foran. Vi vil ofte få forholdsvis store varianser på $\hat{\beta}_j$ som kommer "langt ut" i regresjonen, fordi de er dannet av mange $\hat{p}$-verdier.

Videre kan vi teste hypoteser om enkelte β_j eller om grupper av β_j. Vi viser til A [1976,2] avsnitt 3, der ulike problemstillinger og tester er behandlet.

7.2. Binær regresjon for variable med mer enn to kategorier

7.2.1 Flere enn to kategorier for x-ene

Vi kan fremdeles bruke regresjoner av formen (7.15) om en eller flere av de kategoriske variable i vårt problem har fler enn to kategorier. Vi gjør dette ved å innføre flere binære variable for hver kategorisk variabel. Anta f.eks. at y (helgetur) og x_1 (adgang fritidshus) er binære, mens x_2 står for alder, og er delt inn i K=4 aldersgrupper (som i tabell 4.3.a). Vi innfører da (K-1), her 3, nye binære variable etter skjemaet nedenfor.

Tabell 7.2.a Tilordning av verdier for (K-1) binære variable for en variabel med K>2 kategorier.

Variabel (x_2) kategori nr.	Binære variable z_1	z_2	z_3	....	z_{K-1}
1	0	0	0	....	0
2	1	0	0	....	0
3	0	1	0	....	0
4	0	0	1	....	0
⋮					
K	0	0	0	....	1

Vi kan velge hvilken kategori vi vil som en "basiskategori" der alle $z_j = 0$. Her har vi valgt kategori nr. 1. Poenget er at hver kategori, unntatt "basis", skal være tilordnet verdien 1 på én og bare én av de binære variable.

I vårt eksempel har vi da regresjonen

$$y = \beta_o+\beta_1x_1+\beta_2z_1+\beta_3z_2+\beta_4z_3+\beta_5x_1z_1+\beta_6x_1z_2+\beta_7x_1z_3+ \text{ tilf. avvik} \quad (7.1.6)$$

Vi får ingen ledd med produkter av to z-er, for ingen telleenhet kan være i mer enn én x_2-kategori (aldersgruppe). Derfor er (7.1.6) en full (mettet) regresjon i dette tilfellet. Vi har

$P(y=1\|0000) = \beta_o$	(ikke fritidshus, aldersgruppe 1)
$P(y=1\|0100) = \beta_o+\beta_2$	(ikke fritidshus, aldersgruppe 2)
$P(y=1\|1100) = \beta_o+\beta_1+\beta_2+\beta_5$	(fritidshus, aldersgruppe 2)
----------	----------
$P(y=1\|1001) = \beta_o+\beta_1+\beta_4+\beta_7$	(fritidshus, aldersgruppe 4)

Er det to eller flere x-variable med mer enn to kategorier, lager vi et sett av binære variable for hver x-variabel.

7.2.2 Flere enn to kategorier for y

I eksemplet i tabell 4.3.a har vi også ført inn tallet på helgeturer, gruppert i I=5 kategorier. Vi kan da innføre flere binære variable $y_1,y_2,\ldots,y_I$. I dette tilfellet kan vi innføre én y_i for hver gruppe. Vi må sette opp en regresjonslikning, av formen (7.1.6), for hver y_i, ialt I stykker. Men én av disse er overflødig fordi summen av y-ene alltid må være lik én. Hvis vi estimerer hver av de I regresjonene på vanlig måte, vil summen automatisk bli lik én, jfr. A [1976,1.].

7.2.3 Likhet og forskjell mellom estimatene ved ulike metoder

Så lenge vi estimerer i mettede modeller vil de estimatene vi får for p-ene bli nøyaktig de samme enten vi estimerer ved hyppighetene direkte, eller ved binær regresjon eller ved log-lineær metode. Ved umettede modeller blir resultatene i alminnelighet noe forskjellige, vi får en form for "glattede" $\hat{p}$-verdier når vi bruker binær regresjon, og disse blir noe forskjellige fra dem vi får ved log-lineær estimering.

Selv en mettet modell kan det være interessant å estimere, fordi vi får frem den partielle virkningen av de enkelte høyre-side variablene på estimert forventning av y.

En fordel ved en binær lineær regresjon er at vi ikke behøver å innskrenke oss til hierarkiske modeller. Vi kan, ut fra a priori innsikt, sette hvilke koeffisienter vi ønsker lik null, eller teste om de er null.

7.3. Logistiske modeller, Logit

En velkjent måte å gå frem på for å analysere én kategorisk variabel, y, som en funksjon av en eller flere andre, $x_1, x_2, \ldots, x_m$, er å foreta en såkalt logistisk transformasjon av de betingede sannsynlighetene for y, gitt de ulike kombinasjonene av x-verdier.

Hvis y bare har to kategorier, for i = 1 og i = 2, kan vi sette

$$P(y=1|x_1,\ldots,x_m) = \frac{e^{\varphi(x_1,\ldots,x_m)}}{1+e^{\varphi(x_1,\ldots,x_m)}}, \tag{7.3.1}$$

der φ er en valgt funksjon av x-ene. Vi har da

$$1-P(y=1|x_1,\ldots,x_m) = \frac{1}{1+e^{\varphi(x_1,\ldots,x_m)}}$$

og

$$\log \frac{P(y=1|x_1,\ldots,x_m)}{1-P(1|x_1,\ldots,x_m)} = \varphi(x_1,\ldots,x_m). \tag{7.3.2}$$

Det er uttrykket på venstre side som kalles logit $P(y=1|x_1,\ldots,x_m)$. Spesielt er det vanlig å la φ være en lineær funksjon av x-ene, så vi har en lineær logistisk modell som f.eks.

$$\log \frac{p_{1|x_1,\ldots,x_m}}{p_{2|x_1,\ldots,x_m}} = \beta_o + \beta_1 x_1 + \quad + \beta_m x_m. \tag{7.3.3}$$

Vi ser at "log odds" i avsnitt 6.8 er et eksempel på en slik modell. Vi behøver imidlertid ikke å gå veien om en log lineær modell, men kan sette opp (7.3.3) direkte. I våre tabeller er det naturlig å la y og x-ene være binære variable, men vi ser at modellen også har mening hvis x-verdiene er gitt kvantitativt. Det er vel særlig i slike situasjoner den har vært brukt. Se f.eks. Cox [1970], der også estimering, testing osv. er behandlet, eller BFH avsnitt 10.4.1 - 10.4.4.

En generalisering til situasjonen der I > 2, kan gjøres på ulike vis, idet vi har valg mellom flere mulige nevnere i det logaritmiske utrykket svarende til (7.3.2), f.eks.:

$$\log \frac{p_{i|x_1,\ldots,x_m}}{p_{I|x_1,\ldots,x_m}} \qquad \text{for } i = 1,2,\ldots,I-1,$$

eller

$$\log \frac{p_{i|x_1,\ldots,x_m}}{p_{i+1|x_1,\ldots x_m}} \qquad \text{for } i = 1,2,\ldots,I-1.$$

Se f.eks. Fienberg [1978] avsnitt 6.5.

7.4. Andre analysemetoder. Veiet regresjon.

Grizzle, Starmer and Koch [1969] har vist at ulike former av lineære modeller for sammenhengen mellom de variable i en kontingenstabell kan analyseres ved hjelp av veiet regresjon og tester basert på asymptotisk χ^2-fordelte observatorer. Det gjelder modeller som er lineære i de variable eller lineære i funksjoner av de variable som log-lineære modeller og andre varianter. Det er utviklet dataprogrammer for analyser, f.eks. GENCAT og FUNCAT. Det siste finnes i SAS-pakken.
Testene forutsetter store observasjonsmaterialer og observasjoner i alle ruter i tabellene.

Det finnes i litteraturen forslag til andre metoder for spesielle situasjoner, se f.eks. Cox [1970] og BFH, ch. 10. For problemer med ordnede kategorier kan også vises til Fienberg [1978], avsnitt 4.4 og 6.3.

7.5. Valg av metode

Vi har omtalt flere forskjellige metoder for estimering og testing av parametrene i modeller der vi ser på én kategorisk variabel, y, som funksjon av de øvrige variable. Vi har:

Lineær regresjon i 7.1-.2 og 7.4
Log-lineær resp. logistisk i 6.8, 7.3 og 7.4
Andre mulige varianter i 7.4.

Vårt første valg gjelder selvsagt modellen. Ønsker vi lineær regresjon i kategoriske variable, kan vi bruke 7.1-.2 eller 7.4. Den siste krever stort datamateriale, observasjoner i alle ruter og tilgang til det spesielle EDB-programmet.

Vi kan bruke 7.1-.2 for store, men også for mindre datamaterialer, og regresjonen kan tilpasses situasjoner med strukturelle eller tilfeldige nuller. Vi er ikke bundet til hierarkiske modeller. Vi kan bruke et vanlig regresjonsprogram, supplert med forholdsvise enkle tilleggsberegninger for testing o.l.

Hvis vi har kvantitative variable (regressorer) i tillegg til de kategoriske, er vel metodene i 7.3, de logistiske, mest brukt. Også her bør datamaterialet ha en viss størrelse.

For en log-lineær modell er 6.8 det naturlige valg. Et forholdsvis stort datamateriale og tilgang til f.eks. ECTA-programmet er nødvendig.

Forøvrig må vi, her som ellers, først formulere vårt problem og vår modell, og deretter søke å finne en metode som utnytter data best mulig i den situasjonen vi har.

8. NOEN SPESIELLE PROBLEMSTILLINGER. PARVISE OBSERVASJONER.

Som vi alt har sett, er det ikke nok for å finne en god analysemetode at vi kan stille opp data i en to- eller flerveistabell. Metoden må bestemmes ut fra den problemstillingen vi har og hvordan data er fremkommet. Vi skal her kort skissere noen spesielle problemstillinger og enkelte testmetoder som kan brukes.

8.1. Parvise observasjoner

8.1.1 Parvise utvalg, symmetritest

I blant kan vi redusere usikkerheten i den statistiske analyse ved å foreta parvise observasjoner av kjennetegn vi er interessert i. Anta at vi ønsker å finne ut om tilbøyeligheten til å dra på helgetur er forskjellig hos kvinner og menn. Vi antar at tilbøyeligheten avhenger både av alder og av adgang til fritidshus. Hvis vi har mulighet for å dele inn (stratifisere) populasjonen i aldersklasser og etter adgang/ikke adgang fritidshus, så kunne vi trekke tilfeldige utvalg for hvert av de to kjønn fra hvert av disse stratene slik at vi får "par" der kvinner og menn har samme bakgrunn m.h.p. aldersgruppe og adgang fritidshus. (Vi kan også ha "naturlige par", som mann og kone, eller samboere, eller bror og søster.)

For hvert medlem i et slikt par noterer vi så svaret på spørsmålet om helgetur/ikke helgetur. Vi kan telle opp antallene for de fire mulige svarkombinasjonene og sette dem opp som i tabell 8.1.a.

Tabell 8.1.a. Svar på spørsmål om helgetur for n par.

Kvinner	Menn Ja	Menn Nei	Marginal for kvinner
Ja	n_{11}	n_{12}	n_{1+}
Nei	n_{21}	n_{22}	n_{2+}
Marginal for menn	n_{+1}	n_{+2}	n

Tabell 8.1.b. Sannsynligheten for svarkombinasjoner.

Kvinner	Menn Ja	Menn Nei	Marginal for kvinner
Ja	p_{11}	p_{12}	p_{1+}
Nei	p_{21}	p_{22}	p_{2}^{+}
Marginal for menn	p_{+1}	p_{+2}	1

Hvis det ikke er forskjell på kvinner og menn m.h.t. turgåing, skulle de ha samme sannsynlighet for ja-svar, dvs. $p_{1+} = p_{+1}$. (Dette medfører at også

$p_{2+} = p_{+2}$). Nå er jo $p_{1+} = p_{11} + p_{12}$, og $p_{+1} = p_{11} + p_{21}$. Dette vilsi at vi kan formulere nullhypotesen som

$$H_o: p_{12} = p_{21}.$$

Vi har n_{11} par der begge svarer ja og n_{22} par der begge svarer nei. Disse parene er uinteressante når det gjelder spørsmålet om det er forskjell på de to kjønn. Det må være tallene n_{12} og n_{21} som inneholder informasjon om vår hypotese. Vi søker derfor å finne en betinget test, gitt summen $n_{12}+n_{21}=n'$. Vår H_o kan da uttrykkes ved at

$$H_o' = \frac{p_{12}}{p_{12}+p_{21}} = \frac{p_{21}}{p_{12}+p_{21}} = 0,5.$$

Vi ser at vi har et problem som det vi behandlet i avsnitt 3.2.4, symmetri om en diagonal.

Hvis n' er et lite tall, sammenlikner vi n_{21} med fraktilene i den binomiske fordeling med parametre n' og p=0,5. Det avhenger av den alternative hypotesen hvilke fraktiler i denne fordelingen vi skal sammenlikne n_{21} med for å finne ut om nullhypotesen må forkastes.

Hvis n' er stort nok, kan vi bruke den tilnærmede normaltesten, med observatoren

$$d = \frac{n_{21} - 0,5n'}{0,5\sqrt{n'}} = \frac{n_{21} - n_{12}}{\sqrt{n_{21} + n_{12}}}$$

Vi sammenlikner den observerte verdi av d med fraktilene i den standardiserte normalfordelingen. Denne testen kalles McNemar's test i lærebøkene.

Den blir også angitt ved å forlange forkasting av H_o når

$$d^2 = \frac{(n_{21} - n_{12})^2}{n_{21} + n_{12}} \geq z_{1-\varepsilon}, \text{ dvs. fraktilen i } \chi^2\text{-fordelingen med 1 df.}$$

Dette gir samme resultat som den tosidige normaltesten.

En versjon med (Yates) korreksjon er

$$\frac{(|n_{21} - n_{12}|-1)^2}{n_{21} + n_{12}} \geq z_{1-\varepsilon}$$

Hvis vi har et materiale med fler enn to alternativ, f.eks. som angitt i tabell 8.1.c, så kan vi bruke en lignende test.

Tabell 8.1.c Antall par for de mulige svarkombinasjoner på tre alternativ.

Menn

Kvinner	Alt. 1	2	3	Sum
Alt. 1	n_{11}	n_{12}	n_{13}	n_{1+}
2	n_{21}	n_{22}	n_{23}	n_{2+}
3	n_{31}	n_{32}	n_{33}	n_{3+}
Sum	n_{+1}	n_{+2}	n_{+3}	n

Vi kan teste "symmetrien" i denne tabellen ved en enkel summarisk test.

Vi slår sammen de to "hjørnene" i tabellen, og bruker observatoren

$$d^2 = \frac{(n_{21} + n_{31} + n_{32} - n_{12} - n_{13} - n_{23})^2}{n_{21} + n_{31} + n_{32} + n_{12} + n_{13} + n_{23}}$$

på tilsvarende måte som ovenfor.

Vi kan også utvikle mer inngående tester for sammenlikning av p_{12} med p_{21}, p_{13} med p_{21} og p_{23} med p_{32}. Se Aa II , avsnitt 3.2.

8.1.2 Andre typer av observasjoner som leder til samme testsituasjon som i 8.1.1, kan være:

(i) Panelstudier, der samme individ blir klassifisert etter svar på de samme spørsmålsalternativ på to ulike tidspunkter, f.eks. i valgundersøkelser.

(ii) Panelstudier der to "dommere" klassifiserer individer etter de samme kjennetegn.

(iii) Toveistabeller der de to kategoriske variable har samme inndeling i kategorier, f.eks. turer til fjells mot turer til sjøen, og vi ønsker å vite om det er forskjell på tur-tendensen for fjellet og sjøen.

Også her blir det spørsmål om å teste symmetrien i tabellene, som i 8.1.1.

8.1.3 Andre testmetoder

Det er utviklet flere metoder for disse problemstillingene og varianter av dem. Se f.eks. Everitt [1977], kapittel 5.4, Plackett [1974], kapittel 8 eller BFH, kapittel 8.

I den siste er det også utviklet metoder ut fra log-lineære modeller, noe som gjør det enklere å teste hypoteser for mer komplekse problemstillinger i store utvalg.

8.1.4 Parvise observasjoner med ordningseffekt, Garts test

I en situasjon der hvert individ svarer på to ulike spørsmål, A og B, eller får to ulike behandlinger, A og B (mot f.eks. hodepine), på to forskjellige tidspunkter, kan det hende at rekkefølgen av spørsmålene eller behandlingene spiller en rolle for svarsannsynlighetene.

La oss her anta at svarene er enten ja eller nei på hvert spørsmål (f.eks. om god virkning av behandlingen). I så fall bør vi sikre oss observasjoner der begge rekkefølger finnes, og helst gitt ved randomisering.

Hvis vi antar at rekkefølgen kan spille en rolle, bør vi ikke bruke testen i 8.1.1. Istedenfor å sette opp tallene som i tabell 8.1.a, kan vi bruke formen 8.1.d eller 8.1.e, der vi setter opp fordelingen av antall svar for de $n' = n_{12} + n_{21}$ individene som svarer forskjellig på de to spørsmålene. Vi fordeler nå også etter rekkefølgen av de to spørsmålene, og bruker topp-indeks, A,B for å markere at A kommer først og B som nr. to, og omvendt.

Tabell 8.1.d. Antall individer som svarer ja(1) på ett spørsmål og nei(2) på et annet, fordelt etter rekkefølgen av spørsmålene.

	Ordning av spørsmålene		
Svartype	A,B	B,A	I alt
Ja på 1. spm. Nei på 2. spm.	n_{12}^{AB}	n_{21}^{BA}	$n_{12}^{AB} + n_{21}^{BA}$
Nei på 1. spm. Ja på 2. spm.	n_{21}^{AB}	n_{12}^{BA}	n_{21}^{AB} n_{12}^{BA}
I alt	n_{ulik}^{AB}	n_{ulik}^{BA}	n'

i) Vi kan teste hypotesen om samme svarsannsynlighet på første spørsmål, uansett rekkefølgen. Innfører vi sannsynligheter som svarer til rutene i tabellen med tilsvarende symboler som for hyppighetene, har vi altså:

$$H_o': p_{12}^{AB} = p_{21}^{BA} \quad \text{mot} \quad p_{12}^{AB} \neq p_{21}^{B|A} \quad \text{(evt. > eller < istedenfor } \neq\text{)}$$

Vi har samme situasjon som i 3.1, og kan bruke Fisher-Irwins test (3.1.1) eller den tilsvarende normaltesten (3.1.2).

Vi forkaster nullhypotesen hvis n_{12}^{AB} er mindre enn nedre $\varepsilon/2$-fraktil eller større enn øvre $\varepsilon/2$-fraktil i fordelingen gitt ved:

$$f_1\left(n_{12}^{AB}\right) = \frac{\binom{n_{ulik}^{AB}}{n_{12}^{AB}}\binom{n_{ulik}^{BA}}{n_{21}^{BA}}}{\binom{n'}{n_{12}^{AB} + n_{21}^{BA}}}$$

Hvis vi ikke forkaster H_o', så kan vi ikke påstå at det er forskjell på svarsannsynlighetene på de to spørsmålene. Men uansett sluttresultat kan vi ha:

ii) Forskjellig svarsannsynlighet ved første og ved annet spørsmål. Vi setter opp de samme data som i tabell 8.1.d:

Tabell 8.1.e Antall individer som svarer ja (1) på ett spørsmål og nei (2) på det annet, fordelt etter svar på spørsmålene.

Svartype	Ordning av spørsmålene A,B	Ordning av spørsmålene B,A	I alt
Ja på A/ Nei på B	n_{12}^{AB}	n_{12}^{BA}	n_{12}
Ja på B/ Nei på A	n_{21}^{AB}	n_{21}^{BA}	n_{21}
	n_{ulik}^{AB}	n_{ulik}^{BA}	n'

Vi tester her hypotesen om samme sannsynlighet for ja-svar, uansett om spørsmålet kommer først eller sist, dvs.

$$H_o^2: \; p_{12}^{AB} = p_{12}^{BA} \quad \text{mot f.eks.} \quad p_{12}^{AB} \neq p_{12}^{BA}.$$

Igjen kan vi bruke Fisher-Irwins test (3.1.1). Vi forkaster H_o^2 hvis n_{12}^{AB} er mindre enn nedre $\varepsilon/2$-fraktil eller større enn øvre $\varepsilon/2$-fraktil i fordelingen nå gitt ved

$$f_2\left(n_{12}^{AB}\right) = \frac{\binom{n_{ulik}^{AB}}{n_{12}^{AB}}\binom{n_{ulik}^{BA}}{n_{12}^{BA}}}{\binom{n'}{n_{12}}}$$

Valget av ε er greit hvis vi bare bruker den ene av de to testene. Bruker vi begge, bør vi velge ε noe lavere enn ved bare en testing, men det er neppe nødvendig å gå ned til halvparten av vanlig ε-nivå, fordi de to testresultatene må være nokså avhengige av hverandre.

Kombinasjonen av de to testene kalles Garts test. En kan også utlede andre testmetoder for situasjoner av denne art.

9. TABELLER MED STRUKTURELLE NULLER.

Foran har vi forutsatt at det er observasjoner i alle ruter i tabellene ved de testmetodene for store utvalg som vi har sett på, og der vi bruker χ^2- eller LL-tester. Noen av testene kan modifiseres til bruk selv om det finnes enkelte nuller i tabellene, jfr. avsnitt 6.7. Hvis vi har tilfeldige nuller, kan vi, som antydet i avsnitt 2.6, kanskje finne en ad-hoc utvei for å kunne bruke testene i enkelte tilfeller.

Testene for "små" utvalg, de såkalte eksakte testene, som Fisher-Irwins test o.l., samt testene ved binær regresjon, kan vi bruke selv om vi har nuller i noen ruter, forutsatt at vi vet om vi har strukturelle eller tilfeldige nuller. Ved strukturelle nuller setter vi parametre som svarer til de tomme rutene lik null, f.eks. har vi $p_{24} = 0$ i eksemplet i tabell 9.1.a nedenfor.

Det finnes mange varianter av problemer/tabeller med strukturelle nuller. Vi skal bare ta noen enkle eksempler og forøvrig vise til BFH Chapter 5.

9.1 En toveis tabell med én strukturell null

I en (fiktiv) tidsnyttingsundersøkelse har vi funnet nedenstående tall på personer som oppgir at de har brukt et visst antall timer pr. dag til husarbeid.

Tabell 9.1.a Kvinner og menn fordelt etter antall timer pr. dag brukt til husarbeid (fiktive tall)

Kjønn	i	Under 1 time	1-2 timer	2-3 timer	Over 3 timer	
		j = 1	j = 2	j = 3	j = 4	Sum
Kvinner	1	10	90	200	700	1000
Menn	2	850	140	10	0	1000
I alt		860	230	210	700	2000

Ut fra tidligere erfaring antar vi at nullen i ruten "Over 3 timer" for menn kan betraktes som strukturell for den befolkningsgruppen vi har tatt utvalget fra. En tabell over sannsynlighetene blir derfor som i tabell 9.1.b.

Tabell 9.1.b. Sannsynligheten for eksemplet i tabell 9.1.a når $p_{24} = 0$.

i \ j	1	2	3	4	Marg.
1	p_{11}	p_{12}	p_{13}	p_{14}	p_{1+}
2	p_{21}	p_{22}	p_{23}	0	p_{2+}
Marg.	p_{+1}	p_{+2}	p_{+3}	p_{+4}	1

Forskjellen mellom menn og kvinner i tabell 9.1.a er så tydelig at det ikke skulle være noen grunn til å foreta en statistisk test i dette tilfellet. Men av hensyn til andre problemer der forskjellen ikke er så påfallende, bør vi antyde noen testmetoder.

