DR. PER OTTESTAD

Forelesninger

over

MATEMATIKK

og

STATISTIKK

ved

NORGES LANDBRUKSHÖGSKOLE

III. Sannsynlighetsregning (1. og 2.del)

IV. Statistisk induksjon

D R. P E R O T T E S T A D

F o r e l e s n i n g e r

over

M A T E M A T I K K o g S T A T I S T I K K

ved

NORGES LANDBRUKSHÖGSKOLE

III. Sannsynlighetsregning (1. del).

Skrivemaskinstua
Oslo i jan. 1947

Innhold:

III. Sannsynlighetsregning (1. del).

1. Matematisk sannsynlighet.

Det vi hittil har gjennomgått av statistikkens teori, har vesentlig omfattet beskrivelser av noen av de samletallene som vi har bruk for når vi skal beskrive eller analysere en rekke observasjoner. Før vi kan komme videre i faget, må vi lære litt sannsyhlighetsregning. Det ville da naturligvis ha vært en fordel om vi kunne innlede dette avsnitt med en enkel definisjon av det grunnleggende begrep "matematisk sannsynlighet". Det finnes imidlertid minst to slike definisjoner, og det synes ikke som fagstatistikerne kan bli enige om hvilke av dem en bør foretrekke. Vanskelighetene ved formuleringen henger sammen med at "matematisk sannsynlighet" er nøye knyttet til en almenmenneskelig begrepsdannelse. Med tanke på et eller annet som kanskje kan hende i fremtiden, bruker en i daglig tale meget ofte uttrykk som "rimelig", "sannsynlig", "urimelig", "utrolig", "meget sannsynlig" osv. Går tanken bakover i tiden og søker forklaring på noe som har hendt, karakteriserer vi ofte gitte forklaringer med de samme uttrykkene.

En bruker slike uttrykk som sannsynlig, usannsynlig, rimelig, utrolig osv. når en i daglig tale skal formulere utsagn om noe som en ikke kan uttale seg om med sikkerhet. Når vi så ofte må ty til slike uttrykksmåter, kommer det av at årsak-virkning komplekset bare i meget sjeldne tilfelle er så enkelt og lett forståelig at en har grunnlag for utsagn som gir uttrykk for en rasjonell fast overbevisning.

Leder en på en eller annen måte vannstoff fram til intim berøring med opphetet kobberoksyd, er en før resultatet foreligger, overbevist om at det vil bli dannet kobber og vann ($CuO + 2H = Cu + H_2O$). Derimot vil det alltid foran en fotballkamp mellom to lag P og Q, selv blant de best informerte, være delte meninger om utsikten for f.eks. at P skal seire. I dette tilfelle er det nemlig så mange ukjente faktorer med i spillet, at det er meget vanskelig å gi noe bestemt forhåndsutsagn.

La oss nå tenke oss at en bestemt hending er en av flere hendinger som kan inntreffe under nærmere angitte omstendigheter. "Seir for P" er en av de tre hendinger ("seir for P", "seir for Q" og "uavgjort") som kan inntreffe som resultat av en fotballkamp mellom lagene P og Q. Vi sier at "seir for P" er en av de hendinger som kan inntreffe ved en nærmere angitt begivenhet, nemlig begivenheten kamp mellom P og Q. På samme måte er "jente" en hending som kan inntreffe ved begivenheten "fødsel", "sparkort" en hending som kan inntreffe ved begivenheten "trekning av ett kort av en kortstokk" osv. En slik hending vil vi i det følgende betegne med H.

Hvis vi har tilstrekkelige opplysninger å bygge på, kan vi forut for begivenheten vurdere utsikten for at H skal inntreffe. Den matematiske sannsynlighet (eller bare sannsynligheten) for at H skal inntreffe ved denne begivenheten, er et slikt utsagn om utsikten for at H skal inntreffe. Vi skal senere se at sannsynligheten for H uttrykkes i et tall mellom 0 og 1. Jo nærmere 1 sannsynligheten for H er, jo større utsikt er det for at H skal inntreffe. Er sannsynligheten for H nær lik 0, betyr det at det er liten utsikt for at H skal inntreffe.

Har vi konstatert at hendingen H har inntruffet ved en bestemt begivenhet, er vi naturligvis klar over at årsak-virkning komplekset har vært slik at H nødvendigvis måtte inntreffe ved denne begivenheten. Det at vi forut for begivenheten, ikke kan uttale noe sikkert, men må nøye oss med å formulere utsagnet i sannsynligheter, er en innrømmelse av vår utilstrekkelige innsikt og viten.

La oss nå tenke oss at vi foran en kamp mellom fotballagene P og Q skal prøve å vurdere utsikten for seir for P. Det er klart at et forsøk på å avveie de faktorer som taler til gunst for denne hendingen i forhold til de faktorer som taler til gunst for de andre mulige hendinger som kan inntreffe, ikke kan føre fram. Vi kan naturligvis nevne opp noen av de faktorer som har noe å bety for utfallet av kampen, som f.eks. lagenes trening, kapteinenes lederegenskaper, banens beskaffenhet osv. Men vi kan ikke uttrykke disse faktorers innflytelse på utfallet kvantitativt, og dessuten må vi innrømme at det kan finnes andre faktorer som kan ha noe å bety for utfallet og som vi ikke er oppmerksomme på. Fotballinteresserte vil naturligvis heller ikke vurdere utsikten for hendingen "seir for P" på denne måten. De vil heller benytte seg av opplysninger om utfallet av kamper som de to lagene har spilt tidligere. Og det er umiddelbart innlysende at det beste grunnlag for en vurdering er opplysninger som eventuelt foreligger, om utfallet av kamper som P og Q har spilt mot hverandre. Hvis de to lagene umiddelbart forut for den kampen som vi skal prøve å danne oss en mening om utfallet av, har spilt n kamper mot hverandre og P har vunnet h av disse, vil vi kunne bruke det relative antall seire for P, h/n, som grunnlag for vurderingen av utsikten for at P skal seire.

Det er dette resonnement moderne sannsynlighetsteoretikere legger til grunn for definisjonen av den matematiske sannsynlighet for at hendingen H skal inntreffe ved en begivenhet. Denne definisjonen som vi skal gjengi senere, er imidlertid ikke en definisjon av det almenmenneskelige begrep sannsynlighet.

Det er bare en definisjon av det matematiske begrep som er utkrystalisert av dette, og som danner grunnlaget for sannsynlighetsregningen.

Før vi imidlertid gir oss i kast med denne definisjonen, skal vi prøve å bli fortrolige med to andre begreper, nemlig gjentakelse og univers.

Det er klart at skal vi kunne vurdere utsikten for at en hending H skal inntreffe ved en bestemt begivenhet, må vi ha en beskrivelse av denne begivenheten. En begivenhet kan beskrives ved at en nøvner opp et større eller mindre antall kjennetegn som er karakteristisk for den. Kamp mellom to fotballag er en begivenhet. Skal vi kunne uttale oss om utsikten for hendingen H=seir for P, må det først og fremst være oppgitt hvilket lag P skal spille mot. Det første kjennetegnet på begivenheten er derfor navnet på P's motspiller. Men begivenheten kan være beskrevet mer inngående, f.eks. ved opplysningen om at kampen skal spilles på gressbane og i overskyet oppholdsvær.

La i sin alminnelighet $k_1, k_2, k_3, \ldots\ldots, k_c$ være kjennetegnene på en begivenhet. En kamp som P skal delta i, er en begivenhet som kan være beskrevet ved et større eller mindre antall slike kjennetegn. Den kan være beskrevet:
1) bare ved kjennetegnet k_1 = motspiller Q,
2) ved kjennetegnene k_1 = motspiller Q og k_2 = gressbane,
3) ved kjennetegnene k_1 = motspiller Q, k_2 = gressbane og k_3 = overskyet oppholdsvær.

Det er naturligvis mulig at utsikten for H = seir for P kan være forskjellig ved disse tre begivenhetene. Hvis P spiller bedre på gressbane enn på annen bane, mens banens beskaffenhet ikke har noe å si for motspillerens ytelse, vil utsikten for at H skal inntreffe være større ved begivenheten 2) enn ved begivenheten 1). Vi kan derfor ikke tale om sannsynligheten for en hending H i sin alminnelighet. Sannsynligheten for at H skal inntreffe, er nøye knyttet til en bestemt begivenhet beskrevet ved et bestemt sett kjennetegn.

Hvis nå to begivenheter har nøyaktig samme sett kjennetegn, sier vi at den ene begivenheten er en gjentakelse av den annen. To kamper mellom fotballagene P og Q på gressbane er altså gjentakelser av hverandre. Derimot er begivenheten "kamp mellom P og Q på gressbane" ikke en gjentakelse av begivenheten "kamp mellom P og Q på gressbane i overskyet oppholdsvær" fordi den siste begivenheten har et kjennetegn som den første begivenheten ikke har.

Vi skylder kanskje her å tilføye at de fleste læreboksforfattere som forsøker å definere gjentakelse, tar med et krav utover det at begivenhetene skal ha samme sett kjennetegn. Det stilles også det krav at alle kjennetegn som har noe vesentlig å bety for sannsynligheten for den hending det er tale om,

skal være med i settet. Men så kan en jo spørre: hvilke kjennetegn er det som har vesentlig innflytelse? Dette kan ikke avgjøres på forhånd. Bare erfaring kan vise hvilke kjennetegn det er som har noe å bety for sannsynligheten for en bestemt hending.

Antallet av gjentakelser av en bestemt begivenhet er naturligvis alltid endelig. Men i tanken kan en forutsette at begivenheten kan gjentas uendelig mange ganger. Disse uendelig mange tenkte gjentakelser kaller en et univers. Et univers som naturligvis bare er en forestilling, omfatter altså uendelig mange gjentakelser av en begivenhet med et bestemt sett kjennetegn. En aktuell rekke gjentakelser (begivenheter) med dette sett kjennetegn må oppfattes som et utvalg av universet.

Til de begivenheter (gjentakelser) som danner et utvalg av et univers, må en stille det krav at de skal være uavhengige av hverandre. Med dette menes at resultatet av en gjentakelse ikke har noe å bety for resultatet av en annen gjentakelse. At dette kravet ikke alltid er oppfylt, er lett å innse. To kamper mellom fotballagene P og Q kan således være avhengige gjentakelser av samme begivenhet. Hvis P har tapt kamp nr. 1, kan dette virke til at lagets spillere anstrenger seg mer under kamp nr. 2. Resultatet ved første gjentakelse (kamp nr. 1) kan da ha noe å bety for resultatet av annen gjentakelse (kamp nr. 2).

La oss nå tenke oss at vi har et utvalg på n uavhengige gjentakelser av et univers hvor gjentakelsene har et bestemt sett kjennetegn. For universet vil vi bruke betegnelsen U. La oss videre tenke oss at hendingen H har inntruffet i h(H,U) av disse n gjentakelser. Den matematiske sannsynlighet for at H skal inntreffe i en ny gjentakelse i dette universet, defineres som grensen for h(H,U)/n når n økes over alle grenser. La oss betegne sannsynligheten for H med s(H,U). Da er

$$s(H,U) = \mathrm{Gr}.\frac{h(H,U)}{n} \qquad \text{når } n \to \infty$$

At s(H,U) må være et tall mellom 0 og 1 følger direkte av at h(H,U) er et tall mellom 0 og n (grensene medregnet).

Siden det i praksis ikke er mulig å gjenta en begivenhet uendelig mange ganger, kan vi ikke få bestemt s(H,U) nøyaktig. Men vi kan på grunnlag av resultatene av et endelig antall (n) uavhengige gjentakelser få bestemt en anslagsverdi av denne sannsynligheten. Anslagsverdien er h(H,U)/n, og om denne sier en at den er et estimat (av det engelske substantiv estimate) av s(H,U).

Av det som er forklart foran, vil det fremgå at sannsynligheten for at en hending skal inntreffe ved en begivenhet er nøye knyttet til det univers som denne begivenheten tilhører. I regelen kan imidlertid den samme hendingen også

inntreffe ved begivenheter som er gjentakelse i andre universer. Med det at to universer er ulike menes at gjentakelsene i det ene (U) har et annet kjennetegnsett enn gjentakelsene i det andre (U'). Det er naturligvis ikke noe i veien for at flere kjennetegn kan være felles for de to sett, men det må være minst ett kjennetegn i det ene sett som ikke finnes i det andre. La oss ta for oss et eksempel.

Sett at hendingen H er at en bestemt 30 år gammel person skal dø innen ett år. Sannsynligheten for H kan da ha forskjellig verdi, avhengig av hvilket univers vedkommende person regnes inn under. Vi kan som eksempler nevne følgende universer:

1) et univers U svarende til et utvalg som omfatter alle personer (begivenheter, gjentakelser) med kjennetegnssettet

k_1 = alder 30 år og k_2 = norsk statsborger,

2) et univers U' svarende til et utvalg som omfatter alle personer med kjennetegnssettet

k_1 = alder 30 år, k_2 = norsk statsborger og k_3 = mann,

3) et univers U" svarende til et utvalg som omfatter alle personer med kjennetegnssettet

k_1 = alder 30 år, k_2 = norsk statsborger, k_3 = mann, k_4 = gift og k_5 = industriarbeider,

Sannsynlighetene for hendingen H - $s(H,U)$, $s(H,U')$ og $s(H,U")$ - kan naturligvis være ulike i disse tre ulike universer.

Vi legger merke til at kjennetegnssettet for gjentakelsene i U' i dette tilfelle inneholder hele kjennetegnssettet for gjentakelsene i universet U. Universet U' omfatter altså alle de gjentakelsene i universet U som foruten de kjennetegn som er felles for alle gjentakelsene i U, også har kjennetegnet k_3 = mann. Universet U er derfor mer omfattende, det har større bredde enn U'. En sier at U' er et delunivers eller et subunivers av U. Hvis i sin alminnelighet gjentakelsene i U har kjennetegnssettet $k_1,k_2,k_3, \ldots\ldots k_c$ og gjentakelsene i U' kjennetegnssettet $k_1,k_2,k_3, \ldots\ldots k_c,k_{c+1},k_{c+2} \ldots\ldots$, er U' et subunivers av U. Vi skal senere se at det i mange tilfelle er meget viktig å få brakt på det rene om $s(H,U)$ og $s(H,U')$ er ulike når U' er et subunivers av U. Hvis nemlig disse to sannsynlighetene er ulike, betyr det at det eller de kjennetegn som gjentakelsene i U' har og som ikke er felles for alle gjentakelsene i U, har noe å bety for sannsynligheten for H.

Det følger direkte av definisjonen av $s(H,U)$ at hvis H er en hending som nødvendigvis må inntreffe i hver gjentakelse i U, må $s(H,U) = 1$. Da er nemlig

det relative gjentakelsestall - $h(H,U)/n$ - for H lik enheten hvor stort eller lite utvalget (n) av gjentakelser er. En oppfatter derfor $s(H,U) = 1$ som det sannsynlighetsteoretiske uttrykk for at H må inntreffe i en gjentakelse i U. Er på den annen side H en hending som nødvendigvis ikke kan inntreffe i en gjentakelse i U, må $h(H,U) = 0$ hvor stort utvalget (n) av gjentakelser enn er. Følgelig er $s(H,U) = 0$, En oppfatter derfor $s(H,U) = 0$ som det sannsynlighetsteoretiske uttrykk for at H ikke kan inntreffe i en gjentakelse i U.

Den definisjonen av $s(H,U)$ som vi har gjengitt foran, er ikke samstemmig godtatt av alle som må forutsettes å ha tilstrekkelige kvalifikasjoner til å ta et personlig standpunkt til saken. I de fleste lærebøker brukes derfor en annen definisjon som ofte kalles den klassiske fordi den var den første som ble utformet. Vi skal gjengi denne definisjonen senere. Men først skal vi gjennomgå et par viktige setninger.

La oss anta at vi har visshet for at en av m hendinger som vi vil betegne med $H_1, H_2, H_3, \ldots\ldots H_m$, nødvendigvis må inntreffe ved en bestemt begivenhet. Foran en fotballkamp mellom lagene P og Q har vi visshet for at en av hendingene H_1 = seir for P, H_2 = seir for Q og H_3 = uavgjort må inntreffe. Vi vil videre forutsette at disse m hendingene utelukker hverandre, dvs. at hvis en av dem inntreffer, kan ingen av de andre hendinger inntreffe. Denne begivenheten er nå en gjentakelse i et univers U hvor alle gjentakelser har de kjennetegn som er karakteristisk for vedkommende begivenhet, og ingen andre kjennetegn. Under disse forutsetninger kan det bevises at summen av sannsynlighetene for hver av disse hendinger er lik enheten, eller :

$$\sum_{i=1}^{m} s(H_i,U) = s(H_1,U)+s(H_2,U)+ \ldots\ldots\ldots\ldots+s(H_m,U) = 1.$$

Har nemlig disse m hendinger inntruffet $h(H_1,U), h(H_2,U)\ldots$ og $h(H_m,U)$ ganger i n uavhengige gjentakelser i U, dvs. at $\sum_{i=1}^{m} h(H_i,U) = n$, er

$$\sum_{i=1}^{m} s(H_i,U) = \sum_{i=1}^{m} \text{Gr.}\ \frac{h(H_i,U)}{n} = \text{Gr.}\ \sum_{i=1}^{m} \frac{h(H_i,U)}{n} = \text{Gr.}\ \frac{n}{n} = 1.$$

Dette er en viktig setning som vi skal gjøre bruk av senere.

2. Enten - eller loven.

Det hender meget ofte at en rekke hendinger $H_1, H_2, \ldots\ H_g$ som kan inntreffe ved samme begivenhet, kan innordnes under en felles betegnelse : hendingen A. Vi skal nevne noen eksempler.

1) La oss tenke oss at vi trekker ett kort av en kortstokk. Ved denne be-

givenheten er det 52 ulike hendinger som kan inntreffe, nemlig de 52 forskjellige kortene. Hendingene H_1 = spar 2, H_2 = spar 3, H_3 = spar 4, ... H_{12} = spar konge og H_{13} = spar ess kan da innordnes under fellesbetegnelsen A=spar. Hendingene "spar ess", "hjerter ess", "ruter ess" og "kløver ess" kan innordnes under hendingen "ess".

2) Ved en tvillingfødsel kan følgende hendinger inntreffe : H_1 = 2 jenter, H_2 = 2 gutter og H_3 = 2 barn av ulike kjønn. Hendingene H_1 og H_2 kan da innordnes under hendingen A = 2 barn av samme kjønn.

Vi vil nå forutsette at $H_1, H_2, \ldots. H_g$ kan inntreffe i en gjentakelse i et univers U og at de utelukker hverandre i dette universet. Videre vil vi forutsette at disse g hendinger kan innordnes under hendingen A. Dette betyr da at hvis en av de g hendinger inntreffer, vil også A inntreffe. Vi forutsetter også at A ikke kan inntreffe på annen måte. Hendingen A er da identisk med "enten H_1 eller H_2 eller ..... eller H_g". Oppgaven er å finne sannsynligheten for A - s(A,U) - når vi kjenner sannsynlighetene for $H_1, H_2, \ldots\ldots H_g$. Vi kan bevise at under de nevnte forutsetninger er

$$s(A,U) = s(H_1,U)+s(H_2,U)+\ldots\ldots+s(H_g,U) = \sum_{i=1}^{g} s(H_i,U).$$

La oss anta at vi har et utvalg på n uavhengige gjentakelser i U og at H_1 har inntruffet i $h(H_1,U)$, H_2 i $h(H_2,U)$, ... og H_g i $h(H_g,U)$ av disse gjentakelser. Siden nå A inntreffer hver gang H_1 inntreffer, hver gang H_2 inntreffer, ... og hver gang H_g inntreffer, har A inntruffet i h(A,U) av de n gjentakelser, hvor

$$h(A,U) = h(H_1,U)+h(H_2,U)+ \ldots\ldots + h(H_g,U) = \sum_{i=1}^{g} h(H_i,U).$$

I følge vår definisjon av s(A,U) er da

$$s(A,U) = Gr.\frac{h(A,U)}{n} = Gr.\ \frac{\sum_{i=1}^{g} h(h_i,U)}{n} = \sum_{i=1}^{g} Gr.\ \frac{h(H_i,U)}{n} = \sum_{i=1}^{g} s(H_i,U).$$

Dette er enten-eller loven for hendinger som utelukker hverandre i et univers.

Oppgave 1. Hva er grunnen til at denne setningen ikke kan brukes med mindre de hendinger hvis sannsynligheter skal adderes, utelukker hverandre ?

Oppgave 2. Anta at sannsynligheten for at det ved en tvillingfødsel blir født to jenter er 0,31 og sannsynligheten for at det blir født to gutter er lik 0,34. Finn sannsynligheten for at det ved en tvillingfødsel skal bli født to barn av samme kjønn.

Oppgave 3. La oss tenke oss at vi trekker ett kort av en kortstokk. Da det i kortstokken er 52 ulike kort, kan en av 52 hendinger inntreffe ved denne begivenheten og disse utelukker hverandre. Anta at sannsynligheten for hver av disse hendinger er lik 1/52. Finn sannsynligheten for å trekke et sparkort, sannsyn-

ligheten for å trekke et ess og sannsynligheten for å trekke enten hjerter eller kløver.
Kan en bruke den beviste setningen til å finne sannsynligheten for å trekke enten spar eller konge?

Oppgave 4. Sett at et fotballag R har valget mellom å spille en kamp mot lag P og å spille en kamp mot lag Q. Anta at en på en eller annen måte har funnet at sannsynligheten for at R skal vinne kampen mot P er 3/4 og sannsynligheten for at R skal vinne kampen mot Q er 1/2. Kan en da bruke den beviste setningen til å finne sannsynligheten for at R skal vinne enten kampen mot P eller vinne kampen mot Q?

La oss nå anta at ingen andre hendinger enn $H_1, H_2, \ldots H_m$ kan inntreffe i en gjentakelse i universet U og at disse utelukker hverandre. Da er som vist foran

$$\sum_{i=1}^{m} s(H_i,U) = 1.$$

La oss videre anta at g av disse hendinger kan innordnes under hendingen A. La dette være $H_1, H_2, \ldots H_g$. Hvis en av disse inntreffer, vil altså også A inntreffe. Men hvis en av de andre hendinger - $H_{g+1}, H_{g+2}, \ldots H_m$ - inntreffer, kan A nødvendigvis ikke inntreffe. Disse m-g hendinger kan da innordnes under fellesbetegnelsen hendingen ikke-A eller kort iA. I følge enteneller loven har vi at

$$s(A,U) = \sum_{i=1}^{g} s(H_i,U)$$

og

$$s(iA,U) = \sum_{i=g+1}^{m} s(H_i,U)$$

Følgelig er

$$s(A,U) + s(iA,U) = \sum_{i=1}^{g} s(H_i,U) + \sum_{i=g+1}^{m} s(H_i,U) = \sum_{i=1}^{m} s(H_i,U) = 1$$

Hendingene A og iA kalles motsatte. Summen av sannsynlighetene for to motsatte hendinger i samme univers er altså lik enheten. Eller : summen av sannsynligheten for at en hending skal inntreffe og sannsynligheten for at den ikke skal inntreffe i en gjentakelse i et univers er lik enheten.

Oppgave 5. Forutsett det samme som i oppg. 2. Finn sannsynligheten for at det ved en tvillingfødsel blir født to barn av ulike kjønn.

3. Den klassiske definisjon.

La oss forutsette at ingen andre hendinger enn $H_1, H_2, \ldots H_m$ kan inntreffe i en gjentakelse i universet U og at disse utelukker hverandre. La oss videre anta at disse m hendinger er like sannsynlige. Da summen av sannsynlighetene for hver av dem er lik enheten, er

$$s(H_1,U) = s(H_2,U) = \ldots\ldots\ldots = s(H_m,U) = \frac{1}{m}$$

Er nå A en hending som må inntreffe hvis en av hendingene $H_1, H_2, \ldots H_g$ $(g < m)$ inntreffer og som ikke kan inntreffe på annen måte, er

$$s(A,U) = \sum_{i=1}^{g} s(H_i,U) = \sum_{i=1}^{g} \frac{1}{m} = \frac{g}{m}$$

Vi ser at $s(A,U)$ er lik en brøk hvis nevner er antallet av de like sannsynlige hendinger som kan inntreffe og som utelukker hverandre, og hvis teller er antallet av disse hendinger som medfører at A inntreffer. Eller: sannsynligheten for A er lik forholdet mellom antall gunstige hendinger og antall mulige hendinger forutsatt at alle mulige hendinger er like sannsynlige og utelukker hverandre.

Det er denne definisjonen som er mest brukt i fremstillinger av sannsynlighetsregningen. Den gir anvisning på hvordan sannsynligheten for en hending skal bestemmes når en kan finne antallet av mulige og like sannsynlige hendinger som utelukker hverandre, og antallet av disse som er gunstige for vedkommende hending. Å bestemme disse antall hendinger er imidlertid praktisk talt aldri mulig, og en har derfor i aktuelle tilfelle meget lite nytte av denne definisjonen. På den annen side gir den anvisning på hvordan en skal løse en rekke enkle modelloppgaver som det er nyttig å ta med for å innøve de forskjellige regneregler. Og dessuten er den et gunstig utgangspunkt for fremstillingen av noen av reglene for mer sammensatte regneoperasjoner.

Anta at en pose inneholder 5 svarte og 5 hvite kuler, ialt 10 kuler. Vi tenker oss at vi trekker en kule. Hva er sannsynligheten for å trekke en svart kule? Vi kan løse denne oppgaven under den forutsetning at de 10 kulene er like sannsynlige hendinger når vi trekker en kule. Under denne forutsetning er det nemlig 5 av 10 like sannsynlige hendinger som er gunstige for hengingen "svart kule" og disse 10 hendingene utelukker hverandre. Følgelig er sannsynligheten for å trekke en svart kule lik

$$s(\text{svart kule},U) = \frac{g}{m} = \frac{5}{10} = \frac{1}{2}$$

Universet U omfatter i dette tilfelle uendelig mange gjentakelser av den begivenheten som består i at en trekker en kule av posen.

En bør legge merke til at når vi i dette tilfelle taler om hendingen "svart kule" i stedet for om hendingen A, er det bare for å feste tanken til noe konkret. Skulle vi estimere sannsynligheten for hendingen "svart kule" ved å utføre et antall virkelige trekninger av en pose med 5 svarte og 5 hvite kuler, ville vi ikke ha noen garanti for at våre teoretiske forutsetninger var realisert. Det er to slike forutsetninger. For det første har vi forutsetningen om at de 10 kulene er like sannsynlige hendinger. Det er rimelig at denne forutsetningen kan realiseres tilnærmet ved at vi gjør kulene så like som mulig (størrelse, form, vekt), unntatt fargen naturligvis, og at vi blander dem godt. Den andre forutsetningen er at trekningene skal være uavhengige gjentakelser. For å realisere best mulig denne forutsetningen, måtte vi sørge for at den utrukne kulen ble lagt tilbake i posen og blandet godt med de andre kulene før en ny kule ble trukket.

Vi skal referere resultatet av en slik serie trekninger. I en pose ble det lagt samme antall svarte og hvite kuler og det ble trukket en kule 4096 ganger. Resultatet ble 2030 svarte og 2066 hvite kuler. Det relative antall svarte kuler ble altså lik 2030/4096 = 0,4956 eller omtrent $\frac{1}{2}$.

Oppgave 6. En tenker seg at en trekker ett kort av en vanlig kortstokk. I det det forutsettes at de 52 kortene er like sannsynlige hendinger når det trekkes ett kort, skal sannsynlighetene for følgende hendinger beregnes:
1) sparkort, 2) herrekort, 3) ess, 4) enten konge eller knekt og 5) enten knekt eller dame eller konge eller ess.

Oppgave 7. En kaster "mynt og krone" med et pengestykke 2 ganger. Finn sannsynlighetene for følgende hendinger:
1) krone begge ganger, 2) krone en gang og mynt en gang, 3) krone minst en gang.
Gjør rede for forutsetningene for løsningene.

4. Både - og loven I.

La A og B være to hendinger som ikke utelukker hverandre i en gjentakelse i universet U, de kan altså inntreffe samtidig. Betegner vi hendingen ikke-A med iA og hendingen ikke-B med iB, kan vi skille mellom følgende fire sammensatte hendinger som kan inntreffe i en gjentakelse i U, nemlig:
1) både A og B inntreffer; vi betegner denne hendingen med AB
2) A inntreffer og B uteblir; vi betegner denne hendingen med AiB
3) A uteblir og B inntreffer; vi betegner denne hendingen med iAB
4) både A og B uteblir; vi betegner denne hendingen med iAiB.
Eksempel: ved en enkeltfødsel kan hendingene A = jente og iA = gutt inntreffe. Det kan bli født B = blåøyet barn eller iB = brunøyet barn. Ved en enkeltfødsel

kan derfor følgende fire sammensatte hendinger inntreffe:
1) AB = blåøyet jente, 2) AiB = brunøyet jente, 3) iAB = blåøyet gutt og 4) iAiB = brunøyet gutt.

La oss videre tenke oss at det er m like sannsynlige hendinger som kan inntreffe i en gjentakelse i U og at disse utelukker hverandre. Av disse er g_1 gunstige for AB, g_2 gunstige for AiB, g_3 gunstige for iAB og g_4 gunstige for iAiB. Vi behøver ikke diskutere hvilke hendinger dette er, og det spiller ingen rolle om vi i et aktuelt tilfelle ikke er i stand til å påvise dem.

I følge den klassiske definisjon er da sannsynlighetene for hver av de sammensatte hendinger følgende:

$$s(AB,U) = \frac{g_1}{m}, \quad s(AiB,U) = \frac{g_2}{m}, \quad s(iAB,U) = \frac{g_3}{m}, \text{ og } s(iAiB,U) = \frac{g_4}{m}$$

En har at

$$s(AB,U) + s(AiB,U) + s(iAB,U) + s(iAiB,U) = \frac{g_1 + g_2 + g_3 + g_4}{m} = \frac{m}{m} = 1.$$

Dette stemmer med at de fire sammensatte hendinger utelukker hverandre i en gjentakelse i U.

Vi skal nå se nærmere på s(AB,U). Dette er sannsynligheten for at A og B skal inntreffe sammen i en gjentakelse i U. En sier også at det er sannsynligheten for "både A og B". Vi kan skrive formelen på følgende måte:

$$s(AB,U) = \frac{g_1}{m} = \frac{g_1 + g_2}{m} \cdot \frac{g_1}{g_1 + g_2}$$

Da hendingen A inntreffer både i kombinasjonen AB og kombinasjonen AiB, er det $g_1 + g_2$ av de m mulige og like sannsynlige hendinger som medfører at A inntreffer. Følgelig er

$$\frac{g_1 + g_2}{m} = s(A,U)$$

For å finne ut hva den annen faktor i uttrykket for s(AB,U) betyr, kan vi tenke oss at vi undersöker alle gjentakelsene i U. I noen av disse gjentakelser må A inntreffe og i resten må A utebli. Vi kan derfor ved hjelp av hengingen A dele opp universet U i to subuniverser. I det ene av disse subuniverser som vi vil betegne med UA, tar vi med alle gjentakelsene som foruten de kjennetegn som er felles for alle gjentakelsene i U, også har A som kjennetegn. De m mulige og like sannsynlige hendinger kan vi også dele i to grupper. I den ene gruppen plasserer vi alle de av disse hendinger som medfører at A inntreffer; antallet av disse er $g_1 + g_2$. Dette blir da antallet av mulige og like sannsynlige hendinger som kan inntreffe i en gjentakelse i subuniverset UA. Hvis nemlig en av de andre av de m hendinger inntreffer, en av de som hører til

gruppen $g_3 + g_4$, vil jo A nødvendigvis utebli. Av de $g_1 + g_2$ mulige og like sannsynlige hendinger som kan inntreffe i en gjentakelse i UA, er det g_1 som medfører at B inntreffer. Følgelig er faktoren $\frac{g_1}{g_1 + g_2}$ sannsynligheten for B i en gjentakelse i subuniverset UA, eller:

$$\frac{g_1}{g_1 + g_2} = s(B,UA)$$

Vi har derfor at

$$\underline{s(AB,U) = s(A,U)\cdot s(B,UA)}$$

Oppgave 8. Bevis at vi også kan sette

$$s(AB,U) = s(B,U)\cdot s(A,UB)$$

Sannsynlighetene $s(B,UA)$ og $s(A,UB)$ kalles ofte betingete sannsynligheter. Den første, $s(B,UA)$, er sannsynligheten for at B skal inntreffe i en gjentakelse i U forutsatt, eller betinget av, at A inntreffer. Og $s(A,UB)$ er sannsynligheten for at A skal inntreffe i en gjentakelse i U forutsatt, eller betinget av, at B inntreffer.

