Per Ottestad

STATISTIKK

(Del I)

Utgave 1970

Per Ottestad

STATISTIKK

(Del I)

Utgave 1970

A.NOEN ALMINNELIGE EMNER.

A. NOEN ALMINNELIGE EMNER.

A.1. Innledning.

Dette kurset i statistikk er ment å gi en viss innføring i noen av de viktigste sider av vitenskapelig metodelære. Noen vil kanskje si at emnet angår dem som tar sikte på å bli forskere og at det er overflødig hvis en tenker på mer praktisk betont virksomhet. Litt ettertanke vil imidlertid sikkert overbevise om at slik er det ikke.

Til de fleste som har fått utdanning ved universitet eller høgskole, vil det bli stilt krav om å ta standpunkt i faglige spørsmål. Noen vil da kanskje foretrekke å vende seg til personer som de tror er autoriteter når det gjelder de spørsmål det skal tas standpunkt til. Andre vil heller prøve å gjøre seg opp en mening selv, og har da bruk for innsikt i metodelære.

Empiriske fag er slike hvor de regler og lovmessigheter en regner som kunnskap, er basert på erfaringer. Med erfaringer mener vi da enkeltopplysninger eller data. Ordet erfaring har imidlertid dobbelt betydning. Dels brukes det i betydningen data, dels om regler og lovmessigheter. Uttrykket "jeg har gjort den erfaring at ...." betyr vanligvis at vedkommende jeg har gjort seg opp en mening som han gir uttrykk for i en setning. Vi skal derfor unngå å bruke ordet erfaring. Vi skal bruke betegnelsen observasjon eller datum på faktiske enkeltopplysninger og regel, lovmessighet eller utsagn om det en kommer fram til ved metodisk behandling av observasjonene.

Den tankevirksomhet som fører fra observasjoner til regler, lovmessigheter eller utsagn, kalles induksjon. Det er den tankevirksomhet som er karakteristisk for empirisk forskning. Den alternative form for tenkning tar utgangspunkt i alminnelige set-

ninger og utleder andre setninger fra disse. Dette kalles deduksjon og er karakteristisk for logikken og matematikken.

Historisk sett er naturforskning slik vi kjenner den i dag - og empirisk forskning i det hele tatt - av ny dato. Grunnlaget ble lagt i det 17. århundre av slike betydelige menn som Galileo (1564-1642) og Kepler (1571-1630). Forskere som har behandlet forskningshistoriske emner vil riktignok også ta med som grunnleggere filosofer fra den greske storhetstid. Aristoteles er nevnt som en av disse. Til dette bemerker* den engelske filosof Bertrand Russell at "Aristotle maintained that women have fewer teath than men; although he was twice married, it never occurred to him to verify this statement by examining his wives' mouths."

En kan kanskje si at det å indusere regler eller lovmessigheter på basis av observasjoner er forskerens sak. Men den som skal gjøre bruk av resultatene, bør ikke være ukjent med måten slike induksjoner kommer i stand på. De som skal bruke resultatene, bør kunne øve en viss kritikk.

Det er tre spørsmål en har rett til å få svar på. Det er
1) hvor stammer de observasjonene fra som induksjonen bygger på?
2) hvilken metode er brukt for å skaffe dem til veie?
3) hvilken metode er brukt til bearbeidelsen eller analysen av dem?
Disse spørsmål skal forskeren kunne svare på, også fordi det er nødvendig for ham selv og hans arbeid. Forskeren må alltid sørge for at de resultatene han eventuelt kommer til, kvalifiserer dem som rettesnor for praktisk virksomhet eller grunnlag for videre forskning.

Ikke all forskning tar sikte på resultater som kan bli til nytte for praktisk virksomhet. Det har vært gjort undersøkelser,

*The Impact of Science on Society. London 1952.

og det vil det også bli gjort i fremtiden, som ikke har praktisk siktepunkt. Det er like viktig at slike undersøkelser blir utført. Med et visst forbehold kan vi vel si at alt forskningsarbeid som er planlagt slik at det gi håp om at våre kunnskaper blir utvidet, er nyttige. Og prinsippene for forskning er de samme om målet er resultater som har praktisk nytteverdi eller om det bare er å utvide våre kunnskaper.

Vi har ikke anledning her til å komme nærmere inn på hva det i dypere mening betyr å ha kunnskap. Vi må nøye oss med å si at vi har skaffet oss kunnskap om eller innsikt i en sak når vi har fått svar på spørsmål vi har stilt. Ønsker vi derfor å skaffe oss kunnskap om en sak, må vi stille fornuftige spørsmål og innrette oss slik at vi har håp om å få svar på dem. Det er da viktig at spørsmålene er entydige og at det ikke er for mange av dem. Spør vi dumt eller om for meget om gangen, kan vi ikke regne med å få fornuftige svar.

Utgangspunktet er og må være det vi vet eller tror vi vet om saken. Oftest er det viten om at våre kunnskaper er mangelfulle som reiser nye spørsmål. Er så spørsmålet stilt, blir det vår oppgave å finne ut hvordan vi skal innrette oss slik at vi har håp om å få svar. Dette er vanskelig, men en vet nå nok å meget om hvordan en skal gå fram. I neste avsnitt skal vi ta for oss et enkelt forsøk eller eksperiment og ved hjelp av det som eksempel prøve å forklare noen viktige prinsipper.

Alle som har vært nødt til å ta standpunkt til et vanskelig spørsmål, vet ofte med seg selv at standpunktet kanskje ikke er det riktige. Også i forskningsarbeidet er det å ta standpunkt meget vanskelig. En vil vel alltid føle seg litt usikker på om det resultatet en er kommet til, er det riktige. Noe bevis i

egentlig forstand for riktigheten av regler, lovmessigheter eller utsagn som er kommet i stand empirisk, kan aldri gis. Vi kan aldri bevise at en sort bygg gir større avling enn en annen sort. Vi kan bare si at det er god nok grunn til å mene at den ene sorten gir større avling enn den andre. Det er bare innen matematikk og logikk en kan gi beviser.

Gjelder det empiriske regler eller utsagn, må vi ta standpunkt for eller imot. Det vi da må ta sikte på er å unngå i størst mulig grad å ta standpunkt mot riktige eller treffende utsagn og standpunkt for feilaktige.

A.2. Et eksempel på forsøk.

La oss tenke oss at vi vil foreta en sammenligning mellom to sorter poteter, T_1 og T_2. Det er i regelen flere egenskaper ved sortene en ønsker å sammenligne, slike som produktivitet, sykdomsresistens og smak. La oss tenke oss at vi ønsker å få brakt på det rene om vi kan si at T_1 under visse vekstvilkår kan ventes å gi større (eller mindre) avling enn T_2.

Vi vet at avlingsmengden er avhengig av slike faktorer som temperatur, nedbørsmengde og de mange forskjellige egenskaper som dyrkingsjorda har. Det nytter derfor ikke å dyrke T_1 på en åker og T_2 på en annen åker. Gjør vi nemlig det, vil vi ikke kunne finne ut om den forskjellen forsøket viser det er, beror på ulikheter mellom sortene. Forskjellen kan helt eller delvis skyldes ulikheter i dyrkingsjorda på de to åkrene, kanskje også ulikheter i temperaturen og nedbørsmengden, og vi vil ikke ha noe middel til å avgjøre hvor meget. Vi må derfor dyrke de to sortene side om side på samme åkeren, og det blir da spørsmål om hvordan

vi skal gjøre det. Hensikten er jo å få sammenlignbare observasjoner for de to sortene.

Til en viss grad kan vi også bestemme oss for hvilke vekstvilkår forsøket ønskes utført under. Vi kan f.eks. bestemme oss for jordtype og velge forsøksfelt etter det. Temperatur og nedbørsmengde kan vi naturligvis ikke velge, men vi kan sørge for at vi skaffer oss observasjoner for begge faktorer og ved hjelp av disse gi en beskrivelse av vekstvilkårene.

Et forsøk som dette, utført på et valt forsøksfelt og i ett år, er hva vi kan kalle et lokalt forsøk. Hvis formålet med sammenligning av et antall sorter er å skaffe seg et grunnlag for veiledning om valg av sort, er utfallet av et slikt lokalt forsøk ikke tilstrekkelig. Grunnen til dette skal vil komme tilbake til. Et lokalt forsøk er imidlertid ofte et nødvendig ledd i et større forsøksprosjekt.

I de siste ca. 25 år er det utført et stort og omfattende arbeid med sikte på å finne gode planer for slike forsøk. Vi har derfor nå flere planer å velge mellom. Felles for disse er at dyrkingsfeltet eller forsøksfeltet deles opp i et antall like store småfelter eller ruter, og så fordeles de forsøksleddene en vil sammenligne, f.eks. de to potetsortene, på disse rutene slik at det blir like mange ruter til hvert forsøksledd. Er det k forsøksledd som skal sammenlignes og en vil ha n ruter til hvert av dem, må feltet deles i N = nk ruter. En av de planer som brukes, går under navn av blokkplanen eller "the randomized block design".

La oss tenke oss at det er k = 2 forsøksledd, f.eks. to potetsorter, og at forsøksfeltet er en noe langstrakt rektangulær åker. Ønsker en da å bruke blokkplanen, deles åkeren først i n felter eller blokker. Hver av disse deles så i to like store ruter. På

den måten blir hele feltet delt opp i 2n like store ruter, slik som vist i Fig. A.1. Blokkene er her betegnet med $B_1, B_2, \ldots . B_n$.

Figur A.1.

	B_1	B_2	B_3	. . . . .	B_n
I	T_2	T_1	T_1	. . . . .	T_2
II	T_1	T_2	T_2	. . . . .	T_1

Hver av de to sortene, T_1 og T_2, skal så dyrkes på n av de 2n rutene. Før forsøket settes i gang må vi ta ut eller velge de n rutene som skal brukes til T_1, og spørsmålet blir hvordan dette skal gjøres.

Felles for alle forsøksplaner er at forsøksleddene fordeles på rutene - eller i alminnelighet forsøksenhetene - ved hjelp av en eller annen teknikk for loddtrekning. Vi kaller dette å randomisere*. I vårt enkle tilfelle kan vi randomisere ved at vi først gir hver av rutene et nummer, nr. 1,2,3,... 2n, og at vi så ved loddtrekning tar ut n rutenummer for T_1. Dette går under navn av fri randomisering eller randomisering uten restriksjoner. En innser vel da lett at hvis vi gjør det på denne måten, blir de to sortene fordelt utover feltet på helt tilfeldig måte.

Vi har imidlertid tenkt oss at feltet er delt i n blokker og hver blokk i to ruter. En av rutene i hver blokk skal så brukes til T_1, den andre til T_2. Randomisering i samsvar med dette går ut på at vi trekker lodd for hver blokk for å finne ut hvilken av de to rutene skal brukes til T_1. I dette enkle tilfelle kan loddtrekningen utføres ved å kaste mynt og krone. Vi kan f.eks. bruke

*En fornorskning av det engelske verbet "to randomize". Substantivet er "randomization" som fornorskes til "randomisering". I det følgende kommer vi til å bruke flere slike fornorskede engelske ord.

den regelen at hvis kastet gir "mynt", skal T_1 plasseres på den ruten som ligger i rutebeltet I i fig. A.1. Figuren viser et eksempel på hvordan T_1 og T_2 kan bli plassert ved hjelp av slik teknikk.