9.1.1 Sammenlikning av sannsynligheter

Vi ville gjerne teste en hypotese om at sannsynligheten for å være i gruppe j er den samme for kvinner og menn. Vi har imidlertid forkastet denne a priori, siden vi har $p_{14} > 0$ og $p_{24} = 0$. Det vi kan teste er om sannsynlighetene for menn er lik de tilsvarende betingede sannsynlighetene for kvinner gitt at kvinnen er i en av de tre første gruppene, dvs. at vi har nullhypotesen

$$H_o: \frac{p_{11}}{p_{1+} - p_{14}} = p_{21}; \quad \frac{p_{12}}{p_{1+} - p_{14}} = p_{22} \text{ og } \frac{p_{13}}{p_{1+} - p_{14}} = p_{23}.$$

Vi kan teste dette som i en 2 x 3-tabell, $n' = n - n_{14} = 1300$ observasjoner, og med $n'_{1+} = n_{1+} - n_{14} = 300$.

Vi antar multinomisk fordeling.

Hvis vi har to utvalg, ett for menn og ett for kvinner, bruker vi testene i avsnitt 3.3.1. Vi bytter om i og j, slik at testobservatoren i den første testen i 3.1.3 blir

$$Z_h = \sum_{j=1}^{3} \frac{(n' n_{ij} - n'_{1+} n_{+j})^2}{n'_{1+} n_{2+} n_{+j}} ,$$

med (2-1)(3-1) = 2 df.

Hvis vi har ett utvalg, her regnet med n' = 1300 observasjoner, kan vi bruke metodene i avsnitt 3.3.3.

Se også avsnitt 9.2 nedenfor.

9.1.2 Bruk av log-lineær modell, jfr. kapittel 6

I problemer med strukturelle nuller setter vi de μ-parametrene lik null som svarer til kombinasjoner med sannsynlighet null. I eksemplet i 9.1.a vil vi altså sette $\mu_{24}^{12} = 0$. Har vi fler enn to variable, må vi også sette f.eks. alle $\mu_{24k} = 0$. Dette er forklart i dataprogrammene for log lineære modeller, f.eks. i ECTA-manualen. Der finnes også programmer for testing av kvasiuavhengighet (se avsnitt 9.2 nedenfor). Jfr. BFH, Ch. 5 eller Fienberg [1978] 8.2 og 8.3.

9.1.3 Binær regresjon, jfr. kapittel 7

Som nevnt i avsnitt 7.1.6 så setter vi a priori regresjonskoeffisienten lik null for ledd som ikke kan forekomme.

9.2 Uavhengighet ? Kvasiuavhengighet

Vi ser at det i en tabell som 9.1.a ikke kan være tale om stokastisk uavhengighet mellom de to kategoriske variable, jfr. avsnitt 2.2, spesielt setning 2.2.1. Den betingede fordelingen for j = 4 kan ikke være lik fordelingen for j = 1 og j = 2 og j = 3 (unntatt hvis $p_{21} = p_{22} = p_{23} = 0$, og i så fall er ikke fordelingen for i = 1 lik fordelingen for i = 2, om da ikke alle $p_{ij} = 0$).

Vi kan analysere sammenhengen mellom de variable ved å utelate data, jfr. 9.1.1, eller slå sammen data f.eks. slå j = 3 og j = 4 sammen til én gruppe. Dette kan føre til vilkårlige resultater.
En annen fremgangsmåte er å innføre begrepet kvasiuavhengighet.
Dette betyr at vi setter

$$p_{ij} = a_i b_j \qquad \text{for} \quad i = 1,2 \quad \text{og} \quad j = 1,2,3$$

dvs. for alle $p_{ij} > 0$ unntatt p_{14} som må være lik p_{+4}.
Samtidig skal de marginale sannsynlighetene være som før. Ved kvasiuavhengighet skal altså tabell 9.1.b kunne skrives som i 9.2.a.

Tabell 9.2.a Kvasiuavhengige sannsynligheter

i \ j	1	2	3	4	Marg
1	a_1b_1	a_1b_2	a_1b_3	p_{+4}	p_{1+}
2	a_2b_1	a_2b_2	a_2b_3	0	p_{2+}
Marg	p_{+1}	p_{+2}	p_{+3}	p_{+4}	1

Vi finner i dette eksemplet at vi kan ha

$$a_1 = \frac{p_{1+} - p_{+4}}{1 - p_{+4}}, \qquad a_2 = \frac{p_{2+}}{1 - p_{+4}},$$

$$b_1 = p_{+1}, \quad b_2 = p_{+2}, \quad b_3 = p_{+3}. \qquad (9.2.b)$$

Vi estimerer altså produktene i tabell 9.2.a ved å estimere a_i og b_j v.hj.a. estimatene

$$\hat{p}_{i+} = \frac{n_{i+}}{n} \quad \text{og} \quad \hat{p}_{+j} = \frac{n_{+j}}{n}$$

som vi så setter inn i (9.2.b). Så sammenlikner vi de observerte data med estimatene under nullhypotesen om kvasiuavhengighet, ved hjelp av χ^2-testen eller LL-testen. Antall frihetsgrader blir (I-1)(J-1) - antall strukturelle nuller. I eksemplet altså (2-1)(4-1)-1 = 2.

I dette eksemplet ser vi at kvasiuavhengighet betyr at sannsynligheten p_{ij} skal kunne skrives som produktet av den marginale sannsynligheten p_{+j} og den betinget marginale sannsynligheten

$$P(i|j = 1,2,3) = \frac{p_{i+} - p_{i4}}{1 - p_{+4}} .$$

Vi finner

$$Z = \sum_{i=1}^{2} \sum_{j=1}^{3} \frac{(n_{ij} - n\hat{a}_i\hat{b}_j)^2}{n\hat{a}_i\hat{b}_j} = 3690$$

som selvsagt bekrefter at her tyder ikke data på noen uavhengighet mellom de variable.

I større tabeller og med andre problemstillinger er prinsippet for kvasiuavhengighet det samme, men vi kan ikke i alminnelighet finne eksplisitte løsninger for $a_i b_j$ eller tilsvarende produkter av høyere orden. I dataprogrammene, f.eks. ECTA, er det lagt inn iterasjonsrutiner som fører til estimering av sannsynlighetene under hypotesen om kvasiuavhengighet, og testing av denne hypotesen. Jfr. henvisningen i 9.1.2 ovenfor.

Vi må selvsagt vurdere i hvert enkelt tilfelle hvilken mening begrepet kvasiuavhengighet kan ha i det problemet vi analyserer.

10. KAN VI TREKKE FORNUFTIGE KONKLUSJONER SELV OM VI HAR GALE A PRIORI FORUTSETNINGER ELLER HYPOTESER?

Vi vet at den stokastiske modellen vi bruker for et problem og et observasjonsmateriale aldri er "riktig" i den forstand at den gir en nøyaktig beskrivelse av vårt "virkelige" problem. Vi må nøye oss med å forsøke å få modellen til å omfatte de vesentlige trekk i problemet og så trekke konklusjoner som er holdbare gitt modellen. (Se f.eks. kapitlene 1 og 2 i "utdrag av forelesninger i teoretisk statistikk" av Frisch og Haavelmo [1971].).

Hvor avhengige er nå konklusjonene vi trekker av de forutsetningene vi har gjort ? Kan spesifikasjonsfeil i modellen gjøre hele den statistiske analysen meningsløs ? Dette er selvsagt mulig, det er lett å finne eksempler på det. Kanskje vi bør gjøre færrest mulige forutsetninger om fordelingen av de variable i vårt problem ? Dessuten bør vi vel bruke metoder som er robuste mot spesifikasjonsfeil ? Prisen vi må betale for å sikre oss slik, kan imidlertid være at det ikke blir mulig å trekke konklusjoner ut fra vårt observasjonsmateriale, teststyrken blir altfor svak. For å få sterkere tester forsøker vi derfor å trekke all den a priori viten vi måtte ha om vårt problem, inn i modellformuleringen, så denne kan bli mest mulig spesifisert, med den risiko dette innebærer for feilspesifisering.

Vi kan ikke ta opp en generell diskusjon av hva som kan hende hvis vi har gjort feilaktige forutsetninger. Vi skal bare se på enkelte problemstillinger som dukker opp, og som lar seg håndtere.

10.1 "Gal nullhypotese" ?

I avsnitt 2.3.2 har vi skissert problemstillingen i vanlig hypotesetesting: Vi formulerer en nullhypotese og visse alternativ til denne. I alminnelighet er det alternativene som er interessante. Noen forskere har motforestillinger mot å bruke en null-hypotese som de på forhånd mener å vite er gal. Men poenget med

nullhypotesen er ikke at vi tror den er riktig, den er bare en beskrivelse av hvordan modellen må være hvis en interessant hypotese ikke gjelder. (Se f.eks. Sverdrup [1977] avsnitt 2).
Ved vår testmetode kan vi trekke den konklusjon at alternativet gjelder, hvis vi forkaster nullhypotesen.

10.2 "Utvidet nullhypotese"

Et annet problem med nullhypotesen kan være at vi formulerer den for "skarpt". Vi sammenlikner to binomiske fordelinger ved å teste H_o: $p_1 = p_2$ mot f.eks. $p_1 \neq p_2$. I små eller moderate observasjonsmaterialer vil det ofte være slik at vi ikke kan forkaste H_o. Men i virkelig store observasjonsmaterialer finner vi en tendens til å forkaste H_o selv om forskjellen i de relative hyppighetene vi har observert er svært liten. Spørsmålet er da om vi virkelig er interessert i H_o slik vi har skrevet den med nøyaktig likhet, eller om vi burde ha skrevet p_1 er "praktisk talt" lik p_2, e.l.

Dette fenomen med forkasting av nullhypotesen i store observasjonsmaterialer gjør seg også gjeldende i analyse av flerveistabeller, f.eks. ved uavhengighetstesting.

Et botemiddel er å finne en formulering for "praktisk talt likhet", hvis det er det vi ønsker, istedenfor nøyaktig likhet.

Vi kan f.eks. i det binomiske eksemplet sette

$$H_o: -c < p_1 - p_2 < c$$

der c er et lite valgt tall, og finne et forkastingsområde for vår testobservator som svarer til dette.

Se f.eks. Hodges og Lehman [1954] eller Bjørnstad [1973] om testing av uavhengighet i toveis tabeller.

10.3. Tester som tar hensyn til at vi ikke klarer å formulere en "riktig" nullhypotese

I sin avhandling "Robust inference in contingency tables" (1981), fremhever Harald Goldstein især to vanskeligheter vi ofte møter i analysen av flerveis krysstabeller: Vi vet svært lite a priori om strukturen i tabellen. Vi kan kanskje postulere at den simultane sannsynlighetsfordelingen av de tellevariable er multinomisk (eller Poisson, eller produktmultinomisk), men kan vi forutsette f.eks. en eller annen form for uavhengighet, iallfall av høyere orden,mellom de kategoriske variable, slik at tallet på parametre i sannsynlighetsfordelingen kan reduseres ? Kan vi skrive visse $p_{ijk} = p_{ij+}\, p_{++k}$ (avsnitt 4.3), eller $\mu^{123}_{ijk} = \mu^{12}_{ij} = 0$ (avsnitt 6.5) e.l. ?
Det vi da ofte gjør, jfr. f.eks. "vanlig" bruk av ECTA-programmet, (avsnitt 6.6), er å prøve oss frem. Vi forsøker med ulike måter å redusere antall parametre på, og bruker χ^2-tester eller LL-tester for å se om en eller flere av de "reduserte" modeller "passer" til data (egentlig: om data passer til modellene). Når vi så har valgt ut en muligens redusert modell ut fra dette, begynner vi å analysere det problem som egentlig interesserer oss, og foretar tester eller konstruerer konfidensintervall på grunnlag av den reduserte modellen vi har valgt ut.

Vi bruker altså det samme observasjonsmaterialet til en rekke ulike tester e.l., og kanskje uten å ta hensyn til at resultatet av én test vanligvis vil influere både på valget av test og av sannsynlighetsnivået på de senere stadier. Vi mister rett og slett kontrollen over sannsynlighetsnivået for vår endelige konklusjon.

Til overmål kommer så det at når vi til slutt velger en testmetode, så kan den være valgt ut fra en "gal modell". Anta f.eks. at vi ut fra vår innledende analyse er kommet til at vi vil anta

(10.3.a) $$p_{ij} = p_{i+}p_{+j}$$

i en toveistabell. Deretter vil vi teste:

$$H_o: p_{12} = p_{21} \quad \text{mot} \quad p_{12} \neq p_{21}.$$

Vi kan skrive H_o som

$$p_{1+}\, p_{+2} = p_{2+}\, p_{+1} \quad \text{eller som} \quad p_{1+} = p_{+1}.$$

Vi vil vel her bli ledet til den testen vi fant i avsnitt 3.2.4, og som ikke forutsatte (10.3.a). Denne testen er altså robust mot en feilspesifikasjon av typen (10.3.a). Testen holder selv om (10.3 a) er gal.

Men så enkelt er det ikke alltid. I alminnelighet vil en sterkere spesifisert modell lede til en spesiell test som er forskjellig fra den vi ville brukt uten den sterkere spesifikasjonen. Da kan fordelingen av vår spesielle observator være en annen uten den sterke spesifikasjonen enn med, og dermed kan vårt sannsynlighetsnivå og vår konklusjon være gale. I eksemplet i avsnitt 6.5 er det forskjell på å teste den gitte nullhypotesen: $\mu^{13}_{ik} = 0$ og $\mu^{123}_{ijk} = 0$, og det å teste en nullhypotese: $\mu^{13}_{ik} = 0$ når vi a priori forutsetter at $\mu^{123}_{ijk} = 0$. I siste tilfelle får vi en LL-test med én df, mens vi i 6.5 hadde 2 df.

Goldsteins forslag til fremgangsmåte kan skisseres slik: Når det gjelder det første problem, søking etter en forenklet modell, kan vi lete oss frem med "reduserte" modeller omtrent som omtalt ovenfor. Men vi ser på utvelgingsprosessen som et middel til å komme frem til en tilnærmet riktig modell, ikke som en testprosedyre for "den sanne modell". Vi kan regne ut χ^2-verdier eller LL-verdier som vanlig, men vi tolker "små" verdier (dvs. slike verdier som vi i dag kaller "ikke signifikante") som en indikasjon på at det er så pass god overensstemmelse mellom vår reduserte modell og den sanne (ukjente) modellen, at vi tør bruke den reduserte som grunnlag for vår videre analyse. Men vi påstår ikke at den reduserte modellen er riktig.

Vi går så videre med analysen etter de kjente prinsipper, men vi må muligens velge et annet testkriterium, fordi fordelingen av våre vanlige χ^2- eller LL-kriterier kan være en annen under den generaliserte nullhypotesen enn under en stringent nullhypotese. Vi må finne en robust observator som har samme fordeling enten den stringente formulering er riktig eller ei. H.G. gir eksempler på slike for store datamaterialer.

11. FLERVARIABELPROBLEMER. HVORFOR BØR VI ANALYSERE FLERE VARIABLE SIMULTANT? ER DET IKKE NOK Å SE PÅ KRYSSTABELLER FOR DE VARIABLE PARVIS?

Når vi har et problem med et bestemt antall variable som vi vil analysere simultant, skulle det fremgå av kapitlene foran hvordan en del spesielle problemstillinger kan analyseres. Vi har imidlertid ikke sett noe særlig på den situasjon vi ofte møter i praksis når vi har observert flere (mange) variable: kan vi se på de variable to ad gangen og kanskje 3 ad gangen for å finne ut noe om samvariasjonen ? Toveis- og noen treveistabeller er jo enkle å sette opp og forstå. Men hvilke konklusjoner tør vi trekke av dem ? Vi bør faktisk være ganske forsiktige, viser det seg. En simultan analyse av mer enn tre variable ser imidlertid mer innviklet ut. Vi kan nok teste for uavhengighet og finne at det er avhengighet mellom de variable, men hvordan er egentlig avhengighetsforholdene ?

Her kan de nyere metodene, som omtalt i kapitlene 6-10, hjelpe oss til å utføre en simultan analyse slik at vi kanskje kan få frem mer spesifikk informasjon om hvordan de variable avhenger av hverandre. Vi skal se på et eksempel som viser iallfall noen aspekter av dette.

11.1. Et eksempel med testing av to-veis tabeller og med simultan analyse

Vi har et observasjonsmateriale med fire variable, der vi er interessert i hvordan én av de variable varierer med de tre andre. Data er hentet fra en artikkel av J.E. Higgins og G.G. Koch i International Statistical Review vol. 45 nr. 1, 1977. Vi har et noe annet formål med analysen og bruker en annen metode enn forfatterne gjør. Deres opplegg må vel kalles en form for "dataanalyse" ved hjelp av en viss bruk av regresjoner. Men deres konklusjoner blir omtrent som dem vi kommer til.

Data gjelder en sykdom i åndedrettsorganene som hos oss kalles "bomullsyke", på engelsk "byssinosis". Den opptrer hos arbeidere i bomullsindustrien. For å forenkle eksemplet slår vi sammen en del grupper i det opprinnelige observasjonamateriale, så vi får følgende variable:

$y = 0$ bomullsyke ikke observert
$\quad 1$ ———"—— er observert
$x_1 = 0$ arbeidsplass med lite bomullstøv
$\quad = 1$ ——"——— " meget bomullstøv
$x_2 = 0$ ikke røker
$\quad = 1$ røker

x_3 = 0 under 10 år på samme arbeidsplass
= 1 10 år eller mer på samme arbeidsplass.

Observasjonene er gjengitt i tabell 11.1.a.

Tabell 11.1.a Antall observasjoner for de forskjellige kombinasjoner av 4 variable

Arbeids-tid	Bomull-syke		Lite støv $x_1 = 0$		Meget støv $x_1 = 1$		Sum
			Ikke røker	Røker	Ikke røker	Røker	
x_3		y	$x_2 = 0$	$x_2 = 1$	$x_2 = 0$	$x_2 = 1$	
< 10 år	Nei	0	1004	1340	119	203	2666
$x_3 = 0$	Ja	1	12	14	7	30	63
≥ 10 år	Nei	0	986	1360	81	161	2588
$x_3 = 1$	Ja	1	10	24	11	57	102
Sum			2012	2738	218	451	5419

Ialt 165 observasjoner av bomullsyke blant de 5419 observerte arbeiderne.

Vi ønsker å undersøke samvariasjonen mellom sykdomshyppighet og de tre andre variable.

11.1.1. Testing av toveis tabeller

En gjengs fremgangsmåte er å se på toveistabellene mellom sykdomsvariabelen og hver av de tre andre variable for seg. Vi finner da ved å bruke testen for store utvalg i avsnitt 3.1.2, jfr. 3.2.1 (sammenlikning av to sannsynligheter), eller ved uavhengighetstesten i avsnitt 3.1.3, følgende resultater:

Bomull-syke	Arbeidsplass Lite støv	Meget støv	Sum
Nei	4690	564	5254
Ja	60	105	165
Sum	4750	669	5419

$d = 20,33$ som svarer til $z_h = (d^2) = 413,7$

Bommull-syke	Røking Nei	Ja	
Nei	2190	3064	5254
Ja	40	125	165
	2230	3189	5419

$d = 4,48$ som svarer til $z_h = 20$

Bomull-syke	Arbeidstid < 10 år	≥ 10 år	
Nei	2666	2588	5254
Ja	63	102	165
	2729	2690	5419

$d = 3,17$ som svarer til $z_h = 10$

Med én frihetsgrad er 5-prosentfraktilen for z lik 3,84 og 1-prosentfraktilen er 6,63.

Vi ser at i alle tre tilfellene er det signifikante forskjeller, d.v.s. vi må forkaste hypotesen om uavhengighet mellom de variable i hver av tabellene. Betyr dette at bomullstøv, røking og langvarig arbeid i bomullspinderi hver for seg kan medføre bomullsyke ? Kanskje ville vi også være fristet til å si noe slikt som at meget støvet arbeidsplass betyr mest, så kommer røking, og deretter arbeidstid i denne industrien. Da bruker vi imidlertid et rent "dataanalyse"-synspunkt. Denne rekkefølgen av testobservatorverdiene gjelder for dette materialet, men vil vi

få det samme resultat i et nytt materiale ?

11.1.2. Mettet lineær regresjon

La oss se på en simultan analyse av alle variable. Vi velger metoden i avsnitt 7.1, og uttrykker y som en lineær funksjon av x-ene, med alle samvariasjonsledd inkludert.

$$y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3 + \text{tilf. avvik}$$

(Vanligvis vil vi gi β_j fortløpende nummerering, opp til 7, men en nummerering som her gjør det lettere å huske hvilke variable β_j er knyttet til). Vi finner at den estimerte regresjonen blir

$$\hat{p}(y = 1 | x_1 x_2 x_3) = \underset{}{0{,}0118} + \underset{0{,}02}{0{,}0437 x_1} - \underset{0{,}004}{0{,}0015 x_2} - \underset{0{,}005}{0{,}0018 x_3} + \underset{0{,}02}{0{,}0747 x_1 x_2} + \underset{0{,}04}{0{,}0658 x_1 x_3} + \underset{0{,}006}{0{,}0088 x_2 x_3} + \underset{0{,}05}{0{,}0599 x_1 x_2 x_3}.$$

Standardavvik beregnet som i avsnitt 7.1 står under koeffisientene.
En direkte tolkning av tallene, uten hensyn til usikkerhetene i koeffisientene, må bli som følger:

$\hat{\beta}_o$ tyder på at det er en viss liten risiko for bomullsyke på en arbeidsplass med lite støv for ikke-røkere med kort arbeidstid.

$\hat{\beta}_1$ tyder på at meget støv på arbeidsplassen øker risikoen, selv for ikke-røkere med kort arbeidstid.

$\hat{\beta}_2$ røking øker <u>ikke</u> risikoen ved kort arbeidstid på en arbeidsplass med lite støv.

$\hat{\beta}_3$ risikoen øker heller <u>ikke</u> med arbeidstiden, for ikke-røkere på en arbeidsplass med lite støv.

$\hat{\beta}_{12}$ Røking kombinert med støv øker risikoen ytterligere, ut over støv alene

$\hat{\beta}_{13}$ lenger tid på støvet arbeidsplass øker risikoen ytterligere selv uten røking

$\hat{\beta}_{23}$ Røking kombinert med lenger arbeidstid gir muligens en liten økning av risikoen på en lite støvet arbeidsplass

$\hat{\beta}_{123}$ Røking kombinert med lenger arbeidstid øker risikoen ytterligere på en støvet arbeidsplass.

Vi ser at koeffisientene $\hat{\beta}_2$ og $\hat{\beta}_3$ er små negative tall og ikke signifikant forskjellig fra null. Vi må trygt kunne regne med at $\beta_2 = 0$ og $\beta_3 = 0$. Av de øvrige koeffisientene er de tre siste egentlig ikke signifikante hvis vi ser på dem enkeltvis. På den annen side er de av en viss størrelsesorden sammenliknet med de første, og vi vet at denne estimeringsmetoden gir stor varians på de siste koeffisientene. Ut fra dette bør vi kanskje ikke være for snare til å tro at de ikke betyr noe.