La oss tenke oss at vi trekker ett kort av en kortstokk. Vi vil forutsette at de 52 kort er like sannsynlige hendinger ved denne begivenheten. Hva er da sannsynligheten for å trekke et herrekort i spar, dvs. et kort som er både herrekort og sparkort?

Denne oppgaven kan vi naturligvis løse uten å bruke både-og loven. Av de $m = 52$ mulige og like sannsynlige hendinger som kan inntreffe, er det $g = 3$ som er gunstige for herrekort i spar, nemlig spar-konge, spar-dame og spar-knekt. Altså er

$$s(\text{herrekort i spar},U) = \frac{g}{m} = \frac{3}{52}$$

Men vi kan også bruke både-og loven. Sannsynligheten for å trekke et sparkort er $s(\text{spar},U) = 13/52$. Sannsynligheten for å trekke et herrekort forutsatt at det kort en trekker er et sparkort, er

$$s(\text{herrekort},U\text{spar}) = 3/13$$

Når vi nemlig har trukket et kort og konstatert at det er et sparkort, er det bare 13 mulige og like sannsynlige hendinger som kan inntreffe (nemlig et av de 13 sparkort) og av disse er det 3 (spar-konge, spar-dame og spar-knekt) som er gunstige for hendingen herrekort. Altså er

$$s(\text{herrekort i spar},U) = s(\text{spar},U)\cdot s(\text{herrekort},U\text{spar}) = \frac{13}{52}\cdot\frac{3}{13} = \frac{3}{52}$$

I noen tilfelle er sannsynligheten for B i en gjentakelse i UA lik sannsynligheten for B i en gjentakelse i U, altså $s(B,UA) = s(B,U)$. Da er også $s(A,UB) = s(A,U)$. Vi har nemlig at

$$s(B,U)\cdot s(A,UB) = s(A,U)\cdot s(B,UA)$$

fordi begge disse produktene er lik $s(AB,U)$. Innsettes nå i denne ligningen $s(B,UA) = s(B,U)$, får vi $s(A,UB) = s(A,U)$.

At $s(B,UA) = s(B,U)$ betyr at de to hendingene A og B opptrer helt uavhengig av hverandre. Er f.eks. begivenheten en enkeltfødsel og A = jente og B = blåøyet barn, betyr likheten $s(B,U) = s(B,UA)$ at øyefargen ikke har noen som helst sammenheng med kjønnet. Eller omsatt til vår terminologi: Sannsynligheten for blåøyet barn i subuniverset av nyfødte jenter er den samme som sannsynligheten for blåøyet barn i universet av nyfødte barn.

Kriteriet på at A og B opptrer helt uavhengige av hverandre, er at $s(B,UA) = s(B,U)$ eller $s(A,UB) = s(A,U)$. I aktuelle tilfelle kjenner en ikke disse sannsynlighetene og en må derfor nøye seg med å sammenlikne deres estimater. En må da være oppmerksom på at det praktisk talt alltid er en forskjell mellom estimatene av to sannsynligheter. Først senere kan vi imidlertid vise hvordan en kan avgjøre om forskjellen er så stor at det må sluttes at også de to sannsynligheter er ulike.

Eksempel 1. Ved en bestemt krysning av bananfluen opptrer hos avkommet følgende karakterer:

A = normale børster, iA = reduserte børster
B = normale øyne, iB = reduserte øyne.

Etter en rekke krysninger som ialt ga 2835 avkom, fant en følgende antall avkom med de fire forskjellige karakterkombinasjoner:

AB	AiB	iAB	iAiB
1705	506	489	135

Estimatet av $s(B,U)$ er

$$\frac{1705 + 489}{2835} = \frac{2194}{2835} = 0,7739$$

og estimatet av $s(B,UA)$ er

$$\frac{1705}{1705 + 506} = \frac{1705}{2211} = 0,7711$$

Når en avrunder til to desimaler (0,77) har altså $s(B,UA)$ og $s(B,U)$ samme estimat. Dette beviser ikke, men det tyder på at disse to karakterene (hendingene) opptrer uavhengig av hverandre ved denne krysning.

Eksempel 2. For å få brakt på det rene om en bestemt vaksine beskytter mot angrep av tyfoidfeber, ble det utført et større forsøk. Av 18483 personer som en antok var utsatt for smitte, ble 6815 vaksinerte. Når en bruker betegnelsene

A = vaksinert, iA = ikke vaksinert
B = angrepet, iB = ikke angrepet

kan resultatet av dette forsøket sammenfattes i følgende oversikt:

AB	AiB	iAB	iAiB
56	6759	272	11396

Estimatet av S(B,U) er

$$\frac{56 + 272}{18483} = \frac{328}{18483} = 0{,}0177$$

Estimatet av s(B,UA) er

$$\frac{56}{56 + 6759} = \frac{56}{6815} = 0{,}0082$$

Siden det forekommer angrepne blandt de vaksinerte, er vaksinasjon ikke noen sikker forholdsregel mot sykdommen. Men hvis vi kan legge noen vekt på forskjellen mellom estimatene av s(B,UA) og s(B,U), - dvs. hvis det kan bevises at disse to sannsynligheter er ulike - kan vi slutte at vaksinasjon reduserer faren for angrep. Omsatt til vår terminologi vil dette si at sannsynligheten for angrep i det univers som omfatter alle personer utsatt for smitte, er større enn sannsynligheten for angrep i det univers som omfatter vaksinerte personer.

En slik slutning som dette bør imidlertid ikke godtas uten videre. En påvisning av at s(B,UA) og s(B,U) er ulike (vi forutsetter at dette er påvist for dette eksemplet) betyr ikke alltid at A og B er avhengig av hverandre. Det kan tenkes at subuniverset UA stort sett faller sammen med et subunivers UC og at forskjellen mellom s(B,U) og s(B,UA) i virkeligheten betyr at det er B og C som ikke opptrer uavhengig av hverandre.

Hvis f.eks. forsøket er planlagt og utført på en slik måte at de vaksinerte personer er fortrinsvis utvalt blandt de økonomisk mer velstående personer som må forutsettes å ha bedre anledning til å ta hygieniske forholdsregler enn de økonomisk mindre velstilte, kan forskjellen skyldes dette uheldige valg av forsøkspersoner og ikke ha det aller minste med vaksinasjonen å gjøre. Vi er imidlertid med disse merknader kommet inn på et helt annet område av vitenskapelig metodelære, nemlig forsøksplanleggingen, og må derfor her nøye oss med en kort antydning.

Vi skal senere vise hvordan vi kan avgjøre om en forskjell mellom estimatene av to sannsynligheter betyr at sannsynlighetene selv er ulike.

Oppgave 9. La A = spar og B = ess. En trekker ett kort av en kortstokk. Vis at s(B,UA) = s(B,U). (De 52 kortene forutsettes å være like sannsynlige hendinger når en trekker ett kort.)

Oppgave 10. Beregn på grunnlag av observasjonene i eks. 1 og eks. 2 estimatene av følgende sannsynligheter:

s(reduserte øyne,U)
s(normale øyne,U reduserte børster)
s(reduserte øyne,U reduserte børster)
s(ikke angrepet,U)
s(angrepet,U ikke vaksinert)
s(ikke angrepet,U ikke vaksinert).

Oppgave 11. En trekker to kort av en kortstokk. Dette kan gjøres på to ulike måter:
1) en trekker to kort under ett
2) en trekker først ett kort, noterer fargen, legger kortet tilbake i kortstokken, blander og trekker så kort nr. 2.
Finn for hvert av disse tilfelle sannsynligheten for å trekke 2 spar. (De 52 kort forutsettes å være like sannsynlige hendinger.)

La oss nå tenke oss at det er tre hendinger - A,B og C - som kan inntreffe i en gjentakelse i universet U. Hva er da sannsynligheten for at de alle tre skal inntreffe i samme gjentakelse? Vi kan bevise at

$$s(ABC,U) = s(A,U) \cdot s(B,UA) \cdot s(C,UAB)$$

Vi kan nemlig danne et subunivers i U hvor gjentakelsene foruten ved de kjennetegn som gjelder alle gjentakelsene i U, er karakterisert ved at både A og B har inntruffet. La dette universet være UAB. Sannsynligheten for at C skal inntreffe i en gjentakelse i dette universet er s(C,UAB). Sannsynligheten for den sammensatte hending ABC i en gjentakelse i U er derfor

$$s(ABC,U) = s(AB,U)\cdot s(C,UAB)$$

Men nå er jo

$$s(AB,U) = s(A,U)\cdot s(B,UA)$$

og innsettes dette, får vi nettopp den refererte formel for s(ABC,U).

Hvis nå A og B opptrer uavhengig av hverandre, er s(B,UA) = s(B,U). Og hvis C opptrer uavhengig av den sammensatte hending AB, er s(C,UAB) = s(C,U). Under disse forutsetninger er derfor

$$s(ABC,U) = s(A,U)\cdot s(B,U)\cdot s(C,U)$$

Denne setningen kan uten videre utvides til å gjelde et vilkårlig antall hendinger A,B,C,D,E,..... Opptrer de alle uavhengig av hverandre, er

$$s(ABCDE....,U) = s(A,U)\cdot s(B,U)\cdot s(C,U)\cdot s(D,U)\cdot s(E,U).....$$

Forutsetningen for denne setningen er imidlertid ikke bare den at hendingene skal opptre uavhengig av hverandre parvis. Hver enkelt av dem må opptre uavhengig av en hvilket som helst kombinasjon av de andre. Hendingen C må således - for å nevne et eksempel - opptre uavhengig av den sammensatte hending ABDE.....

Oppgave 12. Skriv opp formelen for s(ABCD,U) når det ikke gjøres noen forutsetning om uavhengighet.

Oppgave 13. Anta at sannsynligheten for at det ved en enkeltfødsel blir født en gutt er 0,52, sannsynligheten for at det blir født et levende barn er 0,98 og at sannsynligheten for at det blir født et blåøyet barn er 0,65. Anta videre at hendingene "gutt", "levende barn", og "blåøyet barn" opptrer som uavhengige hendinger ved en enkeltfødsel. Finn da sannsynlighetene for

1) at det blir født en levende blåøyet gutt
2) at det blir født en levende brunøyet jente
3) at det blir født en dødfødt blåøyet jente
4) at det blir født en dødfødt brunøyet gutt

(Blåøyet og brunøyet er motsatte hendinger. Disse forutsetninger om uavhengighet er visstnok ikke i overensstemmelse med virkeligheten, men det er en annen sak).

Oppgave 14. La A, B og C betegne følgende hendinger:

A = ektefødt barn	iA = uektefødt barn
B = levendefødt barn	iB = dødfødt barn
C = gutt	iC = jente

I Statistisk Årbok for Norge, 58. årg. 1939, side 25, finnes følgende observasjoner for året 1937.

Levende fødte		Derav utenfor ekteskap		Dødfødte		Derav utenfor ekteskap	
gutt	jente	gutt	jente	gutt	jente	gutt	jente
22563	21245	1409	1360	580	423	41	34

Et relativt meget lite antall av disse barn er tvillinger, trillinger osv. Men vi vil anta at alle er født ved enkeltfødsler.

Anta at de sannsynligheter som kan estimeres på grunnlag av disse observasjoner, er eksakt lik estimatene avrundet til 2 desimaler. Undersøk under denne forutsetning om A,B og C opptrer som uavhengige hendinger ved begivenheten enkeltfødsel.

5. B å d e - o g l o v e n II.

I foregående avsnitt har vi funnet sannsynligheten for "både A og B" under den forutsetning at A og B kan inntreffe i en gjentakelse i samme univers. Vi vil nå tenke oss at A kan inntreffe i en gjentakelse i universet U og B i en gjentakelse i universet U'. Vi skal da vise at hvis A og B opptrer uavhengige av hverandre, er sannsynligheten for at A skal inntreffe i en gjentakelse i U og B i en gjentakelse i U' lik produktet $s(A,U)\cdot s(B,U')$. Med dette menes at hvis vi har en gjentakelse i U og en gjentakelse i U', er sannsynligheten for at både A og B har inntruffet lik dette produktet.

For å bevise dette, vil vi benytte vår opprinnelige definisjon av $s(A,U)$ og $s(B,U')$. Vi tenker oss at vi har observert n gjentakelser i universet U og funnet at A har inntruffet i $h(A,U)$ av disse. I universet U' har vi observert n' gjentakelser og funnet at B har inntruffet i $h(B,U')$ av disse. Da er

$$s(A,U) = \text{Gr.}\ \frac{h(A,U)}{n} \qquad \text{når } n \to \infty$$

og

$$s(B,U') = \text{Gr.}\ \frac{h(B,U')}{n'} \qquad \text{når } n' \to \infty$$

Vi vil nå danne et nytt univers hvor hver gjentakelse består av en gjentakelse i U og en gjentakelse i U'. La dette universet være W. Av de n gjentakelser i U og de n' gjentakelser i U' kan det da dannes n.n' gjentakelser i W fordi hver gjentakelse i U skal stilles sammen med hver gjentakelse i U'. Antallet av disse n.n' gjentakelser der A har inntruffet i U og B i U', er $h(A,U)\cdot h(B,U')$. Hver gjentakelse i U der A har inntruffet, skal nemlig stilles sammen med hver gjentakelse i U' der B har inntruffet.

Sannsynligheten for at den sammensatte hending AB skal inntreffe i en ny gjentakelse i W er da i følge definisjonen lik

$$s(AB,W) = \text{Gr.}\ \frac{h(A,U)\cdot h(B,U')}{n\cdot n'} = \text{Gr.}\ \frac{h(A,U)}{n} \cdot \text{Gr.}\ \frac{h(B,U')}{n'} = s(A,U)\cdot s(B,U')$$

Det er forutsatt her at n og n' hver for seg og uavhengig av hverandre skal økes over alle grenser. Videre er det forutsatt at Gr. $\frac{h(B,U')}{n'}$ er uavhengig av $\frac{h(A,U)}{n}$ og at Gr. $\frac{h(A,U)}{n}$ er uavhengig av $\frac{h(B,U')}{n'}$. Vi forutsetter m.a.o. at utvalget på n gjentakelser i U ikke utvelger et subunivers i U' der B forekommer relativt oftere eller sjeldnere enn alminnelig i U', og at utvalget på n' gjentakelser i U' ikke utvelger et subunivers i U der A forekommer relativt oftere eller sjeldnere enn alminnelig i U.

Skal vi derfor bruke denne setningen i et aktuelt tilfelle, må vi ha sikkerhet for at A og B opptrer uavhengig av hverandre. At A har inntruffet i en gjentakelse i U må altså ikke ha noen betydning for sannsynligheten for at B skal inntreffe i en gjentakelse i U', og omvendt. Hvis denne forutsetningen ikke er realisert, må vi søke sannsynligheten for AB i et annet univers enn det univers W som vi har operert med foran. Vi skal ta for oss et eksempel for å forklare dette nærmere.

La oss anta at to nordmenn P og Q fyller år i dag. P er 50 år og Q 20 år. Hva er sannsynligheten for at de begge er døde om ett år? Dødssannsynligheten for P må settes lik dødssannsynligheten for en person som tilhører det univers som svarer til gruppen (utvalget) 50-årige norske menn. Og dødssannsynligheten for Q må settes lik dødssannsynligheten for en person som tilhører det univers som svarer til gruppen 20-årige norske menn. Disse sannsynligheter er omtrent 0,0091 for P og 0,0058 for Q. Sannsynligheten for at både P og Q er døde om ett år er derfor lik

$$0{,}0091 \cdot 0{,}0058 = 0{,}00005$$

Denne beregningen forutsetter at det at P dør ikke endrer dødssannsynligheten for Q og det at Q dør ikke endrer dødssannsynligheten for P. Men sett nå at P og Q er far og sønn. Da er det ikke sikkert at denne forutsetningen er realisert.

Vi nevner dette eksemplet bare for å vise at selv om A i U og B i U' i sin alminnelighet opptrer uavhengig av hverandre, kan det finnes en forbindelse av et eller annet slag mellom den aktuelle gjentakelse i U og den aktuelle gjentakelse i U' som virker slik at A og B i dette bestemte tilfelle ikke er uavhengige av hverandre. Dette må ikke forstås slik at enhver forbindelse mellom de to gjentakelser har denne virkning. Men hvis det eksisterer en slik forbindelse, bør spørsmålet om uavhengighet undersøkes. En må da bestemme sannsynligheten for AB i et annet univers enn det vi har benyttet foran. Gjelder det dødssannsynligheten for både far og sønn, må en bestemme sannsynligheten for denne sammensatte hending i et univers hvor gjentakelsene består av far og sønn. Har en observert n gjentakelser av 50-årige norske fedre og deres 20-

årige sønner og funnet at både far og sønn er døde etter ett år i h av disse gjentakelser, er h/n det riktige estimat av dødssannsynligheten for 50-årig far og 20-årig sønn.

Har en nå bestemt sannsynligheten for AB på denne måten (vi tenker ikke da spesielt på dødssannsynligheten for far og sønn), kan det hende at denne sannsynligheten er lik produktet $s(A,U)\cdot s(B,U')$. Og dette ville da bety at den forbindelse som eksisterer mellom den aktuelle gjentakelse i U og den aktuelle gjentakelse i U', ikke skaper noen avhengighet mellom A og B.

Oppgave 15. Anta at de sannsynligheter som kan estimeres på grunnlag av observasjonene i oppg. 14, er eksakt lik estimatene avrundet til 2 desimaler.

Sett at en gift kvinne P og en ugift kvinne Q har født ett barn hver. Beregn da sannsynlighetene for

1) at begge kvinner har født levende barn
2) at begge kvinner har født dødfødte barn
3) at P har født levende barn og Q dødfødt barn.

Kan det tenkes tilfelle der både-og loven i den utformingen vi har gitt den i dette avsnitt, ikke kan brukes til beregning av sannsynlighetene for disse sammensatte hendingene?

La oss nå tenke oss at A kan inntreffe i en gjentakelse i universet U,B i en gjentakelse i universet U',C i en gjentakelse i universet U", D i en gjentakelse i universet U"' osv. Hvis disse hendinger opptrer uavhengig av hverandre, er sannsynligheten for at den sammensatte hending ABCD..... skal inntreffe i en gjentakelse i et univers W hvor hver gjentakelse består av en gjentakelse i hver av universene U,U',U",U"' ..... , lik

$$s(ABCD....,W) = s(A,U)\cdot s(B,U')\cdot s(C,U")\cdot s(D,U"')\cdot.....$$

Hvis en altså har en bestemt gjentakelse i U, en bestemt gjentakelse i U', en bestemt gjentakelse i U", en bestemt gjentakelse i U"' osv., er sannsynligheten for at både A,B,C,D,..... har inntruffet lik dette produktet. Det forutsettes da at det mellom disse gjentakelser ikke finnes noen slags forbindelser som virker slik at A,B,C,D,..... ikke opptrer uavhengig av hverandre.

Oppgave 16. Bevis at

$$s(ABC,W) = s(A,U)\cdot s(B,U')\cdot s(C,U")$$

når A,B og C opptrer uavhengig av hverandre.

Det er naturligvis ikke noe i veien for at universene U,U',U"..... kan være identiske universer og at hendingene A,B,C,.... kan være samme hending. Da har en:

$$s(AAA...,W) = s(A,U)\cdot s(A,U)\cdot s(A,U)....$$

Sannsynligheten for at en 50-år gammel mann skal dø før han fyller 60 år er omtrent 0,12 (Norge). Sannsynligheten for at 4 50 år gamle menn skal dø før noen av dem fyller 60 år, er derfor lik $0,12^4 = 0,0002$. Forutsetningen for dette er

at det at en av dem dør, ikke endrer dødssannsynligheten for noen av de andre.

Det er klart at 1 gjentakelse i hver av n identiske universer er det samme som n uavhengige gjentakelser i ett av disse universer. Sannsynligheten for at hendingen A skal inntreffe i alle n uavhengige gjentakelser i universet U, er derfor

$$s(\text{A n ganger, W}) = [s(A,U)]^n$$

W er her et univers hvor hver gjentakelse består av n uavhengige gjentakelser i U.

Sannsynligheten for å trekke spar når en trekker ett kort av en kortstokk er $s(\text{spar},U) = 1/4$ (forutsatt at de 52 kort er like sannsynlige hendinger). Sannsynligheten for å trekke $n = 5$ spar når en trekker ett kort av hver av $n = 5$ kortstokker (eller 5 kort av samme kortstokk når det uttrukne kort legges tilbake og blandes med de andre kort før neste trekning) er derfor lik

$$s(5 \text{ spar},W) = [s(\text{spar},U)]^5 = (\tfrac{1}{4})^5 = \frac{1}{1024}$$

Oppgave 17. En trekker ett kort av hver av 5 kortstokker. Finn sannsynligheten for at en skal få spar og ikke-spar i rekkefølgen

spar, spar, ikke-spar, spar, ikke-spar.

Oppgave 18. Anta at sannsynligheten for at det ved en enkeltfødsel blir født en gutt er lik 0,52. Finn sannsynligheten for at en kvinne i 5 enkeltfødsler skal få
1) 5 gutter
2) 5 jenter
3) 2 gutter og 3 jenter i rekkefølgen jente, gutt, jente, gutt, jente.
Kan denne oppgaven løses uten at en tar visse forbehold?

6. Binomialloven I.

La sannsynligheten for at hendingen A skal inntreffe i en gjentakelse i universet U være $s(A,U) = p$. Sannsynligheten for at den motsatte hending iA skal inntreffe er da $s(iA,U) = 1 - s(A,U) = 1 - p = q$. Hva er da sannsynligheten for at A skal inntreffe i z og utebli i n - z av n uavhengige gjentakelser i U ?

Eksempel : la oss tenke oss at vi gjentar på uavhengig måte trekningen av ett kort av en kortstokk $n = 10$ ganger idet vi etter hver trekning legger det uttrukne kort tilbake og blander godt før neste trekning. Hva er da sannsynligheten for at vi skal få $z = 4$ spar ?

La W være et univers hvor hver gjentakelse består av n uavhengige gjentakelser i U. I en gjentakelse i W er det da n + 1 mulige hendinger som kan inntreffe og som utelukker hverandre, nemlig at hendingen $\bar{A}$ inntreffer $z=0, z=1, z=2, \ldots\ldots, z=n$ ganger. Sannsynligheten for at A skal inntreffe i z av n uavhengige gjentakelser i U, er derfor identisk med sannsynligheten

for hendingen z i en gjentakelse i W. Vi skal vise at denne sannsynligheten - s(z,W) - er lik det alminnelige ledd i utviklingen av $(q+p)^n$ etter binomialformelen (se I, punkt 5), altså at

$$s(z,W) = \frac{n!}{z!(n-z)!} p^z q^{n-z}$$

Sannsynligheten for $z = 0$, $z = 1$, $z = 2$, ..., $z = n$, nemlig $s(0,W)$, $s(1,W)$, $s(2,W)$, ... $s(n,W)$, finnes derfor ved etter hver å innsette $z = 0$, $z = 1$, $z = 2$, ... $z = n$ i denne formelen.

Oppgave 19. En trekker ett kort av hver av $n = 3$ kortstokker. Beregn sannsynligheten for å trekke $z = 0$, $z = 1$, $z = 2$ og $z = 3$ spar ? (De 52 kort i hver stokk forutsettes å være like sannsynlige hendinger.)

Vi skal begynne med å utlede denne formelen for $n = 2$ og $n = 3$.

$n = 2$.

I $n = 2$ uavhengige gjentakelser i U, dvs. i en gjentakelse i W hvor hver gjentakelse består av $n = 2$ uavhengige gjentakelser i U, kan følgende sammensatte hendinger inntreffe:

1) iA i første og iA i annen gjentakelse, eller iAiA
2) iA i første og A i annen gjentakelse, eller iAA
3) A i første og iA i annen gjentakelse, eller AiA
4) A i første og A i annen gjentakelse, eller AA.

Etter både-og loven er sannsynlighetene for disse sammensatte hendinger i en gjentakelse i W lik :

$$s(iAiA,W) = s(iA,U)\cdot s(iA,U) = q\cdot q = q^2$$
$$s(iAA,W) = s(iA,U)\cdot s(A,U) = q\cdot p$$
$$s(AiA,W) = s(A,U)\cdot s(iA,U) = p\cdot q$$
$$s(AA,W) = s(A,U)\cdot s(A,U) = p\cdot p = p^2$$

Vi ser derfor at sannsynligheten for $z = 0$ er lik q^2 og sannsynligheten for $z = 2$ er lik p^2. Hendingen $z = 1$ kan inntreffe på to ulike måter: enten som iAA eller som AiA. Etter enten-eller loven er derfor sannsynligheten for $z = 1$ lik summen $s(iAA,W) + s(AiA,W) = 2pq$. Vi har derfor at

$$s(0,W) = q^2$$
$$s(1,W) = 2pq$$
$$s(2,W) = p^2$$

Disse tre sannsynligheter er lik leddene i utviklingen av $(q+p)^2$ etter binomialformelen. Vi har nemlig at

$$(q+p)^2 = q^2 + 2pq + p^2$$

Følgelig er for $n = 2$

$$s(z,W) = \frac{2!}{z!(2-z)!} p^z q^{2-z}$$

Da q+p=1, er $s(0,W) + s(1,W) + s(2,W) = 1$. Dette stemmer med at ingen andre hendinger enn $z = 0$, $z = 1$ og $z = 2$ kan inntreffe i en gjentakelse i W og at disse utelukker hverandre.

$\underline{n = 3}$.

Vi benytter samme fremgangsmåte som i foregående tilfelle og finner:

$$\begin{aligned}
s(iAiAiA,W) &= q\cdot q\cdot q = q^3 \\
s(iAiAA,W) &= q\cdot q\cdot p = pq^2 \\
s(iAAiA,W) &= q\cdot p\cdot q = pq^2 \\
s(AiAiA,W) &= p\cdot q\cdot q = pq^2 \\
s(iAAA,W) &= q\cdot p\cdot p = p^2q \\
s(AiAA,W) &= p\cdot q\cdot p = p^2q \\
s(AAiA,W) &= p\cdot p\cdot q = p^2q \\
s(AAA,W) &= p\cdot p\cdot p = p^3
\end{aligned}$$

I en gjentakelse i W som består av $n = 3$ uavhengige gjentakelser i U, kan en av hendingene $z = 0$, $z = 1$, $z = 2$ og $z = 3$ inntreffe. Ved å benytte enten-eller loven finner vi lett at sannsynlighetene for disse hendinger er:

$$\begin{aligned}
s(0,W) &= q^3 \\
s(1,W) &= 3pq^2 \\
s(2,W) &= 3p^2q \\
s(3,W) &= p^3
\end{aligned}$$

Vi ser at disse sannsynligheter er lik leddene i utviklingen av $(q+p)^3$ etter binomialformelen. Følgelig er sannsynligheten for hendingen z i en gjentakelse i W, dvs. sannsynligheten for at A skal inntreffe i z av $n = 3$ uavhengige gjentakelser i U, lik

$$s(z,W) = \frac{3!}{z!(3-z)!}\, p^z q^{3-z}$$

Når $n = 2$ og $n = 3$ er altså sannsynligheten for at A skal inntreffe i z av n uavhengige gjentakelser i universet U, gitt ved det alminnelige ledd i utviklingen av $(q+p)^n$ etter binomialformelen. La oss anta at dette også er riktig når $n = k$, dvs. at

$$s(z,W) = \frac{k!}{z!(k-z)!}\, p^z q^{k-z}$$

hvor hver gjentakelse i W består av $n = k$ uavhengige gjentakelser i U. Vi kan bevise at hvis denne forutsetningen er riktig, er også

$$s(z,W') = \frac{(k+1)!}{z!(k+1-z)!}\, p^z q^{k+1-z}$$

hvor hver gjentakelse i W' består av $n = k + 1$ uavhengige gjentakelser i U.

Hendingen A kan inntreffe i z av n = k + 1 uavhengige gjentakelser i U på to forskjellige måter:

1) den kan inntreffe i z av de k første gjentakelser og uteblik i den (k+1)'te gjentakelse, eller

2) den kan inntreffe i z-1 av de k første gjentakelser og inntreffe i den (k+1)'te gjentakelse.

Noen andre muligheter foreligger ikke. Og følgelig må sannsynligheten for at A skal inntreffe i z av n = k + 1 uavhengige gjentakelser i U være lik summen av sannsynlighetene for hver av disse to alternativer. I følge våre forutsetninger er sannsynligheten for det første alternativ lik

$$s(z,W)\cdot s(iA,U) = \frac{k!}{z!(k-z)!}\, p^z q^{k-z} \cdot q = \frac{k!}{z!(k-z)!}\, p^z q^{k+1-z}$$

hvor altså en gjentakelse i W består av n = k uavhengige gjentakelser i U. Sannsynligheten for det annet alternativ er etter samme regel lik

$$s(z-1,W)\cdot s(A,U) = \frac{k!}{(z-1)!(k-z+1)!}\, p^{z-1} q^{k-z+1} \cdot p$$

$$= \frac{k!}{(z-1)!(k+1-z)!}\, p^z q^{k+1-z}$$

Følgelig er etter enten-eller loven

$$s(z,W') = s(z,W)\cdot s(iA,U) + s(z-1,W)\cdot s(A,U)$$

$$= \left[\frac{k!}{z!(k-z)!} + \frac{k!}{(z-1)!(k+1-z)!}\right] p^z q^{k+1-z}$$

$$= \frac{(k+1)!}{z!(k+1-z)!}\, p^z q^{k+1-z}$$

Oppgave 20. Bevis at

$$\frac{k!}{z!(k-z)!} + \frac{k!}{(z-1)!(k-z+1)!} = \frac{(k+1)!}{z!(k+1-z)!}$$

Vi har altså bevist at hvis

$$s(z,W) = \frac{n!}{z!(n-z)!}\, p^z q^{n-z}$$

er riktig for n = k, er formelen også riktig for n = k + 1. Vi vet nå at formelen er riktig for n = 3. Følgelig er den også riktig for n = 3 + 1 = 4. Og siden den er riktig for n = 4, er den også riktig for n = 5. Siden den er riktig for n = 5, er den også riktig for n = 6 osv. Formelen er m.a.o. riktig i sin alminnelighet.

Da ingen andre hendinger enn z = 0, z = 1, z = 2, ... z = n kan inntreffe i en gjentakelse i W, skal summen av sannsynlighetene for hver av disse

hendinger være lik enheten. At den funne formel for $s(z,W)$ tilfredsstiller denne fordring er lett å bevise. Vi har nemlig at

$$\sum_{z=0}^{n} s(z,W) = \sum_{z=0}^{n} \frac{n!}{z!(n-z)!} p^z q^{n-z} = (q+p)^n = 1^n = 1$$

Oppgave 21. Anta at sannsynligheten for at det ved en enkeltfødsel skal bli født en gutt er lik 0,52. Finn sannsynlighetene for de forskjellige fordelinger av gutter og jenter i en barneflokk på $n = 5$ barn forutsatt at det ikke forekommer tvillinger, trillinger osv. (Det forutsettes at de $n = 5$ enkeltfødsler er uavhengige gjentakelser.) Gi en grafisk fremstilling (søylediagram) av $s(z,W)$ som funksjon av z.

I en gjentakelse i det univers W hvor hver gjentakelse består av n uavhengige gjentakelser i universet U, kan en av følgende hendinger inntreffe: $z = 0$, $z = 1$, $z = 2$, ..., $z = n$. Vi skal nå stille oss som oppgave å undersøke hvilken av disse n+1 hendinger det er som er mest sannsynlig. La dette være $z = a$. Sannsynligheten for denne hending er $s(a,W)$. At $z = a$ er den mest sannsynlige hending betyr at $z = a$ er den verdi av z som gjør $s(z,W)$ til et maksimum. Det betyr også at $z = a$ er det sannsynligste antall ganger hendingen A skal inntreffe i n uavhengige gjentakelser i universet U.