Vi skal senere forklare noe mer inngående hvorfor vi må bruke randomisering. Her må vi nøye oss med følgende begrunnelse.

Det er en vanlig menneskelig feil at den som stiller et spørsmål, gjør det på en slik måte at svaret blir lagt i munnen på den som blir spurt. Et forsøk er, kan vi si, et spørsmål til naturen. Vi må derfor planlegge og utføre forsøket på en slik måte at vi ikke legger naturen svaret i munnen. Randomisering er nødvendig for å sikre at vi får et svar som er fordomsfritt og selvstendig.

La oss tenke oss at vi har utført et forsøk etter blokkplanen med n = 10 blokker for sammenligning av to forsøksledd T_1 og T_2. Hvis da alt har gått bra, vil vi etter at feltet er høstet ha 10 observasjoner (av f.eks. mengde avling) for T_1 og 10 for T_2. To og to av disse observasjonene er så nær sammenlignbare som det er mulig å få det til på en åker fordi de stammer fra ruter som ligger ved siden av hverandre i samme blokk. Vi kan derfor danne 10 differenser og behandle disse ved hjelp av metoder vi skal forklare senere.

Disse prinsippene er felles for alle forsøk eller eksperimenter. Vi skal merke oss at det materiale forsøket utføres på, forsøksmaterialet, er mer eller mindre uensartet. Vi sier at det er heterogent. Utføres forsøket etter blokkplanen på et felt må en regne med at det er ulikheter i vekstvilkårene mellom ruter som hører med til samme blokk, og at det er ulikheter i vekstvilkårene blokkene imellom. Det er m.a.o. heterogenitet både innen blokkene og mellom blokkene. Slik er det også i andre tilfelle, og prinsippene for planlegging og utførelse er også de samme.

I eldre og ofte også i nyere litteratur vil en finne fremstillinger som bygger på den forutsetning at en kan utføre forsøk under homogene vilkår. Sannheten er imidlertid at homogene forsøksmaterialer ikke eksisterer. Noen vil kanskje mene at laboratorieforsøk er unntak fra denne regelen. Men ved nærmere ettertanke vil en innse at selv om en gjør alt en kan for at heterogeniteten skal bli liten, vil det aldri lykkes å fjerne den helt. Dette henger bl.a. sammen med at vi ikke er i stand til å utføre en handling eksakt likt to ganger.

Forsøk er ikke et brukbart hjelpemiddel innen alle sektorer av empirisk forskning. Undersøkelser av spørsmål som melder seg under studiet av dyre- og plantesamfunn, geologiske undersøkelser, undersøkelser av økonomiske spørsmål osv. ligger i regelen ikke godt til rette for eksperimentell forskning. En må skaffe seg observasjoner på annen måte. I noen tilfelle må en kanskje nøye seg med mer eller mindre tilfeldige funn, men i regelen kan en planlegge observasjonsarbeidet.

Det er også mange tilfelle hvor forskeren må bruke observasjoner som er skaffet til veie av andre. Det kan f.eks. være nødvendig å bruke data fra den offisielle statistikk. Det er da meget om å gjøre at den som bruker slike data, er helt fortrolig med den plan som er benyttet.

A.3. Gjentak og univers.

Vi har kunnskap eller viten av to slag. Vi har kunnskap om enkelte fakta og vi har kunnskap uttrykt i setninger vi har laget oss på grunnlag av fakta.

I tabell A.1. er gjengitt resultatene av et blokkforsøk for sammenligning av to sorter bygg, T_1 og T_2. Tallene er observa-

sjoner av vekten av kornavlingen pr. rute. Av de seks tallene kan vi danne tre differenser (x), og disse representerer da for dette eksemplet de fakta vi har kunnskap om. Vi kan si at vi har kunnskap om fakta i tre enkelttilfelle, nemlig de tre blokkene.

Tabell A.1.

Blokk	T_1	T_2	x (differens)
B_1	74	51	23
B_2	76	47	29
B_3	67	45	22

Bruker vi statistiske metoder på disse tre differensene, vil vi finne at vi er på trygg grunn når vi formulerer en konklusjon som går ut på at sort T_1 gir større kornavling enn sort T_2 under de vekstvilkår forsøket ble utført. Denne konklusjonen representerer kunnskap av et annet slag enn kunnskapen om de tre fakta vi tar utgangspunkt i. Konklusjonen omfatter noe mer enn de tre enkelttilfelle, den har et generelt innhold.

Mange setninger hvis innhold vi regner som kunnskap, har det til felles at innholdet gjelder for alle enkelttilfelle. Eksempler er "alle mennesker er dødelige" og "stål synker i vann". Dette gjelder imidlertid ikke alle setninger. Et antall differenser mellom observerte avlingsmengder kan gi tilstrekkelig grunnlag for den konklusjon at "sort T_1 gir større kornavling enn sort T_2" selv om noen av differensene er negative og noen positive. En slik setning eller konklusjon hevder derfor ikke at innholdet gjelder for alle enkelttilfelle.

La oss tenke oss at en lege påstår at han har funnet en vaksine mot forkjølelse. Ønsker vi da å sette denne påstanden på prøve, må vi vaksinere et antall personer med vaksinen. Hvis da i

det minste en av de vaksinerte blir forkjølet kort tid etter vaksineringen, viser det at vaksinen iallfall ikke gir full beskyttelse. Men oppfinneren av vaksinen vil med rette hevde at det var da heller ikke det han mente. Påstanden gikk bare ut på at vaksinen gir en viss beskyttelse, slik at risikoen for å bli forkjølt er mindre når en er vaksinert enn når en ikke er det.

De aller fleste setninger innen biologi og økonomi, mange også innen teknikk, er av denne typen. De pretenderer ikke å ha gyldighet for alle tilfelle. Likevel representerer de nyttig kunnskap. Hvis det f.eks. av en eller annen grunn er ønskelig å behandle goudaost på samme måte i alle ostelagre i Norge, kan det være nyttig å vite at "behandlingsmåte T_1 gir mindre svinn enn T_2" selv om dette ikke gjelder for alle gjentak eller ostelagre.

I alle typisk empiriske fag kommer kunnskap om slike setninger i stand ved induksjon. Kunnskapen bygger da på det vi har observert i et antall enkelttilfelle eller gjentak. Setningen eller konklusjonen har imidlertid et generelt innhold, den sier ikke noe om de gjentak vi bygger på. Den sier noe om en større mangfoldighet av gjentak. I statistikken går denne større mangfoldigheten av gjentak under navn av universet eller populasjonen.

Universet er i reglen ikke noe konkret, noe å ta og føle på. Det er en tenkt mangfoldighet av gjentak som har en bestemt felles karakteristikk. Det er hva vi pleier å kalle en abstraksjon og må oppfattes som ubegrenset. Men som vi skal se senere, er det ikke alltid slik. Vi har også universer som er endelige i størrelse.

I aktuelle tilfelle har vi bare en del av universet til rådighet, et utvalg eller sampel* av de gjentak universet består

* Fornorskning av det engelske ordet "sample".

av. I eksemplet i tabell A.1. har vi således et sampel på tre gjentak, og dette må vi oppfatte som representant for et abstrakt univers. Å indusere vil si at vi på grunnlag av de observasjoner samplet gir oss, gjør oss opp en mening om et spørsmål. Denne meningen eller oppfatningen uttrykkes så i en setning, og innholdet i setningen representerer så kunnskap om universet.

De metodene vi da gjør bruk av, går under navn av statistiske metoder. I vid forstand er metoder vi kaller statistiske slike som vi bruker når vi skal skaffe oss et tilfredsstillende sampel av gjentak, slike vi bruker til bearbeidelsen av disse observasjonene og endelig metoder vi bruker når det gjelder selve induksjonsprosessen.

A.4. Litt om tankevirksomhet.

All effektiv virksomhet - det er det samme hva den går ut på - forutsetter ordnet tenkning. Dette forutsetter igjen klargjøring av de begreper som nyttes og kjennskap til prinsipper for ordnet tankeinnhold. Disse emnene blant mange andre hører inn under faget logikk som er, kan vi si, et fag som handler om menneskelig tankevirksomhet. Det er ikke mulig for oss her å komme nærmere inn på logiske emner i noen større utstrekning. Vi må nøye oss med et par punkter.

Tenker en over hvordan tenkning begynner, vil en oppdage at tenkning må ha et utgangspunkt, setninger med forståelig tankeinnhold. Vi kaller slike setninger for premisser. Ordnet tenkning går så ut på å komme fram til nye setninger som er nødvendige avledninger av premissene. I sin aller enkleste form - det en kaller en syllogisme - består det hele i en hovedpremisse, en bipremisse og en konklusjon. La f.eks. hovedpremissen og bipremissen være:

"alle pattedyr er hvirveldyr" og "katten er et pattedyr". En nødvendig slutning eller konklusjon må da bli at "katten er et hvirveldyr". Dette sier naturligvis ikke noe mer enn at alle dyr (tilfelle) som etter definisjonen er pattedyr, har "ryggrad" som felles karakteristikk. Katten er et pattedyr og må da også ha denne karakteristikken.

Syllogismen er et enkelt eksempel på den tankevirksomhet som vi kaller deduktiv. Deduksjon er derfor det resonnement vi gjør bruk av når vi tar utgangspunkt i en eller flere setninger og utleder andre setninger av disse. Den eller de setninger som utledes, må da være nødvendige konsekvenser av premissene.

Deduksjonen er naturligvis sjelden så enkel som i vårt eksempel. Tar vi for oss eksempler fra matematikken, vil vi som oftest finne at utledningen av konklusjonene kan være både lang og omstendelig. Vi kan dessverre ikke komme nærmere inn på dette omfattende emne her. Av hensyn til det følgende må vi imidlertid nevne to sider ved deduksjonen som er meget viktige.

Det er for det første ikke en forutsetning for riktig deduksjon at det premissene gir uttrykk for er sant eller riktig. Vi må skille mellom det at konklusjonen er riktig utledet og at innholdet av premisser og konklusjon er noe vi kan akseptere. Ta oss som eksempel ta for oss denne deduksjonen:

a) alle mennesker har fingrer

b) alle fingrer har negler

og konklusjon: alle mennesker har negler.

Vi kan naturligvis ikke akseptere innholdet i noen av disse setningene. Det finnes mennesker uten fingrer, og det finnes fingrer uten negler. Likevel er konklusjonen riktig utledet. Den er en nødvendig konsekvens av premissene.

For det annet kan innholdet i konklusjonen være riktig selv om deduksjonen er feilaktig. Professor Susan Stebbing skriver følgende som det er vel verd å feste seg ved*: "Many unsound arguments have been used to support conclusions that are in fact true. When, however, the argument is unsound, we have not justified our acceptance of the conclusions. Our belief is to that extent unreasonable, although not false".