En annen måte å gå frem på er å sammenlikne de fire relative hyppighetene

$$\hat{p}_{1|110} = \frac{n_{1.110}}{n_{.110}}, \quad \hat{p}_{1|101} = \frac{n_{1.101}}{n_{.101}}, \quad \hat{p}_{1|011} = \frac{n_{1.011}}{n_{.011}} \text{ og } \hat{p}_{1|111} = \frac{n_{1.111}}{n_{.111}}$$

ved en spesiell χ^2-test, se A. [1976],avsnitt 2, for å se om de er signifikant forskjellige. Hvis de er det, vil vi også finne signifikant forskjell ved minst én sammenlikning av de fire siste hyppighetene (dvs. sannsynlighetene) parvis (eller omvendt), når vi ved sammenlikningen bruker $\sqrt{z_{0,95,3}} = \sqrt{7,81} = 2,79$ istedenfor normalfraktilen. I vårt eksempel finner vi for $p_{1|111}$ og $p_{1|101}$ at

$$\frac{\hat{p}_{1|111} - \hat{p}_{1|101}}{\sqrt{\hat{\sigma}^2_{\hat{p}_{1|111}} + \hat{\sigma}^2_{\hat{p}_{1|101}}}} = 3,15.$$

Vi har 3,15 > 2,79, altså signifikant økning (på 5%-nivået) av risikoen for bomullsyke når røking kommer i tillegg til støv og lang arbeidstid.

Vi kan sammenlikne andre par av $\hat{p}$-verdier, men dette er nok til å vise at det er signifikant forskjell mellom de fire $\hat{p}$-ene.
Ut fra dette tør vi vel ikke uten videre sløyfe siste ledd i regresjonen. Vi har brukt en mettet regresjon, dvs. at vi har igrunnen bare sammenliknet de relative hyppighetene av bomullsyke for de ulike kombinasjoner av de tre øvrige variable.

Det kan være et spørsmål om vi bør regne ut en ny regresjon, der vi setter $\beta_2 = \beta_3 = 0$ a priori. Dette innebærer at vi estimerer $p(y|x_1x_2x_3)$-verdiene på en litt annen måte, og de får litt andre varianser enn i den mettede regresjonen, men forskjellen er liten i dette tilfelle. Vi finner

$$p^*(y = 1|x_1x_2x_3) = 0{,}0107 + 0{,}0449x_1 + 0{,}0732x_1x_2 + 0{,}0640x_1x_3 + \\ + 0{,}0067x_2x_3 + 0{,}0621x_1x_2x_3.$$

I begge regresjonene i dette avsnittet vil et vanlig regresjonsprogram gi standardavvik for koeffisientene som ikke gjelder i våre regresjoner, jfr. avsnitt 7.1 om variansene i vårt tilfelle.

11.1.3. Log-lineær analyse av eksemplet

Vi skal nå foreta den simultane analysen ved en log-lineær modell. Ved hjelp av BMD P4F-programmet finner vi estimatene for $\log np_{yx_1x_2x_3}$, jfr. avsnitt 6.4. Vi bruker toppskrift b for bomullsyke-, s for støvet arbeidsplass-, r for røke- og a for arbeidstidsvariabelen. Vi finner

$$\log \hat{np}_{1111} = \hat{\mu}_m + \hat{\mu}_1^b + \hat{\mu}_1^s + \hat{\mu}_1^r + \hat{\mu}_1^a + \hat{\mu}_{11}^{bs} + \hat{\mu}_{11}^{br} + \hat{\mu}_{11}^{ba} + \\ + \hat{\mu}_{11}^{sr} + \hat{\mu}_{11}^{sa} + \hat{\mu}_{11}^{ra} + \hat{\mu}_{111}^{bsr} + \hat{\mu}_{111}^{bsa} + \hat{\mu}_{111}^{bra} + \hat{\mu}_{111}^{sra} + \hat{\mu}_{1111}^{bsra}$$

$$= 4{,}386 - 1{,}587 - 0{,}467 + 0{,}373 + 0{,}052 + 0{,}615 + 0{,}144 + 0{,}129 + \\ + 0{,}168 + 0{,}008 + 0{,}069 + 0{,}091 + 0{,}084 + 0{,}045 - 0{,}026 - 0{,}041.$$

Det estimerte (asymptotiske) standardavviket på hver koeffisient er $\hat{\sigma}_{\hat{\mu}} = 0{,}049$. Koeffisientene for andre kombinasjoner av variablene enn (1111) finner vi ved å skifte fortegn for $\hat{\mu}$-ene i henhold til reglene i avsnitt 6.2, jfr. tabell 6.5.a.

Ser vi på bomullsyke som funksjon av de tre andre variablene, jfr. avsnitt 6.8, finner vi estimatet for logodds eller logit, som

$$\log \hat{\Omega}_{1|111} = 2\hat{\mu}^{b}_{i} + 2\hat{\mu}^{bs}_{11} + 2\hat{\mu}^{br}_{11} + 2\hat{\mu}^{ba}_{11} + 2\hat{\mu}^{bsr}_{111} + 2\hat{\mu}^{bsa}_{111} + 2\hat{\mu}^{bra}_{111} + 2\hat{\mu}^{bsra}_{1111}$$

$$= -3,174 + 1,23 + 0,288 + 0,258 + 0,182 + 0,168 + 0,09 - 0,082.$$

Setter vi $z_1 \begin{cases} = 1 \text{ når } x_1 = 1 \\ = -1 \text{ " } x_1 = 0, \end{cases}$ $z_2 \begin{cases} = 1 \text{ når } x_2 = 1 \text{ og} \\ = -1 \text{ " } x_2 = 0 \end{cases}$ $z_3 \begin{cases} = 1 \text{ når } x_3 = 1 \\ = -1 \text{ " } x_3 = 0, \end{cases}$

kan vi skrive, jfr. avsnitt 6.8,

$$\log \hat{\Omega}_{1|z_1z_2z_3} = -3,17 + 1,23z_1 + 0,29z_2 + 0,26z_3 + 0,18z_1z_2 + 0,17z_1z_3 + 0,09z_2z_3 - 0,08z_1z_2z_3.$$

Vi kan ikke tolke koeffisientene her helt analogt med dem vi fant i den lineære regresjonen. Der er koeffisient nr. 2, 3 og 4 uttrykk for "rene" egenvirkninger av variablene, og de senere for samvirkning mellom to (resp. tre) variable som kommer i tillegg til summen av egenvirkningene.

I den log-lineære modellen er alle koeffisientene uttrykk for gjennomsnitts-virkninger på de ulike plan. Vil vi se på virkningen av røking alene på en lite støvet arbeidsplass og for kort arbeidstid, må vi se på

$$\log \hat{\Omega}_{1|-1\ 1-1} - \log \hat{\Omega}_{1|-1-1-1} = 4\hat{\mu}^{br}_{11} - 4\hat{\mu}^{bsr}_{111} - 4\hat{\mu}^{bra}_{111} + 4\hat{\mu}^{bsra}_{1111} =$$

$$= 0,576 - 0,364 - 0,18 - 0,164 = -0,132.$$

Dette betyr at odds for røking alene estimeres til ca. 0,88 gange odds for ikke røkere når de andre to variablene heller ikke er positive, dvs. samme resultat som regresjonsanalysen ga (med mettet modell i begge tilfelle skal vi jo få de samme resultatene). Vi kan direkte regne ut oddsforholdet fra tabell 3.1.a og finner $\frac{14}{1340} / \frac{12}{1004} = 0,87$ som stemmer bortsett fra avrunding.

Tilsvarende finner vi at

$$\log \hat{\Omega}_{1|111} - \log \hat{\Omega}_{1|1-1\ 1} = 4\hat{\mu}^{br}_{11} + 4\hat{\mu}^{bsr}_{111} + 4\hat{\mu}^{bra}_{111} + 4\hat{\mu}^{bsra}_{1111}$$

$$= 0,576 + 0,364 + 0,18 - 0,164 = 0,956.$$

Dette vil si at når alle tre variable er positive så er odds 2,6 ganger så stor som for en ikke-røker med positive verdier av de to andre variablene. Direkte regning fra tabell 11.1.a gir også $\frac{57}{161}\Big/\frac{11}{81} = 2,6$.
Selve $\hat{p}$-verdiene er $\hat{p}_{1|111} = 0,26$ og $\hat{p}_{1|101} = 0,12$. Den negative verdien av μ^{bsra}_{1111} betyr altså ikke at odds for sykdomstilbøyeligheten går ned ved lang arbeidstid i tillegg til støv og røking, men bare at den økning vi får ved å addere de tre første koeffisientene er noe for høy og må reduseres litt.

Vi ser av estimatet for $\log np_{1111}$ foran, at mange av de siste leddene, f.eks. fra $\hat{\mu}^{sa}_{11}$ og utover, ikke er signifikante ved en vanlig (tilnærmet) normaltest. Tilsvarende ledd i $\log \hat{\Omega}_{1|z_1z_2z_3}$ er de 4 siste. En vanlig fremgangsmåte er da å sløyfe de 7 siste leddene i $\log \hat{np}_{1111}$, og dermed de 4 siste i $\log \hat{\Omega}_{1|z_1z_2z_3}$ og så undersøke om estimeringen gir "god tilpasning" som omtalt i avsnitt 6.5. Vi har ikke funnet det forsvarlig å sløyfe så mange ledd i dette eksemplet. Vi har foretatt en estimering med $\hat{\mu}^{sra}_{111}$, $\hat{\mu}^{bra}_{111}$ og $\hat{\mu}^{bsra}_{1111}$ lik null, dvs. at de to siste ledd i $\log \Omega_{1|x_1x_2x_3}$ settes lik null. Resultatet er

$$\log \overset{*}{\Omega}_{1|x_1x_2x_3} = -3,172 + \overset{b}{1,21}z_1 + \overset{bs}{0,30}z_2 + \overset{br}{0,288}z_3 + \overset{ba}{0,182}z_1z_2 + \overset{bsr}{0,156}z_1z_3 + \overset{bsa}{\ }.$$

LL-test og χ^2-test med 3 frihetsgrader gir begge verdien 2,02 slik at tilpasningen er god nok. Standardavviket på hver av koeffisientene ($2\hat{\mu}^{bs}_{11}$ osv.) er ca. 0,0485 x 2 = 0,097. I dette tilfelle finner vi f.eks.

$$\log \overset{*}{\Omega}_{1|111} - \log \overset{*}{\Omega}_{1|1-11} = 0,6 + 0,364 = 0,964,$$

altså et forholdstall på 2,62, som ikke avviker stort fra det vi hadde i den mettede modellen.

Det gjør ingen forskjell å sløyfe de siste leddene her.

11.2. Konklusjoner

11.2.1. Feilaktig påstand om "virkningen" av de enkelte variable

Ut fra testingen av toveis tabellene alene, kunne vi kanskje tro at mye støv, røking, og lang arbeidstid i bomullsindustrien hver for seg kan øke risikoen for bomullsyke for arbeidere i bomullsindustrien. Men den simultane analysen viser at hverken røking eller lang arbeidstid øker risikoen når det er lite støv på arbeidsplassen. Derimot kan røking og lang arbeidstid hver for seg øke den risikoen som allerede gjelder på en arbeidsplass med mye støv. Og vi kan ikke se bort fra den mulighet at både røking og lang arbeidstid kombinert øker risikoen ytterligere.

I dette eksemplet kommer disse resultatene spesielt tydelig frem i den direkte regresjonsanalysen.

I den loglineære analysen må vi regne videre ved hjelp av de estimerte koeffisientene for å finne disse resultatene.

Ved andre problemstillinger kan tolkingen av den log lineære analysen være enklere.

11.2.2. Feilaktig "vraking" av variable

Hvis vi i vårt eksempel hadde sagt ut fra resultatene i 11.1.1. "arbeidstid betyr lite sammenliknet med støv og røking, vi sløyfer arbeidstid i den videre analysen", så ville vi ikke fått frem at arbeidere med lang arbeidstid har en større risiko på en støvet arbeidsplass eller/og hvis de røker.

Det kan være en betenkelig praksis å velge "forklaringsvariable" ut fra analysen av toveistabeller alene. Resultatet av toveisanalyser kan få en til både å overvurdere og undervurdere spesielle samvariasjoner. Hver av dem avhenger av de øvrige samvariasjoner i materialet.

11.2.3. Ikke ta med for mange variable i analysen

Det som er sagt ovenfor betyr ikke at vi skal ta med flest mulig variable i en simultan analyse. For mange variable kan føre til at vi ikke får frem interessante resultater i det hele tatt.

Det lønner seg å velge ut variable slik at vi ikke får med mer enn én fra en gruppe som vi a priori vet må være sterkt korrelert. Da tar vi med den ene som en "representant" for hele gruppen. I eksemplet ovenfor kan det tenkes at alder som en ny variabel ville vært sterkt korrelert med arbeidstid i industrien, slik at det ville vært vanskelig å tolke koeffisientene

for alder og arbeidstid. Hvis vi vet a priori at det kan være en tendens til økende bomullsyke hyppighet med alderen, uavhengig av arbeidstid i bomullsindustrien, så må vi, om vi sløyfer alder i analysen, regne med at den økende tendens med økende arbeidstid inkluderer en viss økning med alderen.

12. NOEN FÅ ORD OM SPESIELLE PROBLEMSTILLINGER OG METODER SOM IKKE ER NEVNT FORAN

Det finnes mange problemstillinger og en mengde analysemetoder som vi ikke kan få med her. Det skyldes dels plasshensyn og dels at ikke alle er så viktige lenger, etterat nye metoder som det er utviklet gode EDB-programmer for, er kommet i bruk.

Vi skal kort nevne noen mer spesielle problemstillinger og vise til litteratur hvor de er behandlet.

12.1. Estimering av sannsynlighetene i de enkelte ruter, Pseudo-Bayes estimering

Hvis vi har funnet en modell for simultan estimering i en flerveis-tabell, f.eks. log-lineær som i kapitel 6 eller lineær som i avsnitt 7.1, kan vi selvsagt estimere de enkelte sannsynlighetene $p_{ij...g}$ i hele tabellen ved hjelp av denne.

I noen tilfeller ønsker vi kanskje ikke å gå veien om en slik modell. Vi kan da estimere sannsynlighetene direkte ved de relative hyppighetene

$$\hat{p}_{ij...g} = \frac{1}{n} n_{ij...g}.$$

Men har vi tilfeldige nuller eller ruter med små tall, vil de tilsvarende estimatene være dårlige. Vi kan ønske oss en viss "utjevning" av observasjonene før vi estimerer, fordi vi a priori regner med at det må være et visst "mønster" i sannsynlighetene. Det kan f.eks. ikke være så stor forskjell mellom sannsynlighetene i to naboruter, eller vi mener det er større sannsynligheter i visse ruter (hoveddiagonalen f.eks.) enn i de øvrige.

Pseudo-Bayes estimering er utviklet for slike situasjoner. En starter med å konstruere et sett "a priori sannsynligheter" ut fra a priori innsikt eller bare ut fra inspeksjon av data. Så trekkes disse a priori sannsynlighetene inn ved den egentlige estimeringen av $p_{ij...g}$ ut fra data. Metoden er bl.a. beskrevet i BFH, Chapter 12.

12.2. Tidsrekkedata, Markovkjedemodeller

Hvis vi observerer variabelverdier for de samme telleenhetene på to eller flere forskjellige tidspunkter, har vi tidsrekkedata. Et eksempel er om vi i meningsmålinger spør de samme personene måned etter måned hvilket parti de ville stemme på hvis det var valg i vedkommende måned. Vi får data som vi kan sette opp i en flerveistabell. Denne bør øyensynlig analyseres under hensyn til den avhengighet en må anta mellom tallene på de ulike tidspunkter.

Det er utviklet Markovkjedemodeller for slike og lignende problemstillinger. Se f.eks. BFH, Chapter 7 med litteraturhenvisninger, E.B. Andersen [1980], Chapter 7, eller et eksempel på en 2 x 2-situasjon i Cox [1970], avsnitt 5.7.

Demografiske problemer kan også i mange tilfeller uttrykkes ved Markovkjedemodeller.

12.3. Stianalyse, rekursiv analyse av kategoriske variable. Retrospektive analyser

I avsnitt 6.8 og kapitel 7 har vi sett på problemer der én variabel betraktes som funksjoner av de øvrige. Vi kan også ha problemer med simultane likningssystemer mellom de variable, der vi ønsker å finne ut hvordan de variable innvirker på hverandre. For kvantitative variable finnes det analyseteknikker som stianalyse, rekursiv analyse og økonometriske metoder for simultane systemer, se henholdsvis Laake [1977], Asher [1976] og lærebøker i økonometri.

Det er også utviklet en form for stianalyse og rekursiv analyse for kategoriske variable. Se f.eks. Fienberg [1977], Chapter 7 med henvisninger.

I Manski & Mc Fadden [1980] behandles kategoriske variable i økonometriske modeller.

I Fienbergs avsnitt 7.5 er tatt med litt om retrospektive studier, dvs. at en ut fra de foreliggende observasjoner forsøker å gå tilbake og observere andre variable som kanskje kan "forklare" f.eks. ulikheter mellom grupper i primærmaterialet.

12.4. Latent struktur analyse

En problemstilling som ofte opptrer ved psykometriske eller medisinske undersøkelser, men også i andre sammenhenger, er at vi må anta at det kan være individuelle ulikheter mellom de telleenhetene (individene) vi observerer. I blant kan vi postulere at settet av observerte variabelverdier, $x_1, x_2, \ldots, x_g$ hos et individ avhenger av en verdi av en ikke observerbar (eller latent) variabel θ. Hvert individ kan altså ha sin egen verdi av θ. I fordelingen av de variable $(x_1, x_2, \ldots, x_g)$ for ett individ opptrer altså θ-verdien som en ukjent parameter, som varierer fra individ til individ.

Latent strukturanalyse, se f.eks. Andersen [1980], Chapter 6, er utviklet for kategoriske variable. Se programmet GELAST, Anderson & Weinberg [1983].

Mer generelt er latent struktur analyse utviklet fra Lazarsfelds arbeid i begynnelsen av 40-årene. Den har bl.a. ført til klassifiseringsmetoder for problemer der en regner med diskrete latente variable, som latent klasse analyse for dikotome variable, og latent profil analyse for kontinuerlige observerbare variable. Se f.eks. programmet LPA 2, Mårdberg [1977].

12.5. Skalering av responsmønstre

Anta at vi spør n personer om de vil svare ja (1) eller nei (2) på hvert av m spørsmål (items). Vi ønsker å ordne de m spørsmålene etter hvordan svarene faller. Anta at vi f.eks. for m = 4 kan ordne dem slik:

	Ordnede spørsmål nr.			
	1	2	3	4
Svar:	1	1	1	1
	1	1	1	2
	1	1	2	2
	1	2	2	2
	2	2	2	2

Hvis det nå ikke finnes andre svarmønstre, som f.eks. (2 1 1 1) eller (2 1 2 1), så danner spørsmålene en "perfekt Guttman skala". En slik skala kan så brukes til å klassifisere nye individer etter svarmønster o.l.

Vanligvis kommer en ikke uten videre frem til en slik "perfekt" skala. Hvordan et materiale analyseres for å komme så nær som mulig, finner en i lærebøker i f.eks. psykometri. Se også Fienberg [1978], sist i avsnitt 8.4.

12.6. Klassifisering, diskriminantanalyse

Anta at vi har en populasjon (kanskje hypotetisk), der hvert individ kan henregnes til én og bare en av to eller flere klasser A_1, A_2 osv. Vi har et observasjonsmateriale av kategoriske variable for en rekke individer i hver klasse. Ut fra dette materialet vil vi lage en klassifiseringsregel, basert på kombinasjoner av variabelverdier som er karakteristiske for de ulike klassene. Denne regelen vil vi bruke for å plassere et nytt individ (fra en ukjent klasse) i en av klassene ut fra den variabelkombinasjon vedkommende har.

Den klassiske diskriminantanalysen er utviklet for normalt fordelte variable. Det finnes nå også en diskriminantanalyse for kategoriske variable. Se f.eks. Andersen, [1980], avsnitt 5.13.

TREDJE DEL

Dataanalyse

1. HVA MENES MED DATAANALYSE?

Som nevnt i avsnitt 1.4 i innledningen, kan vi ha et datamateriale, kanskje stort og med mange observerte variable, men vi har liten a priori viten om samvariasjonsmønster og eventuelle teoretiske sammenhenger mellom de variable. Da ligger det nær å forsøke å få "tallene selv til å tale". Gamle velkjente måter å forsøke seg frem på, er å dele opp materialet på ulike vis, regne gjennomsnitt, spredningsmål, korrelasjoner, sette opp hyppighetsfordelinger og spredningsdiagram, osv. Med utviklingen av datamaskinene er det etterhvert kommet mange teknikker for å få frem eventuelle mønstre i data, med grupperinger som maskinen leter frem ut fra gitte kriterier. Det er ikke lett å få eller å gi en oversikt over hele dette feltet idag. Det finnes et stort antall ulike fremgangsmåter og dataprogrammer.

Vi skal forsøke å beskrive i store trekk noen få mer kjente fremgangsmåter men vi vil avstå fra å gi detaljerte anvisninger på bruken av dem. En er nødt til selv å sette seg inn i mulighetene og i bruken av det enkelte program, så en kan velge ut fra de data og ønskemål en har.

Det finnes bøker som gir dels noe av "filosofien" bak og dels eksempler på bruk av ulike fremgangsmåter for forskjellige typer av data. Tukeys "Exploratory data analysis" fra 1977, eller "Exploring data analysis", ed. Dixon & Nicholson 1974, er et par eksempler. I "Modern Data Analysis", ed. Launer & Siegel 1982, kan en både se hvor lite strukturert feltet er, og finne beskrivelse av enkelte metoder. Disse bøkene gir imidlertid ikke egentlig innføring i bruken av de metodene som vel er det mest anvendte for kategoriske data og tar ikke spesielt sikte på slike.

Det som hittil har vært mest brukt, er ulike former for klyngeanalyse (cluster analysis), automatisk interaksjonsdeteksjon, AID, multippel klassifikasjonsanalyse, MCA, eller nyere metoder som korrespondanse analyse. De tre førstnevnte er egentlig utviklet for kvantitative variable, men kan brukes for kvalitative variable når kategoriene kan defineres meningsfylt ved tall, f.eks. ved dikotome variable med verdiene 0 og 1. AID og MCA brukes når en vil uttrykke eller "forklare" en av de variable som funksjon av de øvrige, jfr. avsnitt 6.8 og 7.1.

Opprinnelsen til det vi her kaller "nyere metoder", strekker seg ganske langt bakover i tiden, iallfall til 1930-årene, men det er datamaskinene som har satt fart i utviklingen av metoder for store datamateriale med mange variable. En del av metodene kan betraktes som analogier med tilsvarende multivariable analysemetoder for kvantitative variable, som prinsipale komponenter, kanonisk korrelasjon, klassisk multidimensjonal skalering, diskriminant analyse, o.l.

Vi kan se på et problem med n observasjoner av m variable som et problem med n punkter i m-dimensjonalt rom. Metodene går gjerne ut på å søke å gi en tilnærmet beskrivelse av denne punktsvermen i et rom av lav dimensjon, f.eks. i to eller tre dimensjoner. Dette kan en gjøre ved å søke å transformere de variable, slik at en får nye variable som er kombinasjoner av de opprinnelige variable. De nye har den egenskap at en størst mulig del av variasjonen i det opprinnelige "rom" blir fanget opp av noen få, én, to eller tre av de nye variable. Hvis det f.eks. er slik at to nye variable fanger opp det meste av variasjonen, så betyr det at hvis vi kunne "se" i det m-dimensjonale rommet, så ville vi oppdage at punktsvermen ligger "nesten" i et plan mer eller mindre på snei. Det er dette planet vi vil frem til ved transformasjonen av de opprinnelige variable. (jfr. analogien med å se på en regresjonslinje som en "tilnærmelse" til punktsvermen i et spredningsdiagram.) Det finnes flere variasjoner og utviklinger i ulike retninger over dette temaet. Vi skal forsøke å forklare litt om korrespondanseanalyse, som er en spesiell teknikk for analyse av krysstabeller (kontingenstabeller). Andre lignende metoder er beregnet for datamatriser, jfr. 13.1. Disse nyere metodene kombineres ofte med grafisk fremstilling av resultatene, slik at en skal kunne "se" og tolke karakteristiske trekk ved data f.eks. ved å studere hvordan "bildet" av punktsvermen ser ut i et plan.