La oss imidlertid først undersøke hvordan det går med $s(z,W)$ når z gjennomløper tallverdiene $0,1,2,.....,n$. Dette kan vi få rede på ved å undersøke differensen $s(z+1,W)-s(z,W)$. Vi har at

$$s(z+1,W) = \frac{n!}{(z+1)!(n-z-1)!} p^{z+1} q^{n-z-1} = \frac{p(n-z)}{q(z+1)} s(z,W)$$

Følgelig er

$$s(z+1,W)-s(z,W) = \frac{p(n-z)}{q(z+1)} s(z,W)-s(z,W) = \frac{np-q-z}{q(z+1)} s(z,W)$$

Da nå p,q,n,z og $s(z,W)$ er positive tall, ser vi at

$$s(z+1,W) > s(z,W) \quad \text{for alle} \quad z < np-q$$
$$s(z+1,W) < s(z,W) \quad \text{for alle} \quad z > np-q$$
$$s(z+1,W) = s(z,W) \quad \text{for} \quad z = np-q$$

Den siste ligningen har ingen mening med mindre np-q er et helt tall. Men hvis np-q er et helt tall, er det to naboverdier, nemlig np-q og np-q+1 = np+p, som er like sannsynlige og sannsynligere enn alle andre verdier av z. Vi ser videre at når $z < np-q$ er $s(z,W)$ voksende for voksende verdier av z, og når $z > np-q$ er $s(z,W)$ avtakende for voksende verdier av z. Hvis derfor np-q ikke er et helt tall, kan det være bare en verdi av z ($z = a$) som er sannsynligere enn de andre verdier. Vi skal vise at $z = a$ er det ene hele tallet mellom np-q og np+p.

Hvis nemlig $z = a$ er den sannsynligste verdi av z, må

$$s(a-1,W) < s(a,W) > s(a+1,W)$$

Fra før har vi at

$$s(a+1,W) = \frac{p(n-a)}{q(a+1)} s(a,W)$$

og vi finner lett at (bevis dette)

$$s(a-1,W) = \frac{qa}{p(n-a+1)} s(a,W)$$

Følgelig er

$$\frac{qa}{p(n-a+1)} s(a,W) < s(a,W) > \frac{p(n-a)}{q(a+1)} s(a,W)$$

Av venstre ulikhet finner vi at

$$qa < p(n-a+1)$$

eller at

$$\underline{a < np + p}$$

Av høyre ulikhet finnes at

$$q(a+1) > p(n-a)$$

eller at

$$\underline{a > np - q}$$

Til bestemmelse av a (den sannsynligste verdi av z) har vi derfor

$$\underline{np-q < a < np+p}$$

Differensen mellom disse to grensene er

$$(np+p) - (np-q) = p+q = 1$$

Den sannsynligste verdi av z er derfor lik det ene hele tallet mellom tallene np-q og np+p. Er np-q og np+p hele tall, er disse verdier av z som vi allerede har sett, like sannsynlige og sannsynligere enn alle andre verdier av z.

La oss anta at sannsynligheten for at det ved en enkeltfødsel skal bli født en gutt er $p = 0,52$. Det sannsynligste antall gutter i en barneflokk på $n = 5$ barn hvor det ikke forekommer tvillinger eller trillinger osv. er altså det ene hele tallet mellom

$$np-q = 5.0,52 - 0,48 = 2,12$$

og

$$np+p = 5.0,52 + 0,52 = 3,12$$

Det sannsynligste antall gutter er altså $a = 3$ (sml. oppg. 21).

Sannsynligheten for den sannsynligste hending, $z = a$, er naturligvis

$$s(a,W) = \frac{n!}{a!(n-a)!} p^a q^{n-a}$$

Sannsynligheten for det sannsynligste antall gutter, $a = 3$, i en barneflokk på $n = 5$ barn (forutsetninger som ovenfor) er derfor lik

$$s(3,W) = \frac{5!}{3!(5-3)!} 0,52^3 \cdot 0,48^2 = 0,32$$

Når n er et stort tall, kan en ikke bruke formelen for $s(z,W)$ til beregning av

$s(a,W)$. Da blir det nemlig uoverkommelig å beregne n!, a! og (n-a)! En må derfor bruke en tilnærmingsformel. Denne er:

$$s(a,W) = \frac{1}{\sqrt{2\pi npq}}$$

La oss tenke oss at vi trekker ett kort av hver av n = 1000 kortstokker. (Vi forutsetter at de 52 kort er like sannsynlige hendinger når vi trekker ett kort av en kortstokk.) Hva er da det sannsynligste antall (a) sparkort, og hvor stor er sannsynligheten for dette sannsynligste antall ? Vi har at $s(\text{spar},U) = p = \frac{1}{4}$. Det sannsynligste antall spar ligger mellom grensene

$$np-q = 1000 \cdot 0,25 - 0,75 = 249,25$$

og

$$np+p = 1000 \cdot 0,25 + 0,25 = 250,25$$

Det sannsynligste antall sparkort er altså a = 250. Sannsynligheten for dette antall er

$$s(a,W) = s(250,W) = \frac{1}{\sqrt{2\pi npq}} = \frac{1}{\sqrt{2\pi 1000 \cdot 0,25 \cdot 0,75}} = 0,029$$

Vi ser altså at sannsynligheten for det sannsynligste antall er meget liten. Hva er grunnen til det ?

Oppgave 22. La $s(A,U) = p$ være sannsynligheten for at hendingen A skal inntreffe i en gjentakelse i universet U. Finn det sannsynligste antall ganger A vil inntreffe i n uavhengige gjentakelser i U og beregn sannsynligheten for det sannsynligste antall i følgende tilfelle:

1)	$n = 7$	og	$p = 1/2$		4)	$n = 1000$	og	$p = 1/2$
2)	$n = 7$	og	$p = 1/3$		5)	$n = 1000$	og	$p = 1/3$
3)	$n = 7$	og	$p = 1/10$		6)	$n = 1000$	og	$p = 1/10$

Oppgave 23. Ved en bestemt krysning av bananfluen kan avkommet få B = normale øyne eller iB = reduserte øyne. Anta at $s(B,U) = p = 0,75$. Hva er det sannsynligste antall avkom med normale øyne i et kull på n = 500 avkom og hvor stor er sannsynligheten for dette sannsynligste antall ?

7. Noen tilleggsmerknader.

Når en skal prøve å gi en fremstilling av et eller annet vanskelig emne, hender det at de pedagogiske og de rent faglige hensyn kommer i konflikt med hverandre. En kan bli stilt overfor et valg mellom en stringent fremstilling og en fremstilling som først og fremst tar sikte på å gi en mest mulig lettfattelig innføring i emnet. Den fremstilling som er gitt i de foregående avsnitt av sannsynlighetsregningens enkleste setninger, er et resultat av en avveining av disse to motstridende hensyn. Fremstillingen er til stadighet under-

støttet av eksempler og oppgaver. Sett fra et pedagogisk synspunkt er denne metoden sikkert bra, men den har den feil at leseren gjerne fester seg mer ved de spesielle eksempler enn ved den generelle tankegang. Av den grunn er det kanskje nødvendig ved noen alminnelige merknader å rydde bort eventuelle nærliggende misforståelser før vi går videre.

Det er for det første en mulighet for at en har fått det inntrykk at tidsmomentet spiller en rolle i sannsynlighetsregningen. Dette er ikke riktig. Men inntrykket kan ha festnet sdg fordi oppgavene som oftest har fått omtrent følgende form: en hending A kan inntreffe ved en nærmere definert begivenhet, hva er sannsynligheten for at den skal inntreffe? Det er selvsagt mange tilfelle der dette er en naturlig og riktig problemstilling. Men som oftest kan problemet formuleres på flere måter som er like riktige og treffende. Sett at det gjelder trekning av ett kort av en kortstokk og hendingen er sparkort. Vi kan da velge mellom bl.a. følgende formuleringer :

1) Vi trekker ett kort av en kortstokk. Hva er sannsynligheten for å trekke et sparkort?

2) Vi har trukket ett kort av en kortstokk. Hva er sannsynligheten for at det uttrukne kort er sparkort?

Disse to formuleringer gir uttrykk for nøyaktig samme tanke.

Også når det gjelder den betingete sannsynlighet s(B,UA) kan vi velge mellom flere forskjellige formuleringer. Vi må alltid begynne med å definere gjentakelsen i universet U, og deretter kan vi formulere oppgaven f.eks. slik:

1) Hva er sannsynligheten for at B skal inntreffe i en gjentakelse i U forutsatt at A inntreffer?

2) Det er konstatert at A har inntruffet i en gjentakelse i U. Hva er da sannsynligheten for at også B har inntruffet i denne gjentakelse?

3) Det ble konstatert at A var til stede. Hva er sannsynligheten for at også B var til stede?

Disse tre spørsmål med betingelser kan gjerne være spørsmål angående akkurat samme sak. I første tilfelle tenker en seg at begivenheten (gjentakelsen) er fremtidig, i annet tilfelle at den er nåtidig og i siste tilfelle at den er fortidig. Men det kan i alle tre tilfelle være tale om samme begivenhet og samme hending.

Vi har før understreket at en begivenhet er definert ved et antall kjennetegn. Det samme gjelder imidlertid også det vi har kalt en hending. Også denne må i hvert enkelt tilfelle være definert ved et eller flere kjennetegn. I stedet for å tale om begivenheter og hendinger, kunne vi derfor tale om kjennetegnssett. Og da ville det være mulig å formulere oppgavene i mer generelle

uttrykk, f.eks. slik: Hva er sannsynligheten for at de kjennetegn som definerer hendingen B, er (var, kommer til å være) til stede når (dvs. hver gang) de kjennetegn som definerer vedkommende begivenhet er (var, kommer til å være) til stede?

Alle sannsynligheter er derfor betingete sannsynligheter. En kan si at sannsynligheten for en hending er identisk med sannsynligheten for at dens kjennetegn er til stede når det er konstatert at begivenhetens kjennetegn er til stede. Sannsynligheten s(B,U) er altså identisk med sannsynligheten for at B's kjennetegn er til stede når de kjennetegn er til stede som definerer gjentakelsene i universet U. På samme måte er s(B,UA) lik sannsynligheten for at B's kjennetegn er til stede når både de kjennetegn som definerer gjentakelsene i universet U og A's kjennetegn er til stede.

D R. P E R O T T E S T A D

F o r e l e s n i n g e r

over

M A T E M A T I K K o g S T A T I S T I K K

ved

NORGES LANDBRUKSHØGSKOLE

III. Sannsynlighetsregning (2. del).

Skrivemaskinstua
Oslo i mars 1947

Innhold:

III. Sannsynlighetsregning (2. del).

- - -

8. Variable kjennetegn.

Vi har tidligere (II) forklart at det er to slags variable kjennetegn, diskrete og kontinuerlige. Er kjennetegnet diskret, kan det ha bare bestemte atskilte tallverdier (sml. eks. 2, eks. 6 og eks. 7, II). Er det kontinuerlig, kan det ha en hvilken som helst verdi, eventuelt begrenset til et bestemt tallområde.

Vi har også forklart (II,,7) at de observasjoner av et variabelt kjennetegn som en har skaffet seg i et bestemt tilfelle, må oppfattes som et utvalg av et univers som omfatter uendelig mange observasjoner. Å bruke betegnelsen univers i denne betydning kan forsvares, men vi kommer da i konflikt med det universbegrep som vi har benyttet foran i sannsynlighetsregningen. Vi vil derfor med det samme slå fast at også når det gjelder variable kjennetegn, skal universet bestå av begivenheter (gjentakelser). Gjentakelsene er da de enheter hos hvem det variable kjennetegn kan observeres. Utvalget av universet består etter dette ikke av observasjoner, men av enheter hos hvem det variable kjennetegn er observert. Og i overensstemmelse med dette er det de verdier som kjennetegnet kan ha i et univers, som er de hendinger som kan inntreffe i en gjentakelse. Tar vi for oss eks. 7 (II,3) består universet av et uendelig antall planter som alle har kjennetegnet "soleihov", og de hendinger som har inntruffet i utvalget på $n = 281$ gjentakelser er $x = 5$, $x = 6$, $x = 7$, $x = 8$ og $x = 9$.

Vi vil nå först forutsette at kjennetegnet – x – er diskret, og vi vil tenke oss at observasjonene er ordnet i en fordelingsrekke. Frekvensen til kjennetegnsverdien $x = x_i$ har vi för betegnet med h_i, men vi vil i det följgende bruke betegnelsen $h(x_i,U)$ for å markere at de enheter hos hvem kjennetegnet er observert er gjentakelser i et univers U. I det för nevnte eks. 7 er altså $h(5,U) = 223$, $h(6,U) = 45$, $h(7,U) = 6$, $h(8,U) = 4$ og $h(9,U) = 3$. Ifölge vår definisjon er da

$$s(x_i,U) = \text{Gr.}\ \frac{h(x_i,U)}{n} \qquad \text{når } n \to \infty$$

sannsynligheten for $x=x_i$ i en gjentakelse i U. Det er også sannsynligheten for en enhet som har kjennetegnsverdien $x=x_i$. Den relative frekvens $h(x_i,U)/n$ er et estimat av $s(x_i,U)$.

De forskjellige kjennetegnsverdier, altså de forskjellige verdier som x kan ha, har i regelen ulike sannsynligheter for å inntreffe i en gjentakelse. Sannsynligheten $s(x,U)$ vil derfor oftest forandre verdi med x slik

at vi kan oppfatte den som en funksjon av x og sette

$$s(x,U) = f(x)$$

Denne funksjonen kalles fordelingsloven for kjennetegnet x i universet U.

Da vi alltid er henvist til å arbeide med et endelig antall (n) gjentakelser, kan vi ikke få bestemt verdiene av $s(x,U) = f(x)$ eksakt. En har derfor utledet en del typer av fordelingslover ved å ta utgangspunkt i visse nærmere oppgitte sannsynlighetsteoretiske forutsetninger. En kjenner nå en del typer av slike fordelingslover for diskrete kjennetegn. Fordelingsloven er da gitt ved en formel hvor x er den uavhengig variable. Dessuten inneholder formelen et lite antall parametrer. Er verdien av disse kjent, kan en ved å innsette $x = x_i$ beregne $s(x_i,U)$. I regelen må verdien av disse parametrer bestemmes - estimeres - ved hjelp av de observasjoner av kjennetegnet en har i hvert tilfelle. Men det hender også at en har en hypotese om fordelingsloven som er så pass detaljert at også parameterverdiene er gitt.

Vi har ikke her anledning til å beskjeftige oss noe videre med teorien for diskrete fordelingslover. I tillegg til det vi allerede har nevnt vil vi nöye oss med å referere et eksempel. I följende tabell er $h(x,U)/n$ de relative frekvenser i eks. 2 (II,2) og $s(x,U) = f(x)$ er utregnet av binomialloven

$$s(x,U) = f(x) = \frac{k\,!}{x!\,(k-x)\,!}\; p^x q^{k-x}$$

hvor verdiene av parametrene k og p er satt til: $k = 21$ og $p = 0{,}61$.

x	$\frac{h(x,U)}{n}$	f(x)
4		0,0001
5		0,0005
6	0,0016	0,0020
7	0,0058	0,0069
8	0,0199	0,0188
9	0,0556	0,0426
10	0,0798	0,0798
11	0,1249	0,1249
12	0,1601	0,1628
13	0,1653	0,1763
14	0,1585	0,1576
15	0,1228	0,1150
16	0,0672	0,0674
17	0,0262	0,0311
18	0,0100	0,0108
19	0,0016	0,0027
20	0,0005	0,0004
	0,9998	0,9997

I dette universet kan naturligvis også x = 0, x = 1, x = 2, x = 3 og x = 21 inntreffe. Men sannsynlighetene for disse verider er mindre enn 0,0001 og er derfor ikke tatt med i tabellen.

Vi ser at verdiene av $s(x,U) = f(x)$ og de relative frekvenser stemmer meget bra overens. Dette kan da bety at binomialloven med parametrene k = 21 og p = 0,61 er fordelingsloven for kjennetegnet. En må imidlertid være oppmerksom på at hvor bra overensstemmelse det enn er mellom de relative frekvenser og sannsynlighetene utregnet på grunnlag av en teoretisk utledet fordelingslov, har vi ikke noe bevis for at vi har funnet den riktige fordelingsloven. Vi har nemlig aldri noen garanti for at ikke andre fordolingslover vil kunne stemme like godt med de samme frekvensene.

La oss nå tenke oss at kjennetegnet er kontinuerlig variabelt. Vi vil tenke oss at observasjonene av det er ordnet i en fordelingsrekke hvor klassene har klassevidden 2c. Er da $h(x_i,U)$ frekvensen for den klassen som har midtverdien x_i, vil $Gr.h(x_i,U)/n$ når $n\to\infty$ nærme seg til sannsynligheten $s(x_i,U)$ - for i en gjentakelse i U å få en kjennetegnsverdi innen klassen med grensene x_i-c og x_i+c. I eks. 3 (II,2) vil altså de relative frekvensene for voksende n nærme seg til sannsynlighetene for i en gjentakelse i vedkommende univers å få en kjennetegnsverdi innen klassene 11 - 12 kg, 12 - 13 kg, 13 - 14 kg. osv.

Også i dette tilfelle kan vi oppfatte s(x,U) som en funksjon av x. Denne funksjonen er diskontinuerlig fordi vi for x bare kan innsette klassenes midtverdier. Det er imidlertid ikke denne funksjonen som er fordelingsloven i dette tilfelle. Fordelingsloven er en kontinuerlig funksjon - f(x) - definert slik at arealet av den flaten som er begrenset av kurven for funksjonen, x-aksen og de to ordinatene som trekkes fra de to punktene på x-aksen hvis abscisser er x_i-c og x_i+c, er lik $s(x_i,U)$. Kurven i figur 1 representerer en slik fordelingslov. Arealet av den skraverte flaten er lik $s(x_i,U)$.

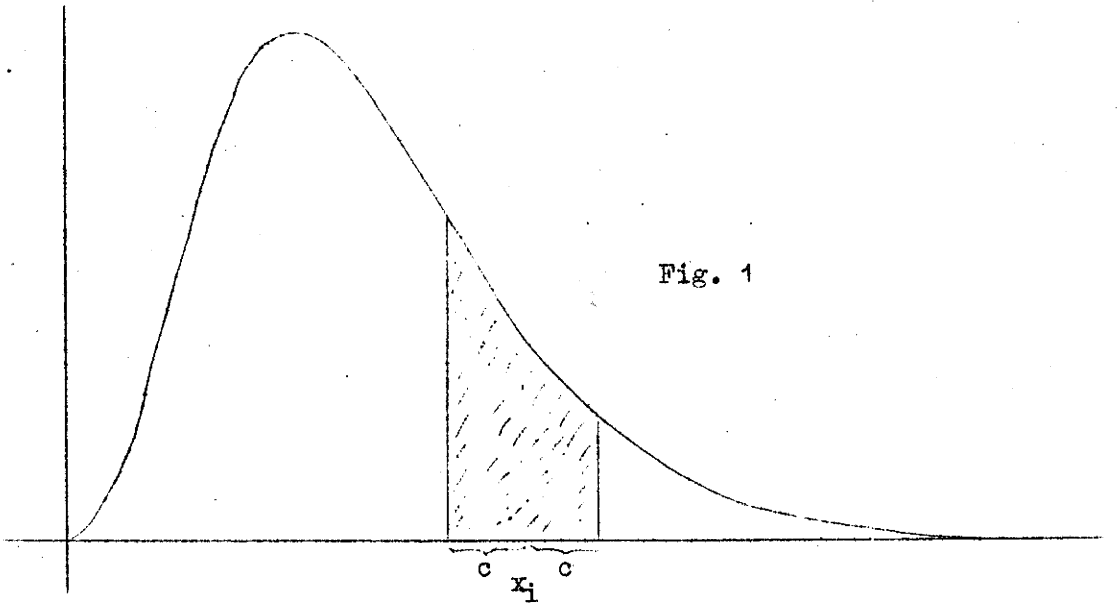

Fig. 1

Også de funksjonene som representerer kontinuerlige fordelingslover er utledet på grunnlag av visse nærmere oppgitte sannsynlighetsteoretiske forutsetninger. Vi skal senere stifte bekjentskap med noen typer av slike fordelingslover.

Vi har tidligere (II) sett hvordan en fordelingsrekke kan karakteriseres ved slike samletall som gjennomsnittet og middelavviket. Lignende størrelser brukes også når det gjelder å beskrive fordelingslover. De to viktigste av disse er forventningen og spredningen. Forventningen svarer til gjennomsnittet og spredningen til middelavviket. For gjennomsnittet og middelavviket har vi foran brukt betegnelsene m og s. For forventningen og spredningen vil vi - for å forhindre forvekslinger - bruke de greske bokstavene μ (my) og σ (sigma).

Da det ikke kan forutsettes at infinitesimalregningens symboler er kjent, vil vi forutsette at kjennetegnet (x) er diskret. Da er forventningen lik summen av produktene av x og $s(x,U) = f(x)$, eller

$$\mu = \Sigma f(x).x$$

Spredningens kvadrat (universets varians) er lik summen av produktene av $f(x)$ og $(x-\mu)^2$, eller

$$\sigma^2 = \Sigma f(x).(x-\mu)^2$$

Under disse summeringer skal en naturligvis ta med alle de verdier av kjennetegnet (x) som kan inntreffe i en gjentakelse i universet.

La oss anta at fordelingsloven er en binomiallov:

$$f(x) = \frac{k!}{x!(k-x)!} p^x q^{k-x}$$

Da kan det bevises at

$$\mu = \Sigma f(x).x = \sum_0^k \frac{k!}{x!(k-x)!} p^x q^{k-x}.x = kp$$

og at

$$\sigma^2 = \Sigma f(x).(x-\mu)^2 = \sum_0^k \frac{k}{x!(k-x)!} p^x q^{k-x}.(x-kp)^2 = kpq$$

Altså er

$$\sigma = \sqrt{kpq}$$

Vi har foran vist at binomialloven med k = 21 og p = 0,61 gir en meget god beskrivels av fordelingsrekken i eks. 2 (II,2). Forventningen og spredningen er altså lik

$$\mu = kp = 21.0,61 = 12,81$$

og

$$\sigma = \sqrt{kpq} = \sqrt{21.0,61.0,39} = \sqrt{4,9959} = 2,24.$$

9. Den normale fordelingslov.

Den normale fordelingslov som også ofte kalles den eksponensielle eller den Gaussiske fordelingslov, er den viktigste av alle kontinuerlige fordelingslover. Vi har ikke anledning til å forklare hvordan den er utledet og må nöye oss med å presentere den ved formelen

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Her er μ og σ forventningen og spredningen og e er grunntallet i det naturlige logaritmestystem ($e = 2{,}71828\ldots.$).

Det grafiske bilde av funksjonen er en kurve med bare et maksimum, nemlig for $x = \mu$ (se figur 2). Fra dette maksimum synker kurven symmetrisk til de to sider og strekker seg utover mot ∞ i begge retninger. Arealet av den flaten som er begrenset av kurven og x-aksen er lik enheten.

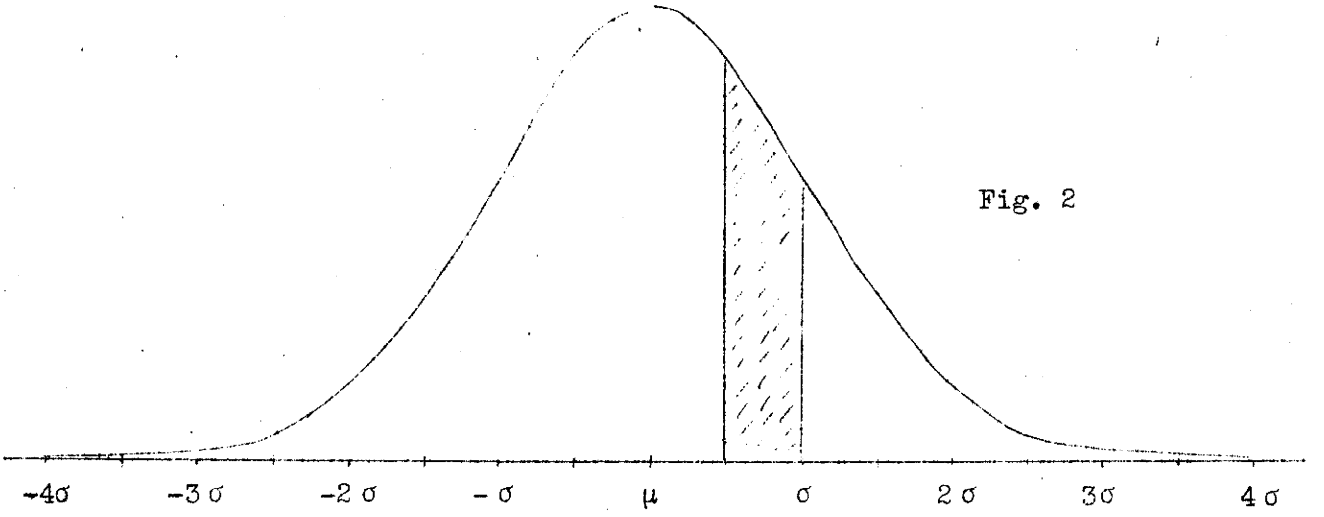

Fig. 2

La oss tenke oss at vi stykker opp x-aksen i et antall like store stykker. La stykkenes lengde være lik $\frac{1}{2}\sigma$ og la oss velge $x = \mu$ som et av delingspunktene. Arealet av den flaten som er begrenset av kurven, x-aksen og ordinatene trukket gjennom endepunktene av et av disse stykker er da lik sannsynligheten for i en gjentakelse i universet å få en kjennetegnsverdi mellom vedkommende stykkes endepunkter. Arealet av den skraverte flaten i gigur 2 er lik 0,15. Dette er da sannsynligheten for en kjennetegnsverdi innen den klassen hvis grenseverdier er $\mu + \frac{1}{2}\sigma$ og $\mu + \sigma$.

I tabell 1 er s(x,U) sannsynlighetene for å få en kjennetegnsverdi innen de oppförte klasser. Når unntas de to ytterste klasser, er klassevidden overalt lik $\frac{1}{2}\sigma$.

Tabell 1

klasser			s(x,U)
$-\infty$	-	$(\mu - 4\,\sigma)$	0,00002
$(\mu - 4\,\sigma)$	-	$(\mu - 3\frac{1}{2}\sigma)$	0,00022
$(\mu - 3\frac{1}{2}\sigma)$	-	$(\mu - 3\,\sigma)$	0,00110
$(\mu - 3\,\sigma)$	-	$(\mu - 2\frac{1}{2}\sigma)$	0,00465
$(\mu - 2\frac{1}{2}\sigma)$	-	$(\mu - 2\,\sigma)$	0,01700
$(\mu - 2\,\sigma)$	-	$(\mu - 1\frac{1}{2}\sigma)$	0,04400
$(\mu - 1\frac{1}{2}\sigma)$	-	$(\mu - 1\,\sigma)$	0,09150
$(\mu - 1\,\sigma)$	-	$(\mu - \frac{1}{2}\sigma)$	0,15000
$(\mu - \frac{1}{2}\sigma)$	-	μ	0,19150
μ	-	$(\mu + \frac{1}{2}\sigma)$	0,19150
$(\mu + \frac{1}{2}\sigma)$	-	$(\mu + 1\,\sigma)$	0,15000
$(\mu + 1\,\sigma)$	-	$(\mu + 1\frac{1}{2}\sigma)$	0,09150
$(\mu + 1\frac{1}{2}\sigma)$	-	$(\mu + 2\,\sigma)$	0,04400
$(\mu + 2\,\sigma)$	-	$(\mu + 2\frac{1}{2}\sigma)$	0,01700
$(\mu + 2\frac{1}{2}\sigma)$	-	$(\mu + 3\,\sigma)$	0,00465
$(\mu + 3\,\sigma)$	-	$(\mu + 3\frac{1}{2}\sigma)$	0,00110
$(\mu + 3\frac{1}{2}\sigma)$	-	$(\mu + 4\,\sigma)$	0,00022
$(\mu + 4\,\sigma)$	-	$+\infty$	0,00002
		S u m =	1,00000

Det er imidlertid en annen tabell som har större interesse for praktisk statistikk. Vi innser lett at sannsynligheten for en kjennetegnsverdi mellom grensene $(\mu - \frac{1}{2}\sigma)$ og $(\mu + \frac{1}{2}\sigma)$ etter enten-eller loven er lik summen 0,1915 + 0,1915 = 0,383. At x faller mellom disse to grensene er imidlertid ensbetydende med at $(x - \mu)$ faller mellom grensen $\div\frac{1}{2}\sigma$ og $+\frac{1}{2}\sigma$. Og detter igjen ensbetydende med at

$$|x - \mu| \leq \tfrac{1}{2}\sigma$$

hvor $|x - \mu|$ betyr tallverdien av $(x - \mu)$. Sannsynligheten for en kjennetegnsverdi som avviker fra forventningen med et belöp som er mindre eller i höyden lik $\frac{1}{2}$ er altså $Q = 0,383$. På samme måte kan vi på grunnlag av tabell 1 finne sannsynligheten for en kjennetegnsverdi som avviker fra forventningen med et belöp som er mindre eller i höyden lik spredningen, d.v.s. sannsynligheten for $|x - \mu| \leq \sigma$. Denne sannsynlighet er lik summen 0,1500 + 0,1915 + 0,1915 + 0,1500 = 0,683. Ved å fortsette på denne måten, kan vi finne sannsynlighetene for

$$|x - \mu| \leq a.\sigma$$

for a = $1\frac{1}{2}$, a = 2, a = $2\frac{1}{2}$ osv. Disse sannsynlighetene er gitt i tabell 2 under betegnelsen Q. I samme tabell er også tatt med verdiene av P = 1 - Q som er sannsynligheten for

$$|x - \mu| \geq a \cdot \sigma$$

d.v.s. sannsynligheten for en kjennetegnsverdi som avviker fra forventningen med et belöp som er lik eller större enn aσ.

Tabell 2.

a	Q	P
0,5	0,383	0,617
1,0	0,683	0,317
1,5	0,866	0,134
2,0	0,954	0,046
2,5	0,988	0,012
3,0	0,9973	0,0027
3,5	0,9995	0,0005
4,0	0,99994	0,00006

Vi ser at sannsynligheten for en kjennetegnsverdi som avviker fra forventningen med et belöp s om er lik eller större en 3σ, er bare 0,0027. Kjennetegnsverdier som avviker fra forventningen med mer enn tre ganger spredningen vil derfor praktisk talt aldri forekomme.

Det har vist seg at et meget stort antall fordelingsrekker for observasjoner av kontinuerlig variable kjennetegn stemmer meget bra med den normale fordelingslov. Når en i et bestemt tilfelle skal sammenligne de relative frekvensene - $h(x_i,U)/n$ - med $s(x_i,U)$, kommer en som oftest opp i vanskeligheter, fordi jo $s(x_i,U)$, er arealer. Dessuten er forventningen og spredningen praktisk talt aldri kjent. En må erstatte dem med gjennomsnittet og middelavviket for de observasjonene en har. Den enkleste framgangsmåten er da å erstatte μ og σ med m og s i tabell 1 for fastsettelsen av klassegrensene og bruke de klasser en finner på denne måten som grunnlag for ordningen av observasjonene i fordelingsrekke. Da er jo verdiene av $s(x_i,U)$ gitt i tabell 1. Denne metoden kan imidlertid ikke brukes hvis observasjonene allerede er ordnet i en fordelingsrekke og en ikke har adgang til den opprinnelige primærlisten over observasjonene. Men da kan en hvis klassevidden ikke er for stor i forhold til s (den bör ikke være större en s), erstatte $s(x_i,U)$ med funksjonsverdiene for klassenes midtverdier. En erstatter altså μ og σ i formelen for fordelingsloven med m og s, setter inn klassenes midtverdier og regner ut. (sml. eks. 3). Disse funksjonsverdier er aldri eksakt lik $s(x_i,U)$, men forskjellen er som oftest ubetydelig.

En kan naturligvis også beregne de eksakte verdier av $s(x_i,U)$ hvis μ og σ er kjent, men de metoder en da må bruke er så pass vanskelige at vi ikke kan redegjöre for dem her.