Innen typisk empiriske fag som biologi har deduksjonen som oftest bare indirekte betydning. Innen disse fag har vi nemlig sjelden setninger med et innhold som vi stoler så fast på, at vi tør bruke dem som premisser for deduksjon av nye setninger. Deduksjonen har størst betydning på følgende måte. Vi sier "la oss anta at" eller "sett at" så og så er tilfelle, og så deduserer vi nye setninger ut fra dette. Premissene har da en hypotetisk karakter. Vi kan f.eks. si: la oss anta at byggsortene T_1 og T_2 under samme vekstvilkår gir samme kornavling. Vi er da klar over at dette kan være feilaktig. Men vi bruker påstanden som en premisse for deduksjon av en konklusjon som er slik at vi kan prøve om innholdet av den er akseptabelt. Prøvingen eller testingen skjer så ved at vi utfører et forsøk for sammenligning av de to sortene. For at de observasjonene vi skaffer oss ved forsøket skal kunne gi oss grunnlag for å ta standpunkt til den hypotetiske premissen, altså ta standpunkt til påstanden om de to sortene, må vi med premissen som utgangspunkt kunne utlede konsekvenser som forteller oss hva vi skal vente av observasjonene dersom premissen er riktig. Oppfyller så observasjonene ikke det vi venter av dem, må vår konklusjon bli at det er noe i veien med premissen. Innholdet i den kan ikke opprettholdes, og konklusjonen må derfor

* Fra hennes bok "Thinking to Some Purpose". Penguin Books Ltd.

bli at det er en eller annen ulikhet mellom de to sortene som gjør at den ene gir større kornavling enn den andre.

I litteraturen finner en ikke sjelden slike uttrykk som "det er statistisk bevist at ..." eller "det er statistisk motbevist at ....". Slike uttrykksmåter er meget uheldige. Det dreier seg ikke om bevis eller motbevis av samme karakter som i matematikk og logikk. I empirien gjelder det standpunkt for eller mot et utsagn. Og da er det meget bedre og mer treffende å bruke ord som akseptere og forkaste. Å akseptere vil si å ta standpunkt for, å forkaste å ta standpunkt mot. Det er også en tredje mulighet, nemlig det å la være å ta standpunkt. En kan kanskje finne at grunnlaget er for svakt, eller at det er gjort feil under planlegging og/eller utførelse av forsøket. I slike tilfelle er naturligvis det tredje standpunktet det riktige.

A.5. Hypotese, regel, lov og teori.

I forskning med praktisk målsetting er hensikten å finne regler som kan tjene som rettledning for praktiske handlinger. Som eksempel kan vi tenke oss et forsøk for sammenligning av et antall valte potetsorter med den målsetting å finne den av sortene som en kan anbefale til dyrkerne innen et geografisk område. Sammenligningen måtte da gjelde alle praktisk relevante egenskaper hos poteter, slike som produktivitet, tidlighet, sykdomsresistens og smak. Ved å utføre et forsøk etter en plan som vi skal beskrive senere, vil det kanskje lykkes å peke ut en av sortene som kan karakteriseres som den beste blant de sortene som er tatt med i forsøket. Vi kan i så fall si at vi har funnet en regel som vi kan bruke som rettledning for en praktisk handling.

Som et annet eksempel kan vi tenke oss at en har to eller flere impregneringsmidler for trematerialer som skal brukes til ytre kledning i bolighus. En kan da også utføre et lignende forsøk for sammenligning av disse midlene. Resultatet kan bli at en kan peke ut et av midlene som, når alt tas i betraktning, kan sies å være det mest fordelaktige blant de midlene som er tatt med i forsøket.

I alle praktiske virksomheter brukes daglig slike regler som rettesnor for handling. Det er karakteristisk for de aller fleste slike regler at det er knyttet vilkår til dem. Det er f.eks. rimelig å tro at det impregneringsmiddel som er best etter resultatet av forsøket, ikke er det beste under alle klimatiske forhold. Det kommer an på hvordan forsøket er planlagt og utført. Er det utført i et distrikt med relativt fuktig klima, kan en ikke uten videre regne med at regelen kan brukes i et distrikt hvor klimaet er relativt tørt. Valget av potetsort vil også avhenge av klimafaktorer og oftest også av karakteren av dyrkingsjorda. Er forsøket planlagt og utført med sikte på et bestemt distrikt, og det har lykkes å peke ut en sort som den beste, kan en ikke uten videre anbefale denne sorten som den beste for et annet distrikt.

Oppdagelsen av slike regler har vært og er av den største betydning både for det enkelte menneske og for samfunnet. Vi kan vel si at materiell fremgang skyldes oppdagelsen av slike handlingsregler. Dessverre er verken forskerne eller brukerne alltid oppmerksomme på at praktiseringen av slike regler kan ha alvorlige negative virkninger.

Et annet motiv for forskning er trangen til dypere innsikt og forståelse, trangen til å utvide erkjennelsen. Et observert

fenomen som kan være en ting, en prosess, en hending osv., føles av mange som en utfordring. Spørsmålet kan da være: hva er fenomenet for noe, er det kanskje noe det kan identifiseres med? Eller det kan være: hva er den umiddelbare årsak til at det har inntruffet? Franklins oppdagelse av at fenomenet "lyn" er et elektrisk fenomen er et eksempel på svar på det første spørsmål. Oppdagelsen av penicillinet er et eksempel på svar på det andre. Svar på slike spørsmål kan naturligvis også føre til regler for praktiske handlinger. Det skyldes Franklins oppdagelse at en kan beskytte seg mot lynnedslag ved å sette opp lynavleder. Bruken av penicillinet i medisinen er også vel kjent.

Resultatet av slik motivert forskning kalles ofte en vitenskapelig lov. Med lov mener en da en regel som har universell gyldighet. Forskjellen mellom en regel og en lov kan imidlertid være nokså utvisket. Betegnelsen lov stammer fra fysikken. Eksempler er loven om fritt fall og Keplers lover for planetenes bevegelser. Selv om vi ikke kan bruke loven om fritt fall uten å kjenne akselerasjonen som er avhengig av breddegraden og høyden over havnivået, er selve loven universell. Slike lover forutsetter uniformitet. I biologiske og sosiale sammenhenger finner vi ikke denne strenge uniformiteten, og da er det kanskje riktigere å oppfatte en lov som uttrykk for en tendens. Liebigs minimumslov kan brukes som eksempel.

Vi vet at for at en plante skal vokse må den ha tilgang på flere forskjellige næringsemner. Minimumsloven sier da at veksten er avhengig av mengden av det næringsemne som det er minst av i forhold til behovet. Undersøkelser har imidlertid vist at den nytten en plante kan ha av et næringsemne kan være avhengig av andre næringsemner, og dessuten at en organisme kan ha evne til å er-

statte med et annet et næringsemne som ikke finnes i tilstrekkelig mengde. Vi kan derfor bare si at minimumsloven er en generell regel som er realisert under visse vilkår.

Når en skal prøve å finne uttrykk for en lovmessighet, må en naturligvis benytte de uttrykksmidler som er funnet opp. En kan bruke ord og begreper bundet sammen i en entydig setning, og i mange tilfelle - som f.eks. i fysikken - kan en ty til matematikken. Men enten en bruker det ene eller det andre hjelpemiddel, blir resultatet som oftest en forenkling. Enkelte mindre betydningsfulle karakteristikker må sløyfes, slik at en lov eller en regel må oppfattes som en abstraksjon. Som eksempel kan nevnes at i det matematiske uttrykk for fallovene er ikke luftmotstanden tatt med.

Den betydning oppdagelsen av en vitenskapelig lov har, er vel først og fremst at vår erkjennelse blir utvidet. Vi vet mer om de krefter som regulerer naturprosessene eller prosesser i det menneskelige samfunn. Vi mener at vi forstår mer. Men oppdagelsen kan også føre til at forskning med praktisk målsetting kan bli mer effektiv. Arvelovene er et eksempel på dette. Også før disse lovene ble oppdaget og kjent lyktes det å skape nye plantesorter og husdyrraser. Oppdagelsen skapte imidlertid muligheter for et mer målbevisst arbeid.

Hvis en vitenskapelig lov eller flere slike lover beskriver eller forklarer et større eller mindre antall fenomener ell r prosesser, bruker en betegnelsen teori. Det er særlig innen de såkalte eksakte naturfag som fysikk og kjemi at teorier er blitt utformet. Dette kommer av at det innen den fenomenkrets disse fagene omfatter, er en utstrakt uniformitet. Men selv om det er meget vanskeligere, har en også innen andre fag gjort forsøk på å knytte sammen regler eller/og lover til en helhet. Resultatet

blir da omtalt som en teori. Men ofte befinner disse teoriene seg på et nokså forberedende stadium, og burde kanskje derfor i mange tilfelle betegnes som hypoteser.

På et forberedende stadium har alle regler, lover og teorier vært hypoteser. Det sies ofte at en hypotese er en foreløpig forklaring. Bedre er det kanskje å si at en hypotese er en tanke eller en ide som melder seg hos en forsker som arbeider med et observasjonsmateriale eller som resultat av ren spekulasjon. Innholdet i en hypotese kan derfor ikke regnes som kunnskap. Den er noe en oppfatter som en mulighet og som en tar sikte på å få bekreftet eller forkastet ved å konfrontere den med observasjoner.

Hvis vi i en gitt situasjon gir tanken helt fritt spill, kan antall hypoteser bli stort. Til en nyttig hypotese må en derfor sette det krav at den ikke inneholder noe element som er i strid med kjente fakta. Vi må også kreve at den blir uttrykt i en entydig setning. Dette er nemlig forutsetningen for at den kan settes på prøve.

Det er ikke mulig å komme inn på her hvordan en hypotese blir til*. Vi skal imidlertid legge merke til at det ikke finnes noen metode som en kan ta i bruk. Hypoteser oppstår hos mennesker med fantasi, kombinasjonsevne, evne til å oppdage nyttige analogier og evne til å oppdage fakta som ikke er vanlige. Alexander Flemings ide om bruken av penicillinet oppstod hos ham ved at han festet seg ved at noen stafylokokkkulturer oppførte seg annerledes enn andre.

I empirisk forskning bruker en et begrep som kanskje noe misvisende er blitt kalt en null-hypotese. Det er en slik null-

* Interesserte vises til de mange eksempler som er beskrevet av W.J.B. Beveridge i boken "The Art of Scientifie Investigations".

hypotese vi tenkte oss brukt i eksemplet med de to byggsortene i avsnitt A.4. Vi tenkte oss da at vi som premisse for deduksjonen påstod at de to sortene gav samme mengde kornavling. En vanlig og egentlig hypotese er en tanke, en ide, eller en konstruksjon som muligens kan være riktig og som derfor kan aksepteres hvis prøvningen faller ut til gunst for den. En nullhypotese kan neppe i noe tilfelle aksepteres. Men vi skal se senere at den kan være meget nyttig fordi den kan gi oss mulighet for påvisning av et eller annet alternativ, f.eks. at byggsort T_1 gir større kornavling enn T_2.

Når en skal gi en muntlig eller skriftlig fremstilling av et emne, må en bruke ord og uttrykk som en bruker til daglig. Det en tar sikte på er naturligvis at det en sier eller skriver, skal oppfattes slik en selv mener at det skal forstås. Men vi blir jo ofte misforstått. Dette kan komme av at andre ikke alltid legger samme mening i ord og uttrykk som vi gjør selv.

Vi har foran brukt uttrykket at en regel eller lov er sann eller riktig. Dette er kanskje ikke heldig fordi de fleste sannsynligvis legger noe endelig og fastslått i slike karakteristikker. Hvis dette er tilfelle, er uttrykksmåten ikke dekkende. Erfaringen viser nemlig at en regel eller lov ofte har kort levetid. Fortsatt forskning viser at det har vært nødvendig med endringer eller modifikasjoner, om ikke forkastelse. Minimumsloven som vi brukte som eksempel foran, ble sannsynligvis tidligere oppfattet som like endelig og fastslått som fallovene i fysikken. Senere forskning har imidlertid vist at modifikasjoner er blitt nødvendige. Det ville derfor kanskje vært en fordel om ord som sann og riktig som karakteristikker på regler, lover eller teorier ble erstattet med noe mer nøytralt, f.eks. treffende.