Når en taler om grafiske metoder, så gjelder det vanligvis grafiske fremstillinger av data etter at visse beregninger er utført. Dette kan gjøres på mange trinn i dataanalysen som et visuelt hjelpemiddel til å "få en oversikt over data", til å vurdere resultater underveis i beregningene, og til å vurdere det endelige resultat.

Selvsagt blir de vanlige statistiske teknikkene vi har omtalt i del II, også brukt for dataanalyse formål, jfr. avsnitt 6.6. En prøver seg da frem med forskjellige teknikker og modeller til en finner noe en synes "passer".

En må være oppmerksom på at hvis en bruker statistiske tester i slik prøving, eller i forbindelse med rene dataanalyseprogrammer, så må dette betraktes som et rent formelt hjelpemiddel. En kan ikke regne med at de vanlige signifikansnivåene gjelder, med mindre en har lagt opp analysen på en spesiell måte som tar vare på dette problem. Jfr. avsnitt 3.4.2 eller Sverdrup [1977], slutten av avsnitt 2.

13.1. Datamatrisen

Beskrivelsen av mange av disse metodene tar utgangspunkt i at vi har data i form av en datamatrise. Dette er en liste over de enkelte observerte enhetene med observert verdi av hver variabel skrevet på en linje for hver enhet, med de variable i samme rekkefølge for hver enhet. Data i tabell 1.1 vil f.eks. være gitt som antydet i tabell 13.1.a. De to variablene kan uttrykkes ved binære variable som antydet.

Tabell 13.1.a. Datamatrise for lønnstakere med observert arbeidstid og yrkesutdanning.

Lønnstaker nr.	Arbeidstid			Yrkes-utdanning	
		x_1	x_2		x_3
1	dag	0	0	uten	0
2	skift	1	0	med	1
3	dag	0	0	uten	0
4	natt	0	1	uten	0
5	dag	0	0	med	1
.	.			.	
.	.			.	
.	.			.	
.	.			.	
.	.			.	
1318	natt	0	1	med	1
1319	skift	1	0	uten	0
1320	dag	0	0	med	1

Generelt ser matrisen ut som i tabell 13.1.b, der x_{ij} står for den verdien en har tillagt variabel nr. j observert for enhet nr. i.

13.1.b. Datamatrise

Enhet nr.	Variabel nr. 1 x_1	2 x_2	3 . . . . x_3	m x_m
1	x_{11}	x_{12}	x_{13}	x_{1m}
2	x_{21}	x_{22}	x_{23}	x_{2m}
3	x_{31}	x_{32}	x_{33}	x_{3m}
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	x_{n1}	x_{n2}	x_{n3} . . .	x_{nm}

I litteraturen om dataanalyse er det ofte en tabell som 13.1.b som blir kalt en toveistabell. Terminologien er altså da en annen enn den vi har brukt i del I og II.

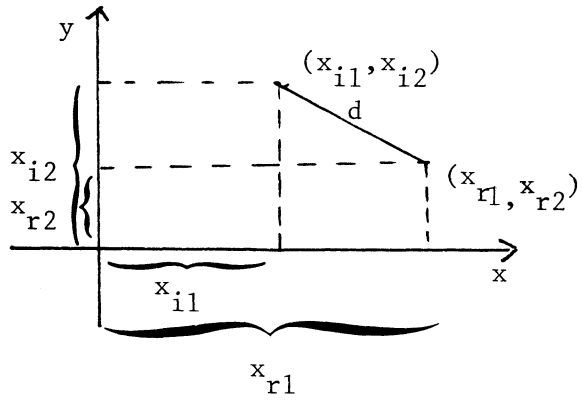
13.2. Klyngeanalyse (cluster analysis, clustering methods)

Klyngeanalyse går ut på å sortere de observerte enhetene i forholdsvis homogene grupper som ikke er definert på forhånd. Enhetene i en gruppe skal "likne" hverandre mhp. de observerte variable på en nærmere definert måte. Enhetene i ulike grupper skal være "forskjellige" fra gruppe til gruppe.

"Likhet" og "ulikhet" defineres ved hjelp av en likhets- eller avstandsfunksjon (eller - mål). Det finnes mange slike funksjoner, og mange måter å foreta sorteringen på. Manualen for TROLL-programmet "CLUSTER ANALYSIS", D 0070Q, angir 10 ulike "proximity measures" som programmet kan bruke.

For "CLUSTAN" programmet er det angitt 40 slike mål. De fleste er definert for kvantitative variable, men det finnes en del alternativ for binære variable.

Et vanlig avstandsmål mellom to observerte enheter, e_i og e_r, er den såkalte Euklid-avstanden. Det er en generalisering til flere dimensjoner av avstanden $d_{i,r}$ mellom to punkter i planet. Når punktet for e_i er definert ved koordinatene $x = x_{i1}$ og $y = x_{i2}$, og tilsvarende for de øvrige , er



$$d^2_{i,r} = (x_{r1} - x_{i1})^2 + (x_{i2} - x_{r2})^2 = (x_{i1} - x_{r1})^2 + (x_{i2} - x_{r2})^2.$$

I m dimensjoner har vi tilsvarende

$$d^2_{i,r} = \sum_{j=1}^{m} (x_{rj} - x_{ij})^2.$$

Hvis "avstanden" skal ha en fornuftig mening bør x_{ij} være kvantitative variable.

De fleste programmene kan brukes til å gi enten en hierarkisk eller ikke-hierarkisk utvelging av grupper. Ved denne siste angir man selv (eller programmet) hvor mange grupper en ønsker. Med utgangspunkt i en mer eller mindre vilkårlig angitt første gruppering undersøker maskinen gjennomsnittspunktet i hver gruppe. Så blir punkter i gruppen som ligger langt fra gjennomsnittspunktet, tatt ut av gruppen, mens andre punkter som ligger nær dette, tas inn. Dette gjentas i flere omganger til en kommer frem til grupper der enhetene ligger nær hverandre, slik at spredningen innen hver gruppe, og summen for alle grupper, er "liten" i en nærmere definert forstand.

Ved aggregert hierarkisk gruppering tas det utgangspunkt i avstandene mellom enhetene. Det lages først grupper med to enheter som ligger nær hverandre. Så samles to grupper som ligger nær hverandre i en nærmere definert forstand. Dette fortsetter så til visse gitte kriterier er oppfylt.

Hierarkisk gruppering kan også begynne med at alle enhetene samles i to grupper, med minst mulig spredning, så deles hver gruppe videre, jfr. avsnitt 13.3.

13.3. AID. Automatic Interaction Detection

AID-programmene er utviklet for å analysere hvordan en bestemt av de observerte variable, y, (den "avhengige variable") samvarierer med de øvrige observerte variable $x_1, x_2, x_3, \ldots, x_m$ (forklaringsvariable, prediktorer) når vi ikke på forhånd kan si noe om hvordan samvariasjonen ser ut for hver telleenhet. Vi antar altså her at vi har observasjoner av (m+1) variable. Vi kan f.eks. ikke forutsette lineær regresjon e.l. AID-programmet er utviklet for kvantitative variable, men kan også brukes med kategoriske forklarings-variable, samt når y er dikotom og kan gis verdiene 0 og 1. Det bør bare brukes for store datamaterialer.

Det finnes et program CHAID som er laget for en kategorisk y med fler enn to kategorier, og som har en del andre modifikasjoner i forhold til AID. Det er beskrevet av G.V. Kass: " An Exploratory Technique for Investigating Large Quantities of Categorical Data", i tidsskriftet "Applied Statistics" Vol. 29, no. 2, 1980, pp. 119-127. CHAID ser ikke ut til å være kommet til Norge enda. Et tidligere program for kategoriske variable, THAID, har visstnok ikke vunnet særlig utbredelse.

AID kan operere med ulike avstandsfunksjoner. Mest brukt er den vanlige kvadratsummen av avvik fra gjennomsnittet for y-ene. For alle n observasjonene har vi

$$d = \sum_{i=1}^{n} (y_i - \bar{y})^2, \text{ der } \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i.$$

(I dataprogram kalles d ofte for TSS, Total Sum of Squares). For en gruppe nr. k har vi tilsvarende

$$d_k = \sum_{i=1}^{n_k} (y_{ik} - \bar{y}_k)^2,$$

der n_k er antall enheter i gruppen og $\bar{y}_k$ gjennomsnittet. Prinsippet er å dele inn enhetene i grupper som gjør $\Sigma\, d_k$ minst mulig (under visse bibetingelser). Dette er det samme som å gjøre $\sum_k n_k(\bar{y}_k - \bar{y})^2$ (ofte kalt BSS) størst mulig. Vi har nemlig sammenhengen

$$d = \sum_k d_k + \sum_k n_k (\bar{y}_k - \bar{y})^2.$$

Gruppene kan dannes ved hierarkisk eller ikke-hierarkisk valg av enheter. I den norske DDPP-pakken (og mange andre) er den hierarkiske prosedyren slik: I første trinn søker programmet å dele inn enhetene i to grupper, nr. 1 og 2, slik at $d_1 + d_2$ gjøres minst mulig. Her søker programmet, for hver forklaringsvariabel etter tur, gjennom ulike inndelinger i to grupper av

kategoriene for den enkelte variabel, å finne den inndelingen som gjør kvadratsummen minst mulig ut fra denne variabelen. Så velges den forklaringsvariabelen og gruppering blant alle de m som gir minst $d_1 + d_2$. I neste skritt deles den gruppen som har størst kvadratsum videre i to ved søking blant de gjenværende forklaringsvariable. Denne prosessen fortsetter skritt for skritt inntil den stopper fordi visse fastlagte skranker for antall grupper, antall enheter pr. gruppe, og størrelsen på kvadratsummene i forhold til den opprinnelige d, er nådd. På denne måten finner en på hvert trinn den forklaringsvariabel som har størst samvariasjon med y i vedkommende gruppe, og dermed det utvalg av forklaringsvariable som ser ut til å "bety mest" for variasjonen i y.

Det blir gjerne beregnet en koeffisient som angir hvor mye den opprinnelige kvadratsummen d (eller den tilsvarende variansen) er redusert ved grupperingen. (Koeffisienten svarer til kvadratet av den multiple korrelasjonskoeffisienten i en regresjon). Det sies ofte populært at "så mye av variasjonen i y forklares av de utvalgte x-ene". Et slikt utsagn er nokså lettvint og kan kritiseres av flere grunner. (Jfr. diskusjon av korrelasjonskoeffisienten s. 9 og avsnitt 4 i A 1980).

Spesielt må vi være oppmerksomme når y er binær. Da vil minimeringen medføre at programmet søker å skille enhetene slik at de som har y = 1 fortrinnsvis kommer i én gruppe mens de som har y = 0 kommer i den annen.

En må forøvrig være klar over diverse begrensninger i muligheten for å få skilt ut de forklaringsvariable som har størst samvariasjon med y. Hvis det f.eks. er to (eller flere) forklaringsvariable som har sterk innbyrdes samvariasjon, så vil de i alminnelighet gi omtrent samme oppsplitting i to grupper. Hvis én av dem kommer ut som en "god" forklaringsvariabel, så vil den (de) andre ha liten sjanse til å komme med senere. Det gjelder her som ellers i multivariabel analyse at det ikke lønner seg å ta inn sett av forklaringsvariable som er sterkt korrelert, jfr. avsnitt 11.2.3.

Ikke-hierarkisk gruppering kan en få ved på forhånd å spesifisere det antall grupper en vil ha frem, og som skal velges ut slik at $\Sigma_k d_k$ blir minst mulig, gitt visse bibetingelser.

En referanse for AID er Sonquist, Baker og Morgan, [1973].

13.4. MCA. Multiple Classification Analysis

Anta at vi har samme problemstilling som i avsnitt 13.3, vi vil se på samvariasjonen mellom en variabel, y, på den ene siden og et sett av "forklaringsvariable" eller "prediktorer" $x_1, x_2, \ldots, x_m$ på den annen. Vi vet lite på forhånd, og vil forsøke å finne de forklaringsvariablene som ser ut til å bety noe, og eventuelt hvor meget de betyr.

MCA er en teknikk for å gjøre dette. Programmet er primært utviklet for kvantitative variable, men kan ta kategoriske forklaringsvariable og en dikotom y. For hver av forklaringsvariablene defineres det en rekke klasser eller kategorier. Vi kan ha kategori nr. $j = 1,2,\ldots,J$ for x_1, $k = 1,2,\ldots,K$ for x_2, $q = 1,2,\ldots,Q$ for x_3, osv.. Så setter vi for enhet nr. i innen kategorien (j,k,q osv.):

$$y_{ijkq\ \ldots} = \mu + a_j + b_k + c_q + \ldots. + \text{tilfeldig avvik.} \qquad (13.4.1)$$

Her er μ forventningen av y_{ijkq} som estimeres ved gjennomsnittet, $\bar{y}$, av samtlige n y-verdier, mens a_j, b_k, c_q osv. er parametre som estimeres ved minste kvadraters metode, dvs. ved å minimere

$$\sum_j \sum_k \sum_q \ldots . \sum_i (y_{ijkq\ \ldots} - \bar{y} - a_j - b_k - c_q \ldots .)^2$$

mhp. a_j-ene, b_k-ene, c_q-ene osv. Det er vanlig å bruke bibetingelsene

$$\sum_{j=1}^{J} a_j = 0, \quad \sum_{k=1}^{K} b_k = 0, \quad \sum_{q=1}^{Q} c_q = 0 \quad \text{osv.}$$

slik som i variansanalyse.

I dette programmet brukes bare den additive sammenhengen (13.4.1) for hver gruppe (j,k,q....). Det tas ikke med noen samvariasjonsledd, slik vi har det f.eks. i (7.1.1), (7.1.5) eller (6.2.1). Det er derfor tilrådelig å unngå å ta med forklaringsvariable som har forholdsvis sterk innbyrdes samvariasjon. Dette kan nemlig føre til feiltolkninger om samvariasjonen mellom y og x-ene. Programforfatterne anbefaler en bestemt bruk av AID-programmet for å plukke ut et passende sett av forklaringsvariable for MCA.

En bør helst ha mange observasjoner for å bruke MCA-programmet. Resultatet kan ellers bli temmelig vilkårlig pga. få eller ingen observasjoner i mange av gruppene. Som "mål" for betydningen av de enkelte forklaringsvariable som kommer med i utskriften av programmet, blir det oppgitt to koeffisienter for hver variabel:

En eta-koeffisient som gir uttrykk for den lineære samvariasjonen mellom y og vedkommende x_j (med den gitte grupperingen) uten hensyn til de andre x-ene i problemet. Eta svarer til en vanlig korrelasjonskoeffisient mellom to kvantitative variable.

En beta-koeffisient som gir et uttrykk for samvariasjonen mellom y og vedkommende x_j når alle de andre x-ene holdes konstante. Beta er analog med en regresjonskoeffisient eller en partiell korrelasjonskoeffisient i en multippel regresjon, men er beregnet på en noe annen måte.

Det blir gjerne antatt at rangordningen av beta-ene for de x-ene som er kommet med, angir den relative betydningen disse har for y. En stor beta angir "stor betydning" og liten beta "liten betydning". Dette kan være misvisende bl.a. hvis det er samvariasjon mellom x-ene.

En bør studere metoden og programmet mer inngående hvis en vil bruke den. Se f.eks. Andrews, Morgan, Sonquist & Klem, [1973].

Det er en bestemt sammenheng mellom betakoeffisientene og de regresjonskoeffisientene en vil få ved å kjøre en multippel regresjon for y mhp. en rekke binære variable som definerer gruppene (j.,k,q ...). Se avsnitt 7.2 om definisjon av binære variable for en kategorisk variabel. Sammenhengen er nærmere forklart hos Andrews & al.

13.5. Korrespondanseanalyse (Correspondence Analysis, Analyse factorielle des correspondances)

13.5.1. Idéen til korrespondanseanalyse tillegges Hirschfeld [1935], men den er utviklet og hittil mest brukt i Frankrike, særlig av Benzécri [1976] og hans medarbeidere. Se f.eks. Lebart, Morineau & Tabard [1977], Greenacre [1981], Sikkel [1979], de to siste på engelsk.

For korrespondanseanalyse er det vanlig å tenke seg at datamaterialet foreligger i en toveis tabell (kontingenstabell) som tabell 1.1 eller 1.2, jfr. avsnitt 2.2.1. Tallene i tabellen, n_{ij}, angir det antall av de n observasjonene som har kombinasjon nr. i for forspaltevariabelen y og j for hodevariabelen x. Hvis y angir bydel i Oslo, og x angir yrkesgruppe (resp. grupper av ikke yrkesaktive), så er n_{35} antall innbyggere i bydel 3 som er i yrkesgruppe 5. Linjesummene $n_{i+} = \sum_{j=1}^{J} n_{ij}$ angir innbyggere i bydel nr. i. Kolonnesummene $n_{+j} = \sum_{i=1}^{I} n_{ij}$ angir antall innbyggere i yrkesgruppe j. Vi har observert innbyggertallet

$$n = \sum_{i=1}^{I} n_{i+} = \sum_{j=1}^{J} n_{+j} = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}.$$

Her er I tallet på linjer (bydeler) og J er tallet på kolonner (yrkesgrupper) i tabellen. Det forutsettes at alle $n_{ij} > 0$.

I korrespondanseanalysen er man interessert i to ting, som i eksemplet svarer til

1) å finne ut likheter/forskjeller mellom bydelene mhp. yrkesfordelingen,

2) å finne likheter/forskjeller i bosettingsmønster (-fordeling) for de ulike yrkesgruppene.

Det er vanlig å se på en tabell over de relative hyppighetene

$$h_{ij} = \frac{n_{ij}}{n}, \text{ der } h_{i+} = \frac{n_{i+}}{n} \text{ og } h_{+j} = \frac{n_{+j}}{n}.$$

Nå kan jo antallene n_{1+}, n_{2+} .... (resp. h_{1+}, h_{2+} ....) være svært ulike, vi kan ha noen bydeler med forholdsvis få innbyggere og noen med mange, så vi kan ikke uten videre sammenligne linjene i tabellen. Det tilsvarende

gjelder for kolonne (yrkesgruppene). Det vi vil sammenligne for linjenes vedkommende er de I betingede hyppighetsfordelingene etter x-gruppe (yrke), jfr. tabell 2.2.b. For linje nr. i, ser vi på fordelingen

$$h_{j|i} = \frac{h_{ij}}{h_{i+}} = \frac{n_{ij}}{n_{i+}} \quad \text{for alle } j.$$

Så kan vi sammenligne disse for de ulike i-verdiene (bydelene).

Tilsvarende vil vi sammenligne de J betingede fordelingene etter bydel for yrkesgruppene, jfr. tabell 2.2.9. Dvs. vi ser på "kolonnene"

$$h_{i|j} = \frac{h_{ij}}{h_{+j}} = \frac{n_{ij}}{n_{+j}} \qquad \text{for alle } i,$$

og sammenlikner disse for ulike j-verdier (yrkesgrupper).

I større tabeller er det ikke så lett å foreta sammenlikningene "på øyemål". Derfor er korrespondanseanalysen utviklet med sikte på å finne hvilke linjer, bydeler, som "likner hverandre" og hvilke som ikke gjør det. Tilsvarende gjøres det for kolonnene.

13.5.2. Sammenlikning av y-kategori (bydeler) mhp. fordeling etter x-(yrkes-)grupper

Vi tenker oss nå hver bydel (linje) "avbildet" i et J-dimensjonalt rom ved ett punkt. Punktet for linje (bydel) nr. i har koordinatene $h_{1|i}$, $h_{2|i}, \ldots, h_{J|i}$. Vi har altså en "sverm" av I slike punkter i det J-dimensjonale rommet. (Punktene vil ligge i et (J-1)-dimensjonalt underrom, siden summen av koordinatene er én.) Som avstandsmål mellom to punkter, nr. i og nr. q, brukes et kji-kvadratmål d(i,q) der

$$d^2(i,q) = \sum_{j=1}^{J} \frac{1}{h_{+j}} (h_{j|i} - h_{j|q})^2 = \sum_{j} \frac{1}{h_{+j}} \left(\frac{h_{ij}}{h_{i+}} - \frac{h_{qj}}{h_{q+}} \right)^2 .$$

Dette er valgt bl.a. fordi det endres lite hvis vi slår sammen (eller deler videre opp i) linjer (bydeler) som har omtrent samme fordeling etter yrke (kolonner).

Vi ser at hvis den betingede fordeling etter yrke er svært lik for to bydeler, f.eks. nr. i og nr. q, dvs. at de to tallene $h_{j|i}$ og $h_{j|q}$ i hvert av de J tallparene er svært like, så blir de enkelte differensene $(h_{j|i} - h_{j|q})$

små, og dermed må d(i,q) bli et lite tall. Dvs. at punkt nr. i og nr. q ligger "nær" hverandre. På den annen side må d(i,q) kunne bli stor hvis noen eller mange differenser er store i tallverdi. "Vektene" $1/h_{+j}$ kommer også inn her, vi ser at yrker med mange utøvere ikke får for stor vekt i forhold til yrker med få utøvere. Punkter med stor verdi av d(i,q) ligger altså "langt fra hverandre".

For å gjøre det lettere å "se" hvordan punktene i punktsvermen ligger i forhold til hverandre, er det neste skritt i analysen å forsøke å finne et rom av lavere dimensjon som er slik at når vi projiserer punktsvermen inn i det, så vil "avstandene" mellom punktene i denne nye svermen gi gode tilnærmelser til "avstandene" d(i,q) i det J-dimensjonale rommet. Hvis vi f.eks. projiserer punktene ned i et to-dimensjonalt rom, et plan, og tegner inn de projiserte punktene i planet, så kan vi se hvilke som ligger "nær" hverandre og hvilke som har "lang avstand" i planet. Hvis punktene i J-rommet faktisk ligger tilnærmet i et plan, og vi har greid å finne nettopp dette, så vil punkter som ligger nær hverandre i vårt projeksjonsplan også ligge nær hverandre i J-rommet (og omvendt).

Ved å bruke minste kvadraters metode finner en (dvs. dataprogrammet) først frem til en koordinatakse for det nye rommet som er slik at punktenes avstandsvariasjon langs denne er størst mulig. Så finner en frem til en akse nr. to som står loddrett på den første og som svarer til en størst mulig del av den resterende del av variasjonen mellom punktene. Et plan med disse to aksene vil nettopp være det planet som "passer best" i punktsvermen. Hvis det fremdeles er mye variasjon som ikke er tatt hensyn til, (det er stor variasjon omkring det funne planet), kan en gå videre og finne en akse nr. tre og kanskje nr. fire. Deretter regner en (dvs. dataprogrammet) ut punktenes koordinater i forhold til de funne aksene.

Det er så vanlig å tegne inn aksene i ett plan med akse nr. en og nr. to, og deretter i et nytt plan med akse nr. to og nr. tre, osv. om nødvendig, se 13.5.4 nedenfor.