Eksempel 3. Hodeskallens største bredde ble målt hos n = 2000 voksne menn. Observasjonene (i mm) ble ordnet slik som følgende fordelingsrekke viser. Gjennomsnittet og middelavviket er m = 156,16 og s = 5,73. $s(x_i,U)$ er satt lik funksjonsverdiene for klassenes midtverdier. For noen av klassene er både frekvensene og verdiene av $s(x_i,U)$ summert. Forventningen og spredningen er satt lik gjennomsnittet og middelavviket.

klasser	x_i	$h(x_i,U)$		$\frac{h(x_i,U)}{n}$	$s(x_i,U)$
120-124	122	1			
125-129	127	0			
130-134	132	0	34	0,017	0,018
135-139	137	2			
140-144	142	31			
145-149	147	173		0,087	0,097
150-154	152	567		0,283	0,268
155-159	157	701		0,351	0,344
160-164	162	390		0,195	0,207
165-169	167	119		0,059	0,058
170-174	172	11			
175-179	177	3	16	0,008	0,008
180-184	182	1			
185-189	187	1			
		2000		1,000	1,000

Vi ser at verdiene av $s(x_i,U)$ stemmer meget bra med de relative frekvensene.

Oppgave 24. Sammenlign $s(x_i,U)$ og $h(x_i,U)/n$ for eks. 3 (II,2) når $s(x_i,U)$ settes lik funksjonsverdien for klassens midtverdi av den normale fordelingslov. Forventning og spredning settes lik gjennomsnitt og middelavvik.

10. Tchebycheffs ulikhet.

Det ville naturligvis være av stor interesse om en kunne stille opp en tabell svarende til tabell 2 som hadde gyldighet for alle fordelingslover. Saken er jo nemlig den at en i svært mange tilfelle - kanskje i de aller fleste tilfelle - ikke kjenner til hva slags fordelingslov kjennetegnet har. Vi har vist tidligere (II,6) ved noen eksempler at det overveiende antall observasjoner faller innenfor et tallområde fra m - 3s til m + 3s. Det er iallfall et relativt lite antall observasjoner som faller utenfor dette område. En vet nå erfaringsmessig at denne regelen har nokså almen gyldighet.

Tabell 2 viser at hvis kjennetegnet har normal fordelingslov, er sannsynligheten for en kjennetegnsverdi som faller utenfor området fra $\mu - 3\sigma$ til $\mu + 3\sigma$ praktisk talt lik null. Dette kan sies å være et spesialtilfelle av en helt almengyldig setning som sier at sannsynligheten for

$$|x - \mu| \geq a.\sigma$$

er lik eller mindre enn

$$P = \frac{1}{a^2}$$

for verdier av a som er större enn enheten. (Beviset for setningen tas ikke med her). I tabell 3 er gitt noen sammenhörende verdier av a og P.

Tabell 3.

a	P
1,5	0,4444
2,0	0,2500
2,5	0,1600
3,0	0,1111
3,5	0,0816
4,0	0,0625
4,5	0,0494
5,0	0,0400
10,0	0,0100

Vi ser at P avtar på samme måte som i tabell 2 for voksende verdier av a. Men den avtar langsommere. Videre ser vi at sannsynligheten for en kjennetegnsverdi som avviker fra forventningen med et belöp som er lik eller större enn $4.\sigma$ er liten, nemlig lik eller mindre enn 0,0626.

Denne setningen har gyldighet for alle mulige fordelingslover. Den gjelder derfor også for alle usedvanlige typer, typer som en sjelden eller kanskje aldri kommer over i praktisk statistisk arbeid. De fleste fagstatistikere synes derfor å være av den oppfatning at sannsynligheten for $|x-\mu| \geq a.\sigma$ er betydelig mindre enn etter tabell 3. De fordelingslover en i alminnelighet har med å gjöre, har meget större likhet med den normale. En pleier derfor gjerne å regne avvikelser fra forventningen på mer enn 3 - 4 ganger spredningen for å være så sjeldne at en setter sannsynligheten for $|x - \mu| \geq 4\sigma$ til praktisk talt lik null.

11. Binomialloven II.

La oss tenke oss at vi har n uavhengige gjentakelser i et univers U og at vi ved å undersöke hver enkelt av disse gjentakelsene har funnet hvor

mange ganger et konstant kjennetegn (enkelt eller sammensatt) eller en verdi av et variabelt kjennetegn har inntruffet i disse n gjentakelser. Dette antall har vi kalt frekvensen for vedkommende kjennetegn eller kjennetegnsverdi. Og vi har brukt slike betegnelser som h(A,U), h(AB,U), h(AiB,U) osv. når kjennetegnet er konstant. Er det variabelt, har vi brukt betegnelsen $h(x_i,U)$ hvor $x = x_i$ er selve kjennetegnets verdi når det er diskret variabelt og klassens midtverdi når det er kontinuerlig variabelt. I eks. 1 (III,4) har vi n = 2835 og h(A,U) = 2211, h(AB,U) = 1705 osv. I eks. 2 (II,2) har vi n = 1905 og h(6,U) = 3, h(7,U) = 11, h(8,U) = 38 osv.

La oss nå anta at vi kjenner sannsynligheten for at vedkommende kjennetegn eller kjennetegnsverdi skal inntreffe i en gjentakelse i U. For slike sannsynligheter har vi brukt betegnelser som s(A,U), s(AB,U) osv. når kjennetegnet er konstant og s(x,U) når det er variabelt. Men vi vil i det følgende bruke p som felles betegnelse for alle disse sannsynligheter og sette q = 1 - p.

Ifölge III,6 er da sannsynligheten for at kjennetegnet (eller kjennetegnsverdien) skal intreffe i z av n uavhengige gjentakelser i U lik

$$s(z,W) = \frac{n!}{z!(n-z)!}\, p^z q^{n-z}$$

hvor hver gjentakelse i W består av n uavhengige gjentakelser i U. Dette er da fordelingsloven for frekvensen for kjennetegnet (eller kjennetegnsverdien) i universet W. Forventningen og spredningen er (se III,8) lik

$$\mu = np \qquad \text{og} \qquad \sigma = \sqrt{npq}$$

La oss anta (som hypotese) at sannsynlighetenfor A i eks. 1 (III,4) i en gjentakelse er s(A,U) = p = 0,75 og at gjentakelsene er uavhengige gjentakelser. En har at n = 2835 og den hypotetiske forventning og spredning er derfor lik

$$\mu = np = 2835 \cdot 0,75 = 2126,25$$

og

$$\sigma = \sqrt{npq} = \sqrt{2835 \cdot 0,75 \cdot 0,25} = 23,06$$

Differensen mellom frekvensen for A og dens forventning er altså

$$h(A,U) - \mu = 2211 - 2126,25 = 84,75$$

og

$$\frac{|h(A,U) - \mu|}{\sigma} = \frac{84,75}{23,06} = 3,68$$

Etter Tchebycheffs ulikhet er sannsynligheten for en avvikelse fra forventningen - avvikelsen tatt iforhold til spredningen - som er lik eller större enn a = 3,68, lik eller mindre enn $(\frac{1}{3,68})^2 = 0,074$. Differensen mellem frekvensen for A og forventninger er derfor i störste laget, men den er ikke så stor at vi kan si at den er urimelig. En hending med sannsynligheten 0,074 vil jo forekomme ikke så rent sjelden. Hvis vi derfor måtte basere vår vurdering på Tchebycheffs ulikhet, måtte vi slutte at det ikke er noen god overensstemmelse

mellom hypotesen og den observerte frekvens. Men differensen er heller ikke så betydelig at vi kan slutte at hypotesen ikke er bekreftet. Hvis vi derimot kunne tillate oss å bruke tabell 2, ville vi bli tvunget til å forkaste hypotesen. Etter denne tabell er nemlig sannsynligheten for en avvikelse fra forventningen - avvikelsen tatt i forhold til spredningen - som er lik eller större enn den vi har funnet ($a = 3{,}68$) mindre enn 0,0005 som er sannsynligheten for en avvikelse lik eller större enn $a = 3{,}5$. Den funne differens er m.a.o. så stor at vi må karakterisere den som meget urimelig. Følgelig må vi slutte at det er et eller annet i veien med hypotesen. Vi måtte forkaste den. Noe bevis i matematisk forstand for denne slutningen har vi naturligvis ikke, for selv de mest usannsynlige hendinger vil jo kunne inntreffe av og til. Det er imidlertid alltid större eller mindre usikkerhetsmomenter knyttet til induktive slutninger. Det er det ikke noe å gjöre ved.

Hvis vi derfor kan tillate oss å bruke tabell 2, må vi slutte at vår hypotese ikke er bekreftet. Det må altså være et eller annet i veien med den. Da vi imidlertid senere skal komme tilbake til dette eksemplet, skal vi ikke hefte oss med noen diskusjon av saken på dette tidspunkt.

Dette eksemplet viser at det ville ha store praktiske fordeler om en i slike tilfelle som dette kunne bruke tabell 2. En ville da i mange flere tilfelle ha anledning til å forkaste feilaktige hypoteser enn om en var henvist til å bruk Tchebycheffs ulikhet som grunnlag. Det har lykkes å vise at når n er et stort tall, kan binomialloven med tilstrekkelig tilnærmelse gjengis av den normale fordelingslov. Når n er et stort tall, kan en derfor erstatte Tchebycheffs ulikhet med sannsynlighetene i tabell 2. Hvor stort tall n må være avhenger av verdien av p. Har p en verdi i nærheten av $\frac{1}{2}$, behöver ikke n være så svært stor. Har derimot p en verdi nær 0 eller nær 1, må en stille större krav til verdien av n.

Beskrivelsen av metoder for statistisk prövning av hypoteser hörer egentlig hjemme i avsnittet om statistisk induksjon, og vi skal derfor her bare ta med noen enkle eksempler og oppgaver som det passer å ta med som övelse.

Vi har i eks. 3 (III,9) sammenlignet sannsynlighetene gitt ved den normale fordelingslov med de relative frekvenser i fordelingsrekken for observasjonene av hodeskallens störste bredde hos voksne menn og fant da at det var en meget bra overensstemmelse. Vi har nå midler til å sammenligne de absolutte frekvenser med deres forventninger og sammenligne differensene med spredningene. De beregninger som da må utföres er demonstrert i följende tabell.

x_i	$h(x_i,U)$		$s(x_i,U)=p_i$	$q_i=1-p_i$	$\mu_i=np_i$	$\lvert h(x_i,U)-\mu_i\rvert$	$\sigma_i=\sqrt{np_iq_i}$	$\frac{\lvert h(x_i,U)-\mu_i\rvert}{\sigma_i}$
122	1							
127	0							
132	0	34	0,018	0,982	36	2	5,95	0,34
137	2							
142	31							
147	173		0,097	0,903	195	21	13,27	1,58
152	567		0,268	0,732	535	32	19,27	1,62
157	701		0,344	0,656	688	13	21,24	0,63
162	390		0,207	0,793	414	24	18,12	1,32
167	119		0,058	0,942	117	2	10,50	0,20
172	11							
177	3	16	0,008	0,992	16	0	3,98	0,00
182	1							
187	1							
2000			1,000		2001			

For å unngå å regne med sannsynligheter som er altfor nær null - d.v.s. for å unngå for store avvikelser fra den normale fordelingslov som fordelingslov for frekvensene - er noen av frekvensene og deres forventninger summert. For enkelhets skyld er dessuten alle forventninger avrunnet til nærmeste hele tall. Av siste kolonne ser en at alle frekvensenes avvikelser fra forventningen i forhold til spredningene ligger innenfor rimelige grenser både etter Tchebycheffs ulikhet og etter tabell 2. Den normale fordelingslov gir altså en god beskrivelse av fordelingsrekken. Det er iallfall ikke noen grunn til å slutte at hypotesen (den normale fordelingslov) ikke er bekreftet.

Oppgave 25. Ved en bestemt krysning av bananfluen opptrer hos avkommet karakteren "normale öyne". Anta (som hypotese) at sannsynligheten for denne karakter er $p = 0,75$ og at gjentakelsene (de enkelte avkom) er uavhengige gjentakelser. Sett at en som resultat av en krysning finner 160 avkom med normale öyne i et kull på 200 avkom. Er da dette resultat i strid med hypotesen?

Oppgave 26. La A og B bety det samme som i eks. 1 (III,4). Anta at $s(A,U) = s(B,U) = 0,75$ (A og B er uavh. hendinger) og at gjentakelsene er uavhengige gjentakelser. Undersök om observasjonene (frekvensene) i eks. 1 er i strid med denne hypotesen.

Oppgave 27. I följende tabell er $h(x,U)$ antall familier på 5 barn med x gutter. Anta at sannsynligheten for gutteödsel ved en enkeltfödsel er lik 0,52 og at de enkelte födsler i en familie er uavhengige gjentakelser. Undersök om observasjonene er i strid med denne hypotesen (sml. oppg. 21, III,6).

x	h(x,U)
0	3429
1	16851
2	36648
3	38665
4	20085
5	4459
	n = 120137

12. Forventning og spredning for enkle funksjoner av ett variabelt kjennetegn.

Det hender ofte at vi har bruk for å transformere observasjonene av et variabelt kjennetegn over til et nytt system med et annet nullpunkt og en annen måleenhet enn det opprinnelige system. La oss f.eks. tenke oss at vi har målt temperaturen (lufttemperaturen, vanntemperaturen e.l.) med et Fahrenheittermometer og at vi har bruk for gjennomsnittet og middelavviket for observasjonene uttrykt i Celsiusgrader. Vi kan da naturligvis gjöre om hver enkelt observasjon fra Fahrenheitskalaen til Celsiusskalaen etter formelen: $C^\circ = (F^\circ - 32).\frac{5}{9}$. For disse nye observasjonene -de transformerte - kan vi så beregne gjennomsnittet og middelavviket på vanlig måte. Denne fremgangsmåten er imidlertid meget arbeidskrevende, og dessuten har vi bruk for alminnelige formler for gjennomsnittet og middelavviket for slike transformerte observasjoner.

La oss derfor tenke oss at vi har n observasjoner av et variabelt kjennetegn hos n enheter: $o_1, o_2, o_3, \ldots o_n$. Vi tenker oss så at vi transformerer hver enkelt av disse over til et nytt system og at observasjonene i det nye system er $o_1', o_2', o_3', \ldots o_n'$ slik at for i = 1, 2, 3, ... n

$$o_i' = a.o_i + b$$

Er o_i gitt i Fahrenheitgrader og o_i' i Celsiusgrader, er $a = \frac{5}{9}$ og $b = +\frac{32.5}{9} = -\frac{160}{9}$.

Betegner vi gjennomsnittet og middelavviket for de opprinnelige observasjoner med m(o) og s(o) og for de transformerte med m(o'), og s(o') har vi

$$m(o') = \frac{1}{n}\sum o_i' = \frac{1}{n}\sum (ao_i + b) = \frac{a}{n}\sum o_i + \frac{1}{n}.nb = a.m(o) + b$$

og

$$s(o')^2 = \frac{\sum [o_i' - m(o')]^2}{n-1} = \frac{\sum [(ao_i + b) - (am(o) + b)]^2}{n-1}$$

$$= \frac{a^2 \ \Sigma [o_i - m(o)]^2}{n-1} = a^2 . s(o)^2$$

Forandrer vi altså nullpunktet og måleenheten for observasjonene etter ligningen $o_i' = a.o_i + b$, forandres gjennomsnittet og middelavviket til

$$m(o') = a.m(o) + b$$

og $$s(o') = a.s(o)$$

Oppgave 28. Sett 1) a = 1 og 2) b = 0 og redegjör for hva transformasjonen betyr i disse to tilfelle.

Oppgave 29. Följende tall er observasjoner av temperaturen målt med et Celsiustermometer. Beregn gjennomsnittet og middelavviket for disse observasjoner transformert til Fahrenheitskalaen.

4,53	5,21	4,82
5,15	4,75	4,93
5,02	4,95	4,93
5,09	4,63	4,79

Oppgave 30. Anta at en har oppgaver over årsinntektene for en bestemt gruppe personer i Norge og at en önsker å sammenligne med inntektene for samme gruppe i U.S.A. En får opplyst at denne gruppen har en årlig gjennomsnittsinntekt på 2000 $ og at middelavviket for inntektene er 100 $. Hva vil disse tallene svare til i norske kroner når en etter undersökelse av prisnivå og andre faktorer som en må ta hensyn til, har funnet at 1 $ svarer til 3 norske kroner?

La x være et diskret variabelt kjennetegn og fordelingsloven for dette kjennetegn i et univers U være $s(x,U) = f(x)$. Forventningen og spredningen for x er da (III,8):

$$\mu(x) = \frac{\Sigma f(x).x}{}$$

og $$\sigma(x) = \sqrt{\Sigma f(x).[x - \mu(x)]^2}$$

La videre x' være en entydig funksjon av x:

$$x' = \varphi(x)$$

Ved forventningen og spredningen for x' forstår vi da

$$\mu(x') = \frac{\Sigma f(x).x'}{}$$

og $$\sigma(x') = \sqrt{\Sigma f(x).[x' - \mu(x')]^2}$$

Siden x' er en entydig funksjon av x, d.v.s. at det til enhver verdi av x svarer bare en verdi av x', må sannsynligheten for $x' = x_i' = \varphi(x_i)$ i en gjentakelse i U være den samme som sannsynligheten for $x = x_i$ nemlig $f(x_i)$.

La oss nå forutsette at transformasjonen er lineær, d.v.s. at

$$x' = \varphi(x) = ax + b$$

Da er

$$\mu(x') = \Sigma f(x).x' = \Sigma f(x).(ax + b) = a \ \Sigma f(x).x + b \ \Sigma f(x) = a\,\mu(x) + b$$

$$\text{og} \quad \sigma(x')^2 = \Sigma f(x).[x' - \mu(x')]^2 = \Sigma f(x).[(ax + b) - (a\,\mu(x) + b)]^2$$

$$= \Sigma f(x).a^2.[x - \mu(x)]^2 = a^2.\sigma(x)^2$$

eller:
$$\sigma(x') = a.\sigma(x)$$

Vi ser at transformasjonsformlene for forventning og spredning er akkurat de samme somfor gjennomsnitt og middelavvik. Formlene er riktige også når x - og dermed x' - er et kontinuerlig variabelt kjennetegn.

Hvis funksjonen $x' = \varphi(x)$ ikke er lineær, er det meget vanskeligere å foreta en slik transformasjon av forventning og spredning. Vi skal derfor ikke komme nærmere inn på saken.

Oppgave 31. Finn forventning og spredning for det relative antall gutter i en familie på k barn, alle født ved enkeltfødsler, under den forutsetning at sannsynligheten for at det ved en enkeltfødsel blir født en gutt er p og at de enkelte fødsler er uavhengige gjentakelser.

Oppgave 32. Utled formlene for forventning og spredning for en relativ frekvens, f.eks. h(A,U)/n.

Oppgave 33. Benytt samme hypotese som i oppg. 26 (III,11) og beregn den hypotetiske forventning og spredning for de fire relative frekvenser i eks. 1 (III,4).

13. To eller flere variable kjennetegn.

La x og y være to variable kjennetegn som i et univers U har fordelingslovene f(x) og g(y). Er de begge diskret variable, er $s(x_i,U) = f(x_i)$ sannsynligheten for $x = x_i$ og $s(y_j,U) = g(y_j)$ sannsynligheten for $y = y_j$ i en gjentakelse i U. Er kjennetegnene kontinuerlig variable, vil vi som foran med $s(x_i,U)$ forstå sannsynligheten for en x-verdi i den klassen som har midtverdien x_i og med $s(y_j,U)$ sannsynligheten for en y-verdi i den klassen som har midtverdien y_j.

La oss nå tenke oss at vi deler opp universet U, i en rekke subuniverser slik at x har en konstant verdi i hvert av disse. Antallet av slike subuniverser er da naturligvis lik antallet av de ulike verdier x kan ha i en gjentakelse i U eller i det kontinuerlige tilfelle antallet av klasser som x-verdiene innordnes i. Disse subuniversene vil vi betegne med Ux. Sannsynligheten for $y = y_j$ i en gjentakelse i Ux_i er $s(y_j,Ux_i)$. Etter både - og loven (III,4) er da sannsynligheten for både $x = x_i$ og $y = y_j$ i en gjentakelse i U lik

$$s(x_iy_j,U) = s(x_i,U).s(y_j,Ux_i)$$

Vi kan naturligvis også sette

$$s(x_i y_j, U) = s(y_j, U).s(x_i, Uy_j)$$

hvor $s(x_i, Uy_j)$ er sannsynligheten for $x = x_i$ i en gjentakelse i Uy_j.

Da vi ikke kan forutsette kjennskap til integralregningen, vil vi i det følgende anta at både x og y er diskrete kjennetegn. De setninger som vi skal referere har imidlertid også gyldighet om begge kjennetegn er kontinuerlig variable og om det ene er kontinuerlig og det andre diskret variabelt.

Fordelingsloven for y i subuniverset Ux vil vi betegne med $g(y/x)$. I denne formelen er da x en alminnelig konstant, den forandrer bare verdi fra det ene subunivers til det annet. På samme måte er $f(x/y)$ fordelingsloven for x i subuniverset Uy. I denne formelen er da y en konstant som forandrer verdi fra subunivers til subunivers. Disse to fordelingslovene kalles betingete fordelingslover.

Det er klart at sannsynligheten for at x skal ha en bestemt verdi og y en bestemt verdi i en gjentakelse i U er avhengig av både x og y. Vi kan derfor oppfatte denne sannsynlighet som en funksjon av både x og y. La denne funksjonen være $F(x,y)$. Da er etter både - og loven:

$$F(x,y) = f(x).g(y/x) = g(y).f(x/y)$$

Sannsynligheten for $x = x_i$ og $y = y_j$ i en gjentakelse i U er derfor lik

$$s(x_i y_j, U) = F(x_i, y_j) = f(x_i).g(y_j/x_i) = g(y_j).f(x_i/y_j)$$

Hvis vi har et antall (n) parobservasjoner av x og y og har ordnet disse i en korrelasjonstabell, kan vi beregne estimatne av $F(x,y)$, $g(y/x)$, $f(x/y)$, $g(y)$ og $f(x)$ for alle forekommende verdier av x og y.

Eksempel 4. I følgende korrelasjonstabell er x antall ribber og y antall presakrale hvirvler hos griser. Presakrale hvirvler omfatter hals-, brust- og lende-hvirvler. n = 4928.

y \ x	14	15	16	17	h(y,U)
27	176	41			217
28	263	1603	69		1935
29		587	1993	8	2588
30			134	54	188
h(x,U)	439	2231	2196	62	4928

For x = 15 og y = 28 har vi at

estimatet av $f(x) = f(15)$ er $\frac{2231}{4928} = 0,45$

" " $g(y) = g(28)$ " $\frac{1935}{4928} = 0,39$

" " $F(x,y)=F(15,28)$" $\frac{1603}{4928} = 0,33$

Estimatet av $g(y/x)$ for $y = 28$ i subuniverset $Ux = U15$ er $\frac{1603}{2231} = 0,72$

Estimatene av $g(y/x)$ i subuniversene Ux er:

y	U14	U15	U16	U17
27	0,40	0,02		
28	0,60	0,72	0,03	
29		0,26	0,91	0,13
30			0,06	0,87
	1,00	1,00	1,00	1,00

Vi ser at estimatene av $g(y/x)$ for hver verdi av y er forskjellige i de fire subuniversene Ux. Hvis dette gjelder også det univers som disse gjentakelsene er et utvalg av, vil det si at de betingete fordelingslovene for y er ulike.

Oppgave 34. Beregn estimatene av $f(x/y)$ for alle verdier av x og y for eks. 4 (III,13). Beregn også estimatene av $g(y/x)$ for $y = 3$ i subuniversene Ux for eks. 13 (II,10).

Forventningen for y i subuniverset Ux er

$$\mu(y/x) = \Sigma\, g(y/x).y$$

hvor en under summeringen skal ta med alle de verdier av y som kan forekomme i en gjentakelse i Ux. På samme måte er

$$\mu(x/y) = \Sigma\, f(x/y).x$$

forventningen for x i subuniverset Uy. Disse to forventningene kalles ofte betingete forventninger. Det er naturligvis like mange betingete forventninger for y som det er subuniverser Ux og like mange betingete forventninger for x som det er subuniverser Uy. Disse betingete forventninger svarer til de betingete gjennomsnitter(II,11) slik at de siste er estimater av de første. I regelen er $\mu(y/x)$ en funksjon av x og $\mu(x/y)$ en funksjon av y. De kurvene som representerer disse to funksjonene i et rettvinklet koordinatsystem, har vi tidligere (II,11) kalt regresjonslinjene.

Hvis vi nå i et bestemt tilfelle kjenner de verdier som x og y kan ha i en gjentakelse i U og vi kjenner $f(x)$ og $g(y/x)$ eller $g(y)$ og $f(x/y)$ kan vi beregne $F(x,y)$ for alle verdier av x og y. Verdiene av $F(x,y)$ må da

stilles opp i et skjema som tilsvarer en korrelasjonstabell.

Oppgave 35. Anta at K er en karakter som nedarves bare til det hanlige avkom og at sannsynligheten for K i en gjentakelse i subuniverset av hanlig avkom er p = 0,25. Anta videre at sannsynligheten for hankjönn og hunkjönn er den samme, P = 0,5. Det forutsettes at de enkelte avkom er uavhengige gjentakelser.

La x være antall hanner og y antall hanner med K i et kull på k = 5 avkom.

a. beregn for alle verdier av x og y $f(x)$ og $g(y/x)$
b. " " " " " $F(x,y)$ og $g(y)$
c. beregn $\mu(x)$ og $\sigma(x)$
d. vis ved beregning at $\mu(y) = kpP$ og $\sigma(y) = \sqrt{kpP(1-pP)}$
e. hva er ligningen for regresjonslinjen for y m.h.p. x?

Dersom $g(y/x = g(y)$ for alle verdier av x og y, d.v.s.at fordelingsloven for y er den samme i subuniversene Ux som i universet U, er også $f(x/y) = f(x)$ fordi

$$f(x).g(y/x) = g(y).f(x/y)$$

Da er $$F(x,y) = f(x).g(y)$$

En sier da at x og y er uavhengige kjennetegn. Sannsynligheten for y = yj i en gjentakelse er da nemlig uavhengig av hvilket verdi x har.
I dette tilfelle er

$$\mu(y/x) = \Sigma g(y/x).y = \Sigma g(y).y = \mu(y)$$

og $$\mu(x/y) = \Sigma f(x/y).x = \Sigma f(x).x = \mu(x)$$

De to regresjonslinjene er m.a.o. i dette tilfelle parallelle med koordinataksene og följelig er x og y ukorrelerte kjennetegn (se II,15). Uavhengige kjennetegn er altså også ukorrelerte kjennetegn. Det er derimot teoretisk sett ikke noe i veien for at ukorrelerte kjennetegn kan være avhengige kjennetegn. Det at $\mu(y/x) = \mu(y)$ er nemlig ingen garanti for at $g(y/x) = g(y)$. Avhengigheten mellom x og y kan f.eks. være slik at den betingete forventning for y - $\mu(y/x)$ - kan være uavhengig av x mens det betingete spredning for y kan være avhengig av x.

Dersom x og y er uavhengige kjennetegn, er ifölge III,11 forventningen for antallet av gjentakelser med $x = x_i$ og $y = y_j$ lik $n.f(x_i).g(y_j)$ og spredningen er $\sqrt{n.f(x_i).g(y_j).[1 - f(x_i).g(y_j)]}$. Påviselig avhengighet mellom x og y ville da gi seg utslag i at differensene mellom den observerte frekvens for kombinasjonen $x_i y_j$ og forventningen, i forhold til spredningen,

iallfall for noen verdier av x og y ville overskride 3 - 4. En foretrekker imidlertid som regel å undersöke dette spörsmål ved hjelp av de metoder som er beskrevet i II.

Vi har forutsatt foran at x og y er kjennetegn i samme univers. La oss nå tenke oss at x er et kjennetegn i Universet U med fordelingsloven f(x) i dette universet. La videre y være et kjennetegn i universet U' med fordelingsloven g(y). Da er $x = x_i$ en hending som kan inntreffe i en gjentakelse i U og $y = y_j$ en hending som kan inntreffe i en gjentakelse i U'. Og forutsetter vi at $x = x_i$ og $y = y_j$ er uavhengig opptredende hendinger, er (se III,5).

$$s(x_iy_j,W) = s(x_i,U).s(y_j,U') = f(x_i).g(y_j)$$

hvor hver gjentakelse i W består av en gjentakelse i U og en gjentakelse i U'. Gjelder forutsetningen om uavhengighet alle verdier av x og y, kan vi mer alminnelig skrive

$$s(xy,W) = F(x,y) = f(x).g(y)$$

Vi sier også i dette tilfelle at x og y er uavhengige kjennetegn. Dette betyr at hvis en har en gjentakelse i U og en gjentakelse i U', har verdien av x i gjentakelsen i U ingen innflytelse på sannsynligheten for $y = y_j$ i gjentakelsen i U'. Dette gjelder alle mulige kombinasjoner av en gjentakelse i U og en gjentakelse i U' og alle verdier av x og y. Disse forutsetningene vil i alminnelighet være realisert. Men en har også eksempler på tilfelle der det finnes forbindelser mellom bestemte gjentakelser i U og bestemte gjentakelser i U' slik at x og y ikke lenger opptrer som uavhengige kjennetegn. Disse eksempler må vel likevel betraktes som unntakelser som vi her ikke har anledning til å komme nærmere inn på. Det kan forresten henvises til III,5.

Hvis vi kan forutsette at x og y er uavhengige kjennetegn og vi kjenner fordelingsloven - f(x) - for x i U og fordelingsloven - g(y) - for y i U', er det meget enkelt å beregne F(x,y) = s(xy,W) for alle mulige kombinasjoner av x-verdier og y-verdier.

Oppgave 36. F er en familie med 3 barn og F' en familie med 4 barn, alle født ved enkeltfödsler. Anta at sannsynligheten for gütt ved en enkeltfödsel er lik 0,5 og at alle födsler er uavhengige gjentakelser. Beregn da sannsynligheten for at F har x gutter og F' y gutter for de forskjellige mulige verdier av x og y.

Både i teoretiske og i praktiske undersökelser har en meget ofte bruk for forventning og spredning for summen (x + y) og differensen (x - y). Vi har at

$$\mu(x + y) = \mu(x) + \mu(y)$$
$$\mu(x - y) = \mu(x) - \mu(y)$$
og $$\sigma(x + y) = \sigma(x - y) = \sqrt{\sigma(x)^2 + \sigma(y)^2}$$

Formlene for forventningen er riktige både når x og y er avhengige og når de er uavhengige kjennetegn. Formlene for spredningen er riktige bare når x og y er uavhengige kjennetegn. Alle formlene er riktige både når begge kjennetegn er diskrete, når det ene er diskret og det andre kontinuerlig og når begge er kontinuerlige.

Oppgave 37. Forutsett det samme som i oppg. 36. Finn fordelingsloven for x + y og vis at forventningen og spredningen for x, for y og for x + y tilfredsstiller de foran gitte formler.

Oppgave 38. På to forsöksgårder er det lagt ut forsöksruter på 25 m^2 for dyrking av en bestemt sort kveite. På F er det n ruter og på F' n' ruter. En får på denne måten n og n' observasjoner av avlingen pr. 25 m^2 (sml. eks. 3, II,2). En kan så beregne gjennomsnitt og middelavvik for begge observasjonsrekker. Disse störrelser er da estimater av forventning og spredning for kjennetegnet = antall kg.korn pr. 25 m^2 i de to universene som de n og n' rutene er utvalg av. Anta at en kjenner disse forventninger og spredninger:

$$\mu = 75 \quad \text{og} \quad \sigma = 3{,}2$$
$$\mu' = 60 \quad \text{og} \quad \sigma' = 2{,}8$$

Beregn forventning og spredning for avlingsdifferensene mellom F og F'.

Setningene om forventningen og spredningen for sum og differens av to kjennetegn kan lett utvides til å omfatte et vilkårlig antall kjennetegn. En har f.eks.:

$$\mu(x+y-z+u-v+\ldots) = \mu(x) + \mu(y) - \mu(z) + \mu(u) - \mu(v) + \ldots.$$
og $$\sigma(x+y-z+u-v+\ldots) = \sqrt{\sigma(x)^2 + \sigma(y)^2 + \sigma(z)^2 + \sigma(u)^2 + \sigma(v)^2 + \ldots}$$

Den siste formelen er riktig bare når kjennetegnene er uavhengig av hverandre innbyrdes og når hvert kjennetegn er uavhengig av enhver kombinasjon av de andre kjennetegn. Formelen for forventningen er riktig også når kjennetegnene er avhengige av hverandre.