A.6. Prognose.

En handlingsregel er noe som forteller oss eller foregir å kunne fortelle oss hva som kommer til å hende dersom vi handler etter den. Den sier m.a.o. noe om fremtiden og er derfor det vi kaller en prognose.

En kokebok inneholder et helt register av slike prognoser, en oppskrift på en matrett er jo ikke noe annet enn en regel som sier hvilke råstoffer vi skal bruke og hvordan disse skal behandles for at vi skal oppnå det vi tilsikter. Men vi vet av erfaring at om vi følger oppskriften pinlig nøyaktig i to tilfelle, vil resultatet ikke bli nøyaktig det samme. Dette kan bl.a. komme av at råstoffene er varierende i kvalitet.

En handlingsregel kan derfor som oftest ikke fortelle oss helt nøyaktig hva som vil hende dersom vi handler etter den. Den endelige prøve på om en handlingsregel er treffende blir derfor om det den sier om fremtiden viser seg å holde stikk med så stor nøyaktighet at vi våger fortsatt å bruke den. Det er som med værmeldingene som jo også er prognoser. Det ville være i strid med erfaringene om vi stoler helt og fast på hva meteorologene sier om været i morgen.

Det er flere slags prognoser. En handlingsregel og en værmelding hører ikke med til samme type. Det er imidlertid ikke mulig å gå nærmere inn på dette emne før vi har gjennomgått metoder for utarbeidelse av prognoser. Vi skal derfor komme tilbake til saken.

A.7. Forskning og samfunnet.

Forskning har ført til at våre kunnskaper på de aller fleste områder er blitt utvidet og fordypet. Spør en derfor om hva forskning betyr for samfunnet, kan en vel trygt svare at den er en forutsetning for fremgang både økonomisk og erkjennelsesmessig. Et viktig resultat av naturvitenskapelig forskning er at menneskene er blitt frigjort fra meget av den usikkerhet og vilkårlighet som skyldes naturfenomener en tidligere ikke kunne forklare på annen måte enn som virkning av overnaturlige krefter. En kan lett forestille seg at Franklins påvisning av at lyn er et elektrisk fenomen måtte kjennes som en befrielse. Et uforklarlig fenomen, for de fleste noe mystisk og farlig, ble med oppdagelsen redusert til noe som hadde liten betydning.

At forskning med praktisk målsetting, bl.a. teknisk forskning, på mange måter har betydning for samfunnet, er det ikke delte meninger om. Det er ikke vår sak her å gi en beskrivelse av den utvikling som skyldes slik forskning*. Det er tilstrekkelig å nevne at evnen til å produsere er økt, at meget av det menneskelige slit er blitt borte og at arbeidstiden er kortet inn. Vi må imidlertid stanse litt ved to meget vesentlige spørsmål. Det første av disse blir vi stilt overfor fordi de midlene som stilles til rådighet for forskning, ikke strekker til for alle forskningsprosjekter og at det derfor er nødvendig å prioritere. Det andre gjelder effekten av tiltak som bygger på handlingsregler funnet gjennom forskning.

* En illustrerende og morsom fremstilling er gitt av Bertrand Russell i boka "The impact of Science on Society".

Hvor meget samfunnet skal bruke av sine inntekter til forskning er et politisk spørsmål som det sannsynligvis alltid vil være strid om. Det er imidlertid en kjennsgjerning at det til stadighet finnes flere forslag om forskningsprosjekter enn de en har midler til å ta opp. Dette betyr at noen må velge ut de prosjekter som skal støttes. Flere av dem som denne boka er skrevet for, vil direkte eller indirekte få med slike saker å gjøre. Det kan derfor bli aktuelt for den enkelte å ta stilling til hvilke kriterier som bør brukes når det skal prioriteres.

De fleste som har hatt eller har med slike avgjørelser å gjøre, går vel oftest ut fra formålet med forskningsprosjektet som det viktigste. Dette er sikkert riktig i mange tilfelle, men det må ikke dominere for sterkt. Og det må alltid og konsekvent forlanges at formålet er klart beskrevet og avgrenset.

Legger en imidlertid for stor vekt på formålet med en undersøkelse, vil en altfor ofte komme til å gi høy prioritet til prosjekter som det er rimelig å tro vil gi resultater på kort tid. En vil da kunne komme i skade for å skyve ut undersøkelser av mer fundamental natur hvor formålet ikke er presentert med stor vekt på umiddelbar nytte, men som likevel også kan gi grunnlag for løsning av praktisk betonte oppgaver av stor betydning.

Et annet utgangspunkt for prioriteringen er den plan for forskningsprosjektet som legges fram. Og dette utgangspunktet er kanskje det viktigste. Det hender at formålet med en undersøkelse er slik at prosjektet bedømt ut fra dette må gis høy prioritet, men at planen er så mangelfull at det er liten mulighet for at undersøkelsen kan føre til et resultat. Vi skal senere komme noe nærmere inn på planleggingen av undersøkelser. Her skal vi der-

for nøye oss med å nevne at dersom det gjelder en empirisk undersøkelse, vil planen bestå av tre hovedledd, nemlig 1) problemstilling, 2) observasjonsarbeid og 3) analyse. Hvert av disse leddene bør kreves beskrevet. Hvor mange detaljer en skal forlange er det ikke lett å si. Men det er f.eks. ikke nok at det blir opplyst at observasjonene er tenkt skaffet til veie ved forsøk, en må kreve at forsøksplanen er skissert. At det er nødvendig eller meget ønskelig at også planen for analysen er gjennomtenkt og beskrevet skal vi gi en begrunnelse for senere.

Det er også andre momenter det må tas hensyn til når det gjelder valget av forskningsprosjekter. Undersøkelser av mange forskere og gjennom lengre tid har vist at en handling, f.eks. et inngrep i naturen, har mer enn en effekt. Blant disse effektene er det noen som vi kan klassifisere som gode, noen er mindre gode og noen er uønsket eller skadelige. I industrien må vi regne med forurensning av luft og vann som skadelige og vi må regne støy som iallfall lite ønskelig.

Dette må tas hensyn til når det gjelder prioritering av forskningsprosjekter. Er det en undersøkelse med praktisk målsetting det er tale om, må det forlanges at observasjonsarbeidet skal omfatte flest mulige effekter. Det kan naturligvis ikke kreves at alle effekter er tatt med fordi en jo som oftest ikke vet hvor mange og hvilke det er. Det som kan og bør forlanges er at denne siden av saken er forsvarlig gjennomtenkt.

Innen enkelte områder er forskerne kommet så langt at innsikt som er vunnet eller som kan regnes med ved fortsatt forskning, er slik at den omsatt i handling kan føre til de verste ulykker. Eksempler kan finnes innen atomfysikk og bakteriologi.

De rent etiske problemer som melder seg i slike sammenhenger og som er meget viktige, har imidlertid liten relevans til det emne vi skal beskjeftige oss med her.

Vi vet at på nesten alle områder i samfunnet har autoriteter og eksperter fått stor innflytelse. I de mest avanserte samfunn kommer dette av at det er blitt så mange spørsmål en skal ta stilling til. Men det kommer også av at mange spørsmål er vanskelige og kompliserte og at en derfor finner det lettvint å spørre ekspertene om hva en bør mene. Dette er uheldig.

I mange tilfelle er svar en skal gi på spørsmål, avhengig av resultater av forskningsvirksomhet og det blir da naturlig nok forskerne en henvender seg til. Ikke alle slike resultater er tilstrekkelig underbygget, og det kan derfor være bra om flest mulige i et samfunn har fått en utdanning som gjør det mulig for dem å møte forskerne på deres eget område. Det det gjelder om i slike tilfelle er at en kan stille forskeren relevante spørsmål som angår måten de resultater som det vises til, er kommet i stand på. Først og fremst gjelder det da at flest mulige har et visst innblikk i forskningsmetode.

Litteratur.

Det finnes mange bøker som handler om slike alminnelige emner som de vi har tatt for oss foran. Det er likevel ikke lett å oppgi videregående litteratur som studentene kan ha nytte av. I tillegg til de bøker som er nevnt i teksten, kan det kanskje være av interesse for noen å stifte bekjentskap med de bøker som er ført opp nedenfor. Tar en for seg en eller flere av disse og lignende bøker, vil en finne at både emnevalg og emnebehandlingen som oftest er nokså subjektiv.

Freedman, Paul: The Principles of Scientific Research. London 1949.

Lastrucci, Carlo L.: The Scientific Approach. Basic Principles of the Scientific Method. Cambridge (Massachusetts) 1967.

Nash, Leonard K.: The Nature of the Natural Sciences. Boston 1963.

Næss, Arne: En del elementære logiske emner. Oslo 1950.

B. Observasjonene og midler til å beskrive dem.

B.1. Random variable.

Vi kan tenke oss at en undersøkelse i en granskog begynner med at en tar ut et utvalg eller sampel av trær. For hvert av disse måler en høyden, brysthøydediameteren, høyden av kvistfri stamme og lignende størrelser. En bestemmer kanskje også alderen. For hvert tre i samplet har en da/et tall for høyden og i samplet på n trær n slike tall. Disse tallene er da våre observasjoner av høyden. Når vi har skaffet oss slike observasjoner, vil vi se at det er en større eller mindre variasjon i dem. Observasjonene varierer fra tre til tre, eller fra gjentak til gjentak.

Disse observasjonene er ikke nøyaktig riktige tall for det som er observert. I praksis er vi nemlig ikke i stand til å måle helt nøyaktig, og våre observasjoner er derfor bare tilnærmet riktige tall for det vi har observert. Disse riktige tallene sier vi er verdier av en random variabel. Høyden av grantrær er altså en random variabel. Det samme gjelder brysthøydediameteren og høyden av kvistfri stamme.

Andre eksempler på random variable er antall kronblader hos soleihov, antall grisunger pr. kull, prosent fett i melk, potetavlingen pr. rute i et feltforsøk osv. En observasjon av en random variabel refererer seg til et gjentak i et sampel av gjentak.

Noen random variable kan ha bare bestemte atskilte tallverdier. Eksempler er antall kronblader og antall grisunger. Disse kalles diskrete random variable. Andre random variable kan ha en hvilken som helst reell tallverdi mellom en nedre og en øvre grense. Eksempler er høyden av grantrær og prosent fett i melk. Disse kalles kontinuerlige random variable.

I noen tilfelle er det ikke en random variabel vi observerer. Det er en enkelt størrelse. Den rettlinjede avstand mellom to punkter i terrenget er et eksempel. Observerer vi en slik størrelse på uavhengig måte et antall ganger, får vi en rekke tall som varierer fra gjentak til gjentak. Variasjonen skyldes her målefeilene eller observasjonsfeilene.

En observasjon av en random variabel er en karakteristikk av det gjentak observasjonen refererer seg til. En person kan således karakteriseres ved observasjoner av f.eks. høyde, skulderbredde, alder og andre random variable. En person kan imidlertid også karakteriseres ved f.eks. kjønnet og ved øyefargen. En rekrutt kan karakteriseres ved resultatet av legeundersøkelsen (udyktig, hjelpedyktig, stridsdyktig). Blomster karakteriseres ved fargen. Slike karakteristikker kalles konstante kjennetegn eller bare kjennetegn.