Matematisk sett er fremgangsmåten vel kjent. Med den gitte metrikk går en ut fra matrisene

$$\underset{IxJ}{H} = [h_{ij}],\ \underset{IxI}{D_I} = \begin{bmatrix} h_{1+} & & 0 \\ & \ddots & \\ & & h_{2+} \\ & \ddots & \\ 0 & & h_{I+} \end{bmatrix} \text{ og } \underset{JxJ}{D_J} = \begin{bmatrix} h_{+1} & & 0 \\ & \ddots & \\ 0 & & h_{+J} \end{bmatrix}$$

Så dannes

$$M_J = H' D_I^{-1} H D_J^{-1},$$

som er symmetrisk og ikke-negativ definit.
En finner egenverdiene for M_J,

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \ldots.$$

og de tilhørende egenvektorene u_1, u_2, u_3,.... for de største egenverdiene. Dermed har en bestemt akse nr. en, nr. to, nr. tre osv. Vi ser her bort fra en eventuell triviell rot $\lambda = 1$.]

13.5.3. Sammenlikning av x-(yrkes-) grupper mhp. fordeling etter y-kategori (bydeler)

I dette tilfellet ser vi på hver yrkesgruppe som et punkt i et I-dimensjonalt rom, med koordinater $h_{1|j}$, $h_{2|j}$,...,$h_{I|j}$ for punkt nr. j, og j er 1,2,...,J. Avstandene mellom to punkter, nr. j og nr. r, i svermen av J punkter, defineres som kji-kvadratmålet d(j,r) der

$$d^2(j,r) = \sum_{i=1}^{I} \frac{1}{h_{i+}} (h_{i|j} - h_{i|r})^2 = \Sigma \frac{1}{h_{i+}} \left(\frac{h_{ij}}{h_{+j}} - \frac{h_{ir}}{h_{+r}} \right)^2.$$

På tilsvarende måte som skissert ovenfor, projiseres så punktene i I-rommet ned i et rom av lav dimensjon, og vi ser på avstandene i dette nye rommet.

[Det er selvsagt bestemte forbindelser mellom løsningene for de to rom. De to matrisene M_J og

$$M_I = H D_J^{-1} H' D_I^{-1}$$

har samme rang og samme egenverdier $\lambda_1 \geq \lambda_2 \geq \lambda_3 \ldots$. Egenvektorene har bestemte relasjoner, f.eks. for den største egenverdien λ_1 har u_1 for M_J og v_1 for M_I sammenhengene

$$u_1 = \frac{1}{\sqrt{\lambda_1}} H' D_I^{-1} v_1 \quad \text{og} \quad v_1 = \frac{1}{\sqrt{\lambda_1}} H D_J^{-1} u_1 .]$$

13.5.4. Avbildninger i de funne plan (rom)

"Analysen" av hver av de to punktsvermene gjøres altså ved at hver av dem avbildes i ett eller flere plan. Det første planet har koordinatakser som svarer til de to største røttene i datamatrisen, dvs. de som gir størst mulig variasjon mellom punktene.

Det er vanlig å avbilde begge svermene (med ulike symboler) i samme plan. Grunnen til dette er at aksene og spredningen langs disse, er de samme. Det kan også være en tendens til at bydeler med høye verdier for visse yrker, vil ha billedpunkter i det område hvor disse yrkespunktene finnes.

Men ellers har det ikke mening å se på avstanden mellom et punkt fra den ene svermen og et punkt fra den andre svermen

Derimot har det mening å finne ut hvilke punkter som ligger nær hverandre og hvilke som er "langt fra hverandre" innen hver sverm. Vi kan se om bydel nr. 1 og bydel nr. 8 ligger nær hverandre, dvs. om de har en yrkesfordeling som er nokså lik. Og vi kan se om yrkesgruppene nr. 3 og nr. 9 er nær hverandre, dvs. om bosettingsfordelingen over de ulike bydelene er omtrent den samme. Men det har vanligvis ingen mening å fortelle at bydel 8 og yrkesgruppe 9 ligger nær hverandre.

Videre kan vi finne grupper av punkter som ligger nær hverandre innen hver sverm, osv.

Hva som kalles for "kort" avstand og hva som er "lang" vil vel avhenge av både av det problemet som blir analysert og av den som analyserer.

Det er også vanlig å "tolke" de funne aksene på tilsvarende måte som f.eks. i faktoranalyse, men vi vil ikke forsøke noe slikt her. En må bli tilstrekkelig fortrolig med hele metoden før en kan bruke den fornuftig.

13.5.5. Multi-korrespondanseanalyse.

Det finnes versjoner av korrespondanseanalyse for flerveistabeller, se f.eks. Lebart et.al. (1977) og Bølviken (1985). Bølviken har skrevet et dataprogram for sin versjon av multi-korrespondanseanalyse.

13.6. Grafiske metoder

Grafiske fremstillinger blir brukt som hjelpemiddel i nesten alle typer dataanalyse, det har vært gjort siden statistikkens barndom. Hensikten har i alminnelighet vært å gi oversikt over data eller å illustrere resultatene av en statistisk analyse. I den senere tid er imidlertid grafisk fremstilling tatt i utstrakt bruk også i selve analysearbeidet. Det lar seg neppe gjøre å sette skille mellom "grafiske metoder" og andre metoder i dataanalyse.

De mest utpregede grafiske metoder idag går vel ut på å sitte ved en EDB-terminal og kalle frem bilder av kurver, spredningsdiagram, projeksjoner osv., osv., på en skjerm, og så la utviklingen av den videre analysen bli influert eller avgjort av det eller de bilder en alt har fått frem. De beregningsmetodene som ligger bak bildene er en del av dataprogrammet og er selvsagt viktig for resultatene. Hensikten med dataprogrammene er gjerne "to bring out hidden facts". For multivariable data finnes det bl.a. "Projection Pursuit Methods" for å finne struktur i datamaterialet. Se f.eks. Friedmann & Stuetzle (1981). I forbindelse med regresjonsberegninger finnes det nå mange varianter av grafiske hjelpemidler, men de hører ikke inn her. Det fremgår av dataprogrambeskrivelsene hva som finnes i hvert tilfelle.

En referanseliste over artikler om grafiske metoder finnes i en oversiktsartikkel av Fienberg i The American Statistician, November 1979.

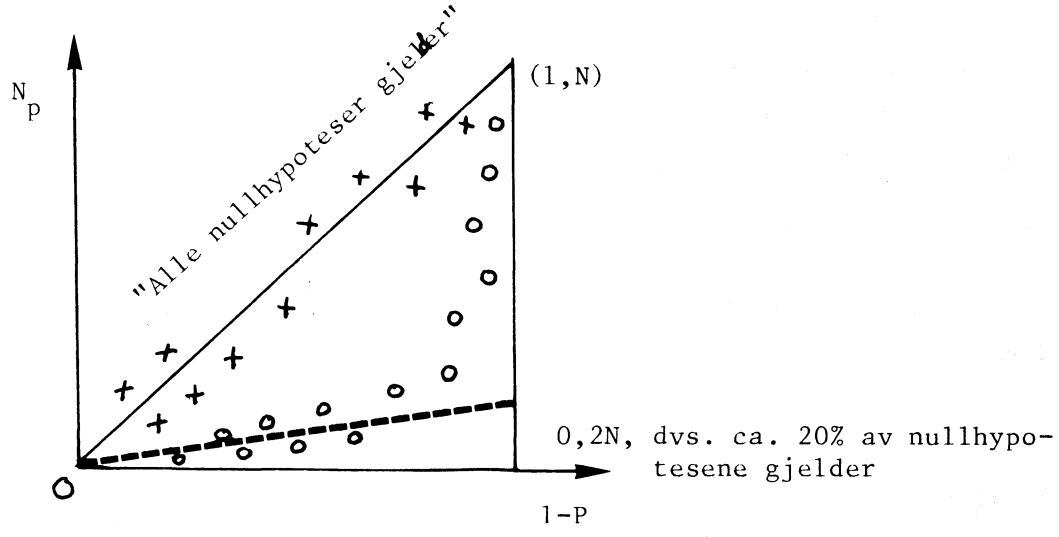
Vi skal gi en skisse av en ny metode som kan brukes også for kategoriske data.

P -plot for multippel testing

Det finnes at stort antall "plotterutiner" for diverse formål. Schweder og Spjøtvold (1982) foreslår plotting av p-verdier ved multippel sammenlikning der en bruker et stort antall enkelttester for delhypoteser. Hver test tillegges en P-verdi som er den observerte signifikanssannsynligheten (dvs. sannsynligheten under nullhypotesen for å få en verdi av observatoren som er like stor eller større enn den faktiske observerte).

Anta f.eks. at vi har en krysstabell (kontingenstabell) for to variable og utfører særskilt testing av alle de 2 x 2-tabellene som kan dannes i tabellen.

La dette være N stykker. Anta også at vi kan teste uavhengighet mellom de variable i hver 2 x 2-tabell ved χ^2-testen i 3.1.3. For hver test observerer vi en z_h-verdi, la oss si $z_{h,j}$ for test nr. j. For test nr. j kan vi regne ut sannsynligheten, P_j, for at en χ^2- fordelt variabel Z (med en frihetsgrad) vil være større enn det observerte tallet $z_{h,j}$ når nullhypotesen om uavhengighet gjelder. Vi kan så ordne de N verdiene $1-P_j$ etter størrelse og danne en kumulert hyppighetsfordeling for dem. Denne tegner vi inn i et diagram med 1-P som abscisse og hyppigheten N_P (av $(1-P_j)$-verdier $\leqq$ 1-P) som ordinat Hvis alle nullhypotesene gjelder, så vil punktene (1-P), N_P ligge pent nær den rette linjen fra origo til punktet (1,N) i diagrammet. Hvis ikke alle de N nullhypotesene gjelder, men f.eks. $N_o < N$ av dem, så vil det være forholdsvis flere små P-verdier, dvs. flere store $(1 - P_j)$ verdier enn det ville være ved uavhengighet. Punktene i venstre del av diagrammet vil ha en tendens til å ligge nær en linje med vinkelkoeffisient N_o, mens resten av punktene stiger raskt til høyre i diagrammet. Vi kan estimere N_o ved å føye en rett linje til punktene til venstre i diagrammet.



Det er ikke helt lett å finne de statistiske egenskapene ved N_p-verdiene eller ved den estimerte N_o. S og S har funnet et (stygt!) uttrykk for var N_p i tilfellet med χ^2-tester.

Ellers kan det være vanskelig å komme med sannsynlighetsutsagn for de slutninger en treffer ut fra denne metoden.

FJERDE DEL

Avhengighetsmål

14. SUMMARISKE MÅL FOR SAMVARIASJON MELLOM TO VARIABLE

Vi skal ganske kortfattet gi en oversikt over noen av de vanligste "avhengighetsmålene" uten å ta med eksempler på bruken. Slike eksempler samt diskusjon om tolkningen av de ulike mål, finner en i BFH, chapter 11. Vi skal holde oss til toveistabeller. Det er ikke uvanlig å ønske seg et enkelt "mål" for samvariasjon eller overensstemmelse mellom to variable. Dels vil en gjerne ha et numerisk uttrykk for styrken i en eventuell samvariasjon, dels vil en trekke sammenlikninger mellom to eller flere situasjoner. Det er imidlertid viktig å være klar over at det er *meget begrenset hva ett enkelt mål kan uttrykke om en kanskje komplisert sammenheng.* Nytten av det enkelte mål vil avhenge av problemstillingen: hvilken sannsynlighetsmodell, hva slags samvariasjon er det tale om ? Bruken av de estimerte målene bør i alle tilfeller suppleres med en test eller et konfidensintervall for den parameteren vi er interessert i. Det kan vel hende at vi heller bør bruke en av metodene i del II, enn bare å regne ut et enkelt "mål".

Det finnes en rekke ulike mål, vi skal kort omtale de vanligste av dem og forsøke å gi en viss presisering av problemstillingene.

14.1. Korrelasjonsmål

En gruppe av mål kan sammenliknes med den vanlige korrelasjonskoeffisienten. Vi definerer den empiriske korrelasjonskoeffisienten for n observasjonspar av kvantitative variable, $(x_1,y_1),\ldots,(x_n,y_n)$ ved

$$r = \frac{m_{xy}}{s_x s_y}, \text{ der } m_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad (14.1.1)$$

er den empiriske kovariansen mellom x og y, mens de empiriske variansene er

$$s_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ og } s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2.$$

Den tilsvarende teoretiske korrelasjonen mellom de variable x og y som (x_i, y_i) er observasjonspar for,er

$$\rho_{xy} = \frac{\text{kovar }(x,y)}{\sigma_x \sigma_y},$$

der vi nå har de teoretiske variansene og kovariansen istedenfor de empiriske.

Ikke så sjelden blir r oppgitt som et mål for avhengigheten mellom x og y. Vi må da presisere:

i) r er et estimat for ρ_{xy}

ii) ρ_{xy} er et uttrykk for styrken i den *lineære* samvariasjonen mellom x og y. Vi har jo

$$\rho_{xy} = 1 \text{ (og } r = 1) \text{ når } y = a + bx \text{ med } b > 0,$$

$$\rho_{xy} = 0 \text{ når kovar } (x, y) = 0, \qquad (14.1.2)$$

$$\rho_{xy} = -1 \text{ (og } r = -1) \text{ når y er } a + bx \text{ med } b < 0.$$

ρ_{xy} er positiv når den teoretiske lineære regresjonen for y mhp. x er stigende med x, og negativ når den er fallende. Med denne bakgrunn kan det ha mening å angi r som et mål for lineær samvariasjon når de variable refererer seg til en *intervall* skala, jfr. avsnitt 1.2.

For variable med ren nominell skala har det ingen mening å regne ut r, resultatet vil jo avhenge av hvordan vi, mer eller mindre tilfeldig, har ordnet hvert av de to settene med variabelverdier. Har vi ordningene 1: blå øyne, 2: grønne øyne, 3: brune øyne og 1: lyst hår, 2: rødt hår, 3: mørkt hår", kan vi få "god korrelasjon", men permuterer vi én av skalaene og ikke den andre, blir den neppe så god.

Har begge variable ordinal skala, så gjelder ikke denne innvendingen. Derimot må vi merke oss at tendens til lineær samvariasjon mellom to *ordningsvariable* ikke kan tolkes som "lineær samvariasjon" mellom de to egenskapene vi egentlig er interessert i, det er her en eventuell monoton samvariasjon som indikeres (f .eks. inntekt og alder angitt ved inntektsgrupper og aldersgrupper). I alle tilfelle er en korrelasjonskoeffisient et meget primitivt "mål for samvariasjon".

Spearmans rangkorrelasjonskoeffisient er den vanlige korrelasjonskoeffisienten (14.1.1) brukt på rangordningsvariable (i enkelte tekster kalles denne empiriske koeffisienten for ρ, rho, men det er ingen grunn til å gi den et særskilt navn her). I det spesielle tilfelle der hver variabel antar verdiene 1,2,...,n i et sampel på n observasjoner (og det ikke finnes sammenfallende observasjoner) blir uttrykket for Spearmans r (eller rho) spesielt enkelt. Utregning og innsetting i r av de uttrykk vi får for $\bar{x}, \bar{y}, s_x, s_y$ og m_{xy} gir

$$r = 1 - \frac{6}{n(n-1)(n+1)} \sum_{i=1}^{n} d_i^2, \text{ der } d_i = x_i - y_i. \qquad (14.1.3)$$

Hvis flere observasjoner har *samme rang* for den ene eller begge variable, *gjelder formelen ikke*. (Vi kan i så fall bruke "midtrangmetoden", dvs. gi hver av de like observasjonene en rang som er gjennomsnittet av de rangplassene de opptar i rekkefølgen. Vi må beregne r for det spesielle tilfellet.)

For å finne ut om det er signifikant samvariasjon mellom de variable, trenger vi fordelingen av Spearmankoeffisienten for $\rho = 0$. Det finnes tabeller for liten n. Med sammenfallende ranger må fordelingen utledes. I begge tilfeller er fordelingen asymptotisk normal, men n skal være ganske stor før approksimasjonen er god. Variansen på r estimeres ved $(1 - r^2)/n$. Til nød kan en for $n > 10$ bruke en t-test som jo egentlig forutsetter at den betingede fordelingen av en av de variable, gitt den annen, er normal,

14.2. Mål basert på rangordning

En rekke samvariasjonsmål tar utgangspunktet i sannsynlighetene for at to telleenheter har samme rangordning i begge variable, og for at de ikke har det.

La (x_1,y_1) og (x_2,y_2) være de to sett av rangordningsvariable for to telleenheter.

Vi sier at enhetene stemmer overens (i rang) hvis $x_1 - x_2$ og $y_1 - y_2$ har samme fortegn. Har de motsatte fortegn, stemmer de ikke overens (engelsk: discordant). La oss innføre sannsynlighetene

$$\pi_s = P((x_1 - x_2)(y_1 - y_2) > 0)$$

$$\pi_d = P((x_1 - x_2)(y_1 - y_2) < 0) \qquad (14.1.4)$$

$$\pi_0 = P((x_1 = x_2) \text{ eller } (y_1 = y_2))$$

Vi har $\pi_s + \pi_d + \pi_0 = 1$.

Disse sannsynlighetene kan vi estimere ved de tilsvarende relative hyppighetene i observasjonsmaterialet.

Definisjonene av de teoretiske versjonene av en rekke mål er:

Kendalls rangkorrelasjonskoeffisient (tau)

$$\tau = \pi_s - \pi_d$$

Kendalls rangkorrelasjonskoeffisient modifisert til kontingenstabeller (tau b)

$$\tau_b = \frac{\pi_s - \pi_d}{\sqrt{\pi_s + \pi_d}} \quad \text{når } \pi_0 \neq 1.$$

Stuarts modifikasjon for en IxJ kontingenstabell (tau c)

$$\tau_c = (\pi_s - \pi_d) \frac{m}{m-1}, \quad \text{der m er min av I og J.}$$

Goodman og Kruskals mål (gamma)

$$\gamma = \frac{\pi_s - \pi_d}{\pi_s + \pi_d} = \frac{\pi_s - \pi_d}{1 - \pi_0} \quad \text{når } \pi_0 \neq 1.$$

Vi ser at hvis $\pi_0 = 0$ (som er tilfelle for kontinuerlige variable), så vil τ, τ_b og γ falle sammen, mens τ_c er litt forskjellig fra de øvrige.

Hvis $\pi_s = 1$ så er $\tau = \tau_b = \gamma = 1$, dvs. når enhetene alltid stemmer overens i rang.

Hvis $\pi_d = 1$ så er $\tau = \tau_b = \gamma = -1$, dvs. når enhetene er diskordante.

Hvis $\pi_s = \pi_d$ så er alle fire mål lik null og vi kan si at x og y er *ordningsuavhengige*.

Egenskapene ovenfor er analoge med (14.1.2). Det kan vises at $\pi_s = \pi_d$ når de to variable er stokastisk uavhengige, men det omvendte behøver ikke holde. De tre siste kan betraktes som "normaliseringer" av τ, i den hensikt å få mål som kan sammenliknes for forskjellig situasjoner. γ kan tolkes som differens mellom to betingede sannsynligheter, og er definert unntatt når $\pi_0 = 1$.

For $\pi_0 > 0$ kan γ anta verdier i hele intervallet $[-1,1]$, mens grenseverdiene -1 og 1 ikke kan oppnås av τ og τ_b.

Samtlige mål estimeres altså ved de tilsvarende relative hyppigheter i observasjonsmaterialet. Opptellingen av disse kan kreve en del arbeid for materialer som ikke er ganske små, men beregningene er lagt inn i mange regnemaskinprogrammer.

Det kan utledes asymptotiske varianser for estimatorene $\hat{\tau}$, $\hat{\tau}_b$, $\hat{\tau}_c$ og $\hat{\gamma}$, og angis konfidensintervall med tilnærmet konfidensgrad for parametrene.

Somers D er et asymmetrisk mål som beregnes særskilt for linje- og for kolonneinndelingen. Produktet av de to er lik τ_b^2.

14.3. Mål basert på "mean square contingency"

Det finnes en serie mål som kan sies å ta utgangspunkt i χ^2-uavhengighetsobservatoren i avsnitt 3.4.3, jfr. (2.4.3).

Den teoretiske verdien som kalles "mean square contingency" eller fi-kvadrat er med betegnelsene fra avsnitt 2.2.1 for to variable:

$$\phi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(p_{ij}-p_{i+}p_{+j})^2}{p_{i+}p_{+j}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{p_{ij}^2}{p_{i+}p_{+j}} - 1.$$

Den er null ved stokastisk uavhengighet. Ved fullstendig avhengighet, dvs. $p_{ij} = 0$ for $i \neq j$ og $p_{ij} = p_{i+} = p_{+j}$ for $i = j$, er $\phi^2 = m - 1$, der $m = \min(I,J)$.
Vi estimerer ϕ^2 ved å sette inn de tilsvarende relative hyppighetene som estimatorer for p-ene, dvs. ved χ^2-formlen.

K. Pearsons kontingenskoeffisient er

$$K = \sqrt{\frac{\phi^2}{1+\phi^2}}.$$

Tschuprows koeffisient er

$$T = \sqrt{\frac{\phi^2}{(I-1)(J-1)}}.$$

Cramérs koeffisient er

$$C = \frac{\phi^2}{m-1} \quad \text{eller} \quad V = \sqrt{\frac{\phi^2}{m-1}}.$$

Alle tre faller i intervallet [0,1] og er null ved stokastisk uavhengighet. Ellers er det ikke så lett å gi en sannsynlighetsteoretisk tolkning av dem. De har imidlertid den fordelen at de er enkle å estimere, idet vi kan estimere ϕ^2 ved χ^2-observatoren.

Det kan også angis (asymptotiske) konfidensintervall for ϕ^2 og dermed for K,T og C.

Steffensens koeffisient, psi-kvadrat,

$$\psi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{p_{ij}(p_{ij}-p_{i+}p_{+j})^2}{p_{i+}(1-p_{i+})p_{+j}(1-p_{+j})}$$

Vi har $\psi^2 = 0$ ved stokastisk uavhengighet. Det kan vises at $0 \leq \psi^2 \leq 1$. Den estimeres ved å estimere p-ene ved de tilsvarende relative hyppighetene. Man kan finne asymptotisk varians og dermed konfidensintervall ved å regne med at $\hat{\psi}^2$ er asymptotisk normalt fordelt.

14.4. Mål basert på kryssproduktforholdet, for 2 x 2-tabeller

I avsnitt 2.2.4 definerte vi kryssproduktforholdet i en 2 x 2-tabell,

$$\alpha = \frac{p_{11}\, p_{22}}{p_{12}\, p_{21}} .$$

Vi så at når de to kategoriske variable er stokastisk uavhengige, så er $\alpha = 1$, og omvendt. Vi ser også at hvis enten $p_{11} = 0$ eller $p_{22} = 0$, så er $\alpha = 0$. Vi har alltid $\alpha \geq 0$. Hvis $p_{12}p_{21} < p_{11}p_{22}$ blir $\alpha > 1$, og større jo mindre sannsynligheten p_{12} og/eller p_{21} er. I prinsippet kan α altså anta verdier mellom 0 og ∞. Den er ikke direkte nyttig hvis en av p-ene er null. Vi viser til BFH, avsnittene 2.2 og 11.2.2 for en grundig diskusjon av α.

Vi estimerer α ved

$$\hat{\alpha} = \frac{n_{11}n_{22}}{n_{12}n_{21}} .$$

Denne er heller ikke så nyttig hvis en $n_{ij} = 0$. Variansen på denne estimatoren er tilnærmet, for store utvalg.

$$\text{var } \hat{\alpha} \approx \frac{\alpha^2}{n}\left(\frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}\right),$$

som estimeres ved

$$\text{est var } \hat{\alpha} = \hat{\alpha}^2 \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right).$$

Flere funksjoner av α blir brukt som assosiasjonsmål.