Siden nå (se III,12)

$$\mu(ax) = a\,\mu(x),\ \mu(by) = b\,\mu(y),\ \mu(cz) = c\mu(z) \ldots\ldots$$
og $$\sigma(ax) = a\,\sigma(x),\ \sigma(by) = b\,\sigma(y),\ \sigma(cz) = c\sigma(z) \ldots\ldots$$

har en uten videre at

$$\mu(ax-by+cz+\ldots.) = a\mu(x) - b\mu(y) + c\mu(z) + \ldots..$$
og $$\sigma(ax-by+cz+\ldots.) \quad \sqrt{a^2\sigma(x)^2 + b^2\sigma(y)^2 + c^2\sigma(z)^2 + \ldots.}$$

Oppgave 39. R = 100 familier har k = 5 barn hver. Det forutsettes at alle barn er født ved enkeltfödsler, at sannsynligheten for gutt ved en enkeltfödsel er p = 0,5 og at alle födsler er uavhengige gjentakelser. Finn da forventning og spredning for antall gutter i alle R = 100 familier til sammen.

Anta at barneantallet i de R = 100 familier fordeler seg på följende måte:

k	h(k)
2	25
3	20
4	15
5	11
6	9
7	7
8	6
9	4
10	3
	100

Beregn under samme forutsetning som foran forventning og spredning for antall gutter i de R = 100 familier til sammen.

14. Univers og utvalg.

I de foregående avsnitt har vi når det var tale om univers og utvalg, bygget på följende forutsetninger:

1. Universet omfatter uendelig mange gjentakelser,
2. gjentakelsene i utvalget har samme sett felles kjennetegn som gjentakelsene i universet og ingen andre kjennetegn har hatt noe å si for valget av gjentakelser, og
3. gjentakelsene i utvalget er innbyrdes uavhengige, d.v.s. at sannsynligheten for at en bestemt hending skal intreffe i en bestemt gjentakelse er uavhengig av utfallet av de andre gjentakelser i utvalget.

Når disse forutsetninger er realisert, sier en at utvalget er et tilfeldig utvalg. Da vi i det följende vil legge disse forutsetninger til grunn for utviklingen av de statistiske metoder som skal tas med i dette kursus, skal vi pröve å gjöre noe nærmere rede for hva for innhold disse forutsetningene har.

Vi har allerede foran pekt på at et univers er en ren tankekonstruksjon. Et univers er altså ikke noe som eksisterer. I enkelte tilfelle er forestillingen om et uendelig univers meget nærliggende og naturlig, nemlig

når gjentakelsene er forsöksgjentakelser. I praksis kan en naturligvis ikke gjenta et forsök mer enn et begrenset antall ganger, men det er ikke vanskelig å tenke seg rekken av gjentakelser fortsatt i det uendelige. I andre tilfelle er det kanskje noe vanskeligere å forestille seg antallet av gjentakelser ökt i det uendelige. Alle ekornene i Norge danner til sammen det en i biologien kaller en populasjon. Et ekorn skulle da være en gjentakelse i et univers og populasjonen et utvalg av dette univers. Hvordan kan dette forklares og begrunnes?

La oss ta vårt utgangspunkt i det mer kjente biologiske begrep "art". Vi sier at alle individer som har de og de kjennetegn, tilhörer den art som disse kjennetegn er en definisjon av. Et dyr tilhörer arten ekorn hvis det har de kjennetegn som definerer arten ekorn. Alle dyr av ekornarten har altså en endelig antall felles kjennetegn. Men disse felles kjennetegn gir ingen beskrivelse av arten, de tjener bare til å avgrense arten innen mangfoldigheten av arter. Heller ikke den mest minutiöse beskrivelse av et enkelt eksemplar vil gi oss en beskrivelse av arten fordi et annet eksemplar av arten ikke vil passe nöyaktig til denne beskrivelsen. Tenk bare på alle de mange forskjellige utforminger av hodeskallen, av tennene, av örene, de forskjellige variasjoner i hårbekledningen osv. Hvis vi hadde en beskrivelse av alle ekornene i Norge i dag og hver beskrivelse var meget inngående og omfattende, ville vi antakelig ikke finne to helt identiske beskrivelser. La oss tenke oss at vi tar ut to av disse beskrivelser. Bytter vi da ut en detalj i den ene med den tilsvarende detalj i den annen, vi vi få en beskrivelse som sannsynligvis ikke passer på noen av de eksisterende ekorn. Vi vil ha en beskrivelse av et dyr som tilhörer ekornarten fordi det har alle de kjennetegn som definerer arten ekorn, men som det ikke finnes noe gjenpart av i den populasjon som eksisterer i Norge i dag. Og tenker vi oss at vi etter hvert bytter om alle de nær sagt uendelig mange samsvarende detaljer fra beskrivelse til beskrivelse av ekorn innen den eksisterende populasjon, vil vi få en mangfoldighet av beskrivelser av ekorn som ikke eksisterer. Noen av disse beskrivelser vil ikke passe på "levende" ekorn og vi kan regne med en reduksjon av mulighetene på grunn av avhengigheten mellom kjennetegnene. Men vi kan til gjengjeld tenke oss manfoldigheten fordoblet mange ganger på andre måter. Sett at vi har kunnet fastslå den minimumsverdi og den maksimumsverdi som et kontinuerlig variabelt kjennetegn har innen arten ekorn i Norge. La minimumsverdien være a og maksimumsverdien være b. Avsetter vi linjestykket fra a til b på en rett

linje og på dette igjen punkter som svarer til de verdier som kjennetegnet har innen den eksisterende populasjon, får vi avsatt et endelig antall atskilte punkter. Men et hvert annet punkt på linjestykket fra a til b representerer verdien av vedkommende kjennetegn hos ekorn som det ikke finnes noen gjenpart av i den eksisterende populasjon.

Når gjentakelsene er forsöksgjentakelser, må vi forestille oss universet som en rekkefölge av gjentakelser uten begynnelse og uten slutt. I andre tilfelle må vi forestille oss universet som en mangfoldighet av muligheter som stadig kan ökes. En populasjon er en virkeliggjörelse av et endelig antall muligheter i en uendelig mangfoldighet som vi kaller et univers.

(Vi har her brukt betegnelsen populasjon slik den brukes i biologien. I engelsk statistisk litteratur brukes betegnelsen "population" eller kanskje oftest "parent population" i samme betydning som vår betegnelse univers).

La oss så ta for oss neste forutsetning. La oss tenke oss at vi har et utvalg på n gjentakelser i et univers U. Hvis da B er et kjennetegn (eller et kjenntegnssett) som ikke er felles for alle gjentakelsene i U og utvalget er tatt ut etter det prinsipp at alle gjentakelsene skal ha B som kjennetegn, er utvalget ikke et tilfeldig utvalg av U. Men hvis forutsetning 3, som vi skal diskutere senere, er realisert, er utvalget et tilfeldig utvalg av subuniverset UB. En pleier likevel å si at utvalget er et tilfeldig utvalg av U med hensyn på et kjennetegn A dersom sannsynligheten for A i en gjentakelse i UB er densamme som sannsynligheten for A i en gjentakelse i U, altså hvis $s(A,UB) = s(A,U)$. En sier også at utvalget er et tilfeldig utvalg av U m.h.p. et variabelt kjennetegn x hvis fordelingsloven for x er den samme i subuniverset UB som i universet U. Det er imidlertid alltid kjennetegn som på en eller annen måte er avhengig av B. Og et utvalg som er tatt ut på en slik måte som forutsatt foran, er derfor ikke et generelt tilfeldig utvalg av U.

Vi har foran (III,11) sammenlignet frekvensen for A-normale börster i eks. 1 (III,4) med forventningen i $n = 2835$ uavhengige gjentakelser i et univers U hvor $s(A,U) = 0,75$. Vi fant at det var så stor forskjell mellom disse at vi ble tvunget til å forkaste hypotesen. La oss nå pröve å finne årsaken til dette. Det ligger da nærmest å undersöke om det utvalg vi har, er et tilfeldig utvalg av vårt hypotetiske univers. Hvis det ikke er det, kan det være dette som er forklaringen på noverensstemmelsen. Vi kan da straks slå fast at det utvalg vi har, ikke er et generelt tilfeldig utvalg. Alle gjentakelsene i utvalget har nemlig et kjennetegn som ikke er felles for alle gjentakelsene i universet,

nemlig kjennetegnet "levende". Det er derfor mulig at hvis vi hadde et utvalg tatt ut på en slik måte at kjennetegnet "levende" ikke utvelger gjentakelsene, ville uoverrensstemmelsen forsvinne. Et slikt utvalg kan vi naturligvis ikke skaffe oss. Men grunnen til uoverensstemmelsen kan altså ligge i det at sannsyn ligheten for A i subuniverset av levende avkom er forskjellig fra sannsynlighete for A i universet av alt avkom, altså både levende avkom og avkom som enten ikke utvikles til levende avkom eller som dör för det stadium er nådd da en kan konstatere om A har inntruffet eller ikke inntruffet. Noen garanti for at dette er grunnen til uoverensstemmelsen har vi naturligvis ikke, men det er iallfall en meget nærliggende forklaring.

La oss tenke oss at vi en sommerdag befinner oss ved en fjord på Sörlandet og at vi iakttar at en makrellstim er på vei inn fjorden. Denne stimen er da et utvalg av et eller annet univers av makrell. Sett at vi så ville skaffe oss et tilfeldig utvalg av dette univers ved hjelp av en snurpenot. Det kan da tenkes at maskene i noten er så store at de minste individer slipper i gjennom. Hvis det er tilfelle er det klart at störrelsen har noe å si for valget av gjentakelser. De minste störrelser blir systematisk underrepresentert. Og utvalget blir derfor ikke et tilfeldig utvalg m.h.p. störrelsen eller m.h.p. kjennetegn som på en eller annen måte er avhengig av störrelsen. Det kan også tenkes at fiskestimen er ordnet etter störrelsen på en eller annen systematisk måte, f.eks. slik at det fortrinnsvis er de störste fisk som går i teten. Hvis vi da stenger i den bakre del av stimen, vil de störste störrelser bli systematisk underrepresentert og utvalget av den grunn ikke et tilfeldig utvalg. Men likevel kan det tenkes at utvalget blir tilfeldig m.h.p andre kjennetegn, nemlig slike som er uavhengige av störrelsen.

Hvis oppgaven går ut på å skaffe tilveie et tilfeldig utvalg av et gitt univers, kan vi lett komme opp i vanskeligheter som en må prøve å overvinne på en eller annen måte. Hvordan en skal overvinne disse vanskelighetene kan det ikke gis noen alminnelige regler for. Undersökelsene av disse spörsmål hörer hjemme i enkeltvitenskapene. På den annen side er det utvalget vi har i et bestemt tilfelle, alltid et tilfeldig utvalg av et eller annet univers. Er det et utvalg av universet U, men ikke et tilfeldig utvalg av U, vil det alltid kunne betraktes som et tilfeldig utvalg av et univers som på en eller annen måte er utsortert av U. Ved å sammenligne utvalget med andre utvalg av samme univers kan en da meget ofte få brakt på det rene at to utvalg m.h.p. et bestemt kjennetegn ikke er tilfeldige utvalg av samme univers og at derfor minst det ene av dem ikke er et tilfeldig utvalg av det univers som begge utvalg er tatt ut av. Men mer om dette senere.

Til slutt noen merknader om den tredje forutsetningen. Et utvalg skal jo danne grunnlaget for den karakteristikk eller beskrivelse som det er vår oppgave å gi av universet. En slutning eller dom om universet på grunnlag av et utvalg er en induksjonsslutning og det er naturligvis meget om å gjöre at slike slutninger blir basert på eksakte vurderinger.

La x være et diskret variabelt kjennetegn med fordelingsloven $s(x,U) = f(x)$ i et univers U. La oss tenke oss at vi har et utvalg på $n = 2$ gjentakelser. Hvis disse da er uavhengige, er sannsynligheten for $x = x_i$ i den ene og $x = x_j$ i den annen gjentakelse lik produktet

$$s(x_i,U).s(x_j,U) = f(x_i).f(x_j)$$

Hvis derimot de to gjentakelsene ikke er uavhengige, er sannsynligheten for $x = x_j$ i annen gjentakelse avhengig av hvilken verdi x har i den förste. Og innflytelsen av slike avhengighetsforhold er det ikke mulig å få med i en alminnelig teori. Hvis derfor gjentakelsene ikke er uavhengige, er det ikke mulig å bygge opp en teori for vurderingen avde relasjoner som eksisterer mellom et utvalg og det univers det er uttatt av. Vi ville f.eks. ikke kunne bruke binomialloven som fordelingslov for frekvensen for et kjennetegn (eller en kjennetegnsverdi) og heller ikke ha noen almengyldig annen fordelingslov å erstatte den med.

Det er imidlertid ikke vanskelig å finne eksempler på tilfelle der det mellom noen av gjentakelsene i et utvalg finnes forbindelser som ikke finnes mellom alle gjentakelsene i universet. La oss tenke oss at utvalget består av et antall 10 år gamle norske gutter. Det er da umiddelbart innlysende at to tvillingbrödre (og enda mindre to eneggete tvillingbrödre) ikke kan betraktes som uavhengige gjentakelser med hensyn på alle kjennetegn. Mellom to tvillingbrödre vil det i regelen finnes forbindelser (særlig av genetisk opprinnelse) som ikke finnes mellom alle 10 år gamle gutter.

En kan da spörre hva nytte en kan ha av en teori som er basert på forutsetninger som en må regne med ikke er realisert. Dette er det lett å svare på. Det er nettopp ved hjelp av en slik teori vi kan ha håp om å avslöre de ekstraforbindelser mellom gjentakelsene som vi hypotetisk forutsetter ikke er til stede.

La oss igjen komme tilbake til eks. 1 (III,4). Vi har konfrontert frekvensen for A-normale börster med en bestemt hypotese. Innholdet i denne hypotesen er:

1. at s(A,U) = p = 0,75, og
2. at det utvalg vi har, er et tilfeldig utvalg, d.v.s. at både forutsetning 2 og forutsetning 3 er realisert.

Vi fant at denne hypotesen måtte forkastes. Hvilke slutninger kan vi trekke av dette? Vi må slutte at minst ett av de elementer som hypotesen er sammensatt av må forkastes. Vi må forkaste enten elementet s(A,U) = p = 0,75 eller forutsetningen om at utvalget er et tilfeldig utvalg eller begge elementene. Vi kan så diskutere hvert av disse elementer. La oss tenke oss at vi först gjör en endring i verdien av s(A,U). Vi ville da finne at hvis vi setter s(A,U) = p = 0,77, ville vi ikke få et resultat som berettiger oss til å forkaste hypotesen (pröv dette). Det er derfor en mulighet for at hypotesen er riktig når vi erstatter elementet p = 0,75 med p = 0,77. Men vi har også sett at det utvalg vi har, ikke er et generelt tilfeldig utvalg fordi det ikke tilfredsstiller forutsetning 2. Grunnen til uoverensstemmelsen kan derfor være at utvalget ikke er et tilfeldig utvalg m.h.p. kjennetegnet A. Vi kunne så undersöke om det kan være forutsetning 3 for et tilfeldig utvalg som må forkastes. Det ligger utenfor rammen av dette kursus å diskutere dette spörsmål. Men resultatet av en overveielse ville være at uoverensstemmelsen ikke kan komme av at forutsetning 3 ikke er realisert. Av dette må vi da ikke slutte at forutsetning 3 er realisert. Det eneste vi kan slutte er at forutsetning 2 ikke er realisert, eller/og at s(A,U) er forskjellig fra 0,75. En statistisk teori som er basert på at utvalget tilfredsstiller de fordringer vi har stilt opp for et tilfeldig utvalg, er altså et slags logisk skjema som anvendt på en riktig måte vil gi oss midler til å få påvist og klarlagt årsakssammenhenger, og det er jo den sentrale oppgave for all naturforskning.

För vi forlater dette emnet må vi tilföye at det i visse tilfelle også kan være spörsmål om utvalg av populasjoner (endelige). Oppgaven er da å skaffe seg utvalg som er representative for vedkommende populasjon. Men de problemer som knytter seg til denne oppgaven har vi her ikke anledning til å beskjeftige oss med.

15. Gjennomsnittets fordelingslov.

Vi vil nå anta at vi har tatt ut et tilfeldig utvalg på n gjentakelser av et univers og at vi har observert et variabelt kjennetegn (x) hos hver enkelt av disse gjentakelser (enheter). Vi har da n observasjoner av x, og

vi kan derfor beregne gjennomsnitt, middelavvik osv. for disse observasjoner. Gjennomsnittet av de n observasjonene av x vil da i regelen ikke være lik forventningen for x og middelavviket for de n observasjonene av x vil da i regelen ikke være lik spredningen for x. Gjennomsnittet er bare et estimat av forventningen og middelavviket et estimat av spredningen.

Har vi observert to variable kjennetegn (x og y), har vi n parobservasjoner av x og y. Og vi kan derfor beregne regresjonskoeffisienter, korrelasjonsforholdet, korrelasjonsindeksen og korrelasjonskoeffisienten. Disse størrelser vil heller ikke være lik de tilsvarende størrelser i universet, men bare estimater av disse.

La oss så tenke oss at vi har r tilfeldige utvalg av dette samme universet og at et eller flere kjennetegn er observert hos hver enkelt enhet innen hvert utvalg. Vi kan da beregne r gjennomsnitter, r middelavvik for hvert observert kjennetegn, r korrelasjonskoeffisienter etc. for hver kombinasjon av to og to kjennetegn. Vi har m.a.o. r uavhengige estimater eller observasjoner av forventningen for x, vi har r uavhengige estimater eller observasjoner av spredningen for x osv. La oss tenke oss at vi har ordnet verdiene av de r gjennomsnittene i en fordelingsrekke med en bestemt klasseinndeling. Tenker vi oss nå at r ökes over alle grenser, vil denne fordelingsrekken hvor vi forutsetter at frekvensene er gitt i relative verdier, nærme seg til en eller annen fordelingslov (se III,8 om fordelingsloven for kontinuerlig variable kjennetegn) og denne fordelingsloven er da fordelingsloven for gjennomsnittet av n observasjoner av kjennetegnet x i et tilfeldig utvalg på n gjentakelser i universet.

På samme måte kan vi forklare hva vi mener med fordelingsloven for middelavviket, fordelingsloven for korrelasjonskoeffisienten osv. Fordelingslovene for disse observasjonskarakteristikker er nokså innviklet og vi skal derfor her innskrenke oss til å behandle fordelingsloven for gjennomsnittet. Karakteren av denne er avhengig av karakteren av fordelingsloven for vedkommende kjennetegn. Derfor er det ikke mulig å stille opp noen generell formel for gjennomsnittets fordelingslov. Men det er mulig å utlede generelle formler for en rekke karakteristikker for den. Vi skal nöye oss med å utlede formelen for forventningen og spredningen.

Vi vil bruke betegnelsene $\mu(m)$ og $\sigma(m)$ for forventningen og spredningen i gjennomsnittets fordelingslov. Hvis da μ og σ er forventningen og spredningen i vedkommende kjennetegns fordelingslov, og det tilfeldige utvalg består av n gjentakelser, er

$$\mu(m) = \mu$$

og

$$\sigma(m) = \frac{\sigma}{\sqrt{n}}$$

Antar vi nemlig at observasjonene som vi som för vil betegne med

$$o_1, o_2, o_3, \ldots\ldots o_n$$

er målinger av kjennetegnsverdiene, kan vi betrakte gjennomsnittet

$$m = \frac{1}{n} \Sigma o_i = \frac{1}{n} o_1 + \frac{1}{n} o_2 + \frac{1}{n} o_3 + \ldots\ldots + \frac{1}{n} o_n$$

som en lineær funksjon av n uavhengige kjennetegnsverdier og da o'ene har samme forventning (μ), er i följge III,13

$$\mu(m) = \frac{1}{n} \mu(o_1) + \frac{1}{n} \mu(o_2) + \ldots\ldots + \frac{1}{n} \mu(o_n)$$

$$= \Sigma \frac{1}{n} \mu(o_i) = \Sigma \frac{1}{n} \mu = n. \frac{1}{n} \mu = \mu$$

På samme måte finnes

$$\sigma(m)^2 = (\frac{1}{n})^2 . \sigma(o_1)^2 + (\frac{1}{n})^2 . \sigma(o_2)^2 + \ldots\ldots + (\frac{1}{n})^2 . \sigma(o_n)^2$$

$$= \Sigma(\frac{1}{n})^2 . \sigma(o_i)^2 = \Sigma(\frac{1}{n})^2 . \sigma^2 = n.(\frac{1}{n})^2 . \sigma^2 = \frac{\sigma^2}{n}$$

Oppgave 40. Ta fram en eller annen bok. Tell antall ord på 3 bokstaver pr. hel linje for 100 linjer. Beregn det gjennomsnittlige antall ord på 3 bokstaver for 10 og 10 linjer. Derved fås 10 gjennomsnitter. Beregn så gjennomsnittet og middelavviket for disse gjennomsnitter på vanlig måte. Beregn også gjennomsnittet og middelavviket for alle 100 observasjoner under ett. Vis så at gjennomsnittet av de 10 gjennomsnittene er lik gjennomsnittet av alle 100 observasjoner og at middelavviket for de 10 gjennomsnittene er tilnærmet lik middelavviket for de 100 observasjonene dividert med $\sqrt{10}$.

I tillegg til den karakteristikk som $\mu(m) = \mu$ og $\sigma(m) = \sigma/\sqrt{n}$ gir av fordelingsloven for gjennomsnittet av n uavhengige observasjoner av et variabelt kjennetegn, er det bevist at denne fordelingsloven for voksende n vil nærme seg til den normale fordelingslov (III,9). Eksakt normal er den ikke med mindre fordelingsloven for kjennetegnet er normal. Men hvis fordelingsloven for kjennetegnet ikke avviker altfor meget fra den normale type, vil tilnærmelsen selv for små verdier av n være så pass tilfredsstillende at en uten noen risiko for feilslutninger kan anta at den er eksakt normal. Följelig vil P i tabell 2 (III,9) angi sannsynligheten for

$$|m - \mu| \geq a \frac{\sigma}{\sqrt{n}}$$

Sannsynligheten for at tallverdien av differensen mellom gjennomsnittet av n uavhengige observasjoner av et variabelt kjennetegn og forventningen for dette kjennetegn, skal overskride a = 3 ganger $\sigma/\sqrt{n}$ er derfor meget liten, nemlig etter tabell 2 lik 0,0027.

Fordelingsloven for kjennetegnet er i alminnelighet ikke kjent,

iallfall ikke kjent på en slik måte at en kan beregne forventningen og spredningen. Men det hender at en har en hypotese om fordelingsloven og hvis da denne hypotesen inneholder så mange opplysninger at en kan beregne forventningen og spredningen, har en anledning til å beregne grenseverdiene for et tallområde der gjennomsnittet av n uavhengige observasjoner av kjennetegnet med en gitt sannsynlighet vil komme til å falle. Setter en etter tabell 2 (III,9) a = 3, vil grenseverdiene for dette område være

$$\mu - 3.\frac{\sigma}{\sqrt{n}} \quad \text{og} \quad \mu + 3.\frac{\sigma}{\sqrt{n}}$$

og den tilsvarende sannsynlighet lik 0,9973. Sannsynligheten for at gjennomsnittet vil falle utenfor dette område er meget liten, nemlig bare 0,0027. Dette gir oss anledning til i visse tilfelle å stille en pröve på en hypotese ved å sammenligne den med gitte observasjoner.

I III,8 har vi framsatt den hypotesen at fordelingsloven for det kjennetegn som vi har observasjoner av i eks. 2 (II,2) er en binomiallov

$$f(x) = \frac{k!}{x!\,(k-x)!}\; p^x q^{k-x}$$

med parameterverdiene k = 21 og p = 0,61. Vi har også beregnet den hypotetiske forventning og spredning (III,8) og funnet $\mu = 12{,}81$ og $\sigma = 2{,}24$. Følgelig er den hypotetiske forventning og spredning i fordelingsloven for gjennomsnittet av de n = 1905 observasjonene lik:

$$\mu(m) = \mu = 12{,}81$$

$$\sigma(m) = \frac{\sigma}{\sqrt{n}} = \frac{2{,}24}{\sqrt{1905}} = 0{,}05$$

Altså er

$$\mu - 3.\frac{\sigma}{\sqrt{n}} = 12{,}81 - 0{,}15 = 12{,}66$$

og

$$\mu + 3.\frac{\sigma}{\sqrt{n}} = 12{,}81 - 0{,}15 = 12{,}96$$

Sannsynligheten for at m skal falle innen dette hypotetiske område er altså 0,9973. Gjennomsnittet av de n = 1905 observasjonene er m = 12,76 og dette ligger innen området. Når det gjelder gjennomsnittet kan en derfor si at hypotesen er bekreftet. Men dette må ikke misforstås derhen at hypotesen i sin helhet er bekreftet og at en kan godta den. Hadde derimot gjennomsnittet falt utenfor dette hypotetiske område, ville en med god grunn kunne slutte at det er noe i veien med hypotesen.

Oppgave 41. Anta at de n = 1905 gjentakelsene i eks. 2 (II,2) er et tilfeldig utvalg av et univers hvor fordelingsloven for det observerte kjennetegn er en binomiallov med parametrene k = 21 og p = 0,5. Undersök om denne hypotesen må forkastes på grunnlag av en sammenligning mellom gjennomsnittet og forventningen for x.

Oppgave 42. Besvar samme spörsmål som i oppg. 27 (III,11) på grunnlag av en sammenlikning mellom gjennomsnittet og forventningen for kjennetegnet x.

D R. P E R O T T E S T A D

F o r e l e s n i n g e r

over

M A T E M A T I K K o g S T A T I S T I K K

ved

NORGES LANDBRUKSHØGSKOLE

IV. Statistisk induksjon

Skrivemaskinstua
Oslo, april 1947

I n n h o l d :

IV. Statistisk induksjon.

1. Induksjon.

Induksjon betyr i logikken en form for slutning som består i at en sammenfatter de erfaringer en har gjort i visse enkelttilfelle i alminnelige setninger. Alle de oppfatninger vi har om økonomiske, biologiske, arbeidstekniske forhold osv. er oppfatninger som i det aller vesentlige er bygget på induksjoner. Noen induksjoner føler vi oss temmelig sikre på, andre stiller vi oss noe skeptiske til, og atter andre vil vi kanskje avvise som altfor mangelfullt underbygget. Selv om vi ikke hadde det minste kjennskap til de fysiske lover, vil vi ikke være det minste i tvil om at det er lettere å trekke en vogn på flat vei enn oppover bakke forutsatt at veidekket er det samme. Vi ville nemlig anse det for utelukket at noe menneske noen gang hadde gjort den motsatte erfaring. Men hvis noen kommer til oss og sier at "professor N.N. er skallet og grinete, følgelig er alle skallete professorer grinete", vil vi naturligvis avvise dette som en løs påstand og det selv om vi personlig kanskje ikke kjenner noen skallet og ikke-grinete professor. Denne slutningen er likevel en induksjonsslutning. Vedkommende har konstatert i et enkelt tilfelle - professor N.N. - at kjennetegnene skallet og grinete opptrer sammen. På grunnlag av denne erfaring har han så dannet seg en oppfatning som uttrykkes i den alminnelige setning "alle skallete professorer er grinete".

Det er i og for seg klart at pålideligheten av en induksjon for en vesentlig del er avhengig av det erfaringsmateriale den grunner seg på. Men det kommer også an på hvilket logisk resonnement som benyttes. La oss f. eks. tenke oss at en gårdbruker har dyrket to sorter kveite og at han for den ene sorten har fått en avling på 300 kg pr. dekar og for den andre 280 kg pr. dekar. På grunnlag av denne erfaring anbefaler han så at en skal bruke den første sorten. Det er klart at vi ikke kan godta denne induksjon i denne generelle form. Det kan nemlig tenkes at forskjellen mellom sortene vil forsvinne eller slå om til det motsatte dersom sortene ble dyrket under andre dyrkningsforhold. Vi vil derfor ha rede på om dyrkningsforsøket var planlagt og utført på en slik måte at den konstaterte avlingsdifferens gir uttrykk for forskjelligheter mellom sortene. Vi vil bl.a. ha rede på om forsøket er utført slik at sortene har vært likt stilt m.h.p. jordbunnsforhold, jordbearbeidelse, gjødsling osv. Kan ikke vedkommende stille oss tilfreds med hensyn til disse spørsmål, har vi ingen grunn til å godta hans induksjon. Men ikke nok med det. Vi vet at det også fins et stort antall ukontrollerbare faktorer som sammen kan øve en betydelig innflytelse på avlingens størrelse. Og vi vil

derfor også kreve beskjed om at det ikke kan være slike faktorer som har frembragt forskjellen mellom sortene.

Til enhver vitenskapelig undersøkelse må en stille det krav at slike spørsmål som vi har antydet med vårt eksempel, kan besvares og besvares på grunnlag av mest mulig eksakte vurderinger. En må alltid finne seg i at det kan reises tvil om pålitelighetен av en induksjon. Men graden av pålitelighet vil være avhengig av hvor eksakte vurderinger det er som ligger til grunn for den.

De erfaringer som en bygger på i enhver statistisk undersøkelse, er de observasjonene en har skaffet seg. Disse er hentet fra et (eller flere) utvalg av gjentakelser, og når en induserer på grunnlag av disse observasjoner, vil det si at en slutter et eller annet om universet (eller universene) som gjentakelsene er utvalg av. En har da prøvet å finne fram til mest mulig eksakte metoder for slike induksjoner. Og den metode en har valt er den som anvises i logikken, nemlig i størst mulig utstrekning å basere induksjonen på deduktive slutninger.

En deduktiv slutning er den motsatte av induksjon. Det er en form for logisk slutning som går ut på at en med utgangspunkt i en mer almen setning formulerer setninger om spesielle forhold. Denne form for slutning er karakteristisk for matematikken.

Vi har tidligere (III, 11 og 14) diskutert observasjonene i eks. 1 (III, 4). Vi sammenlignet dem med den hypotesen at sannsynligheten for kjennetegnet (hendingen) A=normale børster i en gjentakelse er lik $s(A,U)=p=0,75$ og at gjentakelsene er uavhengige. Vi fant da at det var så stor uoverensstemmelse at vi ble tvunget til å slutte at det måtte være noe i veien med hypotesen. La oss nå se noe nærmere på logikken i vårt resonnement.

Det første trinn i resonnementet besto i at vi i tanken konstruerte et hypotetisk univers hvor $s(A,U)=0,75$. Med utgangspunkt i dette universet utledet vi så visse regler for frekvensen for A i et tilfeldig utvalg på n gjentakelser:

1) at fordelingsloven for frekvensen er en binomiallov,
2) at forventningen for frekvensen er np og spredningen $\sqrt{npq}$, og
3) at fordelingsloven for frekvensen når antallet av gjentakelser er så stort som i det foreliggende tilfelle, $n=2835$, er så nær normal at vi kan bruke tabell 2 (III, 9) som basis for vurderingen av forskjellen mellom en observert frekvens og dens forventning.

Dette er konsekvenser som er utledet av hypotesen ved eksakte deduktive slutninger.

Det neste trinn i resonnementet besto i at vi sammenlignet den faktiske frekvens for A (2211) med disse konsekvenser. Derved fikk vi den opplysning at denne frekvens ikke kunne være frekvensen for A i et tilfeldig utvalg på n=2635 gjentakelser i det hypotetiske universet. Konklusjonen måtte derfor bli at hypotesen ikke var bekreftet.

Det første trinn i dette resonnementet besto av et antall eksakte deduktive slutninger. Deretter ble det trukket en slutning som ikke kan sies å ha streng nødvendighet fordi også meget usannsynlige hendinger vil kunne inntreffe av og til. Men i og med at vi kunne fastslå at den hypotetiske sannsynlighet for at A skal inntreffe et antall ganger som avviker fra forventningen med et beløp som er lik eller større enn den funne avvikelse, ikke er større enn 0,0005, hadde vi skaffet oss et tallmessig uttrykk for pålideligheten av induksjonen.