Slike kjennetegn er ikke random variable. Men det som oftest interesserer oss, er det absolutte eller det relative antall gjentak med samme kjennetegn, f.eks. det relative eller prosentiske antall rekrutter med kjennetegnet stridsdyktig, det relative antall gutter med kjennetegnet blå øyne, det relative antall planter med kvite blomster. Disse tallene, både de absolutte og de relative, er random variable. I sammenligning med observasjoner av en random variabel hvor det er variasjon både mellom gjentak innen sampler og mellom sampler, er forskjellen den at antallet gjentak med et bestemt kjennetegn varierer bare sampler imellom.

B.2. Frekvensfordelingen.

Er antallet av gjentak stort, er det vanskelig å ha oversikt over observasjonene. En må derfor ordne dem i det en kaller frekvensfordeling.

I Tab. B.1. er gitt n=29 observasjoner av den dobbelte barktykkelse hos gran. Observasjonene er gitt i hele milimeter, dvs. at det kan enten være avrundede observasjoner eller det kan være observasjoner som er tatt med et redskap som måler med 1 mm nøyaktighet. Den observerte random variable er naturligvis her en av den kontinuerlige typen.

TABELL B.1.

14	20	20	22	25
14	16	20	19	22
17	17	22	21	22
22	17	25	20	24
18	16	24	22	24
18	22	21	20	

I dette tilfelle er antall observasjoner så lite at vi i praksis ikke ville bry oss med å ordne dem. Men vi kan bruke eksemplet til å vise hvordan observasjoner kan ordnes i en frekvensfordeling. Vi ser at noen av observasjonene forekommer flere ganger. Vi har f.eks. at x=14 forekommer i 2 gjentak, x=17 i 3 gjentak og x=20 i 5 gjentak. Frekvensfordelingen er en tabell over observasjonsverdiene (x) og antall gjentak (z) med vedkommende observasjonsverdi. En konstaterer lett at frekvensfordelingen i dette tilfelle er slik som vist i Tab.B.2.

Tabell B.2.

x	z	x	z
14	2	20	5
15	0	21	2
16	2	22	7
17	3	23	0
18	2	24	3
19	1	25	2

n= 29

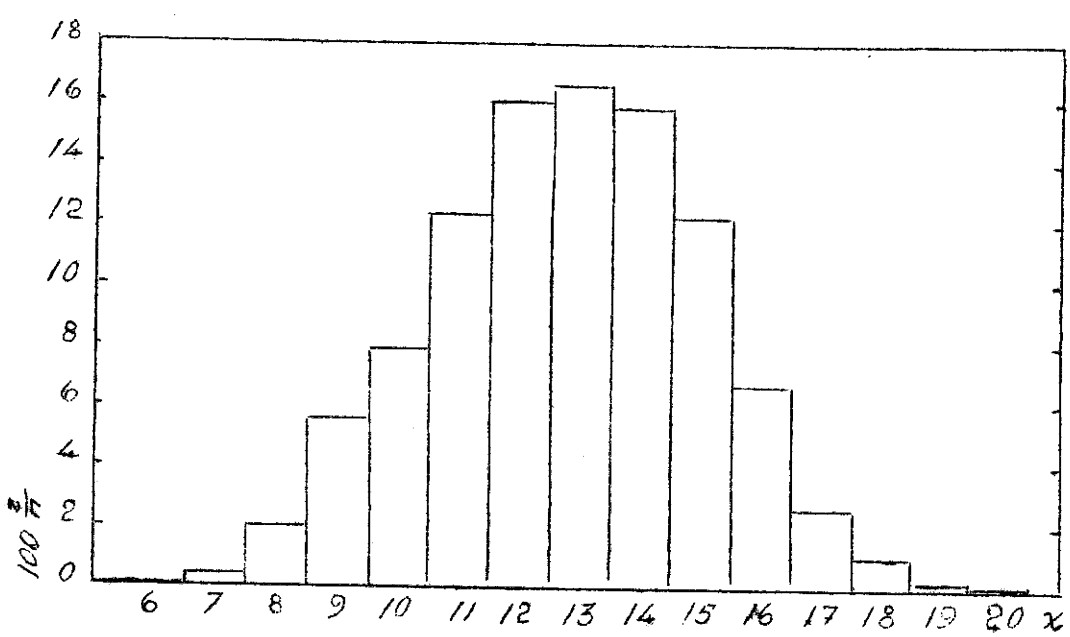
Eksemplet i Tab. B.3. hvor antall gjentak er n=1905, viser hva en kan oppnå ved å ordne observasjonene i en frekvensfordeling. Opprinnelig hadde en her en tabell over 1905 uordnede observasjoner

av en diskret random variabel, nemlig antall arrstråler hos en valmueart. Vi ser at frekvensfordelingen gir en god oversikt. Vi ser at alle observasjonene ligger mellom 6 og 20. Det er få observasjoner ved den nedre og den øvre variasjonsgrensen og en tydelig opphopning omtrent på midten av området.

Tabell B.3.

x	z	$100\frac{z}{n}$	x	z	$100\frac{z}{n}$
6	3	0.16	14	302	15.85
7	11	0.58	15	234	12.28
8	38	1.99	16	128	6.72
9	106	5.56	17	50	2.62
10	152	7.98	18	19	1.00
11	238	12.49	19	3	0.16
12	305	16.01	20	1	0.06
13	315	16.53			
				1905	99.98

En slik frekvensfordeling kan fremstilles grafisk på flere måter. Mest alminnelig er det å bruke et såkalt søylediagram i et



Figur B.1.

rettvinklet koordinatsystem. En avsetter da x på den horisontale aksen og z i prosent av n på den vertikale aksen og tegner søyler med z% som høyde. Fig. B.1. viser et slikt søylediagram for eksemplet i Tab. B.3.

Et lignende eksempel er vist i Tab. B 4. Det er kommet i stand ved at en har foretatt opptelling av antall grisunger i hvert av n = 334 kull. I tabellen er derfor z antall kull med x unger.

Tabell B.4.

x	z	x	z
2	1	10	51
3	1	11	52
4	4	12	39
5	6	13	45
6	17	14	21
7	20	15	7
8	30	16	5
9	35		

n = 334

Er det en kontinuerlig random variabel som er observert, må observasjonene ordnes i klasser. Frekvensfordelingen er da en tabell over disse klassene (eventuelt med klassenes midtverdier) og antall observasjoner innen klassene. Et eksempel er vist i Tab. B.5. En timoteivoll ble delt opp i 240 kvadratiske ruter på 25 kvadratmeter. Avlingen ble så veid for hver rute, og en fikk da n = 240 observasjoner. Ordnes disse i klasser med en klassevidde på ett kg, får en den frekvensfordelingen som er vist i tabellen. Her er da x midtverdien i klassene og z frekvensen.

Tabell B.5.

Klasse	x	z	Klasse	x	z
11-12	11.5	2	19-20	19.5	21
12-13	12.5	6	20-21	20.5	16
13-14	13.5	9	21-22	21.5	12
14-15	14.5	18	22-23	22.5	12
15-16	15.5	30	23-24	23.5	3
16-17	16.5	40	24-25	24.5	2
17-18	17.5	33	25-26	25.5	4
18-19	18.5	30	26-27	26.5	2

n = 240

Eksemplene i tabellene B.3. til B.5. har til felles at stort sett tiltar frekvensene fra den nedre variasjonsgrensen til et maksimum omtrent på midten av området og avtar så mot den øvre variasjonsgrensen. Det er denne typen av frekvensfordelinger som er den mest vanlige. Det er imidlertid også mange avvik fra denne almineligste form. Det finnes eksempler på frekvensfordelinger hvor frekvensene avtar fra den ene variasjonsgrensen til den andre, og det finnes eksempler på fordelinger med mer enn ett maksimum. I Tab. B.6. er vist et eksempel på en helt skjev frekvensfordeling. Den observerte random variable er i dette tilfelle antall kronblader hos soleihov.

Tabell B.6.

x	z
5	223
6	45
7	6
8	4
9	3

n = 281

B.3. Det aritmetiske gjennomsnitt.

Det er to størrelser som nesten alltid blir nyttet i undersøkelser hvor en gjør bruk av statistiske metoder. Det er det aritmetiske gjennomsnitt og middelavviket. Det er mange grunner til at

disse nyttes så meget. Vi må foreløpig nøye oss med å vise hvordan de skal beregnes og å antyde noe av den nytten en har av dem for deskriptive formål. I vitenskapelige meldinger er det ikke vanlig at en gjengir de observasjoner en bygger på. En nøyer seg som oftest med å oppgi verdiene av gjennomsnittet og middelavviket. Forutsetningen for at en kan tillate seg denne forenklingen, er at disse to tallene gir en god nok karakteristikk av observasjonene sett under ett.

Det aritmetiske gjennomsnitt er et middeltall. Middeltall blir brukt som uttrykk for det alminnelige eller karakteristiske. Når det sies at en mann er middels høy, at høsten er som vanlig eller at en gutt har middels evner, er det et slags middeltall som blir nyttet. Det er imidlertid flere slags middeltall, og disse tilfredsstiller ikke alltid samme formål. Det viktigste av dem er det aritmetiske gjennomsnitt. Det er dette middeltallet vi kommer til å bruke i denne boka og skal da korte inn betegnelsen til bare gjennomsnittet.

La oss betegne observasjonene med $x_1, x_2, x_3, \ldots x_n$ eller kort x_i $(i=1,2,3....n)$. Gjennomsnittet er lik summen av observasjonene dividert med antallet (n). La oss betegne det med $\bar{x}$. Vi har da at*

$$\bar{x} = \frac{1}{n} \Sigma x_i$$

For observasjonene i Tab. B.1. er summen lik $\Sigma x_i = 584$ og antallet er n = 29. Gjennomsnittet er derfor $\bar{x} = \frac{584}{29} = 20{,}14$.

* Skulle vi være helt korrekte, måtte summem skrives $\sum_{i=1}^{i=n} x_i$
Men dette er både tungvint og som oftest overflødig.

Er observasjonene av en diskret random variabel ordnet i en frekvensfordeling, kan en bruke denne som grunnlag for beregningen av gjennomsnittet. En må bare huske at summen av observasjonene da er lik summen av produktene av observasjonsverdiene og frekvensene, altså summen av produktene $z_i x_i$. Gjennomsnittet er derfor lik

$$\bar{x} = \frac{1}{n} \Sigma z_i x_i$$

For eksemplet i Tab.B.4. er n=334 og vi finner at $\Sigma z_i x_i = 3462$. Gjennomsnittet er derfor $\bar{x} = \frac{3462}{334} = 10,37$.

Når en skal beregne gjennomsnittet for observasjonene av en kontinuerlig random variabel, bør en helst bruke de uordnede observasjonene. Bruker en frekvensfordelingen, må en erstatte observasjonene med midtverdiene i de klasser som er brukt. Ved å gjøre det, fører en naturligvis inn en feil som helst bør unngås selv om feilen på verdien av gjennomsnittet som oftest ikke er stor. Med det utstyr av regnemaskiner en nå jevnt over har til disposisjon, betyr bruken av frekvensfordelingen ikke noen betydelig innsparing av tid. Dette gjelder også beregning av andre størrelser, f.eks. beregningen av middelavviket som vil bli behandlet i avsnitt B.5.