Yules Q er definert ved

$$Q = \frac{p_{11}p_{22}-p_{12}p_{21}}{p_{11}p_{22}+p_{12}p_{21}} = \frac{\alpha-1}{\alpha+1},$$

med estimatoren

$$\hat{Q} = \frac{\hat{\alpha}-1}{\hat{\alpha}+1}.$$

Denne har estimert tilnærmet varians lik

$$\text{est var } \hat{Q} = \left(\frac{1-\hat{Q}^2}{2}\right)^2 \left(\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}\right).$$

Vi kan regne med tilnærmet normal fordeling av $\hat{Q}$ for store utvalg. Yule foreslo også koeffisienten

$$Y = \frac{\sqrt{\hat{\alpha}}-1}{\sqrt{\hat{\alpha}}+1}.$$

14.5. Prediksjonsmål

Goodman og Kruskal har foreslått to mål λ_A og λ_B som skal brukes for å kunne forutsi hvilken verdi av den ene variable A eller B som vil bli realisert ved en tilfeldig observasjon gitt at vi har en viss informasjon. Se f.eks. BFH, avsnitt 11.3.2.

14.6. Relativ risiko.

I enkelte lærebøker og dataprogrammer opptrer en størrelse som kalles "relative risk", Ψ, i en 2 x 2-tabell. Se f.eks. BFH (2.2-8).
Anta at vi ut fra en tabell som 3.1. vil estimere den betingede sannsynligheten for *ikke* å dra på helgetur, særskilt for dem som har adgang fritidshus og dem som ikke har det, altså

$$\frac{p_{21}}{p_{+1}} \text{ og } \frac{p_{22}}{p_{+2}}.$$

Så ønsker vi å se på forholdet mellom disse to, dvs.

$$\frac{p_{21}}{p_{+1}} / \frac{p_{22}}{p_{+2}} = \frac{p_{21}}{(p_{11}+p_{21})} \frac{(p_{12}+p_{22})}{p_{22}}$$

Hvis nå p_{21} er liten i forhold til p_{11} og p_{22} er liten i forhold til p_{12}, så vil dette forholdstallet være *tilnærmet* lik

$$\Psi = \frac{p_{21}p_{12}}{p_{11}p_{22}}.$$

(Dette er altså *ikke* den Ψ som Steffensen er far til.) Denne estimeres ved

$$\hat{\Psi} = \frac{n_{21}n_{12}}{n_{11}n_{22}}.$$

Vi ser at Ψ er den inverse av kryssproduktforholdet, og kan tolkes som $\frac{1}{\alpha}$. (For 3.1 får vi estimatene 0,40 og 0,31 for de to uttrykkene, men her er ikke $\hat{p}_{22}$ så liten i forhold til $\hat{p}_{12}$.) Vi kan også estimere et konfidensintervall for Ψ ved å gå ut fra $\log \hat{\Psi}$ som har estimert tilnærmet varians

lik

$$\hat{\sigma}^2_{\log \hat{\Psi}} = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$$

(log angir naturlig logaritme). Ved å anta tilnærmet normalitet for $\log \hat{\Psi}$ kommer vi frem til konfidensintervallet ved å ta antilogaritemen til grensene

$$\log \hat{\Psi} - z_{\frac{\varepsilon}{2}} \hat{\sigma}_{\log \hat{\Psi}} \quad \text{og} \quad \log \hat{\Psi} + z_{1-\frac{\varepsilon}{2}} \hat{\sigma}_{\log \hat{\Psi}}$$

i intervallet for $\log \Psi$. Her $z_{\frac{\varepsilon}{2}}$ og $z_{1-\frac{\varepsilon}{2}}$ nedre og øvre $\varepsilon/2$-fraktil i standard-normalfordelingen.

L I T T E R A T U R

Albrecht, P., (1980) On the Correct Use of the Chi-Square Goodness of Fit Test, Scandinavian Actuarial Journal, pp. 149-160.

A I: Amundsen, H.T., (1972) Statistisk metodelære. En elementær innføring. Johan Grundt Tanum.

A II: Amundsen, H.T., (1978) Statistisk metodelære II. Tolking av data, modeller og metoder. Tanum-Norli.

Amundsen, H.T., (1974) Binary Variable Multiple Regressions, Scandinavian Journal of Statistics, Vol. 1, pp. 59-70.

Amundsen, H.T., (1976,1) Binary Regressions for a Polytomeous regressand, Scandinavian Journal of Statistics, Vol. 3, pp. 39-41.

Amundsen, H.T., (1976,2) Analysis of Qualitative Variables by Means of Binary Regression, Memorandum fra Sosialøkonomisk institutt, 12. november 1976.

Amundsen, H.T., og Ljøgodt, H., (1979): Small Sample Tests Against an Ordered Set of Binomial Probabilities, Scandinavian Journal of Statistics, Vol. 6, 81-85.

Do. (1981): Correction Note, Scandinavian Journal of Statistics, Vol. 8, 56.

A 1980: Amundsen, H.T., (1980): Korrelasjonskoeffisienten - enda engang. Arbeidsnotat fra Statistisk Sentralbyrå. IO 80/30.

Andersen, E.B., (1980): Discrete Statistical Models with Social Science Applications, North-Holland Publ. Co.

Andrews, F.M., Morgan, J.N., Sonquist, J.A., Klem, L. (1973): Multiple Classification Analysis, University of Michigan, Institute for Social Research, Ann Arbor.

Asher, H.B., Causal Modelling. Sage University Paper series on quantitative applications in the Social Sciences. Series No. 07-003. Beverley Hills/London.

Benzćri, J.P. (1973, 1976): L'Analyse des Données. Dunod.

BFH: Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., (1975): Discrete Multivariate Analysis. Theory and Practice. MIT Press.

B.I: Bjørnstad, J.F., (1973 I): Inferensteori i kontingenstabeller. Memorandum fra Sosialøkonomisk institutt, 20. september 1973.

B.II: Bjørnstad, J.F., (1973 II): A Multiple Test Procedure for a Series of Binomial Distributions with Non-Decreasing Probabilities. Memorandum fra Sosialøkonomisk institutt, 5. oktober 1973.

Bølviken, Erik (1985): Reduction of Multiway Tables via Bivariate Association: Multicanonical Analysis as Scaling Methodology. Scandinavian Journal of Statistics, Vol. 12.

Cox, D.R., (1970): Analysis of Binary Data. Metuen.

Cox, M.A.A., and Plackett, R.L. (1980): Small Samples in Contingency Tables. Biometrika, Vol. 67, 1, pp. 1-13.

Darroch, J.N., (1974): Multiplicative and Additive Interaction in Contingency Tables. Biometrika, Vol. 61, pp. 207-214.

Dixon, W.J., and Nicholson, W.L., Ed. (1974): Exploring Data Analysis, University of California Press.

Everitt, B.S. (1977): The Analysis of Contingency Tables. Chapman and Hall (Halsted Press).

Fienberg, S.E., (1970): An Iterative Procedure for Estimation in Contingency Tables. Ann. Math. Stat., Vol. 41, No. 3, pp. 907-917.

Fienberg, S.E., (1978): The Analysis of Cross-Classified Categorical Data. MIT Press. Ny utgave: (1980).

Fienberg, S.E., (1979): The Use of Chi-Squared Statistics for Categorical Data Problems. J. Roy. Statist. Soc. B, Vol. 41, No. 1, pp. 54-64.

Fienberg, S.E.(1979b): Graphical Methods in Statistics. The American Statistician, Vol. 33, No. 4, pp. 165-178.

Friedmann, J.H. and Stuetzle, W., (1981): Projection Pursuit Methods for Data Analysis. In: Modern data analysis, se Launer and Siegel.

Friedström, L., (1980): Lineære og log-lineære modeller for kvalitative avhengige variable. Rapporter fra Statistisk Sentralbyrå 80/26.

Frisch, R. og Haavelmo, T. (1971) (red. H.T. Amundsen): Utdrag av forelesninger i teoretisk statistikk. Universitetsforlaget/Sosialøkonomisk institutt.

Goldstein, H.E., (1981): Robust Inference in Contingency Tables, Stensil, Sosialøkonomisk institutt. (Ny versjon av memorandum av 27. februar 1980.)

Goodman, L.A., (ed. J. Magidsen) (1978): Analyzing Qualitative/ Categorial Data. Addison-Wesley Publishing Co.

Greenacre, M.J. (1981): Practical correspondance analysis. Ch. 7 in Vic Barnet (ed.): Interpreting multivariate data. Wiley.

Grizzle, J.E., Starmer, C.F. and Koch, G.G. (1969): Analysis of Categorical Data by Linear Models. Biometrics, Vol. 25, 489-504.

Haberman, S.J., (1973): The Residuals in Cross-classified Tables. Biometrics, Vol. 29, 205-220.

Haberman, S.J., (1978): Analysis of Qualitative Data, Vol. 1. Academic Press.

H.I.: Haldorsen, T., (1977): Testing i tabeller. Arbeidsnotater fra Statistisk Sentralbyrå. IO 77/41.

H.II.: Haldorsen, T., (1977): Om log-lineær analyse av flerveistabeller. Arbeidsnotat fra Statistisk Sentralbyr, IO 77/46.

Hellevik, O., (1971): Forskningsmetode i sosiologi og statsvitenskap. Universitetsforlaget.

Herigstad, H., (1981): Helgeturer 1978/79. Rapporter fra Statistisk Sentralbyrå 81/16.

Higgins, J.E. and Koch, G.G. (1977): Variable Selection and Generalized Chi-Square Analysis of Categorical Data Applied to a Large Cross-Sectional Occupational Health Survey. International Statistical Review, Vol. 45, No. 1, 51-62.

Hirschfeld, H.G. (1935): A Connection between Correlation and Contingency Proc. Camb. Phil. Soc., Vol. 31, pp. 520-524.

Hodges, J.L., and Lehmann, E.L., (1954): Testing the Approximate Validity of Statistical Hypotheses. J. Roy. Statist. Soc. B., Vol. 16, pp. 261-268.

Hodges, J.L., og Lehmann, E.L. (1970): Basic Concepts of Probability and Statistics. Holden-Day Inc.

Kass, G.V. (1980): An Exploratory Technique for Investigating Large Quantities of Categorical Data. Applied Statistics, Vol. 29, No. 2, pp. 119-127.

Klotz, J., (1980): A Modified Cochran-Friedman Test with Missing Observations and Ordered Categorical Data. Biometrics, Vol. 36, No. 4, pp. 665-670.

Lancaster, H.O. (1951): Complex Contingency Table, treated by the partition of χ^2. J. Roy. Statist. Soc. B., Vol. 13, pp. 242-249.

Lancaster, H.O. (1969): The Chi-Squared Distribution. Wiley.

Launer, R.L. and Siegel, A.F. ed. (1982): Modern Data Analysis. Academic Press.

Lebart, L., Morineau, A., Tabard, N. (1977): Techniques de la description statistique. Dunod.

Lewis, B.N. (1962): On the Analysis of Interaction in Multidimensional Contingency Tables. J. Roy. Statist. Soc. A., Vol.125, pp. 88-117.

Lewontin, R.C. and Felsenstein, J. (1965): The Robustness of Homogeneity Tests in 2 x N Tables. Biometrics, Vol. 21, pp. 19-33.

Lillestøl, J., (1982): Sannsynlighetsregning og statistikk med anvendelser. Bedriftsøkonomens Forlag A/S.

Laake, P. og Andersen, A.S. (1977): Stianalyse i sosiologisk forskning: Et eksempel fra levekårsundersøkelsen 1973. Arbeidsnotater fra Statistisk Sentralbyrå IO 77/7.

Manski, C.F. and Mc Fadden, D., eds. (1981): Structural Analysis of Discrete Data with Econometric Applications. MIT Press.

Metodegruppens arbeidsprogram for perioden 1981-85. Stensil 1980, Statistisk Sentralbyrå.

Mårdberg, B., (1974): LPA2 (A Clustering Program). Psyk. inst., Universitetet i Bergen, 5. årg. nr. 2.

Plackett, R.L., (1974): The Analysis of Categorical Data. Griffin, Statist. Mem. No. 35.

Schweder, T. and Spjøtvold, E., (1982): Plots of P-values to Evaluate Many Tests Simultaneously. Biometrika, Vol. 69, No. 3, pp. 493-502.

Sikkel, D., (1979): Analysis of Two-Way Tables by Correspondence Analysis. Stencil, Department of Statistical Methods, Centraal Bureau vor de Statiestik, Voorburg, Nederland.

Sonquist, J.A., Baker, E.L. and Morgan, J.N. (1973): Searching for Structure. Institute for Social Research, University of Michigan, Ann Arbor.

Statistisk årbok 1979. Statistisk Sentralbyrå.

Sverdrup, E. (1975): Multiple Comparisons by Binary and Multinary Observations. Artikler fra Statistisk Sentralbyrå, nr. 75.

Sverdrup, E. (1976): Significance Testing in Multiple Statistical Inference. Scandinavian Journal of Statistics, Vol. 3, pp. 73-78.

Sverdrup, E. (1977): The Logic of Statistical Inference: Significance Testing and Decision Theory. Bull. Int. Statist. Inst., Vol. 47, no. 1, pp. 573-606.

S.I.: Sverdrup, E. (1978): Statistical Analysis of Crossclassified Multinomial trials. Stensilmemo no. 6, 31.1.78, Statistisk Sentralbyrå.

S.II: Sverdrup, E. (1978): An Example of Analyses of Binomial Observations Based on a Regression Model. Stensilmemo no. 8, 12.4.78, Statistisk Sentralbyrå.

S.III.: Sverdrup, E. (1977): Exact Statistical Analysis of Multinomial Trials, Stensilmemo, no. 3, 27.9.77, Statistisk Sentralbyrå.

Swafford, M., (1980): Three Parametric Techniques for Contingency Table Analysis: A Nontechnical Commentary. Am. Sociol. Rev., Vol. 45, No. 4, pp. 664-669.

Thomsen, I., (1976): Bruk av superpopulasjonsmodeller ved innsamling og analyse av data fra utvalgsundersøkelser. Arbeidsnotat IO 76/28, Statistisk Sentralbyrå.

Thomsen, I., (1977): Prinsipper og metoder for Statistisk Sentralbyrås utvalgsundersøkelser. SØS. 33. Statistisk Sentralbyrå.

Tukey, J.W., (1977): Exploratory Data Analysis. Addison-Wesley.

Weiss, H. R. (1978): Approximative und exakte Tests zur Analyse mehr-dimensionaler Kontingenztafeln. Physica Verlag, Würzburg.

AA I: Aaberge, R., (1979): Eksakt analyse av 2 x 2 tabellar. Rapporter fra Statistisk Sentralbyrå, 79/20.

AA II: Aaberge, R., (1980): Eksakte metoder for analyse av to-vegstabellar. Rapporter fra Statistisk Sentralbyrå, 80/22.

T A B E L L E R O G D A T A P R O G R A M M E R

Biometrika Tables for Statisticians, Vol. I. Ed.: Pearson, E.S. and Hartley, H.O. (1962 eller senere), Cambridge University Press.

BMDP: Biomedical Computer Programs. Dixon, W.J. ed. (1981): BMDP Statistical Software. The regents of the University of California.

Bølviken, Erik: Multincanonical Analysis. Program i arbeid.

CLUSTAN: Wishart, D. (1978): CLUSTAN User's Manual. Inter-University/ Research Council Series Report, No. 47. Edinburgh University.

DDPP: Programpakke for samfunnsvitenskapelig databehandling. Jacobsen, P.H. (1981): Håndbok i DDPP. Universitetsforlaget.

ECTA: Programpakke for log-lineære modeller. Se f.eks. Alt, J. and Sanders, D. (1981) Program Library for Essex University.

GELAST: Andersen, E.B. og Weinberg (1983).

GENCAT: Landis, Stanish, Freeman and Koch (1976): A Computer Program for the Generalized Chi-Square Analysis of Categorical Data using Weighted Least Squares. Biostatistics Techn. Report no. 8, University of North Carolina.

LPA 2 : Mårdberg, B. (1974). A Clustering Program. Psyk. inst., Universitetet i Bergen, 5 årg. nr. 2.

SAS: USER'S GUIDE, 1982 EDITION. SAS Institute Inc., Box 8000 Cary, North Carolina.

SCSS: Nie, N.H. et al. (1980): SCSS: A User's Guide to the SCSS Conversational system. Mc Graw Hill.

SPSS: Nie, N.H. et al. (1983) A Complete Guide to the SPSS Language and Operations. Mc Graw Hill.

TROLL EXPERIMENTAL PROGRAMS (1975) National Bureau of Economic Research, Inc. (I Norges Banks datasenter) CCREMS ved MIT.

Appendiks A

Noen grunnbegreper i sannsynlighetsregningen

Den statistiske teori og metodene vi bruker i dette heftet, spesielt i kapitlene 2-12, bygger på sannsynlighetsregningen. Vi skal gi en kortfattet oversikt over noen nødvendige begreper i denne. For en mer utførlig innføring i sannsynlighetsregning og statistikk på et elementært plan, viser vi til lærebøker som A I eller Lillestøl (1978).

Vi skal definere sannsynlighetsbegrepet rent matematisk, men med referanse til situasjoner hvor vi får bruk for det.

Vi tenker oss en situasjon der en observasjon kan gi ett og bare ett av k ulike resultater, la oss kalle dem C_1, C_2, ..., C_j, ..., C_k. Vi kan f.eks. la C-ene være yrkesgrupper som er avgrenset slik at ingen kan tilhøre mer enn én gruppe og dessuten slik at en observert enhet er nødt til å tilhøre en av gruppene. I dette eksemplet svarer C-ene altså til k kategorier for én variabel, men C-ene kan også stå for kombinasjoner av kategorier for hver av to eller flere variable, som vi skal se senere.

Til hvert resultat tenker vi oss knyttet et tall, $P(C_1)$, $P(C_2)$, ..., $P(C_j)$, ..., $P(C_k)$. Hvert av tallene er større eller lik null og mindre eller lik én, dvs.

$$(A1) \qquad 0 \leq P(C_j) \leq 1 \qquad \text{for } j = 1, 2, \ldots k.$$

Videre skal tallene være slik at om vi ønsker å se to av resultatene, f.eks. C_i og C_j, under ett, som ett resultat (vi slår sammen to yrkesgrupper), og sier at vi har resultatet "enten C_i eller C_j", så kan vi sette

(A2) $\quad P$ (enten C_i eller C_j) $= P(C_i) + P(C_j)$.

Det tilsvarende skal gjelde om vi vil se på fler enn to resultater under ett. Dette er addisjonssetningen i sannsynlighetsregningen for resultater som utelukker hverandre. Vi skal også ha

(A3) $\quad P(C_1) + P(C_2) + \cdots + P(C_k) = 1.$

Vi kaller $P(C_j)$ *sannsynligheten for* C_j. Kravet (3) sier at sannsynligheten for i det hele tatt å få et resultat, skal være lik én. Vi forlanger at sannsynligheten for et *sikkert* resultat, la oss kalle det S, må settes lik én også i andre situasjoner, altså

(A4) $\quad P(S) = 1.$

Vi ser at hvis k = 2, slik at

$$P(C_1) + P(C_2) = 1,$$

så har vi

$$P(C_2) = 1 - P(C_1).$$

Dette kan også uttrykkes som

(A5) $\quad P$ (ikke å få C_1) $= 1 - P(C_1)$.

Hvis vi har et tilfelle der $P(C_1) = 1$, dvs. at C_1 er et sikkert resultat,

så må følgelig

$$P\ (\text{ikke} - C_1) = 1 - 1 = 0.$$

Her er altså (ikke - C_1) et umulig resultat. Vi kan si at sannsynligheten for et umulig resultat, U, er lik null,

(A6) $\qquad P(U) = 0.$

Vi har definert sannsynlighetene, $P(C_j)$, rent matematisk, men det er ikke vanskelig å se analogier med bruken av ordet sannsynlighet i dagligtale.

Anta at vi har en situasjon med gjentatte observasjoner, der sannsynligheten for å få resultatet C er lik P(C) ved hver observasjon, uavhengig av de andre resultatene. Da kan vi vise, i teoretisk forstand, at grensen for den relative hyppigheten av C vil være P(C) når antall observasjoner vokser over alle grenser (Bernoullis teorem). Sammen med det foregående får dette oss til *å tolke verdier P(C) nær 1 som det vi i daglig tale kaller stor sannsynlighet for resultatet C og verdier av P(C) nær 0 som liten sannsynlighet for C.* "Økende sannsynlighet" svarer også til økende P-verdier.

Det finnes flere forskjellige "skoler" som definerer eller tolker sannsynlighetsbegrepet erkjennelsesteoretisk på ulik vis. Vi skal ikke diskutere denne siden av saken, men vi kan understreke at *regnereglene* vi bruker, er oppfylt for alle "skolene".

I et *konkret problem har vi lov til å velge eller tenke oss P-verdier akkurat som vi har lyst, bare vi passer på at kravene (A1) - (A6) er oppfylt.* Gjør vi det, kan vi regne med P-ene våre etter regnereglene ovenfor, og etter dem vi skal utlede etterhvert.

Hvis vi mener at to kjennetegn, C_1 og C_2, er *like sannsynlige*, må vi

sette $P(C_1) = P(C_2)$. I et kast med et kronestykke må det "at begge sider har samme sjanse for å komme opp ved et kast" kunne skrives

$$P\ (\text{mynt}) = P\ (\text{krone}).$$

Setter vi den felles verdi lik p og bruker (A3), finner vi

$$1 = P\ (\text{mynt}) + P\ (\text{krone}) = p + p = 2p,$$

altså må vi ha

$$p = \frac{1}{2}\ .$$

Dermed er sannsynlighetsmodellen fullstendig bestemt i dette tilfellet.

Tilsvarende ser vi at hvis vi har ialt m mulige resultater som utelukker hverandre og som alle skal ha samme sannsynlighet p, så får vi

$$p = \frac{1}{m}\ .$$

Et resultat, C som varer til at vi har ett av g forskjellige av de alternative resultatene, får sannsynligheten

$$P(C) = \underbrace{\frac{1}{m} + \frac{1}{m} + \ldots + \frac{1}{m}}_{\text{(g addender)}} = \frac{g}{m}\ .$$

Det kjennetegner problemene i dette notatet at vi observerer to eller flere kjennetegn (kategorier for to eller flere variable) for hver observasjonsenhet. Hvert resultat er altså en kombinasjon av en kategori for

hver av de variable i problemet, og vi kan tenke oss de mulige resultatene av en observasjon satt opp i en tabell i to eller flere dimensjoner.

Vi kan innføre symbolet P(A,B) for sannsynligheten for kombinasjonen av kategori A for den ene variable og kategori B for den annen. Vi kan si at *P(A,B) er sannsynligheten for både A og B.* Tilsvarende kan vi bruke *P(A,B,C,D) som symbol for sannsynligheten for både A og B og C og D ved 4 variable, osv.*

De mulige resultatene av to kast med et kronestykke (eller ett kast med to kronestykker) er

mynt i første kast, mynt i annet kast

krone i første kast, mynt i annet kast

mynt i første kast, krone i annet kast

krone i første kast, krone i annet kast

Bruker vi symbolet 0 for mynt (dvs. krone null ganger) og symbolet 1 for krone (krone én gang) så kan vi sette opp en tabell over de fire resultatene og en tabell over de tilsvarende sannsynlighetene slik:

Første kast	Annet kast: mynt 0	krone 1
Mynt, 0	0,0	0,1
Krone, 1	1,0	1,1

Første kast	Annet kast: 0	1
0	P(0,0)	P(0,1)
1	P(1,0)	P(1,1)

Vi kan innføre de litt enklere betegnelsene $p_{00} = P(0,0)$, $p_{01} = P(0,1)$, $p_{10} = P(1,0)$ og $p_{11} = P(1,1)$. Vi må ha $p_{00} + p_{01} + p_{10} + p_{11} = 1$ ifølge setning (A3). Ved å bruke addisjonssetningen (A2) finner vi *den marginale sannsynligheten* for mynt i første kast, (uten hensyn til hva vi får i annet kast):

$$p_{0+} = p_{00} + p_{01}$$

Tilsvarende for krone:

$$p_{1+} = p_{10} + p_{11}.$$

(Vi setter et + på fotindeksplassen for den variable vi summerer over.) De marginale sannsynlighetene for henholdsvis mynt og krone i annet kast (uten hensyn til resultatet i første kast) er tilsvarende:

$$p_{+0} = p_{00} + p_{10} \quad \text{og} \quad p_{+1} = p_{01} + p_{11}.$$

De marginale sannsynlighetene oppfyller kravet i (A3), idet vi har

$$p_{0+} + p_{1+} = p_{00} + p_{01} + p_{10} + p_{11} = 1, \quad \text{og} \quad p_{+0} + p_{+1} = 1.$$

I tabellform kan vi sette opp de simultane og de marginale sannsynlighetene som i tabell A.a.