Vi fant at

$$\frac{|h(A,U)-\mu|}{\sigma} = 3,68$$

La oss tenke oss at vi hadde funnet 2,68. Sannsynligheten for en avvikelse lik eller større enn denne, er etter tabell 2 (III, 9) noe mindre enn 0,012. Vi ville kanskje også i det tilfelle ha forkastet hypotesen, men det er klart at i så fall ville induksjonen være basert på et mindre overbevisende grunnlag. Den ville være mindre pålitelig.

La oss også tenke oss at vi i eks. 1 (III, 4) hadde funnet h(A,U) ikke lik 2211, men la oss si 2130. Da ville vi ha

$$\frac{|h(A,U)-\mu|}{\sigma} = \frac{|h(A,U)-np|}{\sqrt{npq}} = \frac{2130-2126,25}{23,05} = \frac{3,75}{23,05} = 0,16$$

Hva kunne vi slutte av et slikt resultat? Vi måtte naturligvis si at det var en meget bra overensstemmelse mellom frekvensen for A og den hypotetiske forventning. Men dette betyr ikke uten videre at vi kan godta hypotesen. For å kunne trekke denne slutningen måtte vi også skaffe en garanti for at ingen andre hypoteser ville kunne gi en like god eller bedre overensstemmelse. Og en slik garanti har vi aldri. Det meste vi kunne vise var at denne hypotesen stemmer bedre med frekvensen enn alle andre hypoteser vi i øyeblikket kan finne på. Men det kan jo være andre muligheter enn de vi kan tenke oss.

En aldri så god overensstemmelse mellom konsekvensene av en hypotese og observasjonene, må derfor aldri oppfattes som bevis for hypotesen. Dermed er det imidlertid ikke sagt at en slik overensstemmelse ikke er en positiv opplysning. Hvis en nemlig sammenligner konsekvensene

av en hypotese med stadig nye observasjoner og hver gang finner bra overensstemmelse, er det naturlig at en etterhvert kommer til å oppfatte dette som bevis på at hypotesen gir en treffende forklaring på den sak en er interessert i. En kan iallfall ikke avvise denne form for logisk resonnement.

Det kan ikke i noe tilfelle bli tale om å slutte at hypotesen er "riktig". En hypotese er jo bare en foreløpig forklaring som er oppstilt til prøvning og prøvningen består i at en sammenligner med de observasjoner en kan skaffe seg. Men etterhvert som iakttakelsesmetodene forfines, blir flere og flere detaljer brakt fram i dagen og forklaringene (hypotesene og teoriene) må oftest revideres. Dette kommer tydelig til syne når en studerer vitenskapenes historie.

Det vil forhåpentlig fremgå av det foregående hvor meget mer pålitelig en induksjon kan bli når vi følger skjemaet:

1) formulering av en eksakt hypotese,
2) eksakt deduktiv utledning av dens konsekvenser,
3) sammenlikning med observasjonene, og
4) induksjon som går ut på at en forkaster hypotesen hvis den hypotetiske sannsynlighet for det resultat sammenlikningen har gitt er tilstrekkelig liten.

Vi skal ta for oss et nytt eksempel. En råne med 15 ribber ble parret med purker med 15 ribber og med purker med 16 ribber. Ribbeantallet (x) ble bestemt for hvert enkelt avkom. Fordelingsrekkene for x er gitt i følgende tabell.

x	Frekvens for x	
	Mødre 15 ribb.	Mødre 16 ribb.
14	3	1
15	53	14
16	42	39
17	0	1
n	98	55
m	15,39	15,73
s	0,54	0,53

Hensikten med denne undersøkelsen var å få bragt på det rene om ribbeantallet er genetisk betinget. Vi konstaterer at det gjennomsnittlige antall ribber for avkom av mødre med 16 ribber er noe større enn det gjennomsnittlige antall ribber for avkom av mødre med 15 ribber. Dette er imidlertid ikke noe bevis. For det kan jo tenkes at forskjellen mellom

de to gjennomsnittene ville forsvinne eller endog slå om til det motsatte fortegn dersom en skaffet seg et større antall observasjoner. Et svar på det oppstilte spørsmål kan vi bare få hvis det på grunnlag av de foreliggende observasjonene kan påvises at forventningen for x ikke er den samme i de to universene som disse gjentakelsene er utvalg av.

La oss bruke følgende betegnelser:

	For avkom av mødre med	
	16 ribber	15 ribber
Antall obs.	n_1	n_2
Gjennomsnitt	m_1	m_2
Middelavvik	s_1	s_2
Forventning	μ_1	μ_2
Spredning	σ_1	σ_2

Ifølge III,15 er

$$\mu(m_1) = \mu_1 \qquad \text{og} \qquad \sigma(m_1) = \sigma_1/\sqrt{n_1}$$

$$\mu(m_2) = \mu_2 \qquad \text{og} \qquad \sigma(m_2) = \sigma_2/\sqrt{n_2}$$

Vi vet dessuten at fordelingsloven for et gjennomsnitt er meget nær normal (se III,15), og det kan bevises (beviset tas ikke med her) at også fordelingsloven for differensen mellom to uavhengige gjennomsnitter er normal. Ifølge III,13 er

$$\mu(m_1-m_2) = \mu(m_1) - \mu(m_2) = \mu_1 - \mu_2$$

og

$$\sigma(m_1-m_2) = \sqrt{\sigma(m_1)^2 + \sigma(m_2)^2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Følgelig er sannsynligheten for

$$\frac{|(m_1-m_2)-(\mu_1-\mu_2)|}{\sigma(m_1-m_2)} \geq a$$

gitt ved tabell 2 (III,9).

La oss nå hypotetisk anta at $\mu_1 = \mu_2$. Da er sannsynligheten for

$$\frac{|m_1-m_2|}{\sigma(m_1-m_2)} \geq a$$

også gitt i tabell 2. Denne hypotesen kan imidlertid ikke prøves fordi vi ikke kjenner σ_1 og σ_2. og derfor heller ikke $\sigma(m_1-m_2)$. Men la oss erstatte σ_1 med s_1 og σ_2 med s_2. Da er

$$\sigma(m_1-m_2) \approx \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} = \sqrt{\frac{0,53^2}{55}+\frac{0,54^2}{98}} = 0,09$$

altså er

$$\frac{|m_1-m_2|}{\sigma(m_1-m_2)} \approx \frac{15,73-15,39}{0,09} = \frac{0,34}{0,09} = 3,8$$

Hvis nå σ_1 og σ_2 hadde vært eksakt lik s_1 og s_2, kunne vi avlese av tabell 2 at sannsynligheten for et resultat som er lik eller større enn det vi har funnet, er noe mindre enn 0,0005 som svarer til a=3,5. Det funne resultat måtte derfor betraktes som så urimelig at vi ble tvunget til å forkaste hypotesen $\mu_1 = \mu_2$. Følgelig måtte vi slutte at μ_1 og μ_2 er ulike, og denne konklusjon er jo svar på det spørsmål som vi hadde stillet oss som oppgave å undersøke.

Det er imidlertid klart at tabell 2 ikke gjelder helt strengt når spredningene erstattes av de beregnete middelavvik. Men vi skal vise senere at når antallet av observasjoner er så stort som i det foreliggende tilfelle, er tabell 2 en tilstrekkelig god tilnærmelse slik at vi kan bruke den uten noen risiko for feilslutninger på grunn av at vi erstatter σ med s.

Oppgave 1.

I 1935 ble n_1=18 høyprøver undersøkt m.h.p. innhold av fosfor. Observasjonene (% P) er:

0,183	0,150	0,123	0,163	0,138	0,123
0,100	0,145	0,138	0,131	0,160	0,146
0,140	0,154	0,117	0,131	0,131	0,145

I 1936 ble også n_2=18 prøver analysert. Observasjonene (%P) er:

0,136	0,126	0,127	0,134	0,132	0,121
0,131	0,118	0,123	0,132	0,111	0,129
0,126	0,173	0,130	0,139	0,122	0,124

Undersøk om det er en reell forskjell mellom høyets P-innhold i de to år.

I de følgende avsnitt skal vi prøve å forklare noen av de mest alminnelig brukte statistiske induksjonsmetodene. Dessverre vil fremstillingen få noe av kokebokens preg over seg. De grunnleggende matematiske deduksjonene kan vi nemlig ikke forklare fordi det for forståelsen av disse kreves langt større matematisk innsikt enn vi kan forutsette.

2. Sann verdi, feil og konfidensintervall.

La oss anta at vi har målt en og samme fysiske størrelse n ganger. La observasjonene være $o_1, o_2, o_3, \ldots o_n$. Vi vil gå ut fra at den målte fysiske størrelsen har en konstant verdi C som vi kaller den sanne verdi. Dette er naturligvis en antakelse som må betraktes som en tilnærmelse. Ingen fysiske størrelser er nemlig konstante i absolutt forstand. Vi vet således at avstanden mellom to fine streker på en stålstav er avhengig av temperaturen, og heller ikke noen andre fysiske størrelser har en helt uforanderlig verdi. Når vi derfor forutsetter at det fins en sann verdi, må dette oppfattes i relativ forstand, nemlig i den forstand at ubestemtheten ved størrelsen er av helt underordnet betydning i forhold til de feil vi gjør når vi maler den. I noen tilfelle har en riktignok funnet opp så nøyaktige målemetoder at en kan si at målefeilene er av omtrent samme størrelsesorden som ubestemtheten ved den størrelsen som måles. Men de problemer som melder seg i slike tilfelle ligger helt utenfor rammen av det vi skal beskjeftige oss med her. Tenker vi på slike størrelser som måles f. eks. i landmålingen, kan vi trygt forutsette at de er konstante i forhold til størrelsen av målefeilene. Det kan være slike størrelser som den rettlinjete avstand mellom to punkter, nivådifferensen mellom to punkter, vinkelen mellom to punkter sett fra et tredje punkt osv.

Vi vil forutsette at målingene (observasjonene) er utført så nøyaktig som mulig og at en har gjort sitt beste for å gjøre alle observasjonene like nøyaktige. For alt vi vet er da den ene observasjon en like god bestemmelse av C som enhver av de andre. Vi sier at observasjonene er like gode. Videre vil vi forutsette at målingene er utført på uavhengig måte slik at det resultat en har fått ved en måling ikke på noen måte har noe å si for resultatet av en annen måling.

Målefeilene $(o_i - C)$ skyldes i regelen årsaker av to ulike slag. Av den grunn oppfatter en feilen som en sum av to feil. Den ene av disse kalles den systematiske eller den regelmessige feil. Den virker i samme retning og i mange tilfelle også med samme styrke på alle observasjonene til å gjøre disse enten systematisk for små eller systematisk for store. Den andre feilen virker helt uregelmessig, den skifter både størrelse og fortegn fra observasjon til observasjon. Vi kaller denne feil den tilfeldige feil.

Den systematiske feil skyldes selve observasjonsmetoden, feil

ved instrumentene o.l. Ytre faktorer som f. eks. lysbrytning, jordoverflatens krumning osv. kan også i noen tilfelle være av betydning. Denne feilen kan som oftest bestemmes på en eller annen måte og deretter kan den elimineres av observasjonene. Tilbake blir da den tilfeldige feil som alltid skyldes årsaker som en ikke har noe middel til å kontrollere. I teorien antar en den skyldes et stort antall selvstendig virkende årsaker slik at en kan oppfatte den som en sum av et stort antall av hverandre uavhengige elementærfeil. Årsakene til den tilfeldige feil er ikke alltid lett å påvise i hvert enkelt tilfelle. Men tenker en på en måling som utføres med en teodolitt, kan en lett finne noen av årsakene: underlaget som instrumentet er plasert på er aldri i absolutt ro, en skjelver på hånden osv. At den tilfeldige feil skyldes årsaker som ikke kan kontrolleres, betyr ikke at en ikke kan redusere virkningen av dem. Ved å arbeide nøyaktig og samvittighetsfullt og ved å bruke gode instrumenter kan en naturligvis skaffe seg bedre observasjoner, dvs. observasjoner med mindre tilfeldig feil, enn de observasjoner en får ved unøyaktig arbeid og mindreverdige instrumenter.

Oppgaven går naturligvis ut på å bruke observasjonene til å bestemme den sanne verdi C. Vi vil i det følgende forutsette at alle feil av systematisk art er fjernet fra observasjonene slik at deres avvikelse fra C utelukkende skyldes tilfeldige feil.

For å komme fram til en metode til bestemmelse av C må vi tenke oss et univers av uendelig mange uavhengige gjentakelser av enkeltmålinger av C. Det kjennetegn som observeres er da lik C plus sumvirkningen av de tilfeldige, ukontrollerbare feilårsaker. Dette kjennetegn (x) tilhører de kontinuerlig variable kjennetegn. En antar nå at dette kjennetegn har normal fordelingslov

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{(x-C)^2}{2\sigma^2}}$$

hvor altså C er forventningen og σ spredningen. Forutsetter en nemlig at den tilfeldige feil er sum-virkningen av et meget stort antall (teoretisk uendelig antall) av hverandre uavhengig virkende årsaker, kan det bevises at fordelingsloven for x er normal. De erfaringer som en har gjort ved å gjenta målingen av en fysisk størrelse et stort antall ganger, tyder også på normal fordelingslov.

Hvis vi nå danner gjennomsnittet av n observasjoner av x

$$m = \frac{\Sigma\, o_i}{n}$$

vet vi (III,15) at også dette har normal fordelingslov med forventningen C og spredningen $\sigma/\sqrt{n}$. Sannsynligheten for at m skal avvike fra C med et beløp som er lik eller mindre enn a ganger $\sigma/\sqrt{n}$ kan derfor finnes i tabell 2 (III,9). Sannsynligheten for en avvikelse lik eller mindre enn a=3 ganger $\sigma/\sqrt{n}$ er altså praktisk talt lik enheten (0,9973). Det er m.a.o. praktisk talt sikkert at gjennomsnittet av n observasjoner som ikke har noen systematiske feil, vil falle et eller annet sted mellom grensene

$$C - 3.\ \sigma/\sqrt{n} \qquad \text{og} \qquad C + 3.\ \sigma/\sqrt{n}$$

Det er imidlertid ikke m som skal bestemmes ved C men C ved m. Vi må derfor så å si snu helt opp ned på hele resonnementet. Dette byr på en del teoretiske vanskeligheter som vi ikke kan forklare her. I virkeligheten er vel heller ikke fagstatistikerne helt enige om saken. Men det er iallfall enighet om at det beste en kan foreta seg er å avgrense et intervall fra

$$m - 3.\ \sigma/\sqrt{n} \qquad \text{til} \qquad m + 3.\ \sigma/\sqrt{n}$$

og si at det er praktisk talt sikkert at C befinner seg innen dette intervall.

Grensene for dette intervall kan imidlertid ikke bestemmes fordi spredningen (σ) ikke er kjent. I praksis tillater en seg derfor å erstatte denne med middelavviket for observasjonene

$$s = \sqrt{\frac{\Sigma(o_i - m)^2}{n-1}}$$

og sette

$$m - 3.s/\sqrt{n} \leq C \leq m + 3.s/\sqrt{n}$$

Dette intervall kalles et konfidensintervall for C.

Blant annet fordi s ikke er eksakt lik σ, kan en ikke si at sannsynligheten for at C befinner seg innen dette intervall er lik 0,9973. Men en antar at sikkerheten for at C skal falle innen dette intervall er tilstrekkelig for praktiske formål. I neste avsnitt skal vi imidlertid se at vi kan forbedre teorien meget nettopp på dette punkt.

Det kan naturligvis diskuteres om det er nødvendig å sette a=3. Etter tabell 2 (III,9) vil en jo oppnå en meget bra sikkerhet også ved å sette a=2,5 eller bare a=2. En pleier derfor når en skal referere resultatet av en måleserie å oppgi gjennomsnittet og middelavviket hver for seg. En pleier også ofte å sette

$$C = m \pm s/\sqrt{n}$$

For observasjonene i eks. 5 (II,3) har vi før funnet m=901,453 og s=0,0397. Altså er $s/\sqrt{n} = 0,0397/\sqrt{20} = 0,009$. Konfidensintervallet for C har derfor i dette tilfelle grensene

$$m - 3.s/\sqrt{n} = 901,426 \text{ og } m + 3.s/\sqrt{n} = 901,480$$

En kan også sette

$$C = m \pm s/\sqrt{n} = 901,453 \pm 0,009$$

Oppgave 2.

En bue ble målt på uavhengig måte n=10 ganger med samme metode og under like betingelser ellers. Observasjonene er:

85° 42',20	85° 42',00	85° 41',97
85° 42',07	85° 42',10	85° 42',15
85° 42',05	85° 42',13	85° 41',90
85° 41',93		

Bestem konfidensintervallet for buens sanne verdi.

3. Student's fordelingslov.

I forrige avsnitt ble det vist hvordan en kan bestemme et konfidensintervall for en ukjent sann verdi når en tillater seg å erstatte den ukjente spredning (σ) med middelavviket for observasjonene. Middelavviket er imidlertid bare et estimat av spredningen og når vi derfor erstatter σ med s, gjør vi oss skyldige i en unøyaktighet som naturligvis helst bør unngås. Vi skal nå vise at vi kan unngå denne unøyaktighet ved å gjøre bruk av en fordelingslov som kalles Student's fordelingslov. Den ble utledet av en engelsk statistiker som publiserte sine arbeider under psevdonymet "Student".

La oss anta at x er et kjennetegn som har normal fordelingslov med forventningen μ og spredning σ. La oss videre anta at vi har observert x i et tilfeldig utvalg på n gjentakelser og derved fått n observasjoner av x : $o_1, o_2, \ldots o_n$. La gjennomsnittet og variansen for disse observasjonene være m og $V=s^2$. Etter det vi har forklart i III,15 må vi da betrakte m som et estimat av μ og V som et estimat av σ^2.

La oss nå av m og V danne en ny størrelse, nemlig

$$t = \frac{|m - \mu|}{\sqrt{V/n}}$$

Nevneren i denne formelen er estimatet av spredningen i fordelingsloven for gjennomsnittet (m).

De n uavhengige gjentakelser i det univers hvor x er et kjennetegn, er nå til sammen en gjentakelse i et annet univers og t er observasjonen av et kjennetegn u i denne gjentakelse. Det kan da bevises (beviset tas ikke med her) at fordelingsloven for t i dette universet - når en forutsetter at x har normal fordelingslov - er

$$f(u) = \frac{K}{(u^2+f)^{\frac{1}{2}(f+1)}}$$

Her er f=n-1 og K er en konstant som er uavhengig av μ og σ og bare avhengig av f. I fordelingsloven for t er det m.a.o. bare en parameter som forekommer, nemlig f=n-1. Følgelig kan en beregne for valte verdier av f sannsynligheten (P) for $t \geq a$ hvor a er et valt tall. Denne sannsynlighet er lik arealet av den flaten som er begrenset av kurven for f(u), abscisseaksen (u-aksen) og ordinaten til u=a. Setter vi f. eks. f=19 og a=2,861 (sml. eks. 5,II,3) fins P=0,01.

En har beregnet en rekke sammenhørende verdier av P,f og a. Disse kunne da stilles opp i en tabell over P for valte verdier av f og a. Av praktiske grunner har en imidlertid ordnet disse sammenhørende verdier på en annen måte. Den tabellen som brukes i praksis er en tabell over verdiene av a for valte verdier av P og f, altså en tabell med a som avhengig variabel og P og f som uavhengig variable. Denne tabellen er gjengitt i Tab. I bakerst i boken.

Vi ser at Tab. I er oppstilt for f=1 til f=30. Dessuten er det tatt med f=40, f=60, f=120 og $f \to \infty$. Sammenligner vi nå a-verdiene for hver verdi av P i de fire nederste rekkene, ser vi at det er praktisk talt ingen forskjell mellom disse rekkene. Dette kommer av at når $f \to \infty$, går fordelingsloven for t - f(u) - over til den normale fordelingslov, og tilnærmelsen til denne er praktisk talt fullkommen allerede fra f=30. Tallene i nederste rekke ($f \to \infty$) er verdiene av a svarende til valte verdier av P for den normale fordelingslov. Det er altså bare tabell 2 (III,9) om igjen på en annen måte. I tabell 2 er det jo a som er uavhengig variabel og P som er avhengig variabel, i Tab. I er rollene byttet om.

Nå er jo P i Tab. I sannsynligheten for $t \geq a$. Følgelig er Q=1-P sannsynligheten for $t \leq a$, dvs. for

$$|m-\mu| \leq a\sqrt{\frac{V}{n}} = a.\frac{s}{\sqrt{n}}$$

eller for

$$\mu - a.s/\sqrt{n} \leq m \leq \mu + a.s/\sqrt{n}$$

Gjelder det å bestemme konfidensintervallet for en ukjent sann verdi $C = \mu$ ved hjelp av gjennomsnittet (m) og middelavviket (s) for n uavhengige observasjoner, setter vi

$$m - a.s/\sqrt{n} \leq C \leq m + a.s/\sqrt{n}$$

hvor vi for a velger den verdi som etter Tab. I svarer til f=n-1 og den P-verdi vi ønsker å bruke. For eks. 5 (II,3) har vi m=901,453, $s/\sqrt{n}$ = 0,009 og f=n-1=19. Nøyer vi oss med en sannsynlighet på P=0,01, dvs. Q=0,99, skal vi etter Tab. I sette a=2,861. Det til P=0,01 svarende konfidensintervall for den ukjente sanne verdi har derfor grensene:

$$m - a.s/\sqrt{n} = 901,453 - 2,861.0,009 = 901,427$$

og $$m + a.s/\sqrt{n} = 901,453 + 2,861.0,009 = 901,479$$

Oppgave 3.

I oppg. 2 (forr. avsnitt) er det oppgitt n=10 observasjoner av en ukjent bue. Finn konfidensintervallene for denne buen svarende til P=0,05, P=0,01 og P=0,001.

Student's fordelingslov kan også brukes til løsning av mange andre oppgaver. Vi skal i det følgende ta for oss to av disse.

Vi har før (II,7) referert resultatet av en undersøkelse som tok sikte på å få brakt på det rene om kvelstoffinnholdet i sauegjødsel forandres når gjødsla konserveres i kloroform og oppbevares i isskap i noen døgn. For å få brakt dette på det rene ble hver av n=16 prøver delt i to prøver. Den ene av disse ble analysert m.h.p. N med det samme, mens gjødsla var frisk, den andre ble konservert i kloroform, oppbevart i isskap i 5 døgn og deretter analysert. En fikk på denne måten n=16 uavhengige observasjoner av endringen i N under oppbevaringen i isskap. Disse observasjonene (O_i) er gitt i eks. 10 (II,7).

La oss nå anta hypotetisk at disse observasjonene er observasjoner av et kjennetegn som har normal fordelingslov med forventningen $\mu = 0$. Da er sannsynligheten for $t \geq a$, hvor

$$t = \frac{|m-\mu|}{s/\sqrt{n}} = \frac{m-0}{s/\sqrt{n}} = \frac{m}{s/\sqrt{n}}$$

gitt som P i Tab. I.

Vi har at

$$m = \frac{\Sigma O_i}{n} = \frac{0,3744}{16} = 0,0234$$

Videre finnes:

$$V = \frac{\Sigma(o_i - m)^2}{n-1} = \frac{0,003379}{15} = 0,000225$$

og

$$\frac{s}{\sqrt{n}} = \sqrt{\frac{V}{n}} = \sqrt{\frac{0,000225}{16}} = 0,0038$$

Følgelig er

$$t = \frac{m}{s/\sqrt{n}} = \frac{0,0234}{0,0038} = 6,15$$

Av Tab. I ser vi at til f=n-1=15 og P=0,001 svarer det a=4,073. Vi har funnet en verdi av t som er betydelig større enn denne a. Og vi må derfor slutte at den oppstilte hypotesen ($\mu = 0$) ikke er bekreftet, vi må forkaste den. Følgelig må vi godta den alternative hypotesen og slutte at det skjer en endring i gjødslas N-innhold under oppbevaringen.

Vi må riktignok ta et forbehold her. Saken er jo nemlig den at vi har antatt at observasjonene er observasjoner av et kjennetegn med normal fordelingslov. Tab. I er jo oppstilt under denne forutsetning. Det er derfor en mulighet for at vi opererer med en feilaktig fordelingslov for t. Dette spørsmål er imidlertid undersøkt, og resultatene av disse undersøkelser viser at selv meget store avvikelser fra den normale fordelingslov bare medfører endringer i fordelingsloven for t som ikke har noen praktiske konsekvenser. I slike tilfelle som det vi nettopp har behandlet, kan vi iallfall trygt gå ut fra at fordelingsloven for kjennetegnet avviker så lite fra den normale at det ikke medfører noen som helst risiko for feilslutning om vi bruker Tab. I.

Oppgave 4.

I følgende tabell er oppgitt vektene (i gram) av 10 par 7 uker gamle kyllinger. De to kyllingene som tilhører samme par har samme foreldre. Under o_i er oppgitt vektene av kyllinger oppvokset innadørs og under o_i' vektene av kyllinger oppvokset i en utendørs hønsegård.

Foreldre nr.	o_i	o_i'	Foreldre nr.	o_i	o_i'
1	270	240	6	300	320
2	510	450	7	330	330
3	420	330	8	390	300
4	390	330	9	390	420
5	450	270	10	450	420

Kan en på grunnlag av disse observasjonene slutte noe m.h.t. betydningen av oppvekststedet for kyllingenes vekst?

Oppgave 5.

I følgende tabell er for en rekke år oppgitt frøavlingen i kg frø pr. dekar (Vollebekk) for to sorter erter. Kan en på grunnlag av disse observasjonene påvise at det er en reell forskjell mellom de to sortene m.h.p. frøavlingen?

År	sort	
	Onsrud	Norsk grå
1933	244	149
1934	253	218
1935	242	211
1936	185	158
1937	255	279
1938	177	223
1939	32	66
1940	249	203
1941	127	123
1942	137	153
1943	220	147
1944	204	233

Vi har foran (IV,1) vist hvordan en ved å sammenlikne gjennomsnittene av observasjonene av et kjennetegn for to utvalg av gjentakelser, kan bedømme om disse utvalg er tilfeldige utvalg av to universer som er ulike m.h.p. forventningen for kjennetegnet. Vi benyttet oss da av tabell 2 (III,9) og

$$\frac{|m_1 - m_2|}{\sigma(m_1 - m_2)}$$

hvor m_1 er gjennomsnittet for det ene og m_2 gjennomsnittet for det annet utvalg. Det betenkelige ved den fremgangsmåten som vi benyttet oss av, er at vi for å beregne $\sigma(m_1-m_2)$ må erstatte spredningene med de beregnete middelavvik. Særlig betenkelig må jo dette være når antall gjentakelser i de to utvalgene er små tall. Vi skal nå se at vi kan rette på dette ved å bruke Student's fordelingslov.

La $o_1, o_2, \ldots$ være observasjonene av et kjennetegn i et utvalg på n_1 gjentakelser og $o_1', o_2', \ldots$ observasjonene av det samme kjennetegn i et annet utvalg på n_2 gjentakelser. La gjennomsnittene være m_1 og m_2.

La oss nå hypotetisk anta at kjennetegnet har identiske normale fordelingslover i de to universene som de to utvalg er tilfeldige utvalg av. Det kan da bevises (beviset tas ikke med her) at

$$t = \frac{|m_1 - m_2|}{s} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}$$

hvor $$s^2 = \frac{\Sigma(o_i - m_1)^2 + \Sigma(o_i' - m_2)^2}{n_1 + n_2 \div 2}$$

har Student's fordelingslov til fordelingslov når vi setter $f = n_1 + n_2 - 2$. Sannsynligheten for $t \geq a$ er m.a.o. lik P i Tab. I.

Vi skal ta for oss et eksempel. En dyrket havre i kar under varierte forsøksbetingelser. En brukte 3 parallellkar for hvert forsøksledd, dvs. at $n_1 = n_2 = 3$. Vi skal her ta med resultatene ved ulike vatning. Forsøksbetingelsene (grunnmedium, gjødsling osv.) var ellers like for alle kar. Forskjellen i faktoren vatning var

A: 30 % av full metning
B: 60 % av full metning

Resultatene er gitt i følgende tabell. Observasjonene er gram tørrstoff pr. kar i loavlingen

A			B		
o_i	$o_i - m$	$(o_i - m_1)^2$	o_i'	$o_i' - m_2$	$(o_i' - m_2)^2$
22,22	+ 0,58	0,3364	30,16	+ 0,99	0,9801
23,39	+ 1,75	3,0625	28,97	- 0,20	0,0400
19,31	- 2,33	5,4289	28,38	- 0,79	0,6241
64,92	0,00	8,8278	87,51	0,00	1,6442

Herav finnes

$$m_1 = \frac{64,92}{3} = 21,64 \quad \text{og} \quad m_2 = \frac{87,51}{3} = 29,17$$

Altså er

$$|m_1 - m_2| = 29,17 - 21,64 = 7,53$$

Videre finnes:

$$s^2 = \frac{8,8278 + 1,6442}{3+3-2} = \frac{10,4720}{4} = 2,6180$$

og $$s = \sqrt{2,6180} = 1,62$$

Følgelig er

$$t = \frac{|m_1 - m_2|}{s} \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} = \frac{7,53}{1,62} \sqrt{\frac{3 \cdot 3}{6}} = 5,69$$

Her er nå $f=n_1+n_2-2 = 4$ og av Tab. I vil en se at sannsynligheten for $t \geq a = 4{,}604$ er $P = 0{,}01$. Vi har funnet en t-verdi som er betydelig større enn denne a og slutter derfor at vår hypotese ikke er bekreftet, vi forkaster den. Fordelingslovene for kjennetegnet er m.a.o. ikke identiske normale fordelingslover.

Dette resultatet er naturligvis meget lite opplysende. Men for det første kan vi som nevnt foran sløyfe "normal" og si at fordelingslovene for kjennetegnet ikke er identiske i de to universene. Forskjellen kan da bero på ulikhet i forventningene, spredningene og andre karakteristikker. Teoretisk sett kan en naturligvis tenke seg at to slike fordelingslover er forskjellige selv om forventningene er like. Men slike forskjelligheter forekommer neppe i virkeligheten. En slutter derfor at forventningene for kjennetegnet er ulike i de to universene. Som vi skal vise i neste avsnitt kan en foreta en uavhengig sammenlikning mellom middelavvikene - eller rettere variansene - for de to rekker observasjoner for å prøve om det kan påvises noen forskjell mellom spredningene.

Oppgave 6.

I følgende tabell betyr o_i observasjoner av % aske i prøver av kull fra en kullgruve og o_i' observasjoner av samme kjennetegn i prøver av kull fra en annen gruve. Kan en på grunnlag av disse observasjonene påvise at det er forskjell mellom askeinnholdet i kull fra de to gruvene?

o_i	o_i'
24,3	18,2
20,8	16,9
23,7	20,2
17,4	16,7
21,3	

Oppgave 7.

I IV,1 har vi diskutert om ribbeantallet hos griser er genetisk betinget. Benytt de samme observasjonene og undersøk saken ved hjelp av Student's fordelingslov.

Oppgave 8.

Benytt Student's fordelingslov som grunnlag for undersøkelsen av det spørsmål som er oppstilt i oppg. 1 (IV,1).

4. V a r i a n s a n a l y s e n.

I foregående avsnitt har vi vist hvordan Student's fordelingslov kan brukes når det gjelder å sammenlikne to rekker observasjoner av et kjennetegn. Vi skal nå vise at vi også kan sammenlikne flere rekker under ett. I følgende tabell er under A og B oppgitt observasjonene fra foregående avsnitt av tørrstoff (i gram pr. kar) i loavlingen av havre dyrket i kar vatnet til 30 % metning (A) og 60 % metning (B). Under C er så oppgitt observasjonene av samme kjennetegn under forsøksbetingelsen 90 % metning. Bortsett fra graden av vannmetning var forsøksbetingelsene ens for de tre forsøksledd. I de tre nederste rekkene er oppgitt gjennomsnittet (m), middelavviket (s) og variansen ($V=s^2$).

	A	B	C
	22,22	30,16	30,66
	23,39	28,97	30,61
	19,31	28,38	31,77
m:	21,64	29,17	31,01
s:	2,10	0,91	0,66
V:	4,4139	0,8221	0,4301

Vi ser at gjennomsnittet tiltar og middelavviket avtar fra A til C. Dette tyder på at den varierte faktoren (vannmetningen) har innflytelse på verdien av det observerte kjennetegn. Men vi har naturligvis ikke lov til å slutte at det er slik med mindre vi kan bygge på resultatet av en eksakt statistisk prøvning. Vi må derfor fremsette en eksakt hypotese til sammenlikning med observasjonene. Foreløpig vil vi imidlertid nøye oss med en ikke eksakt formulering og si at hypotesen går ut på at den varierte forsøksfaktoren ikke har noen som helst betydning for verdien av det observerte kjennetegn.