Det aritmetiske gjennomsnitt er en så enkel størrelse at det ikke skulle være grunn til å gå noe nærmere inn på bruken av det. Det gir kan vi si, tyngdepunktet i observasjonsmassen. Det er imidlertid grunn til å advare mot å bruke det uten at det vises til observasjonene selv eller til andre karakteristikker av observasjonene. Meget ulike observasjoner kan nemlig ha samme eller omtrent samme gjennomsnitt. Sett f.eks. at vi får oppgitt at den gjennomsnittlige årsinntekt for voksne menn i en bygd er 15 tusen kroner. Vi må da ikke uten videre bruke dette som uttrykk for inntektsforholdene i bygda. Gjennomsnittet $\bar{x} = 15$ tusen kan nemlig være gjennomsnitt av inntekter

mellom f.eks. 10 tusen og 30 tusen kroner. Men det kan også være gjennomsnitt av inntekter mellom f.eks. 5 tusen og 20 tusen kroner og en eller noen få meget store inntekter.

B.4. Andre middeltall.

Det aritmetiske gjennomsnitt eller bare gjennomsnittet er det mest brukte middeltall eller, om en vil, den mest brukte sentralverdi. Det finnes imidlertid også andre middeltall som blir nyttet i visse sammenhenger. Det er det aritmetiske gjennomsnitt vi skal bygge på i dette kurset, men noen merknader om andre middeltall kan kanskje være på sin plass.

Vi har et middeltall som går under navnet typetallet. Dette er den verdi av den observerte random variable som forekommer oftest i samplet, eller den verdi av x som har den største frekvens. Typetallet for eksemplet i Tab.B.3. er $x = 13$ med frekvensen $z = 315$, mens gjennomsnittet er $\bar{x} = 12{,}76$. Det er lett å innse at hvis frekvensfordelingen er noenlunde symmetrisk, kan ikke forskjellen mellom disse to middeltallene være stor. Det behøver den heller ikke være om frekvensfordelingen er skjev. For eksemplet i Tab.B.6. finner vi således at typetallet er $x = 5$ og gjennomsnittet $\bar{x} = 5{,}29$.

Typetallet kan være nyttig til karakterisering av en frekvensfordeling fordi det er den observasjonsverdi som forekommer oftest. Det gir i noen tilfelle en mer verdifull karakteristikk enn gjennomsnittet. Det kan f.eks. være nyttigere å kjenne den alminneligste inntekt enn den gjennomsnittlige. Typetallet må imidlertid ikke brukes med mindre frekvensfordelingen har et tydelig maksimum. Det er derfor nødvendig at en har et nokså stort antall observasjoner.

Det geometriske gjennomsnitt er lik n'te roten av produktet av de

n observasjonene:

$$G = (x_1 x_2 x_3 \ldots x_n)^{1/n}$$

Vi ser da at $\log G = \frac{1}{n} \Sigma \log x_i$, dvs. at logaritmen til det geometriske gjennomsnitt er lik det aritmetriske gjennomsnitt av observasjonenes logaritmer. Det kan vises at $G \leq \bar{x}$, men forskjellen mellom G og $\bar{x}$ er vel som oftest ikke stor. For eksemplet i Tab.B.6. finner en at G = 5,25, dvs. praktisk talt lik det aritmetiske gjennomsnitt.

Det geometriske gjennomsnitt brukes helst i tilfelle hvor det av en eller annen grunn ville være naturlig å bruke log x i stedet for x selv. Vi setter da

$$y_i = \log x_i$$

som kalles en transformasjon. For denne har vi at $\log G = \bar{y}$. Denne transformasjonen kan naturligvis ikke brukes i andre tilfelle enn der hvor alle observasjonene (x) er positive tall. Det er imidlertid ikke noe i veien for å velge andre logaritmiske transformasjoner, f.eks. sette

$$y_i = A + B \log x_i$$

hvor A og B er gitte konstanter.

Enda et middeltall skal nevnes. Det er midttallet eller medianen M. Dette er et middeltall som deler samplet av gjentak i to like store deler. I den ene delen er alle observasjonene (x_i) større enn M, i den andre mindre.

Vi kan ikke komme nærmere inn på bruken av disse middeltallene her. Vi må nøye oss med disse få linjene og ellers vise til litteratur hvor de er nærmere omtalt.

B.5. Varians og middelavvik.

Gjennomsnittet gir oss det vi kan kalle tyngdepunktet i massen av observasjoner. I tillegg til dette har vi bruk for en størrelse som kan fortelle noe om i hvor stor grad observasjonene varierer omkring denne sentralverdien. Til dette brukes oftest en størrelse som vi kaller middelavviket. Vi skal foreløpig nøye oss med å definere det og å vise hvordan det skal beregnes.

Differensene mellom de enkelte observasjonene og gjennomsnittet er $(x_i-\bar{x})$. Noen av disse differensene er positive og noen negative. Summen av dem er lik null. Vi har nemlig at

$$\Sigma(x_i-\bar{x}) = \Sigma\, x_i - \Sigma\, \bar{x} = n.\bar{x} - n.\bar{x} = 0$$

I praksis blir denne summen som regel ikke eksakt lik null, fordi vi da oftest må bruke en avrundet verdi for $\bar{x}$.

Variansen er lik summen av kvadratene av disse differensene dividert med (n-1), eller

$$V = \frac{1}{n-1}\,\Sigma\,(x_i-\bar{x})^2$$

og middelavviket er kvadratroten av variansen, $s = \sqrt{V}$.

For eksemplet i Tab.B.1. har vi fra før at $\bar{x} = 20{,}14$, og vi finner at $\Sigma\,(x_i-\bar{x})^2 = 267{,}4483$. Siden $n = 29$, finner vi så at

$$V = \frac{267{,}4483}{28} = 9{,}5517 \quad \text{og} \quad s = 3{,}09$$

I likhet med gjennomsnittet har middelavviket samme enhet som den enkelte observasjon. Middelavviket $s = 3{,}09$ for vårt eksempel har derfor enheten millimeter.

Er observasjonene av en diskret random variabel ordnet i en frekvensfordeling, kan variansen beregnes ved formelen

$$V = \frac{1}{n-1}\,\Sigma\, z_i(x_i-\bar{x})^2$$

hvor z_i er frekvensen.

Den formelen for variansen vi har gitt foran, er ikke godt egnet til bruk i praksis. Den er tungvint. Gjennomsnittet er jo oftest et tall med flere desimaler. Det vil derfor som oftest være meget arbeidskrevende å beregne alle differensene $(x_i-\bar{x})$ og dessuten skal jo disse differensene kvadreres og kvadratene summeres. I praksis bruker en derfor helst en indirekte metode.

La c være et valt tall. Vi danner alle differensene (x_i-c). Summen av disse differensene er

$$\Sigma\,(x_i-c) = \Sigma\, x_i - n.c = n.\bar{x} - n.c$$

og følgelig er

$$\bar{x} = c + \frac{1}{n}\,\Sigma(x_i-c)$$

Gjennomsnittet er altså lik det valte tallet c pluss gjennomsnittet av differensene (x_i-c).

Vi bruker også disse differensene til beregning av variansen. Vi har nemlig at

$$\Sigma(x_i-c)^2 = \Sigma(x_i-\bar{x} + \bar{x} - c)^2$$

$$= \Sigma(x_i-\bar{x})^2 + 2(\bar{x}-c)\,\Sigma(x_i-\bar{x}) + n(\bar{x} - c)^2$$

Siden nå $\Sigma(x_i-\bar{x}) = 0$ og $\Sigma(x_i-\bar{x})^2 = (n-1)\,V$, finner vi at

$$V = \frac{1}{n-1}\left\{\Sigma(x_i-c)^2 - \frac{1}{n}[\Sigma(x_i-c)]^2\right\}$$

Når vi bruker denne fremgangsmåten, kan vi velge c slik det passer best i hvert enkelt tilfelle. Vi må naturligvis da ta sikte på å velge c slik at differensene (x_i-c) blir enklest mulige tall. I mange lærebøker blir det anbefalt å velge for c et tall som en antar ligger i nærheten av gjennomsnittet. Men i mange tilfelle er det bedre å sette c lik den minste observasjonsverdi eller lik et tall som er noe mindre enn denne. Da oppnår vi nemlig noe som er regnetekniskt fordelaktig, nemlig at alle differensene (x_i-c) er positive tall. I andre tilfelle, og det er vel det vanligste nå, kan det lønne seg å sette $c = 0$.

For eksemplet i Tab.B.1. er den minste observasjonsverdien lik 14, og det kan derfor være hensiktsmessig å sette c = 14. Vi finner da at $\Sigma(x_i-c) = 178$ og $\Sigma(x_i-c)^2 = 1360$. . Følgelig er

$$\bar{x} = 14 + \frac{178}{29} = 20,14$$

og

$$28\ V = 1360 - \frac{178^2}{29} = 1360 - 1092,5517 = 267,4483$$

dvs. at $$V = \frac{267,4483}{28} = 9,5517$$

Velges c = 0, vil vi naturligvis ha at

$$\bar{x} = \frac{1}{n} \Sigma\, x_i$$

og at $$V = \frac{1}{n-1} \left\{ \Sigma\, x_i^2 - \frac{1}{n} [\Sigma\, x_i]^2 \right\}$$

Er observasjonene av en diskret random variabel ordnet i en frekvensfordeling, vil vi finne at

$$\bar{x} = c + \frac{1}{n} \Sigma\, z_i(x_i-c)$$

og $$V = \frac{1}{n-1} \left\{ \Sigma\, z_i(x_i-c)^2 - \frac{1}{n} [\Sigma\, z_i(x_i-c)]^2 \right\}$$

For vårt eksempel i Tab.B.4. kan vi f.eks. sette c = 10. Vi finner da at $\Sigma\, z_i(x_i-c) = 122$ og $\Sigma\, z_i(x_i-c)^2 = 2318$. Altså er

$$\bar{x} = 10 + \frac{122}{334} = 10,37$$

$$V = \frac{1}{333} \left\{ 2318 - \frac{122^2}{334} \right\} = 6,8271$$

og

$$s = \sqrt{V} = 2,61$$

B.6. Middelavviket som karakteristikk av observasjonene.

Det er nevnt foran at middelavviket brukes som målestokk for størrelsen av variasjonen. Dette kommer vi tilbake til senere på annen måte. Men til en foreløpig orientering skal vi her vise ved noen eksempler at hvis en ved hjelp av gjennomsnittet og middelavviket avgrenser et område med grensene $\bar{x} - 3s$ og $\bar{x} + 3s$, vil en finne alle

eller praktisk talt alle observasjonene innenfor dette.

For eksemplet i Tab.B.1. har vi $\bar{x} = 20,14$ og $s = 3,09$. Altså er $\bar{x} - 3s = 10,87$ og $\bar{x} + 3s = 29,41$. Den minste observasjonsverdi er 14 og den største 25. I dette tilfelle ligger derfor alle observasjonene innenfor området.

For eksemplet i Tab.B.3. er $\bar{x} = 12,76$ og $s = 2.24$. Altså er $\bar{x} - 3s = 6,04$ og $\bar{x} + 3s = 19,48$. Vi ser da at det er 3 observasjoner lik 6 og en observasjon lik 20 som så vidt faller utenfor. Resten av de n=1905 observasjonene faller innenfor området.