Tabell A.a.

Sannsynlighetene for mynt og krone (jfr. tabell 2.1.b.) i to kast.

Første kast	Annet kast 0	Annet kast 1	Marginal sanns. for 1. kast
0	p_{00}	p_{01}	p_{0+}
1	p_{10}	p_{11}	p_{1+}
Marginal sanns. for 2. kast	p_{+0}	p_{+1}	(total) 1

Vi kan spørre: hva er sannsynligheten for å få krone *minst* én gang i to kast, dvs. ett av resultatene (0,1), (1,0), (1,1)? Den er ifølge (A2) lik $p_{01} + p_{10} + p_{11}$ som også er lik $1 - p_{00}$. Men vi kan også skrive den som

$p_{1+} + p_{+1} - p_{11}$, (Idet $p_{10} + p_{11} + p_{01} + p_{11} - p_{11} = p_{01} + p_{10} + p_{11}$), dvs. summen av sannsynlighetene for krone i 1. kast og i 2. kast, minus sannsynligheten for å få krone i begge (fordi denne er kommet med to ganger hvis vi bare tar $p_{1+} + p_{+1}$, jfr. parentesen ovenfor). Dette er et spesialtilfelle av den mer generelle *addisjonssetningen for to resultater, A og B, som ikke utelukker hverandre:*

(A7) $\quad P$ (enten A eller B) = P(A) + P(B) - P(A,B).

I avsnitt 2.2 og senere har vi eksempler på toveistabeller med mer enn to mulige kategorier. Sett at vi har r kategorier (som utelukker hverandre) for variabel A, f.eks. yrke, og s kategorier for variabel B, f.eks. utdannelse, da vil en tabell over de tilhørende p-ene svare til tabell A.b. Vi bruker her symbolet p_{ij} for sannsynligheten $P(A_i, B_j)$.

Tabell A.b.

Sannsynlighetene i en rxs-tabell.

A-kategori	B-kategori B_1	B_2	$B_3 \cdots B_j \cdots B_s$	Marginal for A (linjesum)
A_1	p_{11}	p_{12}	$p_{13} \cdots p_{1j} \cdots p_{1s}$	p_{1+}
A_2	p_{21}	p_{22}	$p_{23} \cdots p_{2j} \cdots p_{2s}$	p_{2+}
.	.	.	. . .	.
A_i	p_{i1}	p_{i2}	$p_{i3} \cdots p_{ij} \cdots p_{is}$	p_{i+}
A_r	p_{r1}	p_{r2}	$p_{r3} \cdots p_{rj} \cdots p_{rs}$	p_{r+}
Marginal for B (kolonnesum)	p_{+1}	p_{+2}	$p_{+3} \cdots p_{+j} \cdots p_{+s}$	1

Å undersøke om det er *samvariasjon* mellom to variable, må være det samme som å finne ut hva slags "mønster" det kan være i p-verdiene. Med et

"skikkelig" kronestykke i myntkasteksemplet er det lett å tenke seg at alle de fire kombinasjonene i tabell A.a er like sannsynlige, dvs. at

$$p_{00} = p_{01} = p_{10} = p_{11} = p = 0{,}25,$$

siden summen av de fire p-ene skal være én. Dessuten blir $p_{0+} = p_{1+} = 0{,}5$ og $p_{+0} = p_{+1} = 0{,}5$. I yrkes-utdannelseseksemplet vil det neppe være slik. Med r = 4 og s = 3, kan vi f.eks. ha et mønster som i tabell A.c.

Tabell A.c.

Konstruert eksempel på sannsynligheter for to variable med samvariasjon.

	B-kategori			
A-kategori	B_1	B_2	B_3	Marginal (sum)
A_1	0,40	0,10	0	0,5
A_2	0,15	0,075	0,025	0,25
A_3	0,04	0,095	0,015	0,15
A_4	0,01	0,03	0,06	0,1
Marginal (sum)	0,6	0,3	0,1	1

Betinget sannsynlighet

I avsnitt 2.2 regnet vi ut *betingede* hyppigheter i en hyppighetstabell. Det tilsvarende begrep for sannsynligheter er *betinget sannsynlighet* for f.eks. A_i, gitt B_j, som vi skriver:

$$(A8.1) \qquad P(A_i|B_j) = \frac{P(A_i, B_j)}{P(B_j)} = \frac{p_{ij}}{p_{+j}}, \qquad \text{når } P(B_j) > 0.$$

Tilsvarende har vi

$$(A8.2) \qquad P(B_j|A_i) = \frac{p_{ij}}{p_{i+}} .$$

I mynt-eksemplet gir dette

$$P\ (\text{mynt i 2. kast}\,|\,\text{mynt i 1. kast}) = \frac{p_{00}}{p_{0+}} = \frac{0,25}{0,5} = 0,5$$

og

$$P\ (\text{mynt i 2. kast}\,|\,\text{krone i 1. kast}) = \frac{p_{10}}{p_{1+}} = \frac{0,25}{0,5} = 0,5$$

De betingede sannsynlighetene er her like store, og lik den marginale sannsynligheten. Dette er en typisk situasjon der de variable er *stokastisk uavhengige* av hverandre, dvs. at sannsynligheten for en kategori av den ene variable ikke blir influert av hvilken kategori den annen variable har.

Stokastisk uavhengighet mellom A_i og B_j kan vi definere ved at

$$(A9.1) \qquad P(A_i|B_j) = P(A_i).$$

Ifølge (A8) kan vi alltid skrive $P(A_i,B_j) = P(B_j)\ P(A_i|B_j)$.
Det følger at ved stokastisk uavhengighet, dvs. når (A9.1) er oppfylt, så blir

$$(A9.2) \qquad P(A_i,B_j) = P(B_j)\ P(A_i) = P(A_i)\ P(B_j).$$

Av dette og av (A8.2) følger det at vi også har

$$(A9.3) \qquad P(B_j|A_i) = \frac{P(A_i)\,P(B_j)}{P(A_i)} = P(B_j)$$

i dette tilfelle.

Hver av de tre likhetene i (A.9) følger altså av hver av de andre to, og vi kan definere stokastisk uavhengighet mellom A_i og B_j ved hvilken vi vil av dem.

Vi ser at i mynt-eksemplet er resultatet i et kast stokastisk uavhengig av resultatet i det annet. Men i yrkesutdannelseseksemplet er det ikke slik. Vi har f.eks.

$$P(A_1|B_1) = \frac{0,4}{0,6} = \frac{2}{3}$$

$$P(A_1|B_2) = \frac{0,1}{0,3} = \frac{1}{3}$$

$$P(A_1|B_3) = \frac{0}{0,1} = 0$$

mens $P(A_1) = 0,5$.

Det er *stokastisk avhengighet* mellom A_1 og B_j-kategoriene. Vi ser at dette også gjelder for A_3 og A_4, men ikke A_2. Vi skal komme tilbake til dette nedenfor, i avsnittet om simultane sannsynlighetsfordelinger.

Sannsynlighetsfordelinger, stokastiske variable

Anta at vi har en variabel x (med intervall skala) som kan anta de ulike verdiene $x_1, x_2, \ldots, x_i, \ldots, x_r$, og ingen andre verdier. Vi knytter sannsynligheter, $P(x_1), P(x_2), \ldots, P(x_r)$ til disse verdiene, etter reglene (A1 - A3). Vi kaller da x for en *stokastisk* (tilfeldig) variabel, og samlingen

av P-verdier for *sannsynlighetsfordelingen* for den variable.

Vi bruker ofte et vanlig funksjonssymbol f(x) for sannsynligheten i dette tilfelle, slik at

(A10) $f(x_i) = P(x_i)$ for $i = 1,2, \ldots, r$

$f(x) = 0$ for alle andre x-verdier.

Vi kaller f(x) *sannsynlighetsfunksjonen* for den stokastiske variable, x.

(A11) Hvis vi lar x være antall ganger vi får krone i to kast, så vil x med forutsetningene om at $p_{00} = p_{01} = p_{10} = p_{11} = 0,25$, ha sannsynlighetsfunksjonen

$f(0) = 0,25$

$f(1) = 0,50$

$f(2) = 0,25$

$f(x) = 0$ for alle andre x-verdier.

Med to variable, x og y, kan vi på tilsvarende måte angi den *simultane sannsynligheten* for variabelkombinasjonen (x_i, y_j) ved en funksjon i de to variable, f(x,y). Vi har $f(x_i,y_j) = P(x_i,y_j)$ for $i = 1,2, \ldots, r$ og $j = 1,2, \ldots, s$. $f(x,y) = 0$ ellers.

Tilsvarende for flere variable, x, y, ..., z, med henholdsvis r, s, ..., t mulige verdier.

$$f(x_i,y_j, \ldots, z_q) = P(x_i,y_j, \ldots, z_q) \text{ for } \begin{array}{l} i = 1,2, \ldots, r \\ j = 1,2, \ldots, s \\ \hline q = 1,2, \ldots, t, \end{array}$$

$f(x,y, \ldots, z) = 0$ ellers.

Når vi må angi de ulike resultatene for de variable ved *kategorier*, jfr.

eksemplet knyttet til tabell A.b og A.c, så vil samlingen av alle sannsynlighetene $p_{ij} = P(A_i,B_j)$ stadig angi den simultane sannsynlighetsfordelingen for de variable, men det har vanligvis liten hensikt å innføre andre betegnelser.

Marginal fordeling og betinget fordeling

Når vi har en simultan sannsynlighetsfordeling i to eller flere variable, vil vi ofte også være interessert i fordelingen for de enkelte variable hver for seg. Gitt $f(x_i,y_j)$ for alle kombinasjoner av i og j, så kan vi finne sannsynligheten for x_i, $f_x(x_i)$, ved å summere $f(x_i,y_j)$ for alle y_j-verdiene, jfr. summen til høyre i tabell A.b og A.c. Vi har jo at x_i kan forekomme sammen med y_1, med y_2, osv. til y_s, og da er ifølge (A2)

$$(A12.1) \quad f_x(x_i) = P(x_i) = f(x_i,y_1) + f(x_i,y_2) + \ldots + f(x_i,y_s).$$

Dette gjelder for i = 1,2, ..., r. Vi finner at

$$f_x(x_1) + f_x(x_2) + \ldots + f_x(x_2) = 1,$$

og sier at $f_x(x_i)$-verdiene angir den *marginale sannsynlighetsfunksjonen for x*.

For kategoriske variable har vi tilsvarende

$$(A12.2) \quad P(A_i) = P(A_i,B_1) + P(A_i,B_2) + \ldots + P(A_i,B_s) \text{ for } i = 1,2, \ldots, r,$$

slik at $P(A_1)$, $P(A_2)$, ..., $P(A_r)$, med sum én, angir den marginale sannsynlighetsfordelingen for A.

I (A8) definerte vi den betingede sannsynligheten $P(A_i|B_j)$ for A_i gitt B_j. Tilsvarende definerer vi *den betingede sannsynlighetsfordelingen for A gitt* B_j *ved*

$$(A13.1) \quad P(A_i|B_j) = \frac{P(A_i,B_j)}{P(B_j)} \qquad \text{for } i = 1,2, \ldots, r.$$

Det finnes i alt s slike fordelinger, én for hver B_j. Likedan har vi den betingede fordelingen for B gitt A_i som

$$(A13.2) \quad P(B_j|A_i) = \frac{P(A_i,B_j)}{P(A_i)} \qquad \text{for } j = 1,2, \ldots, s.$$

Her er det r fordelinger, én for hver A_i. For intervall skala variable har vi tilsvarende de betingede sannsynlighetsfunksjonene

$$(A13.3) \quad f_x(x_i|y_j) = \frac{f(x_i,y_j)}{f_y(y_j)} \qquad \text{for } i = 1,2, \ldots, r.$$

og

$$(A13.4) \quad f_y(y_j|x_i) = \frac{f(x_i,y_j)}{f_x(x_i)} \qquad \text{for } j = 1,2, \ldots, s.$$

Med tallene i tabell A.c finner vi de betingede og marginale fordelingene i tabell A.d og A.e.

Tabell A.d.

Betingede sannsynlighetsfordelinger for A, gitt B-kategori, samt marginal fordeling.

A-kategori	B-kategori B_1	B_2	B_3	Marginal fordeling
A_1	0,67	0,33	0	0,50
A_2	0,25	0,25	0,25	0,25
A_3	0,07	0,32	0,15	0,15
A_4	0,02	0,10	0,60	0,10
Sum	1,01	1,00	1,00	1,00

Tabell A.e.

Betingede sannsynlighetsfordelinger for B, gitt A-kategori, samt marginal fordeling.

A-kategori	B-kategori B_1	B_2	B_3	Sum
A_1	0,80	0,20	0	1,00
A_2	0,60	0,30	0,10	1,00
A_3	0,27	0,63	0,10	1,00
A_4	0,10	0,30	0,60	1,00
Marginal fordeling	0,60	0,30	0,10	1,00

I (A9) definerte vi stokastisk uavhengighet mellom kategoriene A_i og B_j. Vi sier at vi har *stokastisk uavhengighet mellom de variable* x og y (eller A og B) når (A9) gjelder *for alle kombinasjoner* $x_i(A_i)$ av x og $y_j(B_j)$ av y.

For intervall skala variable skal altså sannsynlighetsfunksjonen f(x,y) være slik at

$$(A14.1) \quad f(x_i|y_j) = f_x(x_i) \qquad \text{for } i = 1,2, \ldots, r \text{ og for } j = 1,2, \ldots, s.$$

Dette innebærer

$$(A14.2) \quad f(x_i,y_j) = f_x(x_i)f_y(y_j) \qquad \text{for } i = 1,2, \ldots, r \text{ for } j = 1,2, \ldots, s$$

og

$$(A14.3) \quad f(y_j|x_i) = f_y(y_j) \qquad \text{for } j = 1,2, \ldots, s \text{ og } i = 1,2, \ldots, r.$$

Vi ser at de variable i tabellen A.c-A.e *ikke* er stokastisk uavhengige. Det er ikke nok at den *ene* betingede fordelingen, for B-ene gitt A_2, er lik den marginale.

I myntkast-eksemplet har vi derimot $p_{00} = p_{0+}p_{+0} = 0,5 \cdot 0,5 = 0,25$ og tilsvarende for p_{10}, p_{01} og p_{11}, dvs. stokastisk uavhengighet mellom resultatene i første og annet kast.

For krysstabeller med fler enn to variable kan vi definere marginale og betingede fordelinger analogt med ovenstående. Det blir bare flere variable å summere over, eller flere variabelverdier å kombinere eller betinge med hensyn på. Se f.eks. avsnitt 4.2.3 og 5.2.1.

Forventning, varians, kovarians

Når vi har observasjoner, $x_1, x_2, \ldots, x_n$ av en stokastisk variabel, x, regner vi ofte ut noen størrelser som vi synes gir oss en viss oversikt over observasjonene. Spesielt kan dette være *det aritmetiske gjennomsnittet:*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{1}{n} (x_1 + x_2 + \ldots + x_n).$$

Hvis x antar k ulike verdier, $x_1, x_2, \ldots, x_k$, med absolutte hyppigheter $n_1, n_2, \ldots, n_k$, så kan vi regne ut gjennomsnittet ved

$$\bar{x} = \frac{1}{n} \sum_{j=1}^{k} x_j n_j = \sum_{j=1}^{k} x_j h_j, \quad \text{der } h_j = \frac{n_j}{n}, \text{ er relativ hyppighet av } x_j.$$

Videre har vi *den empiriske variansen:*

$$s_x^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n} \sum_{j=1}^{k} n_j (x_j - \bar{x})^2 = \sum_{j=1}^{k} (x_j - \bar{x})^2 h_j$$

og *standardavviket,*

$$s_x = \sqrt{s_x^2}.$$

Hvis vi har observasjonspar (x_1, y_1), (x_2, y_2) ... (x_n, y_n) av to variable, kan vi regne ut den *empiriske kovariansen* mellom x og y som

$$m_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}) (y_i - \bar{y}).$$

Den *empiriske korrelasjonskoeffisienten* er

$$r_{xy} = \frac{m_{xy}}{s_x s_y}.$$

Alle disse størrelsene har sine teoretiske motstykker.

(A15) *Forventningen*, Ex, av en diskret stokastisk variabel x med verdier $x_1, x_2, \ldots, x_k$ og sannsynlighetsfunksjon f(x), er

$$Ex = \sum_{j=1}^{k} x_j f(x_j).$$

(A16) Den teoretiske variansen for x er tilsvarende

$$\sigma_x^2 = \sum_{j=1}^{k} (x_j - Ex)^2 f(x_j).$$

Det teoretiske standardavviket, σ_x, er kvadratroten av variansen.

(A17) Den teoretiske kovariansen mellom x og y er, med de betegnelsene vi har brukt ovenfor:

$$\sigma_{xy} = \sum_{i=1}^{r} \sum_{j=1}^{s} (x_i - Ex)\,(y_j - Ey)\, f(x_i, y_j).$$

(A18) Den teoretiske korrelasjonskoeffisienten ρ eller ρ_{xy}, er

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

App. B Noen spesielle sannsynlighetsfunksjoner for diskrete variable

De fleste statistiske metodene som er utviklet for analyse av krysstabeller i to og flere variable, bygger på visse forenklende forutsetninger om observasjonsmaterialet, jfr. avsnitt 2.2.1. Vi kan f.eks. postulere at for alle de n telleenhetene (observasjonsenhetene), så fremkommer hvert observasjonsresultat, (A_i,B_j), ved at *den samme* sannsynlighetsfordelingen, $P(A_i,B_j)$, jfr. tabell A.b, gjelder. Sannsynligheten for å få resultatet (A_i,B_j) for den m-te telleenheten skal være $P(A_i,B_j)$ uansett hvilke resultater som fremkommer for alle de andre telleenhetene, d.v.s. at *resultatene fra telleenhet til telleenhet* er stokastisk uavhengige av hverandre. Antall telleenheter, n_{ij}, som får resultatet (A_i,B_j), blir en verdi av en stokastisk variabel som kan anta verdiene $0,1,2,\ldots,n$. I avsnitt 2.2.7 innførte vi navnet *tellevariabel* for en slik variabel i en hyppighetstabell, for å skille den fra de egentlige variable i problemet (som har kategorier som $A_1,A_2,\ldots,A_r$, resp. $B_1,B_2,\ldots,B_s$). Sannsynlighetsfordelingen for de tellevariable kan da utledes ut fra den "underliggende fordelingen" når forutsetningene er gitt. Vi skal referere noen av de vanligste sannsynlighetsfunksjonene som er i bruk. For disse kan sannsynlighetsfunksjonen angis ved et matematisk uttrykk, der det ved siden av verdiene x,y osv. av de variable, inngår visse *parametre*, d.v.s. faste tallstørrelser som karakteriserer fordelingen.

Vi trenger symboler for noen ofte forekommende matematiske uttrykk:

(B1) x-fakultet

Vi kaller produktet av alle de naturlige tallene til og med x for *x-fakultet*, som skrives: x! Altså

$$x!=x(x-1)(x-2)\ldots3.2.1.$$

Vi har

$$1!=1,\quad 2!=2.1=2,\quad 3!=3.2.1=6,\quad 4!=24 \text{ osv.}$$

Dessuten definerer vi av bekvemhetshensyn

$$0!=1$$

for å kunne skrive visse formler på en grei måte.

(B2) Binomialkoeffisienten.

Vi bruker symbolet $\binom{n}{x}$ for *binomialkoeffisienten*, dvs. for uttrykket

$$\binom{n}{x}=\frac{n!}{x!(n-x)!}\ ,\ \text{for } x=0,1,2,\ldots,n.$$

Her er n et helt, positivt tall.

I følge (B1) har vi

$\binom{n}{0} = \frac{n!}{0!n!} = 1$, $\binom{n}{1} = \frac{n!}{1!(n-1)!} = n$, $\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2}$,

osv., $\binom{n}{n} = \frac{n!}{n!0!} = 1$.

(B3) Binomisk fordeling

Anta at vi foretar n observasjoner. Ved hver observasjon kan resultatet bli C eller ikke-C. Sannsynligheten for C er p ved hver observasjon. C kan da forekomme x ganger ved de n observasjonene, og sannsynlighetsfunksjonen for x er den binomiske:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x=0,1,2,\ldots,n.$$

Her er altså n og p de to parametrene i fordelingen.

Forventningen av x er $Ex=np$. Det teoretiske *standardavviket* er $\sigma_x = np(1-p)$.

I myntkasteksemplet i appendiks A har vi $n=2$, $p=1-p=0,5$ slik at

$$f(x) = \binom{2}{x}(0,5)^x (0,5)^{2-x} = \binom{2}{x} 0,25,$$

d.v.s. nettopp tallene vi fant i eks. (A11).

Hvis vi i en toveis *krysstabell* ser på hyppigheten n_{ij} av kombinasjonen (A_i, B_j) med sannsynligheten p_{ij}, så er n_{ij} binomisk fordelt. Vi har

$$f(n_{ij}) = \binom{n}{n_{ij}} p_{ij}^{n_{ij}} (1-p_{ij})^{n-n_{ij}} \text{ for } n_{ij}=0,1,2,\ldots,n.$$

Dette er altså den marginale sansynlighetsfunksjonen for en av de telle-variable i krysstabellen.

(B4) Trinomisk fordeling

Vi skal foreta n observasjoner. Ved hver observasjon kan vi få resultatet C_1 med sannsynlighet p_1 eller resultatet C_2 med sannsynlighet p_2 eller resultatet "hverken C_1 eller C_2" med sannsynlighet $1-p_1-p_2$. De tre resultatene utelukker hverandre og det er uavhengighet fra observasjon til observasjon.

Resultatet C_1 får vi x_1 ganger,
resultatet C_2 får vi x_2 ganger, og
resultatet "hverken C_1 eller C_2" får vi $n-x_1-x_2$ ganger i de n observasjonene.

Da er den simultane sannsynlighetsfunksjonen for x_1 og x_2 gitt ved

$$f(x_1,x_2)= \frac{n!}{x_1!x_2!(n-x_1-x_2)!} p_1^{x_1} p_2^{x_2} (1-p_1-p_2)^{n-x_1-x_2},$$

for $x_1=0,1,2,\ldots,n$. $x_2=0,1,2,\ldots,n$ og slik at $x_1+x_2 \leq n$.

Vi har i den trinomiske fordelingen:

$$Ex_1=np_1, \quad Ex_2=np_2,$$

$$\sigma_{x_1}=\sqrt{np_1(1-p_1)}, \quad \sigma_{x_2}=\sqrt{np_2(1-p_2)}$$

$$\text{kovar } (x_1,x_2) = - np_1p_2.$$

Vi har mest bruk for en generalisering av denne fordelingen til flere alternativ, nemlig den multinomiske situasjon.