I det foreliggende tilfelle er forsøksfaktoren kvantitativ. Vi kan si at de tre grupper av gjentakelser atskiller seg fra hverandre i verdien av et variabelt kjennetegn, nemlig % vannmetning. Vi kunne derfor gjerne angripe spørsmalet om vannmetningenes eventuelle innflytelse på tørrstoffavlingen ved hjelp av en regresjons-korrelasjonsundersøkelse (sml. eks. 16,II,13). En slik undersøkelse måtte da kombineres med statistisk prøvning av en eller flere eksakte hypoteser. Dette skal vi imidlertid ikke komme inn på i denne sammenheng.

Det hender imidlertid meget ofte at gjentakelsene ordnes i grupper etter alternative konstante kjennetegn. I en eksperimentalundersøkelse av hvilke faktorer det er som har betydning for kornavlingen, kan

gjentakelsene være ordnet i grupper etter de sorter en vil sammenlikne eller etter kvalitativt variert gjødsling. I et foringsforsøk med sauer kan gjentakelsene være ordnet i grupper etter kvalitativt variert foring osv. I slike tilfelle kan en naturligvis ikke benytte seg av regresjons-korrelasjonsmetoder.

La oss anta at vi har observert et variabelt kjennetegn hos N gjentakelser som er ordnet i k grupper enten etter verdier av et variabelt kjennetegn eller etter k alternative konstante kjennetegn. Vi gir hver gruppe et nummer j og betegner antallet av gjentakelser i den j'te gruppe med n_j. I vårt eksempel er N=9, k=3 og $n_1=n_2=n_3=3$. Observasjonene av kjennetegnet betegner vi med o_{ij} hvor fotskriften j betyr at vedkommende observasjon tilhører den j'te gruppe. Gjennomsnittet og variansen for observasjonene i den j'te gruppe vil vi betegne med m_j og V_j. Altså er

$$m_j = \frac{\sum_i o_{ij}}{n_j} \qquad \text{og} \qquad V_j = \frac{\sum_i (o_{ij}-m_j)^2}{n_j-1}$$

Vi kan naturligvis også beregne gjennomsnittet av alle N observasjonene. La oss kalle det totalgjennomsnittet. Dette er

$$m = \frac{1}{N}\sum_{ji}\sum o_{ij} = \frac{1}{N}\sum_j n_j \cdot \frac{\sum_i o_{ij}}{n_j} = \frac{1}{N}\sum_j n_j m_j$$

Totalgjennomsnittet er altså også et gjennomsnitt av gruppegjennomsnittene, nemlig det gjennomsnitt vi får når en tillegger hvert av disse gjennomsnitter en vekt lik antallet av observasjoner (n_j) som det er beregnet av. Hvis $n_1=n_2= \ldots . =n_k=n$, er naturligvis $N = n.k$, og da er

$$m = \frac{1}{n.k}\sum_j n.m_j = \frac{1}{k}\sum_j m_j$$

Vi skal nå vise hvordan vi ved hjelp av gruppegjennomsnittene og gruppevariansene kan beregne to nye varianser som vi kan bruke til en eksakt prøvning av den hypotesen som vi foran har gitt en foreløpig og ueksakt formulering av. Disse to variansene kalles V (innen) og V (mellom).

På samme måte som totalgjennomsnittet er et veiet gjennomsnitt av gruppegjennomsnittene, er den første av disse variansene - V (innen) - et veiet gjennomsnitt av gruppevariansene. Som vekter bruker en n_j-1. Og siden nå

$$\sum_j (n_j-1) = \sum_j nj - k = N - k$$

setter en

$$V\ (\text{innen}) = \frac{1}{N-k} \sum_j (n_j - 1).V_j$$

En kan også sette

$$V\ (\text{innen}) = \frac{1}{N-k} \sum_j (n_j - 1) \cdot \frac{\sum_i (o_{ij} - m_j)^2}{n_j - 1} = \frac{1}{N-k} \sum_{ji} \sum (o_{ij} - m_j)^2$$

Den andre variansen er

$$V\ (\text{mellom}) = \frac{1}{k-1} \sum_j n_j.(m_j - m)^2$$

Hvis antall gjentakelser i hver gruppe er det samme, $n_j = n$, blir formlene for V (innen) og V (mellom) forenklet til:

$$V\ (\text{innen}) = \frac{1}{k} \Sigma V_j$$

og

$$V\ (\text{mellom}) = \frac{n}{k-1} \Sigma (m_j - m)^2$$

I vårt eksempel har vi $n_j = n = 3$ og følgelig er

$$V\ (\text{innen}) = \frac{4,4139 + 0,8221 + 0,4301}{3} = \frac{5,6661}{3} = 1,8887$$

Totalgjennomsnittet er

$$m = \frac{1}{k} \Sigma m_j = \frac{81,82}{3} = 27,27$$

V (mellom) beregnes så slik:

m_j	$m_j - m$	$(m_j - m)^2$
21,64	- 5,63	31,6969
29,17	+ 1,90	3,6100
31,01	+ 3,74	13,9876
81,82		49,2945

og

$$V\ (\text{mellom}) = \frac{3}{2}\ 49,2945 = 73,9418$$

Når disse variansene er beregnet, beregner en forholdet (F) mellom dem. En må da alltid passe på at en setter F lik den største varians dividert med den minste. For vårt eksempel har vi altså at

$$F = \frac{73,9418}{1,8887} = 39,15$$

Som allerede antydet foran går den hypotesen vi vil konfrontere med observasjonene ut på at det observerte kjennetegn har identisk samme fordelingslov i de k universene som de k grupper av gjentakelser er tilfeldige utvalg av. I dette ligger da ikke bare det at disse fordelingslovene skal være av samme type, men at også parametrene i disse fordelingslovene skal ha nøyaktig samme verdier. En antar derfor også hypotetisk at forventningen og spredningen for kjennetegnet skal ha samme verdier i de k universene. For å kunne gjennomføre den matematiske deduksjonen som er nødvendig, har en dessuten måttet forutsette hypotetisk at denne felles fordelingsloven er normal.

Forutsetter en nå at de k grupper av gjentakelser er tilfeldige utvalg av universer med identisk samme normale fordelingslov for det kjennetegn en observerer, kan det bevises (dette bevis er vanskelig og tas derfor ikke med her) at fordelingsloven for F er

$$f(u) = A.\frac{u^{\frac{1}{2}(f_1-2)}}{[f_1u + f_2]^{\frac{1}{2}(f_1+f_2)}}$$

I denne formelen forekommer det bare to parametrer, nemlig f_1 og f_2. Disse kalles antall frihetsgrader. Setter en

$$F = \frac{V\ (\text{mellom})}{V\ (\text{innen})} \text{ er } f_1 = k-1 \quad \text{og } f_2 = N-k$$

og setter en

$$F = \frac{V\ (\text{innen})}{V\ (\text{mellom})} \text{ er } f_1 = N-k \text{ og } f_2 = k-1$$

Faktoren A er en størrelse som vi ikke behøver å bry oss med. Den kan beregnes når f_1 og f_2 er gitt. Det er derfor mulig å beregne en gang for alle for enhver valt verdi av f_1 og enhver valt verdi av f_2 sannsynligheten P for $F \geq a$ hvor a er et valt tall. Hvis vi skulle tabulere verdiene av P for valte verdier av f_1, f_2 og a, måtte vi bruke et tabellsystem med tre innganger. Vi måtte altså bruke en bunke av tabeller, hver tabell med to innganger. I et slikt system ville da P opptre som avhengig variabel og f_1, f_2 og a som uavhengig variable. Av praktiske grunner har en imidlertid valt å tabulere de sammenhørende verdier av disse fire størrelser slik at en har a som avhengig variabel og P, f_1 og f_2 som uavhengig variable. Slike tabeller er offentliggjort for P=0,2, P=0,05, P=0,01 og P=0,001. En har altså ialt fire tabeller over sammenhørende verdier av a (som avhengig variabel) og f_1 og f_2 (som uavhengig variable). Vi

gjengir i Tab. II bakerst i boken den av disse som svarer til P=0,01. Denne tabellen har da inngang over f_1 og f_2, og den er ordnet slik at tallene i hodet av tabellen er f for den største varians og tallene i venstre kolonne er f for den minste varians.

For vårt eksempel har vi funnet V (mellom) > V (innen). For V (mellom) har vi f=k-1=3-1=2 og for V (innen) har vi f=N-k=9-3=6. Vi går følgelig inn i Tab. II med f=2 i tabellens hode og med f=6 i venstre kolonne og avleser a=10,92. Dette betyr da at den hypotetiske sannsynlighet for F > 10,92 er P=0,01. Denne funne verdi av F (39,15) er betydelig større, og derav slutter vi at vår hypotese må forkastes.

Vår hypotese går ut på at fordelingslovene for det observerte kjennetegn ikke bare er identiske i alle de k universene som gruppene er tilfeldige utvalg av, men at de er identisk normale. Har en derfor i et aktuelt tilfelle fått en F som er så stor at en må forkaste denne hypotesen, betyr det at de k gruppene av gjentakelser ikke er tilfeldige utvalg av universer med identisk samme <u>normale</u> fordelingslov for det observerte kjennetegn. Dette resultat er i de aller fleste tilfelle av meget liten verdi. Ofte vil en kunne trekke denne slutningen uten alle disse beregningene. Vi må derfor se nærmere på saken.

For det første er det da viktig å bringe på det rene om en i konklusjonen kan sløyfe "normal" og si at gruppene ikke er tilfeldige utvalg av universer med identisk samme fordelingslov for det observerte kjennetegn. For å få rede på dette har en undersøkt om a-verdiene for samme f_1, f_2 og P blir vesentlig forskjellige fra dem som fins i Tab. II når en hypotetisk forutsetter at de k gruppene er tilfeldige utvalg av universer med identisk samme ikke-normale fordelingslov. En har prøvet en rekke ulike typer av ikke-normale fordelingslover. Resultatene av disse undersøkelser viser at fordelingsloven for F er meget lite påvirket av endringer i typen av fordelingslov for det observerte kjennetegn. En kan iallfall gå ut fra at fordelingslovene for de kjennetegn en vanligvis har utsikt til å arbeide med i praksis (iallfall når det gjelder biologiske undersøkelser), avviker så lite fra den normale at en ikke risikerer å komme til feil resultater når en bruker den utledete fordelingsloven for F og dermed Tab. II. Men en bør naturligvis være oppmerksom på at en i visse tilfelle kan komme til å arbeide med kjennetegn som har en så avvikende fordelingslov at en må være meget forsiktig.

Dette er imidlertid ikke nok. Vi må prøve å skaffe oss nærmere rede på hva det betyr at de k gruppene ikke er tilfeldige utvalg av universer med samme fordelingslov for det observerte kjennetegn. For-

skjellen mellom fordelingslovene kan jo bety både det ene og det annet. Det kan bety at forventningen for kjennetegnet er forskjellig i de k universene eller at spredningene er forskjellige eller at det er andre karakteristikker som er forskjellige. Det kan naturligvis også bety at både forventningene og spredningene og andre karakteristikker er forskjellige i de k universene.

For å få nærmere rede på dette skal vi undersøke noen konsekvenser av vår hypotese som vi kan utlede uten å måtte ty til vanskelige matematiske hjelpemidler. For enkelthets skyld vil vi da forutsette at gruppene består av det samme antall gjentakelser, altså at $n_1=n_2= \ldots . = n_k = n$. La nå μ og σ være den felles forventning og den felles spredning for kjennetegnet i de k universene. Ifølge (III,15) er da forventningen og spredningen for gjennomsnittet (m_j) av $n_j=n$ observasjoner av kjennetegnet lik

$$\mu(m_j) = \mu$$

og

$$\sigma(m_j) = \sigma/\sqrt{n}$$

Variansen i gjennomsnittets fordelingslov er altså lik σ^2/n.

Vi har nå k gjennomsnitter og følgelig kan vi beregne (estimere) denne variansen direkte. Vi har at

$$\frac{\sigma^2}{n} \approx \frac{\Sigma(m_j-m)^2}{k-1}$$

og derfor er

$$\sigma^2 \approx \frac{n}{k-1} \Sigma (m_j-m)^2 = V(\text{mellom})$$

V(mellom) er m.a.o. et estimat av variansen σ^2 når gruppene er tilfeldige utvalg av universer med samme forventning og spredning for kjennetegnet. Vi kan vise at under samme forutsetning er også V(innen) et estimat av σ^2. V_j er jo variansen for $n_j=n$ observasjoner av kjennetegnet i et tilfeldig utvalg av gjentakelser. Følgelig må V_j også være et estimat av σ^2. Det samme må imidlertid gjelde gjennomsnittet av disse varianser, dvs. at

$$\sigma^2 \approx \frac{1}{k} \Sigma V_j = V(\text{innen})$$

En konsekvens av vår hypotese er derfor at både V(innen) og V(mellom) er estimater av σ^2. Følgelig skal vi vente (altså som en

konsekvens av hypotesen) at disse to variansene har omtrent samme verdi og at forholdet mellom dem er omtrent lik enheten.

Hvis da V(mellom) i et aktuelt tilfelle er så stor i forhold til V(innen) at vi må forkaste hypotesen, betyr det at gruppegjennomsnittene atskiller seg fra hverandre i vesentlig sterkere grad enn vi skulle vente dersom kjennetegnet hadde samme forventning og spredning i de k universene som gruppene er tilfeldige utvalg av. Vi kan også si at det betyr at vi ved grupperingen av gjentakelsene har dannet tilfeldige utvalg av subuniverser i et felles univers slik at forventningen for det observerte kjennetegn er forskjellig i disse subuniversene.

Hvis derimot V(innen) er så stor i forhold til V(mellom) at vi må forkaste hypotesen, betyr det at gruppegjennomsnittene atskiller seg fra hverandre i vesentlig mindre grad enn vi skulle vente dersom kjennetegnet hadde samme forventning og spredning i de k universene som gruppene er tilfeldige utvalg av. Utslag i denne retning forekommer nokså sjelden og årsaken er ikke alltid lett å finne. Et slikt utslag må komme av at det fins en eller annen forbindelse mellom gruppene som virker til å utviske forskjellen mellom gruppegjennomsnittene. Da dette resultat forekommer så sjelden og har så liten betydning for praktisk statistikk, skal vi ikke ta opp saken til noen mer inngående drøftelse.

Oppgave 9.

Følgende karakterer er gitt i kjemi for 1. klasse M i årene fra 1938 til 1944. Undersøk om det er foregått noen endring i karakternivået i disse årene.

1938	1939	1940	1941	1942	1943	1944
2,0	2,5	1,8	1,8	1,8	1,5	1,8
3,8	2,3	3,0	1,0	3,0	1,8	2,5
1,3	2,0	1,5	2,8	3,3	3,5	2,5
1,5	1,8	1,8	3,5	2,3	1,3	2,8
2,8	1,3	2,8	1,5	1,8	4,5	2,0
2,8	2,3	1,8	2,3	2,5	1,5	1,5
2,8	1,5	2,3	1,0		2,0	1,8
		2,5				2,5

Oppgave 10.

I eks. 11 (II,8) er oppgitt vekten av avlingene (røtter) for tre betesorter. Kan det påvises at de tre sortene gir ulike avlinger, og må det i tilfelle dette kan påvises, tas noe forbehold m.h.t. slutningens almengyldighet?

Oppgave 11.

I enkelte distrikter på Vestlandet fant en en påfallende liten vektøkning hos lammene etter at de var satt på inneforing om høsten. For å få rede på grunnen til dette, ble det i 1940-41 utført følgende

forsøk. N=32 lam ble delt (på tilfeldig vis) i k=4 grupper, hver gruppe på n=8 individer. Disse gruppene ble satt på ulike foring etter følgende plan:

gruppe A: vanlig foring uten tilskudd av andre mineralstoffer enn koksalt
gruppe B: som A med tilskudd av Ca og P
gruppe C: som A med tilskudd av mikronæringsstoffer
gruppe D: som B med tilskudd av mikronæringsstoffer

Den gjennomsnittlige vektøkning (i gram pr. dag) ble så bestemt for hvert lam. Observasjonene er:

A	B	C	D
78	204	151	152
97	165	155	178
- 22	169	158	186
158	77	196	172
- 12	50	125	150
189	- 51	160	117
8	42	40	131
113	21	114	115

Undersøk om forskjellen i foringen har noen innflytelse på vektøkningen.

Vi har foran gitt en beskrivelse av bruken av variansanalysen i det aller enkleste tilfelle, nemlig når gruppene er dannet etter en enveis klassifisering. I oppg. 9 er gruppene dannet ved klassifisering etter årene, i oppg. 10 etter sortene og i oppg. 11 etter variert foring. Men en har meget ofte to-veis eller mange-veis klassifiseringer. I et feltforsøk med korn kan en f. eks. i samme forsøk klassifisere etter sort, etter variert gjødsling, etter variert jordbearbeidelse osv. Variansanalysen kan også brukes i slike tilfelle. Men det vil føre for langt her å gi en beskrivelse av den fremgangsmåten som da må brukes.

Vi skal imidlertid nevne en annen viktig anvendelse av variansanalysen. La oss anta at vi har observasjoner av et variabelt kjennetegn i to utvalg av gjentakelser. På grunnlag av disse kan vi da beregne to gjennomsnitter og to middelavvik. Ved hjelp av Student's fordelingslov (Tab. I) kan vi så undersøke om forventningene for kjennetegnet er ulike i de to universene som utvalgene representerer. En t-verdi som er større enn den a-verdi i Tab. I som svarer til P=0,01, vil i alminnelighet være et tilstrekkelig bevis for at disse forventningene er ulike. Det kan imidlertid også i mange tilfelle ha stor interesse å sammenlikne middelavvikene for de to observasjonsrekkene for å bringe på det rene om det kan sluttes at spredningene for kjennetegnet i de to universene er forskjellige.

La oss anta at antallet av observasjoner i den ene rekken er n_1 og i den andre n_2 og la middelavvikene være s_1 og s_2. Det kan da be-

vises at vi kan sammenlikne variansene $V_1=s_1^2$ og $V_2=s_2^2$ ved hjelp av Tab. II. Vi beregner forholdet mellom disse to variansene (alltid den største dividert med den minste) og går inn i Tab. II med $f=n_2-1$ for V_2 og $f=n_1-1$ for V_1. Den hypotesen som prøves på denne måten går ut på at fordelingslovene for kjennetegnet i de to universene som utvalgene representerer, er normale med samme spredning. En forutsetter altså ikke hypotetisk at forventningene er like.

For det eksemplet vi har brukt foran (gram tørrstoff i loavlingen av havre i kar under ulike vannmetning) har vi funnet:

gruppe A: $V_1 = 4,4139$ med $f=n_1-1 = 2$

gruppe C: $V_2 = 0,4301$ med $f=n_2-1 = 2$

Altså er

$$F = \frac{4,4139}{0,4301} = 10,26$$

Av Tab. II sees at denne F-verdi er betydelig mindre enn den a-verdi som svarer til $f_1=f_2=2$ og P=0,01. Denne er nemlig a=99,01. Det er m.a.o. ingen påviselig forskjell mellom spredningene for kjennetegnet i de to universer som gruppene er tilfeldige utvalg av. Men av dette må en naturligvis ikke slutte at disse to spredningene er like.

Oppgave 12.

I oppg. 1 (IV,1) er gitt to rekker observasjoner. Undersøk om det er en påviselig forskjell mellom spredningene for det observerte kjennetegn i de to universene som utvalgene er uttatt av.

Ennå en sak må nevnes før vi forlater dette emnet. Det kan kanskje virke noe forvirrende at når en skal sammenlikne to rekker observasjoner bruker en t og Tab. I og når en skal sammenlikne tre eller flere rekker, bruker en F og Tab. II. Setter vi imidlertid

$$F = \frac{V(\text{mellom})}{V(\text{innen})}$$

og k=2, dvs. at f for V(mellom) er lik f=k-1=1, er det meget lett å vise at

$$F = t^2$$

Settes nemlig k=2, finner en

$$V(\text{mellom}) = \frac{1}{k-1}\sum_j n_j(m_j-m)^2 = n_1(m_1-m)^2 + n_2(m_2-m)^2$$

og da i dette tilfelle

$$m = \frac{n_1 m_1 + n_2 m_2}{n_1 + n_2}$$

finner en lett at

$$V(\text{mellom}) = (m_1-m_2)^2 \frac{n_1 \cdot n_2}{n_1+n_2}$$

Videre finnes at

$$V(\text{innen}) = \frac{1}{N-k} \sum_j \sum_i (o_{ij}-m_j)^2 = \frac{\Sigma(o_{i1}-m_1)^2 + \Sigma(o_{i2}-m_2)^2}{n+n-2} = s^2$$

idet $N=n_1+n_2$.
Altså blir

$$F = \frac{V(\text{mellom})}{V(\text{innen})} = \frac{(m_1-m_2)^2}{s^2} \frac{n_1 \cdot n_2}{n_1+n_2} = t^2$$

Når k=2 er derfor F-prøven og t-prøven i virkeligheten den samme. Det er naturligvis likegyldig for saken om en bruker t og Tab. I eller om en bruker $F=t^2$ og Tab. II. Det er imidlertid mest alminnelig at en når k=2 bruker t-prøven. Og som vi allerede har sett (IV,3) har jo Student's fordelingslov andre anvendelser hvor F og fordelingsloven for F ikke kan brukes.

Endelig skal det til slutt nevnes at opprinnelig da variansanalysen først ble utformet som forskningsmetode, brukte en ikke F men

$$z = \tfrac{1}{2} \cdot \log.\text{nat. } F$$

Enkelte forskere bruker fremdeles z og fremstillingen av variansanalysen er i enkelte lærebøker fremdeles basert på bruken av z og ikke F.

5. K j i - k v a d r a t m e t o d e n.

La sannsynligheten for at et kjennetegn eller en hending A skal inntreffe i en gjentakelse i universet U være $s(A,U)=p$. Da er sannsynligheten for at A skal inntreffe i z av n uavhengige gjentakelser lik

$$s(z,W) = \frac{n!}{z!\,(n-z)!} p^z q^{n-z}$$

hvor q=1-p og W er et univers hvor hver gjentakelse består av n uavhengige gjentakelser i U (se III,11). Dette er altså fordelingsloven for frekvensen for A i universet W. Forventningen og spredningen er $\mu=np$ og $\sigma = \sqrt{npq}$.

I det følgende vil vi forutsette at np er et så stort tall at en kan regne med at binomialloven kan erstattes av den normale fordelings-

lov (se III,11). Som en praktisk regel kan en da si at np helst ikke bør være mindre enn 10 og slett ikke mindre enn 5.

Sett nå at vi har observert at A har inntruffet i h(A,U)=h av n gjentakelser. Og la oss hypotetisk anta at s(A,U)=p og at de n gjentakelsene er uavhengige gjentakelser i et univers U. Vi kan da som vist i III,11 sette en prøve på denne hypotesen ved å beregne

$$\frac{|h - np|}{\sqrt{npq}}$$

og bruke tabell 2 (III,9). Det har imidlertid mange fordeler å bruke kvadratet på denne størrelsen og sette

$$\chi^2 = \frac{(h - np)^2}{npq}$$

Forutsetter en at fordelingsloven for frekvensen for A er normal med forventningen np og spredningen $\sqrt{npq}$, kan det bevises (beviset tas ikke med her) at fordelingsloven for χ^2 er av typen

$$f(u) = A.u^{\frac{1}{2}(f-2)}\, e^{-\frac{1}{2}u}$$

Her er e grunntallet i det naturlige logaritmesystem, A er en konstant og f er antallet av de såkalte frihetsgrader. Vi skal senere prøve å forklare hvordan f fremkommer. Foreløpig må vi nøye oss med å si at den er en parameter i f(u) som i bestemte tilfelle har en bestemt tallverdi. I det foreliggende tilfelle er f=1.

Faktoren A kan beregnes når f er gitt. I fordelingsloven for χ^2 er derfor f den eneste parameter som forekommer. Det kan bevises at forventningen for χ^2 er $\mu(\chi^2) = f$. En kan derfor for hver valt verdi av f beregne sannsynligheten P for $\chi^2 \geq a$ hvor a er et valt tall. Denne sannsynligheten er lik arealet av den flaten som er begrenset av kurven for f(u), abscisseaksen (u-aksen) og ordinaten til u=a. De sammenhørende verdier av f,a og P kan så ordnes i en tabell med to innganger. Denne tabellen er gjengitt i Tab. III bakerst i boken. Den er som Tab. I ordnet slik at det er f og P som er uavhengig variable og a som er avhengig variabel. Vi ser at den omfatter f-verdier fra f=1 til f=30. For større verdier av f har $\sqrt{2\chi^2}$ tilnærmet normal fordelingslov med forventningen $\sqrt{2f-1}$ og spredningen 1. For f-verdier større enn 30 beregner en derfor

$$\sqrt{2\chi^2} - \sqrt{2f-1}$$

og bruker Tab. I for $f \to \infty$.

La oss ta for oss et eksempel. I eks. 1 (III,4) er n=2835 og frekvensen for A=normale børster er h(A,U)=h=2211. I III,11 har vi hypotetisk satt s(A,U)=p=0,75 og funnet

$$\frac{|h - np|}{\sqrt{npq}} = 3,68$$

Følgelig er

$$\chi^2 = \frac{(h-np)^2}{npq} = 3,68^2 = 13,54$$

I dette tilfelle er f=1. Av Tab. III ser vi at til f=1 og P=0,001 svarer a=10,827. Vi har funnet et χ^2 som er betydelig større enn denne a, og konsekvensen må da bli at vi i overensstemmelse med resultatet av den undersøkelsen vi foretok i III,11 slutter at hypotesen ikke er bekreftet, vi forkaster den.

En meget viktig egenskap ved fordelingsloven f(u) som gjør χ^2 særlig egnet for praktisk statistiske undersøkelser er følgende. La χ_1^2, χ_2^2, χ_3^2 .... χ_k^2 være k uavhengige kji-kvadrater med antall frihetsgrader lik f_1, f_2, f_3, ... f_k. Det kan da bevises at også

$$\chi^2 = \sum_{i=1}^{k} \chi_i^2$$

har f(u) til fordelingslov når en setter $f = \Sigma f_i$. Summen av en rekke uavhengige kji-kvadrater er altså selv et kji-kvadrat med antall frihetsgrader lik summen av frihetsgradene for addendene.

Vi skal ta for oss et eksempel for å demonstrere bruken av denne setningen. I følgende tabell er n_i antall avkom i k=10 familier av Gammarus og $h_i=h_i(A,U)$ er antallet av disse avkom som har A-røde øyne. La oss nå anta hypotetisk at s(A,U)=p=0,25 og at de enkelte avkom er uavhengige gjentakelser. Vi kan sette en prøve på denne hypotesen ved å beregne et χ^2 for hver familie slik som vist i tabellen. Deretter summerer vi de k=10 kji-kvadrater og får derved et χ^2 for hele materialet.

n_i	h_i	n_ip	h_i-n_ip	$(h_i-n_ip)^2$	n_ipq	χ_i^2
93	14	23,25	- 9,25	85,5625	17,4375	4,91
151	31	37,75	- 6,75	45,5625	28,3125	1,61
30	6	7,50	- 1,50	2,2500	5,6250	0,40
146	29	36,50	- 7,50	56,2500	27,3750	2,05
79	17	19,75	- 2,75	7,5625	14,8125	0,51
99	20	24,75	- 4,75	22,5625	18,5625	1,22
78	12	19,50	- 7,50	56,2500	14,6250	3,85
56	11	14,00	- 3,00	9,0000	10,5000	0,86
75	14	18,75	- 4,75	22,5625	14,0625	1,60
77	13	19,25	- 6,25	39,0625	14,4375	2,71
884	167					χ^2 =19,72

Hvert enkelt χ_i^2 har i dette tilfelle $f_i=1$ frihetsgrad og følgelig har sum $\chi^2 = 19,72$ f=10 frihetsgrader og forventningen f=10. Av Tab. III ser vi at til P=0,01 og f=10 svarer det a=23,209. Vi har funnet et χ^2 som er noe mindre enn denne verdi, men det er større enn den a som svarer til samme f og P=0,05. Det er derfor ingen god overensstemmelse mellom hypotesen og observasjonene, men det funne χ^2 er likevel ikke så stort at en med støtte i bare dette kan forkaste hypotesen.

I et tilfelle som dette har naturligvis hvert enkelt χ_i^2 sin selvstendige betydning. Hvert enkelt av dem har $f_i=1$ frihetsgrad, og av Tab. III ser en at til f=1 og P=0,01 svarer det a=6,635. Alle χ_i^2 i vårt eksempel er mindre enn denne verdi.

La oss tenke oss at vi har et utvalg på k=100 kji-kvadrater. I dette utvalg skulle vi da vente at et av disse kji-kvadrater skulle være større enn den a som svarer til P=0,01. Selv i et mindre utvalg, f. eks. på k=20, ville det ikke være noe urimelig i at et av kji-kvadratene var større enn denne verdi. En må derfor være ytterst forsiktig med å tillegge addendene i et sum-χ^2 for stor selvstendig betydning. Er sum-χ^2 mindre enn den a som svarer til P=0,01 for det antall frihetsgrader som gjelder for summen, kan det godt være at en eller flere av addendene er større enn den a som svarer til P=0,01 og antallet av frihetsgrader for den enkelte addend uten at vi kan slutte at dette betyr noe. Likevel må en ikke overse det faktum at hver addend kan ha sin særlige interesse. Overser en dette, kan en komme til å overse uoverensstemmelser med den oppstilte hypotese som kan danne utgangspunkt for nye fruktbare undersøkelser.

Av stor interesse er det å legge merke til at fortegnet for differensene (h_i-n_ip) ikke har noen betydning for χ_i^2 og dermed heller ikke for summen. Det er uten betydning for sum-χ^2 om alle disse differensene er negative, alle positive eller noen negative og andre positive. Men når det gjelder å ta standpunkt til hypotesen er dette ikke likegyldig. I vårt eksempel er alle disse differensene negative og dette tyder naturligvis på at den hypotetiske p=0,25 er noe for stor. Det er derfor nødvendig å undersøke om vi ikke kan beregne et χ^2 som er mer virkningsfullt enn vårt sum-χ^2.

La oss da ta for oss summene $\Sigma n_i = 884$ og $\Sigma h_i = 167$. Det er naturligvis ikke noe til hinder for at vi kan si at vi har n=884 gjentakelser og at kjennetegnet A har inntruffet i h=167 av disse og sammenlikne disse tallene med hypotesen s(A,U)=p=0,25 og at alle 884 gjentakelsene er uavhengige. Vi får da et χ^2 som er lik

$$\chi^2 = \frac{(h-np)^2}{npq} = \frac{(167-884.0,25)^2}{884.0,25.0,75} = \frac{(167-221)^2}{165,75} = 17,59$$

Dette χ^2 har f=1 frihetsgrad og det er betydelig større enn den a som etter Tab. III svarer til f=1 og P=0,001. Denne verdi er nemlig a=10,827. Vi må derfor forkaste hypotesen.

Grunnen til at det siste χ^2 i dette tilfelle er mer virkningsfullt enn sum-χ^2, er nettopp det at fortegnene for (h_i-n_ip) ikke har noe å bety for sum-χ^2, men øver en avgjørende innflytelse på χ^2 beregnet på grunnlag av $n=\Sigma n_i$ og $h=\Sigma h_i$. En bør derfor i slike tilfelle som dette beregne begge kji-kvadrater.

Oppgave 13.