Ved hjelp av eksempler og også på annen måte kan en vise at alle eller nesten alle observasjonene finnes innenfor et område fra $\bar{x}-a.s$ til $\bar{x}+a.s$ når vi for a velger en verdi på 3-4. Denne regelen gjelder for alle typer av frekvensfordelinger, altså også for typer som en sjelden kommer bort i. For de vanligste typer kan en nok trygt regne med at regelen holder for a=3. Dette vil da si at variasjonsbredden, som også er en størrelse som brukes i praksis, er omtrent lik 6s. Vi kan derfor si at gjennomsnittet og middelavviket sammen gir en ganske god karakteristikk av observasjonene sett under ett. To frekvensfordelinger med samme gjennomsnitt og samme middelavvik kan imidlertid ha nokså ulikt utseende. I noen tilfelle kan det derfor være av interesse å bruke også andre karakteristikker. Særlig vil en kanskje være interessert i om fordelingen er symmetrisk eller om den er skjev og i tilfelle hvor sterk skjevheten er. Vi har størrelser som gir uttrykk for slike egenskaper ved fordelingen, men vi må her nøye oss med å vise til annen litteratur.

B.7. Om årsaker til variasjonen.

Det er i regelen mange årsaker som ligger til grunn for variasjonen i observasjonene. Hva for årsaker det er, vil avhenge av hva det er som er observert. Er det en bestemt størrelse som er observert, f.eks. den rettlinjede avstanden mellom to punkter i terrenget, vil variasjonen skyldes målefeil. Er det en random variabel som observeres, vil observasjonsfeil bidra noe til variasjonen. Men i slike tilfelle må en regne med at variasjonen i observasjonene i alt vesentlig beror på variasjon i den random variable selv. For de observasjoner som er gitt i Tab.B.5. spiller observasjonsfeil en viss rolle. Det er ikke til å unngå at størrelsen av ruten blir noe varierende og at det gjøres feil både under slåtten, under bergingen og under veiingen av høyet. Men i dette tilfelle er det sikkert ulikheter i jorda og at temperaturen og nedbøren har vært noe varierende over feltet, som er hovedårsaker til variasjonen. Og variasjonen i jorda er igjen avhengig av mange faktorer.

I mange eksempler fra biologien må en regne med at variasjonen skyldes både genetiske ulikheter gjentakene imellom og miljøfaktorer. For eksemplet i Tab.B.4. er det rimelig å anta at det er genetiske ulikheter som er den viktigste årsak til variasjonen, i andre tilfelle kan det være miljøfaktorer som virker sterkest. I Tab.B.7. er gjengitt en fordelingsrekke for antall eksemplarer av sandlilje (Antheri-cum Liliago) funnet innen ruter på 0,25 m^2 på et felt. I dette tilfelle er det rimelig å anta at det er miljøfaktorer (jordsmonn, temperatur, nedbør) og kanskje også konkurranse fra andre arter som har vært avgjørende for variasjonen

Tabell B.7.

x.	z
0	8
1	31
2	27
3	13
4	8
5	1
6	2

n = 100

Tar vi for oss eksempler fra økonomien, er det også der lett å forstå at det i regelen er mange årsaker til variasjonen. Til eksempel er nettoinntekten på gardsbruk betinget av en rekke forskjellige faktorer som bruksstørrelsen, beliggenhet og driftsmåten.

B.8. Sampel og univers.

I avsnitt A.3. har vi forklart at det samplet av gjentak vi har skaffet oss i et bestemt tilfelle, må oppfattes som en representant for et univers. I regelen er dette universet å oppfatte som en abstraksjon.

Tenker vi oss så at den random variable x er observert i hvert gjentak i universet, vil vi også for disse observasjonene kunne operere med et gjennomsnitt. Er universet abstrakt, kan vi naturligvis ikke skaffe oss alle disse observasjonene og da heller ikke beregne gjennomsnittet. Vi kan bare forestille oss den operasjonen vi måtte utføre, vi kan ikke realisere den.

Vi skal senere se hvordan vi skal definere det gjennomsnittet for x som gjelder for universet. Her må vi nøye oss med å si at det er en størrelse som er knyttet til universet og kalles forventningen for x. Forventningen betegnes ofte med μ eller med $E(x)$. Gjennomsnittet av observasjonene i samplet (altså $\bar{x}$) må vi så oppfatte som en representant for $E(x)$.

På samme måte kan vi også tenke oss en størrelse knyttet til universet som svarer, kan vi si, til middelavviket s. Denne størrelsen som vi også skal definere senere, kalles standardavviket og betegnes med σ. I det følgende vil vi også bruke betegnelsen var(x) i stedet for σ^2.*

La oss ta for oss kvadratsummen $\Sigma(x_i - \mu)^2$. Vi betegner estimatoren av μ med $\hat{\mu}$** og setter

$$S = \Sigma (x_i - \hat{\mu})^2$$

hvor i = 1,2,3,...n.

Når observasjonene (x_i) er gitt, er S en funksjon av $\hat{\mu}$, og vi kan da ta for oss som oppgave å bestemme den verdi av $\hat{\mu}$ som gjør S til et minimum. Denne oppgaven kan vi løse ved å sette den deriverte av S m.h.p. $\hat{\mu}$ lik null. Vi finner at

$$\frac{dS}{d\mu} = \Sigma\, 2(x_i - \hat{\mu}).(-1) = -2\,\Sigma(x_i - \hat{\mu})$$

Settes den deriverte lik null, finner vi at

$$\hat{\mu} = \frac{1}{n}\Sigma\, x_i = \bar{x}$$

Gjennomsnittet ($\bar{x}$) er derfor den verdi av $\hat{\mu}$ som gjør S til et minimum.

Den metoden vi har benyttet, går under navn av minste kvadraters metode. Vi kommer senere til å bruke den også i andre tilfelle.

Slike størrelser som μ eller E(x) og σ som altså er knyttet til universet, kalles parametrer. I matematikken forstår en ved en parameter en konstant i en matematisk funksjon, en konstant som vi kan gi en rekke verdier. I statistikken er det nødvendig å utvide begrepet noe, slik at vi med en parameter forstår en størrels knyttet til et

* Forskjellig betegnelse på s og σ er ikke alltid gjennomført. I engelsk litteratur er det alminnelig at både s og σ går under betegnelsen "standard deviation".

** Det er vanlig å betegne parameteren og estimatoren med samme bokstav. For å skille mellom dem settes tegnet ^(hatt) over bokstaven når den brukes som betegnelse for estimatoren.

univers. Som vi skal se senere kan den da også opptre som en konstant i en matematisk funksjon, men det behøver ikke nødvendigvis være slik.

En av de aller viktigste oppgaver i empirisk forskning går ut på å finne tilnærmede verdier av slike parametrer. Vi sier da at vi estimerer en parameter. Vi skal se senere at $\bar{x}$ kan brukes til estimering av $E(x)$, og vi sier da at $\bar{x}$ er en estimator av $E(x)$. Når vi så i et konkret tilfelle har beregnet verdien av $\bar{x}$, har vi skaffet oss et estimat av $E(x)$.

Denne forskjell mellom estimator og estimat er kanskje noe vanskelig å oppfatte, men den er nødvendig. Vi kan si at estimatoren forteller oss hvordan vi skal gå fram for å skaffe oss et estimat. Estimatoren er derfor oppskriften, mens estimatet er resultatet i et bestemt tilfelle.

Å finne gode estimatorer, og dermed gode estimater, har vist seg å være en vanskelig oppgave. Flere prinsipper eller metoder er i bruk, blant disse minste kvadraters metode. Som vi har sett foran, fører denne metoden til $\bar{x}$ som estimator av $E(x)$. Men dette betyr ikke uten videre at $\bar{x}$ er en god estimator.

B.9. Samvariasjon og regresjon.

Undersøkelser over samvariasjon mellom flere random variable er en meget viktig del av empirisk forskning. I en undersøkelse i granskog kan en således være interessert i høyden, diameteren, alderen osv. samtidig. Og en vil da ha observasjoner av flere random variable for hvert gjentak. I Tab.B.8. er gjengitt observasjonene av årringbredde (x_i) og prosent kvist (x_o) for et sampel på $n = 15$ grantrær.

Det kan være mange grunner til at en observerer flere random vari-

able samtidig. En av grunnene er at en er interessert i om det finnes en eller annen sammenheng, dvs. om det kan påvises at f.eks. en av dem til en viss grad varierer i takt med en eller flere av de andre.

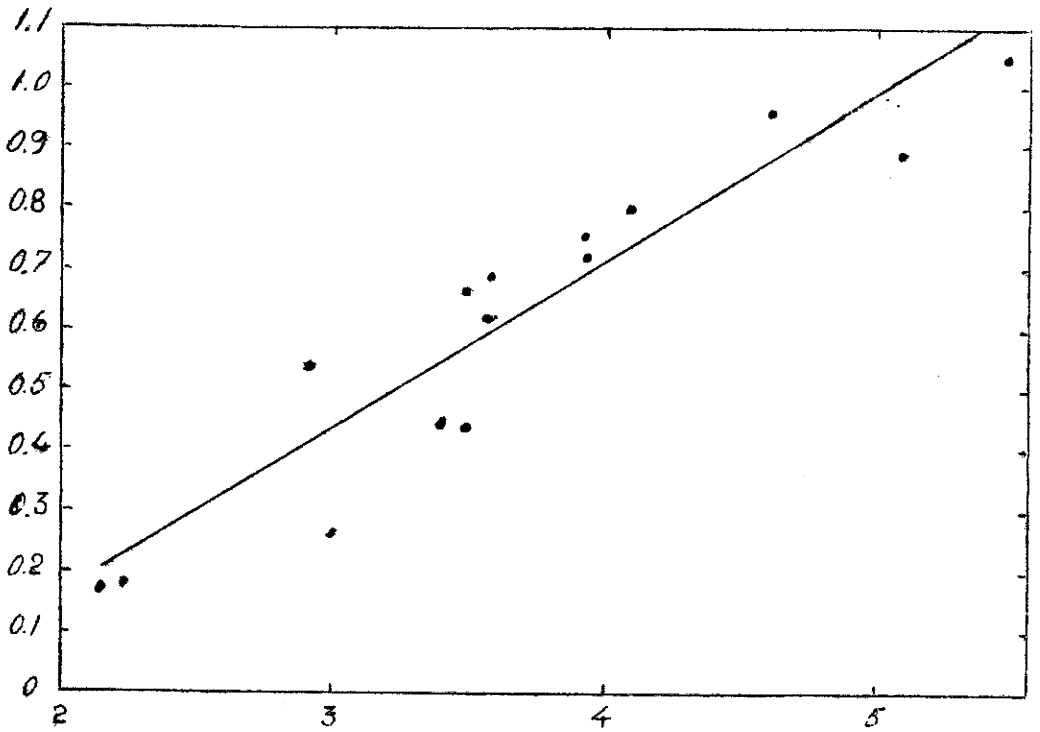
Tabell B.8.

x_1	x_o	x_1	x_o
2.3	0.19	3.9	0.74
3.0	0.26	3.6	0.69
2.2	0.18	3.6	0.60
2.9	0.53	3.5	0.67
3.5	0.44	4.6	0.96
3.4	0.45	5.6	1.05
3.9	0.71	5.1	0.88
4.1	0.79		
	Sum	55.2	9.14

Vi ser av Tab.B.8. at det synes å være en viss sammenheng mellom de to random variable slik at de ikke varierer helt fritt av hverandre. Vi ser nemlig at store x_o-verdier forekommer fortrinnsvis sammen med store x_1-verdier og små x_o-verdier fortrinnsvis sammen med små x_1-verdier. Det ser m.a.o. ut til at trær med stor årringbredde gjennomgående har større kvistmengde enn trær med liten årringbredde. Nå er jo samplet i dette tilfelle meget lite, bare n=15 trær er tatt med i undersøkelsen. Derfor kan den samvariasjon vi synes å kunne konstatere, tenkes å forsvinne dersom samplets størrelse blir økt. Og det er slett ikke sikkert at denne samvariasjonen eksisterer i det universet som samplet representerer i egenskap av et random sampel. Men hvis den regel vi finner for samplet, også gjelder for universet, sier vi at det er korrelasjon mellom to random variable. For eksemplet i Tab.B.8. vil det da bety positiv korrelasjon.