(B5) Multinomisk fordeling

Situasjonen er den samme som i (B4) men vi har ved hver observasjon k mulige, ulike resultater som utelukker hverandre. Resultatene er

C_1 med sannsynlighet p_1 og forekomst x_1 ganger

C_2 med sannsynlighet p_2 og forekomst x_2 ganger

C_{k-1} med sannsynlighet p_{k-1} og forekomst x_{k-1} ganger

C_k med sannsynlighet p_k og forekomst x_k ganger.

Vi har $p_1+p_2+\ldots+p_k=1$, slik at $p_k=1-p_1-p_2-\ldots-p_{k-1}$.

Videre er $x_1+x_2+\ldots+x_k=n$, slik at $x_k=n-x_1-x_2-\ldots-x_{k-1}$.

Den simultane sannsynlighetsfunksjonen for x_1, x_2, ..., x_{k-1} er da

$$(B5.1) \qquad f(x_1,x_2,\ldots,x_{k-1}) = \frac{n!}{x_1!x_2!\ \ldots\ x_{k-1}!\ x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k},$$

der hver x_j kan ha verdiene 0,1, ... n, men slik at summen er n.

Den marginale fordelingen for hver enkelt x_j er binomisk, med $Ex_j = np_j$ og $\sigma_j = \sqrt{np_j(1-p_j)}$.

Hvis vi har en krysstabell med kategorier A_1, A_2, ..., A_r for den ene variable og B_1, B_2, ..., B_s for den annen, så vil de rs resultatene (A_i,B_j) for i = 1,2, ..., r og j = 1,2, ..., s utelukke hverandre. Hvis nå (A_i,B_j) har sannsynlighet p_{ij} ved hver observasjon, og forekommer n_{ij} ganger blant de n observasjonene, så har vi en multinomisk situasjon, og den simultane sannsynlighetsfunksjonen for de (rs-1) tellevariablene n_{11}, n_{12} osv. blir: (En av tellevariablene, f.eks. n_{rs}, er gitt som n minus summen av de øvrige.)

$$(B5.2) \qquad f(n_{11},n_{12},\ldots,n_{ij},\ldots n_{r,s-1}) = \frac{n!}{n_{11}!\ n_{12}!\ \ldots n_{ij}!\ \ldots n_{rs}!} p_{11}^{n_{11}} p_{12}^{n_{12}} \ldots p_{rs}^{n_{rs}}.$$

Dette er sannsynlighetsfunksjonen for en mettet modell som vi refererer til i avsnitt 2.3.1 om sannsynlighetsmaksimeringsmetoden. For krysstabeller i 3 eller flere dimensjoner får vi tilsvarende.

Hvis vi har restriksjoner om uavhengighet mellom A-variabelen og B-variabelen, jfr. 2.3.1, slik at $p_{ij} = p_{i+}p_{+j}$ for alle kombinasjoner (A_i,B_j), så må vi sette dette inn i uttrykket for f. Når vi trekker sammen for alle p_{i+}, resp. p_{+j} med samme fotindeks, blir resultatet

$$(B5.3) \quad f \text{ uavh. } (n_{11},n_{12},\ldots,n_{r,s-1}) = \frac{n!}{n_{11}!\ n_{12}!\ \ldots n_{rs}!} p_{1+}^{n_{1+}} p_{2+}^{n_{2+}} \ldots p_{r+}^{n_{r+}} \ldots p_{+s}^{n_{+s}}$$

Ved å maksimere denne m.h.p. p-ene for gitte n-verdier, får vi estimater for

p_{i+} og p_{+j} som angitt i avsnitt 2.3.1.

(B6) Produktmultinomisk fordeling

I tabeller som ikke er rene krysstabeller, men kan være rent komparative, eller blanding av komparativ og krysstabell, vil vi ofte postulere én multinomisk fordeling for hver komparativ tabell, og uavhengighet mellom disse innbyrdes. Da vil den simultane sannsynlighetsfunksjonen for *alle* de tellevariable i hele tabellen bli et produkt av funksjonene fra de enkelte komparative tabeller.

Ser vi f.eks. på en 2 x 2 komparativ tabell, som i avsnitt 3.1, der n_{11} er binomisk fordelt, likeså n_{12}, og vi postulerer stokastisk uavhengighet mellom dem, så får vi ifølge Appendiks A den simultane sannsynlighetsfunksjonen

$$f(n_{11},n_{12}) = \binom{n_{+1}}{n_{11}} p_{11}^{n_{11}} p_{21}^{n_{21}} \binom{n_{+2}}{n_{12}} p_{12}^{n_{12}} p_{22}^{n_{22}}$$

for de mulige kombinasjonene av n_{11} og n_{12}.

Vi har her satt inn p_{21} for $(1-p_{11})$, n_{21} for $(n_{+1}-n_{11})$ osv., slik at det binomiske uttrykket i (B3) blir skrevet analogt med det multinomiske (B5.2).

(B7) Poisson-fordeling

Iblant har vi en situasjon som minner om den binomiske, idet vi undersøker hvor mange ganger et resultat C forekommer, men vi kan ikke sette noen begrensning n på antall enheter som i prinsippet inngår i undersøkelsen. I en trafikkontroll finner vi f.eks. x biler som har mangler ved lyset, men det noteres ikke hvor mange biler som passerer i løpet av kontrollperioden (en vet

bare at det er *mange*!). Da kan det hende at en kan bruke sannsynlighetsfunksjonen i en Poisson-fordeling:

$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!} \quad \text{for } x = 0,1,2, \ldots \text{ (ubegrenset).}$$

Her er det positive tallet λ en parameter i fordelingen. Disse sannsynlighetene er svært lik dem vi får i en binomisk fordeling med liten p og stor n, der $np = \lambda$. I Poisson-fordelingen er $Ex = \lambda$ og $\sigma_x = \sqrt{\lambda}$.

For en r x s krysstabell der antall observasjoner, n, *ikke* er gitt på forhånd og vi kan postulere at hver kombinasjon (A_i,B_j) forekommer uavhengig av de øvrige, og ifølge en Poisson-fordeling med parameter λ_{ij}, kan vi skrive den simultane sannsynlighetsfunksjonen for de rs tellevariable som produktet

$$f(n_{11},n_{12},\ldots,n_{ij},\ldots n_{rs}) = e^{-\lambda_{11}}\frac{\lambda_{11}^{n_{11}}}{n_{11}!}e^{-\lambda_{12}}\frac{\lambda_{12}^{n_{12}}}{n_{12}!}\ldots e^{-\lambda_{ij}}\frac{\lambda_{ij}^{n_{ij}}}{n_{ij}!}\ldots e^{-\lambda_{rs}}\frac{\lambda_{rs}^{n_{rs}}}{n_{rs}!}$$

Nå viser det seg at om vi utleder den *betingede* fordelingen for tellevariablene *med summen n gitt*, så blir denne en multinomisk fordeling, og de statistiske metodene vi utleder ut fra denne siste kan brukes også i en Poisson-situasjon, jfr. avsnitt 2.2.7.

(B8) Hypergeometriske fordelinger

Vi tar et rent tilfeldig utvalg, jfr. avsnitt 2.1.i, på n telleenheter fra en populasjon på N telleenheter, der M har kjennetegnet C og resten ikke-C. Vi finner x enheter med C i utvalget. Sannsynlighetsfunksjonen for x er da, når både M og N-M er større enn n,

$$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \quad \text{for } x = 0,1,2, \ldots n.$$

Vi har $Ex = n\frac{M}{N}$ og $\sigma_x^2 = \frac{N-n}{N-1} n \frac{M}{N}\left(1-\frac{M}{N}\right)$. For store N-verdier kan sannsynlighetene være svært lik de tilsvarende binomiske med $p = \frac{M}{N}$.

Vi har en simultan hypergeometrisk fordeling i flere variable hvis vi kan få resultatene $C_1, C_2, \ldots, C_j, \ldots, C_k$ som utelukker hverandre, og forekommer hhv. M_1, M_2, ..., M_j, ..., M_k ganger i populasjonen og x_1, x_2, ..., x_j, ..., x_k ganger i utvalget. Vi får da sannsynlighetsfunksjonen

$$f(x_1, x_2, \ldots, x_{k-1}) = \frac{\binom{M_1}{x_1}\binom{M_2}{x_2} \ldots \binom{M_k}{x_k}}{\binom{N}{n}},$$

der hver x_j kan ha verdiene 0, 1, 2, ..., n, men slik at $x_1 + x_2 + \ldots + x_k = n$. Videre er $M_1 + M_2 + \ldots + M_k = N$.

På tilsvarende måte som i avsnitt (B5) kan vi uttrykke en simultan hypergeometrisk sannsynlighetsfunksjon for de tellevariable $n_{11} \ldots n_{ij} \ldots n_{rs}$ i en toveis krysstabell, variablene n_{ijk} i en treveistabell osv.
Når parameterverdiene M_1, ..., M_k, N og n er store tall, vil sannsynlighetene i den hypergeometriske fordelingen ligge svært nær de tilsvarende i en multinomisk fordeling, slik at en også her kan bruke statistiske metoder som utledes for den multinomiske situasjonen. Det vil avhenge av den problemstillingen vi har om vi må utlede spesielle metoder for den hypergeometriske situasjon.

(B9) Den normale sannsynlighetsfunksjonen

Når vi opererer med diskrete variable, kan vi få bruk for enkelte kontinuerlige sannsynlighetsfordelinger som hjelpemidler for å regne ut tilnærmede verdier av visse sannsynligheter. Det gjelder især den normale fordeling og

χ^2 (kji-kvadrat) fordelingen. I kontinuerlige fordelinger kan vi ikke angi sannsynligheter for de enkelte x-verdier (de er faktisk null), men vi regner ut sannsynligheten for at den variable ligger i et visst *intervall* ved hjelp av en sannsynlighets*tetthet* for den variable.

For en normalt fordelt variabel er tettheten

$$(B9.1) \qquad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \text{for } -\infty < x < \infty.$$

Her er μ forventningen av x, σ er standardavviket og som vanlig er $\pi = 3{,}14 \ldots$. Tettheten har maksimum for $x = \mu$, og er symmetrisk om dette punktet.

For $\mu = 0$ og $\sigma = 1$ har vi spesielt den standardiserte normale sannsynlighetstettheten

$$(B9.2) \qquad \varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \qquad \text{for } -\infty < z < \infty.$$

I tabeller over denne fordelingen angis sannsynligheter av typen $P(z \leq z_o)$ for en rekke positive verdier av z_o, og/eller $P(z > z_o)$. Vi har f.eks.

$$P(z > 1{,}64) = 0{,}05, \quad P(z > 1{,}96) = 0{,}025 \text{ osv.}$$

På grunn av symmetrien er

$$P(z < -z_o) = P(z > z_o).$$

Vi ser at fraktilen $z_{0,95} = 1,64$, $z_{0,975} = 1,96$, $z_{0,05} = -1,64$, $z_{0,025} = -1,96$ osv. For en rekke fordelinger, f.eks. den binomiske, vil en "standardisert variabel z, bestemt ved

$$z = \frac{x-Ex}{\sigma_x},$$

være tilnærmet normalt fordelt som (B9.2). Dette bruker vi til å regne ut tilnærmede sannsynligheter eller fraktiler ved hjelp av normaltabellen. Vi har f.eks. at

$$0,05 = P(x > x_{0,95}) = P\left(\frac{x-Ex}{\sigma_x} > \frac{x_{0,95}-Ex}{\sigma_x}\right) \approx P(z > z_{0,95}) \text{ i normalfordelingen,}$$

slik at

$$\frac{x_{0,95}-Ex}{\sigma_x} \approx 1,64,$$

dvs.

$$x_{0,95} \approx Ex + 1,64\,\sigma_x.$$

(B10) χ^2-fordelingen

Hvis vi har en rekke stokastisk uavhengige variable, $z_1, z_2, \ldots, z_n$, hver med tetthet (B9.2), så vil summen av kvadratene av dem, $z = z_1^2 + z_2^2 + \ldots + z_n^2$, ha den såkalte χ^2-fordelingen med n "frihetsgrader", dvs. parameter n.

Tettheten er

$$f(z) = \frac{1}{2^{\frac{n}{2}}\left(\frac{n}{2} - 1\right)!} z^{\frac{n}{2} - 1} e^{-\frac{z}{2}}, \qquad \text{for } z > 0.$$

For odde verdier av n, er

$$\left(\frac{n}{2} - 1\right)! = \left(\frac{n}{2} - 1\right)\left(\frac{n}{2} - 2\right) \ldots \frac{3}{2} \cdot \frac{1}{2}\sqrt{\pi}.$$

Denne fordelingen er tabulert ved at fraktiler som $z_{0,01}$, $z_{0,025}$, $z_{0,05}$, $z_{0,10}$ osv., og $z_{0,90}$, $z_{0,95}$, $z_{0,975}$ osv. er angitt for parameteren n = 1, n = 2, n = 3, osv.

Det kan vises at en rekke av de kjikvadratobservatorene vi får bruk for ved testing av hypoteser i to- og flerveistabeller er *tilnærmet* χ^2-fordelt.

Appendiks C

Sannsynlighetsnivå ved testing etter å ha "kikket på data".

Et enkelt eksempel.

I avsnitt 2.3.5 nevner vi litt om faremomentene ved å la de observerte data virke inn ved oppstilling av hypoteser som vi så tester ved hjelp av de samme data. Vi skal ta et eksempel på hvordan dette kan virke inn på sannsynlighetsnivået før testen.

Vi vil undersøke om det er noen forskjell mellom kvinner og menn i tilbøyeligheten til å dra på helgetur når de ikke disponerer fritidshus (jfr. data i tabell 2.2 c når vi slår sammen tallet på helgeturer fra 1 og oppover til én gruppe). Før vi ser på data, har vi ingen bestemt mening om hvilken vei en eventuell forskjell går, vi vil teste nullhypotesen $p^k = p^m$ mot p^k forskjellig fra p^m, der p^k er sannsynligheten for at en kvinne skal dra på helgetur og p^m er tilsvarende for en mann. Vi bruker en "tilnærmet normaltest", jfr. 3.1.2, for å sammenlikne de relative hyppighetene av å dra på helgetur som vi finner i tabellen, nemlig $h^m = 486/725 = 0{,}67$ og $h^k = 530/779 = 0{,}68$. Vi vil forkaste nullhypotesen om samme "tursannsynlighet" for de to kjønn hvis den standardiserte forskjellen d (se 3.1.2) mellom h^k og h^m er mindre enn - 1,96 eller større enn 1,96. Da har vi en test med nivå tilnærmet 0,05.

Men sett at vi kikker på data og sier: "Her er hyppigheten større for kvinner enn for menn, da velger jeg å teste $p^k = p^m$ mot $p^k > p^m$. Følgelig forkaster jeg nullhypotesen hvis $d > 1{,}64$, dvs. "øvre 5 prosentfraktil i normalfordelingen". Dette gir imidlertid *ikke* en test med nivå 0,05, men med nivå 0,10.

Den fullstendige testen i dette tilfelle er nemlig: Sett at det virkelig er slik at $p^k = p^m$. Da vil det være rent tilfeldig om jeg observerer $h^k > h^m$

eller $h^k < h^m$. I første tilfelle forkaster jeg H_o hvis $d > 1{,}64$, i det annet forkaster jeg H_o hvis $d < 1{,}64$. Dvs. at nivået for testen er $0{,}05 + 0{,}05 = 0{,}10$.

Vi kan også si det slik: Hvis jeg tar som gitt at $d > 0$, så er fordelingen av d ikke lenger normal, men den er gitt ved *den høyre halvdelen* av normalfordelingen med alle sannsynligheter multiplisert med 2 (for å gi en fullstendig fordeling med sum (integral) lik 1). Følgelig er $P(d > 1{,}64 \mid d > 0) = 2 \cdot 0{,}05 = 0{,}10$.

Utkommet i serien Rapporter fra Statistisk Sentralbyrå (RAPP) - ISSN 0332-8422

Trykt 1983

Nr. 83/1 Naturressurser 1982 Foreløpige nøkkeltall fra ressursregnskapene for energi, mineraler, skog, fisk og areal Sidetall 62 Pris kr 15,00 ISBN 82-537-1837-3

- 83/2 Totalregnskap for fiske- og fangstnæringen 1978 - 1981 Sidetall 39 Pris kr 12,00 ISBN 82-537-1882-9

- 83/3 Therese Hunstad: Forbruk av fisk og fiskevarer i Norge 1979 En undersøkelse av fiskeforbruket i Norge i 1979 med bakgrunn i materialet fra momskompensasjonsordningen for fisk og fiskevarer Sidetall 25 Pris kr 12,00 ISBN 82-537-1904-3

- 83/4 Atle Martinsen og Hogne Steinbakk: Planregnskap for Rogaland 1981 - 1992 Hovedresultater Sidetall 42 Pris kr 12,00 ISBN 82-537-1902-7

- 83/5 Anne Mickelson og Hogne Steinbakk: Planregnskap for Akershus 1981 - 1992 Hovedresultater Sidetall 48 Pris kr 18,00 ISBN 82-537-1903-5

- 83/6 Asbjørn Aaheim: Norske olje- og gassreserver Nåverdiberegninger og inndeling i kostnadsklasser Sidetall 28 Pris kr 12,00 ISBN 82-537-1911-6

- 83/7 Roar Bergan: Behandlingen av oljevirksomheten i Byråets makroøkonomiske årsmodeller Sidetall 30 Pris kr 12,00 ISBN 82-537-1918-3

- 83/8 Arbeid og helse 1982 Sidetall 101 Pris kr 18,00 ISBN 82-537-1927-2

- 83/9 Radio- og fjernsynsundersøkelsen Februar 1983 Sidetall 118 Pris kr 18,00 ISBN 82-537-1928-0

- 83/10 Petter Frenger: On the Use of Laspeyres and Paasche Indices in a Neoclassical Import Model Om bruken av Laspeyres og Paasche indekser i en neoklassisk importmodell Sidetall 49 Pris kr 18,00 ISBN 82-537-1931-0

- 83/11 Øystein Olsen: MODAG-RAPPORT Etterspørselsfunksjoner for arbeidskraft, energi og vareinnsats Sidetall 38 Pris kr 12,00 ISBN 82-537-1935-3

- 83/12 Karl-Gerhard Hem: Energiundersøkelsen 1980 Sidetall 47 Pris kr 12,00 ISBN 82-537-1949-3

- 83/13 Jan Byfuglien og Ole Ragnar Langen: Grunnkretser, tettsteder og menigheter Dokumentasjon 1980 Sidetall 57 Pris kr 18,00 ISBN 82-537-1952-3

- 83/14 Even Flaatten: Barnevernsklienter og sosial bakgrunn Sidetall 61 Pris kr 18,00 ISBN 82-537-1989-2

- 83/15 Skatter og overføringer til private Historisk oversikt over satser mv. Årene 1970 - 1983 Sidetall 77 Pris kr 18,00 ISBN 82-537-1961-2

- 83/16 Erik Biørn og Morten Jensen: Varige goder i et komplett system av konsumetterspørselsfunksjoner - En modell estimert med norske kvartalsdata Sidetall 93 Pris kr 18,00 ISBN 82-537-1962-0

- 83/18 Jon Inge Lian: Fylkenes bruk av helseinstitusjoner Oversikt 1980 og forsøk på framskriving Sidetall 89 Pris kr 18,00 ISBN 82-537-1969-8

- 83/19 Redigert av Kjell Roland og Paal Sand: MODIS IV Dokumentasjonsnotat nr. 17 Endringer i utgave 80-1, 81-1 og 82-1 Sidetall 62 Pris kr 18,00 ISBN 82-537-1974-4

- 83/21 Arne S. Andersen og Rolf Aaberge: Analyse av ulikhet i fordelinger av levekår Sidetall 130 Pris kr 18,00 ISBN 82-537-1988-4

- 83/22 Asbjørn Aaheim: Kostnader ved ulike utbyggingsrekkefølger av vassdragsutbygginger En metodestudie Sidetall 27 Pris kr 12,00 ISBN 82-537-1986-8

- 83/23 Vidar Otterstad og Hogne Steinbakk: Planrekneskap for Sør-Trøndelag 1981 - 1992 Hovudresultat Sidetall 43 Pris kr 12,00 ISBN 82-537-1983-3

- 83/24 Otto Carlson: Pasientstatistikk 1981 Statistikk fra Det økonomiske og medisinske informasjonssystem Sidetall 70 Pris kr 18,00 ISBN 82-537-1991-4

- 83/25 Aktuelle skattetall 1983 Current Tax Data Sidetall 46 Pris kr 12,00 ISBN 82-537-1990-6

- 83/26 Konsumprisindeksen Sidetall 57 Pris kr 18,00 ISBN 82-537-1998-1

Utkommet i serien Rapporter fra Statistisk Sentralbyrå (RAPP) (forts.) - ISSN 0332-8422

Trykt 1983

Nr. 83/27 Erik Biørn: Gross Capital, Net Capital, Capital Service Price, and Depreciation: A Framework for Empirical Analysis Sidetall 69 Pris kr 18,00 ISBN 82-537-1995-7

\- 83/28 Jens-Kristian Borgan: Kohort-dødeligheten i Norge 1846 - 1980 Cohort Mortality in Norway Sidetall 200 Pris kr 18,00 ISBN 82-537-1997-3

\- 83/29 Nils Martin Stølen: Etterspørsel etter arbeidskraft i norske industrinæringer Sidetall 68 Pris kr 18,00 ISBN 82-537-2001-7

\- 83/30 Erling Siring: To notater om sammenlikning av data fra Fruktbarhetsundersøkelsen 1977 med data fra registre Sidetall 40 Pris kr 18,00 ISBN 82-537-2006-8

83/31 Knut Fredrik Strøm: Varestrømmer i engros- og detaljhandel Sidetall 89 Pris kr 18,00 ISBN 82-537-2008-4

\- 83/32 Tor Skoglund og Knut Ø. Sørensen: Regionale strukturendringer belyst ved sysselsettingstall Sidetall 52 Pris kr 18,00 ISBN 82-537-2003-3

\- 83/33 Nils Martin Stølen: Importandeler og relative priser En MODAG-rapport Sidetall 62 Pris kr 18,00 ISBN 82-537-2010-6

\- 83/34 Totalregnskap for fiske- og fangstnæringen 1979 - 1982 Sidetall 39 Pris kr 12,00 ISBN 82-537-2013-0

\- 83/35 Holdninger til norsk utviklingshjelp 1983 Sidetall 81 Pris kr 18,00 ISBN 82-537-2014-9

Trykt 1984

Nr. 84/1 Naturressurser og miljø 1983 Foreløpige nøkkeltall fra ressursregnskapene for energi, mineraler, skog, fisk og areal Sidetall 100 Pris kr 18,00 ISBN 82-537-1993-0

\- 84/2 Torstein Bye: Energisubstitusjon i næringssektorene i en makromodell Sidetall 47 Pris kr 12,00 ISBN 82-537-2042-4

\- 84/4 Jon Age Vestøl: Kommunale avfallsbehandlingsanlegg Miljøstandard Sidetall 78 Pris kr 18,00 ISBN 82-537-2062-9

\- 84/7 Tiril Vogt: Social Indicators and Environmental Dimensions Sidetall 33 Pris kr 12,00 ISNB 82-537-2060-2

\- 84/9 Herdis Thorén Amundsen: Statistiske metoder for analyse av samvariasjon i kategoriske data Sidetall 228 Pris kr 24,00 ISBN 82-537-2074-2

Pris kr 24,00

Publikasjonen utgis i kommisjon hos H. Aschehoug & Co. og Universitetsforlaget, Oslo, og er til salgs hos alle bokhandlere.

ISBN 82-537-2074-2
ISSN 0332-8422