I følgende tabell er n_i antall avkom i k=5 kull av bananfluen og h_i er antall avkom med karakteren "sepia". Anta hypotetisk at sannsynligheten for denne karakter (kjennetegn) er p=0,25 og at de enkelte avkom er uavhengige gjentakelser. Undersøk hvordan denne hypotesen stemmer med observasjonene.

n_i	h_i
886	204
447	106
1271	302
1231	318
1202	303
5037	1233

La oss nå igjen tenke oss at et kjennetegn A har inntruffet i h(A,U) av n gjentakelser og at vår hypotese går ut på at s(A,U)=p og at gjentakelsene er uavhengige. Vi skal da vise at χ^2 kan beregnes på en annen måte enn slik vi har beskrevet foran. Når A har inntruffet i h(A,U) av n gjentakelser, har naturligvis den alternative hending (kjennetegn) iA inntruffet i h(iA,U)=n-h(A,U) gjentakelser. Og når s(A,U)=p, er s(iA,U)=1-p=q. Følgelig er den hypotetiske forventning for frekvensen for A lik np og forventningen for frekvensen for iA lik n-np = n(1-p) = nq. Det er da lett å vise at

$$\frac{[h(A,U)-np]^2}{npq} = \frac{[h(A,U)-np]^2}{np} + \frac{[h(iA,U)-nq]^2}{nq}$$

Skriver vi nemlig til avkortning $h(A,U)=h_1$ og $h(iA,U)=h_2=n-h_1$, har vi at

$$\frac{(h_1-np)^2}{np} + \frac{(h_2-nq)^2}{nq} = \frac{q(h_1-np)^2+p(n-h_1-n+np)^2}{npq}$$

$$= \frac{q(h_1-np)^2 + p(h_1-np)^2}{npq} = \frac{(p+q).(h_1-np)^2}{npq} = \frac{(h_1-np)^2}{npq}$$

Vi kan derfor i dette tilfelle beregne χ^2 på to måter. Og enten vi nå bruker den ene eller den annen fremgangsmåten, har det χ^2 vi får f=1 frihetsgrad.

Når vi interesserer oss bare for ett kjennetegn (A), kan n gjentakelser deles opp i bare to grupper. I den ene gruppen plaseres de gjentakelsene som har A og i den andre de som har iA. Hvis vi interesserer oss for to kjennetegn, A og B, kan gjentakelsene deles opp i fire grupper, nemlig AB,AiB,iAB og iAiB. Eks. 1 (III,4) er et eksempel på en slik oppdeling. I dette tilfelle er de kjennetegn som karakteriserer gruppene sammensatte konstante kjennetegn. Men det kan også hende at vi har flere alternative enkle kjennetegn. Eksempel 1 er et eksempel på dette. Her er gjentakelsene (n=97) oppdelt i grupper etter tre alternative enkle kjennetegn.

Eksempel 1.

Etter en krysning av rød og elfenbensfarget torskemunn fant en i F_2-generasjonen både A_1=røde, A_2=lyserøde og A_3=elfenbensfargete individer. Antall gjentakelser (individer) var n=97 og disse var gruppert slik:

grupper	frekvens
A_1	22
A_2	52
A_3	23
	n = 97

La oss tenke oss at vi har n gjentakelser og at disse kan grupperes i c grupper etter c alternative konstante (enkle eller sammensatte) kjennetegn $A_1, A_2, A_3, \ldots A_c$ som utelukker hverandre. La frekvensene være $h(A_1,U)=h_1$, $h(A_2,U)=h_2$, $h(A_3,U)=h_3$, ... og $h(A_c,U)=h_c$. Da er naturligvis

$$h_1+h_2+h_3+ \ldots + h_c = \Sigma h_i = n$$

La videre $s(A_1,U)=p_1$, $s(A_2,U)=p_2$, ... , $s(A_c,U)=p_c$. Da A'ene utelukker hverandre og det ikke fins andre alternativer, er

$$p_1+p_2+ \ldots + p_c = \Sigma p_i = 1$$

Forutsetter en nå at de n gjentakelsene er uavhengige gjentakelser, kan en bevise (beviset tas ikke med her) at forventningen for frekvensen for A_i (i=1,2,3, ... c) er np_i og at

$$\chi^2 = \sum_{i=1}^{c} \frac{(h_i - np_i)^2}{np_i}$$

har f(u) til fordelingslov når en setter f=c-1. Dette er den definisjon av χ^2 som er alminneligst brukt i lærebøker i statistikk.

La oss anvende dette på eks. 1. Vi vil anta hypotetisk at $s(A_1,U) = p_1=0,25$, $s(A_2,U)=p_2=0,50$ og $s(A_3,U)=p_3=0,25$ og dessuten at gjentakelsene er uavhengige. Vi beregner da χ^2 på følgende måte.

Gruppe	h_i	p_i	np_i	h_i-np_i	$(h_i-np_i)^2$	$\frac{(h_i-np_i)^2}{np_i}$
A_1	22	0,25	24,25	- 2,25	5,0625	0,21
A_2	52	0,50	48,50	+ 3,50	12,2500	0,25
A_3	23	0,25	24,25	- 1,25	1,5626	0,06
	97	1,00				$\chi^2 = 0,52$

I dette tilfelle er c=3 og derfor f=c-1=2. Av Tab. III ser en at det funne χ^2 er noe mindre enn den a som svarer til P=0,50 og noe større enn den a som svarer til P=0,90. Det er derfor ingen grunn til å forkaste hypotesen.

Oppgave 14.

I eks. 1 (III,4) er gjentakelsene ordnet i c=4 grupper (AB, AiB, iAB og iAiB). Anta hypotetisk at s(A,U)=s(B,U)=0,75, at A og B opptrer uavhengig av hverandre og at gjentakelsene er uavhengige gjentakelser. Undersøk hvordan denne hypotesen stemmer med observasjonene.

Vi har hittil tenkt oss at grupperingen av gjentakelsene foretas etter alternative konstante kjennetegn. Oppdelingen kan imidlertid like ofte foretas etter verdiene av et variabelt kjennetegn. Gruppefrekvensene er da simpelthen frekvensene for det variable kjennetegns verdier. Har vi da en hypotese om fordelingsloven for dette kjennetegn som er så pass detaljert at vi kan beregne s(x,U) for hver verdi av x, kan vi naturligvis sammenlikne de observerte frekvenser for x og forventningene for disse frekvenser ved χ^2 beregnet etter samme skjema som når grupperingen foretas etter alternative konstante kjennetegn.

Oppgave 15.

Vi har tidligere (III, 8, s. 29) sammenliknet frekvensene i fordelingsrekken i eks. 2 (II,2) med deres forventninger beregnet ved hjelp av binomialloven med parametrene k=21 og p=0,61. Undersøk ved hjelp av χ^2 hvordan denne hypotesen stemmer med observasjonene (frekvensene for x).

Oppgave 16.

Hva blir svaret på oppg. 27 (III,11) når en bruker χ^2?

I begynnelsen av dette avsnitt lovte vi å prøve å forklare hvordan f=antall frihetsgrader for χ^2 fremkommer. Det er imidlertid ikke lett å forklare dette uten å gjennomgå hele det matematiske bevis som deduksjonen av fordelingsloven f(u) bygger på. Likevel vil kanskje følgende merknader være av interesse.

Det grunnlaget en bygger på er jo den setningen som sier at fordelingsloven for frekvensen for et kjennetegn A i et univers W hvor hver gjentakelse består av n uavhengige gjentakelser i U, er den binomiallov vi refererte i begynnelsen av dette avsnitt. Frekvensen for A er derfor en verdi av det variable kjennetegn z i en gjentakelse i W. En må derfor kreve at n er et fast konstant tall. Hvis n varierer fra den ene gjentakelse i W til den annen, er ikke lenger binomialloven fordelingsloven for frekvensen for A.

Bruker vi nå bare frekvensen for A til beregning av χ^2 og setter

$$\chi^2 = \frac{[h(A,U)-np]^2}{npq}$$

har dette χ^2 som nevnt f=1 frihetsgrad. Antall frihetsgrader er derfor i dette tilfelle lik antallet av frekvenser eller observasjoner som er benyttet til beregning av χ^2. Beregner vi imidlertid det samme χ^2 ved formelen

$$\chi^2 = \frac{[h(A,U)-np]^2}{np} + \frac{[h(iA,U)-nq]^2}{nq}$$

bruker vi to frekvenser, men fremdeles er f=1. Dette henger sammen med at de to frekvensene skal tilfredsstille en absolutt betingelse, nemlig den som uttrykkes i ligningen

$$h(A,U) + h(iA,U) = n$$

Krever vi nemlig ikke denne betingelsen oppfylt, betyr det et brudd med de forutsetninger vi gikk ut fra da vi utledet fordelingsloven for frekvensen for A.

Den regelen vi skal benytte til bestemmelse av f kan uttrykkes i ligningen

$$f = c - b$$

hvor c er antallet av frekvenser som er benyttet til beregningen av χ^2 og b er antallet av de betingelsesligninger som disse frekvensene skal tilfredsstille. Vi innser lett at denne regelen er riktig for vårt eksempel. Bruker vi nemlig bare frekvensen for A, er c=1 og b=0. Følgelig

er f=c-b=1-0=1. Og bruker vi både frekvensen for A og frekvensen for iA, er c=2 og b=1. Altså er f=c-b=2-1=1. Vi finner også lett at denne regelen holder stikk for de andre eksemplene vi har gjennomgått. Foretar vi gruppering av gjentakelsene i c grupper etter de alternative kjennetegn $A_1, A_2, \ldots, A_c$, bruker vi c frekvenser til beregningen av χ^2 og disse skal tilfredsstille b=1 ligning, nemlig

$$\Sigma\, h(A_i,U) = n$$

Følgelig er f=c-b=c-1.

Oppgave 17.

I en undersøkelse av blomsterfargen hos erter fikk en i F_2-generasjonen det resultat som er gjengitt i følgende tabell. Her betyr A_1=purpur, A_2=fiolett, A_3=rosa, A_4=lys purpur og A_5=hvit. Anta hypotetisk at $s(A_1,U)=27/64$, $s(A_2,U)=9/64$, $s(A_3,U)=9/64$, $s(A_4,U)=3/64$ og $s(A_5,U)=16/64$ og at gjentakelsene er uavhengige. Undersøk hvordan denne hypotesen stemmer med observasjonene.

Forsøk nr.	n_i	Frekvens for				
		A_1	A_2	A_3	A_4	A_5
1	135	45	23	24	6	37
2	356	141	53	46	23	93
3	562	233	74	83	22	150
Sum	1053	419	150	153	51	280

Vi har forutsatt foran at den hypotesen som vi sammenlikner med observasjonene er gitt a priori i alle detaljer, dvs. at den er oppstilt helt uavhengig av de observasjoner den sammenliknes med. Men la oss nå ta for oss på nytt eks. 2 (III,4). I dette tilfelle går oppgaven ut på å undersøke om sannsynligheten for angrep av tyfoidfeber er forskjellig for vaksinerte og ikke-vaksinerte personer. Den hypotesen vi må sammenlikne med observasjonene går da ut på at angrepssannsynligheten (p) er den samme i de to universene uten at vi fastsetter noen hypotetisk verdi for den. For å kunne beregne χ^2 må vi imidlertid bruke et estimat av p og det hypotetisk mest rimelige estimat er da estimatet av s(B,U) i universet av alle personer. Dette er

$$p' = \frac{h(B,U)}{n} = \frac{328}{18483} = 0{,}017746$$

Deretter beregner vi χ^2 på samme måte som vist foran, dvs. at vi utfører beregningene som om p' var inkludert i hypotesen.

n_i	h_i	n_ip'	h_i-n_ip'	$(h_i-n_ip')^2$	$n_ip'q'$	$\frac{(h_i-n_ip')^2}{n_ip'q'}$
6815	56	120,94	- 64,94	4217,2036	118,7994	35,50
11668	272	207,06	+ 64,94	4217,2036	203,2950	20,73
18483	328	328,00				$\chi^2 = 56,23$

Når vi så skal bruke Tab. III blir det spørsmål om hvilken verdi f har. Dette er i virkeligheten et meget vanskelig spørsmål som vi ikke kan ta opp til drøftelse her. Vi må nøye oss med å referere at en har funnet at f skal bestemmes ved ligningen

$$f = c - b - d$$

hvor c er antall frekvenser som er brukt til beregningen av χ^2 , b er antall betingelsesligninger som disse frekvensene skal tilfredsstille og d er antall parametrer som er estimert ved de observasjoner (frekvenser) som en sammenligner hypotesen med. Dette er ikke eksakt riktig, men det har vist seg at en får en meget bra tilnærmelse til fordelingsloven f(u) for χ^2 når en bestemmer f på denne måten.

I det foreliggende tilfelle er c=2,b=0 og d=1. Følgelig er f=1. Vi ser da at det funne χ^2 er betydelig større enn den a-verdi som svarer til f=1 og P=0,001. Vi må derfor forkaste hypotesen. Konklusjonen må altså bli den at sannsynligheten for angrep i universet av vaksinerte personer er forskjellig fra sannsynligheten for angrep i universet av ikke-vaksinerte personer. Kjennetegnene A og B er m.a.o. ikke uavhengige.

Vi skal ta for oss et annet eksempel. Vi har i III,11 foretatt en sammenligning mellom frekvensene i fordelingsrekken i eks. 3 (III,9) med deres forventninger beregnet under den hypotetiske forutsetning at fordelingsloven for kjennetegnet (hodeskallens største bredde) er normal. I denne hypotesen var imidlertid ikke inkludert noen verdier av de to parametrene μ og σ i den normale fordelingslov. Disse ble estimert ved at vi satte μ lik gjennomsnittet (m) og σ lik middelavviket (s) for observasjonene av kjennetegnet. Dette må vi ta hensyn til når vi bruker Tab. III, men beregningen av χ^2 skal utføres slik som vi har beskrevet det foran. Til avkortning skriver vi $h(x_i,U)=h_i$.

x_i	h_i	np'_i	$(h_i-np'_i)^2$	$\frac{(h_i-np'_i)}{np'_i}$
122-142	34	36	4	0,11
147	173	195	441	2,26
152	567	535	1024	1,93
157	701	688	169	0,25
162	390	414	576	1,40
167	119	117	4	0,03
172-187	16	16	0	0,00
	2000	2001		$\chi^2 = 5,98$

Vi har i dette tilfelle brukt c=7 frekvenser til beregning av χ^2. Disse skal tilfredsstille b=1 betingelsesligning (frekvenssummen lik n=2000) og d=2 parametrer er estimert ved observasjonene. Følgelig er f=c-b-d =7-1-2=4. Av Tab. III ser en at det funne $\chi^2 = 5,98$ er praktisk talt lik den a som svarer til f=4 og P=0,2. Vi har derfor fått en bekreftelse på riktigheten av det resultat vi er kommet til tidligere, at den normale fordelingsloven gir en bra beskrivelse av fordelingsrekken.

Oppgave 18.

I følgende tabell er n_i antall levendefødte barn i Norge i årene fra 1933 til 1937 og h_i antall levendefødte gutter. Anta hypotetisk at alle disse barn er født ved enkeltfødsler, at alle fødsler er uavhengige gjentakelser og at sannsynligheten for at det ved en enkeltfødsel blir født en gutt er uforandret fra år til år. Undersøk hvordan denne hypotesen stemmer med observasjonene.

År	n_i	h_i
1933	42114	21834
1934	41833	21589
1935	41321	21145
1936	42240	21708
1937	43808	22563
Sum	211316	108839

Oppgave 19.

Løs oppg. 28 (III,11) når en estimerer sannsynligheten p for at det ved en enkeltfødsel blir født en gutt ved $p'=m/k$ hvor m er det gjennomsnittlige antall gutter pr. familie og k=5.

Oppgave 20.

La C og B bety det samme som i oppg. 14 (III,4). Anta hypotetisk at disse to kjennetegn opptrer uavhengig av hverandre. Undersøk hvordan denne hypotesen stemmer med de observasjoner som er gitt i oppg. 14 (III,4).

Til slutt en liten tilleggsmerknad. Vi har sett foran at forventningen for χ^2 er $\mu(\chi^2) = f$. Den verdi av χ^2 vi skal vente ved en

helt fullkommen overensstemmelse mellom hypotese og observasjoner er derfor ikke null men f. Av dette følger at meget små χ^2-verdier kan være like urimelige som meget store. En liten χ^2-verdi kan nemlig være en like usannsynlig avvikelse fra forventningen som en stor χ^2-verdi. Setter vi f. eks. f=10, er forventningen for χ^2 lik 10, og vi ser av Tab. III at sannsynligheten for et χ^2 lik eller mindre enn a=2,558 er lik Q=1-P=1-0,99=0,01. Et χ^2 lik eller mindre enn 2,558 er altså for f=10 like usannsynlig som et χ^2 lik eller større enn 23,209. Hvis en derfor i et aktuelt tilfelle finner et χ^2 som er mindre enn den a som etter Tab. III svarer til P=0,99 og den f en skal bruke i vedkommende tilfelle, må en slutte at den hypotesen som har gitt dette resultat ikke er bekreftet, vi må forkaste den. Slike tilfelle forekommer meget sjelden. Kommer en over et slikt tilfelle, bør en først og fremst undersøke om ikke resultatet skyldes regnefeil eller om det er brukt et feilaktig formelsystem. Utslag i denne retning vil imidlertid også forekomme når en sammenlikner en hypotese med observasjoner som er utsorterte med sikte på å demonstrere en meget god overensstemmelse.

Oppgave 21.

En krysset maisplanter med røde+melne korn og planter med hvite +glasne korn. På avkom-plantene fant en da korn som hadde kjennetegnene A_1=røde+melne, A_2=røde+glasne, A_3=hvite+melne og A_4=hvite+glasne. I følgende tabell er oppgitt antall korn pr. plante (n_i) og frekvensene for disse fire kjennetegn. Anta hypotetisk at $s(A_1,U)=9/16$, $s(A_2,U)=s(A_3,U)=3/16$ og $s(A_4,U)=1/16$ og at de enkelte korn er uavhengige gjentakelser. Undersøk hvordan denne hypotesen stemmer med observasjonene.

Plante nr.	n_i	Frekvens for			
		A_1	A_2	A_3	A_4
1	176	102	32	31	11
2	144	84	27	25	8
3	176	99	33	31	13
4	128	80	21	21	6
5	144	82	27	24	11
Sum	768	447	140	132	49

6. Korrelasjonskoeffisienten.

I korrelasjonslæren (II) har vi lært å beregne regresjonskoeffisienter, korrelasjonsforholdet, korrelasjonsindeksen og korrelasjonskoeffisienten. Disse størrelser er naturligvis estimater av tilsvarende kærakteristikker av fordelingsloven for de observerte kjennetegn i det

univers-som gjentakelsene er tilfeldig utvalg av. Vi må her nøye oss med å si noen ord om korrelasjonskoeffisienten.

La oss tenke oss at vi har n parobservasjoner av to variable kjennetegn x og y og at vi har beregnet korrelasjonskoeffisienten (r). Selv om x og y er uavhengige kjennetegn, vil vi i regelen få en verdi for r som avviker noe fra null. Spørsmålet er da om verdien av r avviker så meget fra null at vi har rett til å slutte at også universets korrelasjonskoeffisient (ϱ) som r er et estimat av, er forskjellig fra null. Dette er et meget viktig spørsmål. Hvis nemlig ϱ er forskjellig fra null, betyr det at x og y er avhengige kjennetegn.

Forutsetter en at x og y er uavhengige kjennetegn (og dermed at $\varrho = 0$) og at begge kjennetegn har normal fordelingslov, kan en bevise at

$$t = \frac{|r|}{\sqrt{1-r^2}} \sqrt{n-2}$$

følger Student's fordelingslov når en setter f=n-2.

For eks. 12 (II,9) har vi funnet r=+0,77 (se oppg. 18,II,18). Altså er

$$t = \frac{0,77}{\sqrt{1-0,77^2}} \sqrt{25-2} = 5,78$$

Av Tab. I ser vi at til f=n-2=23 og P=0,001 svarer det a=3,767. Vi har funnet en t som er betydelig større enn denne a-verdi. Følgelig må vi forkaste hypotesen $\varrho = 0$ og slutte at de to kjennetegn (vekt og % kvelstoff i byggkorn) er avhengige kjennetegn.

For at en skal slippe å beregne t har en omregnet Tab. I til en tabell som viser sammenhørende verdier av f,a og P hvor P er sannsynligheten (den hypotetiske) for $r \geq a$. Denne er gjengitt i Tab. IV bakerst i boken. Vi ser av denne tabell at hvis vi har n=4 parobservasjoner, dvs. f=n-2=2, må vi ha $r \geq 0,99$ for å kunne slutte at de to kjennetegn er avhengige, i hvert fall hvis vi krever en sikkerhet som uttrykkes ved P=0,01. Har vi n=10, dvs. f=8, må vi ha $r \geq 0,7646$. For vårt eksempel har vi f=23. Denne f-verdi fins ikke i Tab. IV, men vi ser at vi kan bruke f=20 i stedet. Til f=20 og P=0,01 svarer det a=0,5368. Vi har funnet r=0,77 som er betydelig større enn denne a-verdi. Og derav slutter vi at de to kjennetegnene er avhengige kjennetegn.

Oppgave 22.

Undersøk ved hjelp av Tab. IV om de korrelasjonskoeffisienter som vi har beregnet i oppg. 18 (II,18) avviker så meget fra null at vi kan slutte at kjennetegnene er avhengige kjennetegn.

Etterskrift.

Det er naturligvis bare en meget begrenset mengde stoff en har anledning til å gjennomgå i en forelesningsrekke på ca. 70 timer. Kurset har derfor heller ikke blitt noe mer enn en første elementær innføring i statistisk forskningsmetode. Det er søkt lagt vekt på å forklare de grunnleggende prinsippene. Tildels har dette imidlertid støtt på uovervinnelige vansker fordi studentene ikke behersker høyere matematisk analyse.

Viktige deler av statistikkens teori er ikke nevnt i det hele tatt eller bare så vidt omtalt. Tidsrekkeanalysen er således helt forbigått og utjevningslæren ved minste kvadraters metode bare så vidt nevnt.

De fleste oppgaver og eksempler er originale. Materialet er i det vesentlige hentet fra resultatene av forsøksvirksomheten ved Norges Landbrukshøgskole og Det Norske Skogforsøksvesen. Dessuten er benyttet noe materiale fra Statistisk Årbok for Norge.

Til fortsatt studium kan anbefales følgende bøker. De første fem forutsetter ikke større matematiske kunnskaper. Nr. 6 er atskillig vanskeligere, men viktige avsnitt kan studeres uten at en behersker høyere matematisk analyse.

1) G.U. Yule and M.G. Kendall: An Introduction to the Theory of Statistics. London 1946.

2) F.C. Mills: Statistical Methods. New York 1938.

3) R.A. Fisher: Statistical Methods for Research Workers. London 1936.

4) R.A. Fisher: The Design of Experiments. London 1937.

5) K. Mather: Statistical Analysis in Biology. London 1946.

6) M.G. Kendall: The Advanced Theory of Statistics. London 1945.

Tab. I. Fordelingsloven for t.

f	P									
	0,9	0,7	0,5	0,4	0,3	0,2	0,05	0,02	0,01	0,001
1	0,158	0,510	1,000	1,376	1,963	3,078	12,71	31,82	63,66	636,6
2	0,142	0,445	0,816	1,061	1,386	1,886	4,303	6,965	9,925	31,60
3	0,137	0,424	0,765	0,978	1,250	1,638	3,182	4,541	5,841	12,94
4	0,134	0,414	0,741	0,941	1,190	1,533	2,776	3,747	4,604	8,610
5	0,132	0,408	0,727	0,920	1,156	1,476	2,571	3,365	4,032	6,859
6	0,131	0,404	0,718	0,906	1,134	1,440	2,447	3,143	3,707	5,959
7	0,130	0,402	0,711	0,896	1,119	1,415	2,365	2,998	3,499	5,405
8	0,130	0,399	0,706	0,889	1,108	1,397	2,306	2,896	3,355	5,041
9	0,129	0,398	0,703	0,883	1,100	1,383	2,262	2,821	3,250	4,781
10	0,129	0,397	0,700	0,879	1,093	1,372	2,228	2,764	3,169	4,587
11	0,129	0,396	0,697	0,876	1,088	1,363	2,210	2,718	3,106	4,437
12	0,128	0,395	0,695	0,873	1,083	1,356	2,179	2,681	3,055	4,318
13	0,128	0,394	0,694	0,870	1,079	1,350	2,160	2,650	3,012	4,221
14	0,128	0,393	0,692	0,868	1,076	1,345	2,145	2,624	2,977	4,140
15	0,128	0,393	0,691	0,866	1,074	1,341	2,131	2,602	2,947	4,073
16	0,128	0,392	0,690	0,865	1,071	1,337	2,120	2,583	2,921	4,015
17	0,128	0,392	0,689	0,863	1,069	1,333	2,110	2,567	2,898	3,965
18	0,127	0,392	0,688	0,862	1,067	1,330	2,101	2,552	2,878	3,922
19	0,127	0,391	0,688	0,861	1,066	1,328	2,093	2,539	2,861	3,883
20	0,127	0,391	0,687	0,860	1,064	1,325	2,086	2,528	2,845	3,850
21	0,127	0,391	0,686	0,859	1,063	1,323	2,080	2,518	2,831	3,819
22	0,127	0,390	0,686	0,858	1,061	1,321	2,074	2,508	2,819	3,792
23	0,127	0,390	0,685	0,858	1,060	1,319	2,069	2,500	2,807	3,767
24	0,127	0,390	0,685	0,857	1,059	1,318	2,064	2,492	2,797	3,745
25	0,127	0,390	0,684	0,856	1,058	1,316	2,060	2,485	2,787	3,725
26	0,127	0,390	0,684	0,856	1,058	1,315	2,056	2,479	2,779	3,707
27	0,127	0,389	0,684	0,855	1,057	1,314	2,052	2,473	2,771	3,690
28	0,127	0,389	0,683	0,855	1,056	1,313	2,048	2,467	2,763	3,674
29	0,127	0,389	0,683	0,854	1,055	1,311	2,045	2,462	2,756	3,659
30	0,127	0,389	0,683	0,854	1,055	1,310	2,042	2,457	2,750	3,646
40	0,126	0,388	0,681	0,851	1,050	1,303	2,021	2,423	2,704	3,551
60	0,126	0,387	0,679	0,848	1,046	1,296	2,000	2,390	2,660	3,460
120	0,126	0,386	0,677	0,845	1,041	1,289	1,980	2,358	2,617	3,373
∞	0,126	0,385	0,674	0,842	1,036	1,282	1,960	2,326	2,576	3,291

Tab. II. Fordelingsloven for F. P=0,01.

f for minste varians	f for største varians: 1	2	3	4	5	6	8	12	24	∞
2	98,49	99,01	99,17	99,25	99,30	99,33	99,36	99,42	99,46	99,50
3	34,12	30,81	29,46	28,71	28,24	27,91	27,49	27,05	26,60	26,12
4	21,20	18,00	16,69	15,98	15,52	15,21	14,80	14,37	13,93	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,27	9,89	9,47	9,02
6	13,74	10,92	9,78	9,15	8,75	8,47	8,10	7,72	7,31	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,84	6,47	6,07	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,03	5,67	5,28	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,47	5,11	4,73	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,06	4,71	4,33	3,91
11	9,65	7,20	6,22	5,67	5,32	5,07	4,74	4,40	4,02	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,50	4,16	3,78	3,36
13	9,07	6,70	5,74	5,20	4,86	4,62	4,30	3,96	3,59	3,16
14	8,86	6,51	5,56	5,03	4,69	4,46	4,14	3,80	3,43	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,00	3,67	3,29	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	3,89	3,55	3,18	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,79	3,45	3,08	2,65
18	8,28	6,01	5,09	4,58	4,25	4,01	3,71	3,37	3,00	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,63	3,30	2,92	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,56	3,23	2,86	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,51	3,17	2,80	2,36
22	7,94	5,72	4,82	4,31	3,99	3,76	3,45	3,12	2,75	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,41	3,07	2,70	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,36	3,03	2,66	2,21
25	7,77	5,57	4,68	4,18	3,86	3,63	3,32	2,99	2,62	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,29	2,96	2,58	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,26	2,93	2,55	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,23	2,90	2,52	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,20	2,87	2,49	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,17	2,84	2,47	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	2,99	2,66	2,29	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,82	2,50	2,12	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,66	2,34	1,95	1,38
∞	6,64	4,60	3,78	3,32	3,02	2,80	2,51	2,18	1,79	1,00

Tab. III. Fordelingsloven for χ^2.

f	P							
	0,99	0,90	0,50	0,20	0,05	0,02	0,01	0,001
1	0,000	0,016	0,455	1,642	3,841	5,412	6,635	10,827
2	0,020	0,211	1,386	3,219	5,991	7,824	9,210	13,815
3	0,115	0,584	2,366	4,642	7,815	9,837	11,341	16,268
4	0,297	1,064	3,357	5,989	9,488	11,668	13,277	18,465
5	0,554	1,610	4,351	7,289	11,070	13,388	15,086	20,517
6	0,872	2,204	5,348	8,558	12,592	15,033	16,812	22,457
7	1,239	2,833	6,346	9,803	14,067	16,622	18,475	24,322
8	1,646	3,490	7,344	11,030	15,507	18,168	20,090	26,125
9	2,088	4,168	8,343	12,242	16,919	19,679	21,666	27,877
10	2,558	4,865	9,342	13,442	18,307	21,161	23,209	29,588
11	3,053	5,578	10,341	14,631	19,675	22,618	24,725	31,264
12	3,571	6,304	11,340	15,812	21,026	24,054	26,217	32,909
13	4,107	7,042	12,340	16,985	22,362	25,472	27,688	34,528
14	4,660	7,790	13,339	18,151	23,685	26,873	29,141	36,123
15	5,229	8,547	14,339	19,311	24,996	28,259	30,578	37,697
16	5,812	9,312	15,338	20,465	26,296	29,633	32,000	39,252
17	6,408	10,085	16,338	21,615	27,587	30,995	33,409	40,790
18	7,015	10,865	17,338	22,760	28,869	32,346	34,805	42,312
19	7,633	11,651	18,338	23,900	30,144	33,687	36,191	43,820
20	8,260	12,443	19,337	25,038	31,410	35,020	37,566	45,315
21	8,897	13,240	20,337	26,171	32,671	36,343	38,932	46,797
22	9,542	14,041	21,337	27,301	33,924	37,659	40,289	48,268
23	10,196	14,848	22,337	28,429	35,172	38,968	41,638	49,728
24	10,856	15,659	23,337	29,553	36,415	40,270	42,980	51,179
25	11,524	16,473	24,337	30,675	37,652	41,566	44,314	52,620
26	12,198	17,292	25,336	31,795	38,885	42,856	45,642	54,052
27	12,879	18,114	26,336	32,912	40,113	44,140	46,963	55,476
28	13,565	18,939	27,336	34,027	41,337	45,419	48,278	56,893
29	14,256	19,768	28,336	35,139	42,557	46,693	49,588	58,302
30	14,953	20,599	29,336	36,250	43,773	47,962	50,892	59,703

Tab. IV.

f	P			
	0,1	0,05	0,02	0,01
1	0,98769	0,996917	0,9995066	0,9998766
2	0,90000	0,95000	0,98000	0,990000
3	0,8054	0,8783	0,93433	0,95873
4	0,7293	0,8114	0,8822	0,91720
5	0,6694	0,7545	0,8329	0,8745
6	0,6215	0,7067	0,7887	0,8343
7	0,5822	0,6664	0,7498	0,7977
8	0,5494	0,6319	0,7155	0,7646
9	0,5214	0,6021	0,6851	0,7348
10	0,4973	0,5760	0,6581	0,7079
11	0,4762	0,5529	0,6339	0,6835
12	0,4575	0,5324	0,6120	0,6614
13	0,4409	0,5139	0,5923	0,6411
14	0,4259	0,4973	0,5742	0,6226
15	0,4124	0,4821	0,5577	0,6055
16	0,4000	0,4683	0,5425	0,5897
17	0,3887	0,4555	0,5285	0,5751
18	0,3783	0,4438	0,5155	0,5614
19	0,3687	0,4329	0,5034	0,5487
20	0,3598	0,4227	0,4921	0,5368
25	0,3233	0,3809	0,4451	0,4869
30	0,2960	0,3494	0,4093	0,4487
35	0,2746	0,3246	0,3810	0,4182
40	0,2573	0,3044	0,3578	0,3932
45	0,2428	0,2875	0,3384	0,3721
50	0,2306	0,2732	0,3218	0,3541
60	0,2108	0,2500	0,2948	0,3248
70	0,1954	0,2319	0,2737	0,3017
80	0,1829	0,2172	0,2565	0,2830
90	0,1726	0,2050	0,2422	0,2673
100	0,1638	0,1946	0,2301	0,2540