I andre tilfelle kan regelen være at store x_o-verdier fortrinnsvis forekommer sammen med små x_1-verdier og små x_o-verdier fortrinnsvis sammen med store x_1-verdier. Hvis denne regelen gjelder for universet, sier vi at det er negativ korrelasjon mellom de to random variable.

For å skaffe oss bedre oversikt kan vi fremstille observasjonene i et rettvinklet koordinatsystem. Til hvert observasjonspar, eller hvert gjentak, svarer et punkt som til abscisse har den observerte verdi av x_1 og til ordinat den observerte verdi av x_0. Til n observasjonspar eller gjentak svarer altså n punkter, det vi kan kalle en punktsverm. I Figur B.2. er inntegnet de n=15 punktene for eksemplet i Tab.B.8. Vi ser at den regelen vi kunne lese ut av tabellen, kommer tydelig fram i diagrammet.



Figur B.2.

I figuren er det også trukket opp en rett linje som skjærer gjennom punktsvermen. Hvordan denne er bestemt skal vi komme inn på senere.

Vi har allerede nevnt (avsnitt B.7.) at variasjonen i observasjonene av en random variabel og variasjonen i den random variable selv i regelen skyldes en rekke forskjellige årsaker. To eller flere random variabler vil derfor i mange tilfelle være avhengig av årsaker som i større eller mindre grad er felles for dem. La oss til eksempel tenke oss at x_1 er lengden av høyre og x_0 lengden av venstre lårben hos voksne menn. Det er vel da innlysende at de to random variable må være påvirket av et antall årsaker som virker noenlunde likt på de to lårbens vekst. Det er genetiske årsaker og det er ernæringsforholdene under oppveksten. En vil da også finne at lengden av høyre og lengden av venstre lårben hos voksne menn er positivt korrelerte. Punktdiagrammet for et sampel vil vise punkter som ligger ganske tett samlet omkring en rett linje.

Korrelasjon mellom to random variabler kan imidlertid også bero på at den ene variable er en av de årsaker som er bestemmende for den andre. I de fleste tilfelle kommer en imidlertid ikke lenger enn til en påvisning av at det er samvariasjon, mens oppklaringen av de årsaksforhold som betinger samvariasjonen må oppgis. Dette er kanskje mest alminnelig innen biologien hvor årsaksforholdene er særlig kompliserte. Men selv om en må stoppe opp med dette, er påvisning av samvariasjon av meget stor interesse.

Har vi nå for hvert gjentak observasjoner av to random variabler, er det nødvendig - som når det gjelder en random variabel - å skaffe seg noen få karakteristikker. Betegner vi observasjonene med x_{oi} og x_{1i} (i=1.2.3...n), har vi først de to gjennomsnittene $\bar{x}_o$ og $\bar{x}_1$. Dessuten har vi de to middelavvik s_o og s_1. I tillegg vil vi nå tenke oss at vi legger en kurve gjennom punktsvermen slik at punktene som representerer observasjonsparene (x_{oi}, x_{1i}) fordeler seg noenlunde

jevnt omkring den. Denne kurven som i det enkleste tilfelle er en rett linje slik som vist i Fig. B.2, vil vi så oppfatte som en representant for en funksjon av x_1.

Spørsmålet om hva denne funksjonen står for, må vi utsette til senere. Det samme gjelder spørsmålet om hvilken form for funksjonen vi bør velge. Tenker vi oss her at vi av en aller annen grunn kan gå ut fra at funksjonen er lineær, kan vi skrive den slik

$$r(x_1) = \beta_o + \beta_{o1}\, x_1$$

Den dobbelte fotskriften på β_{o1} betyr at vi oppfatter x_o som avhengig variabel og x_1 som uavhengig variabel el. forklaringsvariabel.

Ved hjelp av funksjonen kan vi lage en modell for x_o. Denne modellen blir :

$$x_{oi} = r(x_{1i}) + e_i = \beta_o + \beta_{o1}\, x_{1i} + e_i$$

dvs. at x_o er en sum av en lineær funksjon av x_1 og en random variabel e. Og vi ser da at e_i står for variasjonen omkring den rette linjen.

Funksjonen $r(x_1)$ går under navn av regresjonsfunksjonen, dvs. her regresjonsfunksjonen for x_o m.h.p. x_1. Vi skal senere komme inn på hva den står for og hva den brukes til. Her må vi nøye oss med å si at den gir den verdien av x_o vi venter å finne for en valt verdi av x_1. En presis beregning av x_o for en valt verdi av x_1 er imidlertid ikke mulig. På grunn av e_i vil vi måtte finne oss i en feil som er desto større jo større variasjonen i e_i er. Dessuten kjenner vi jo i regelen heller ikke verdiene av konstantene eller parametrene β_o og β_{o1} og må derfor i praksis bruke estimater av dem.

Den metoden en har slått inn på for estimering av konstantene er også her minste kvadraters metode. Betegnes estimatorene med $\hat{\beta}_o$ og $\hat{\beta}_{o1}$, skal en estimere parametrene slik at

$$S = \Sigma(x_{oi} - \hat{\beta}_o - \hat{\beta}_{o1}\, x_{1i})^2$$

blir minst mulig.

For å forenkle litt skal vi føre inn gjennomsnittet for x_1. Vi skriver da funksjonen slik

$$\hat{r}(x_1) = \hat{\alpha}_o + \hat{\beta}_{o1}(x_1 - \bar{x}_1)$$

og vi ser at sammenhengen mellom α_o, β_o og β_{o1} er

$$\hat{\alpha}_o = \hat{\beta}_o + \hat{\beta}_{o1}\bar{x}_1$$

Vi setter så

$$S = \Sigma[x_{oi} - \hat{\alpha}_o - \hat{\beta}_{o1}(x_{1i} - \bar{x}_1)]^2$$

Ved å sette de to partielle deriverte av S m.h.p. α_o og β_{o1} lik null, d.v.s.

$$\frac{\partial S}{\partial \hat{\alpha}_o} = \frac{\partial S}{\partial \beta_{o1}} = 0$$

finner vi at estimatorene av α_o og β_{o1} er :

$$\hat{\alpha}_o = \bar{x}_o$$

$$\hat{\beta}_{o1} = \frac{\Sigma(x_{1i} - \bar{x}_1)(x_{oi} - \bar{x}_o)}{\Sigma(x_{1i} - \bar{x}_1)^2}$$

Den siste ($\hat{\beta}_{o1}$) betegnes oftest med b_{o1}.

I formelen for estimatoren b_{o1} har vi i nevneren den vanlige kvadratsummen for x_1, og denne summen beregnes da som vist i avsnitt B.5. Telleren som kalles kovarianssummen, er summen av produktene av observasjonenes avvik fra gjennomsnittene. I praksis beregnes den slik:

$$\Sigma(x_{1i} - \bar{x}_1)(x_{oi} - \bar{x}_o) = \Sigma x_{1i}x_{oi} - \frac{1}{n}\Sigma x_{1i}\Sigma x_{oi}$$

For eksemplet i Tab.B.8. (Fig.B.2.) finner vi følgende summer

$$\Sigma x_{1i} = 55,2 \qquad \Sigma x_{oi} = 9,14$$
$$\Sigma x_{1i}^2 = 215,28 \qquad \Sigma x_{oi}^2 = 6,5724$$
$$\Sigma x_{1i} x_{oi} = 36,887$$

Den estimerte regresjonsfunksjonen* blir da

$$\hat{r}(x_1) = 0,6093 + 0,2678\ (x_1 - 3,68)$$
$$= 0,2678\ x_1 - 0,3762$$

Det er denne funksjonen som er gjengitt grafisk i Fig.B.2.

De to gjennomsnittene, de to middelavvik og regresjonsfunksjonen $\hat{r}(x_1)$ gir sammen en ganske god total karakteristikk av observasjonene. Men det er noe viktig som mangler. I to tilfelle kan disse karakteristikkene være noenlunde de samme. Samtidig kan imidlertid de punktene som representerer observasjonene (se Fig.B.3.) i det ene tilfelle ligge tett inn til funksjonen, i det andre tilfelle mere spredt. I tillegg til de karakteristikkene vi har nevnt, har vi derfor bruk for en størrelse som sier noe om hvor stor spredningen omkring regresjonsfunksjonen er. Den størrelsen som brukes til dette, går under navn av korrelasjonskoeffisienten.

La oss tenke oss at vi måler avstanden mellom hvert punkt i punktdiagrammet og den linjen (regresjonslinjen) som er det grafiske bilde av $\hat{r}(x_1)$. Avstandene vil vi da måle i en retning som er loddrett på x_1-aksen. Vi skal betegne disse avstandene med d_i, og har at

* Tegnet $\hat{}$ (hatt) over r betyr her at det er en estimert funksjon vi har med å gjøre. Denne måten å betegne en estimator på, kunne ha vært gjennomført for alle estimatorer. Som betegnelse på estimatoren av $\mu = E(x)$ kunne vi brukt f.eks. $\hat{\mu}$. Men betegnelsen $\bar{x}$ som vi bruker som estimator i dette tilfelle, er så innarbeidet nå at det ville skaffe leseren vansker om en bryter med denne betegnelsesmåten. Dette gjelder også mange andre estimatorer.

$$d_i = (x_{oi}-\bar{x}_o) - b_{o1}(x_{1i}-\bar{x}_1)$$

Siden $\Sigma(x_{oi}-\bar{x}_o) = \Sigma(x_{1i}-\bar{x}_1) = 0$, er også $\Sigma d_i = 0$ og $\bar{d} = 0$. Summen Σd_i^2 er derfor en kvadratsum av samme karakter som den som brukes til beregning av middelavviket for observasjonene av en random variabel. Denne summem er naturligvis en sum av positive tall som kan være lik null, og derfor er $\Sigma d_i^2 \geq 0$.

For denne summen finner vi så at

$$\Sigma d_i^2 = \Sigma[(x_{oi}-\bar{x}_o)-b_{o1}(x_{1i}-\bar{x}_1)]^2$$
$$= \Sigma(x_{oi}-\bar{x}_o)^2 + b_{o1}^2\Sigma(x_{1i}-\bar{x}_1)^2 - 2b_{o1}\Sigma(x_{1i}-\bar{x}_1)(x_{oi}-\bar{x}_o)$$

Siden $\Sigma(x_{1i}-\bar{x}_1)(x_{oi}-\bar{x}_o) = b_{o1}\Sigma(x_{1i}-\bar{x}_1)^2$

finner vi at

$$\Sigma d_i^2 = \Sigma(x_{oi}-\bar{x}_o)^2 - b_{o1}^2\,\Sigma(x_{1i}-\bar{x}_1)^2$$

og vi ser at

$$\Sigma d_i^2 \leq \Sigma(x_{oi}-\bar{x}_o)^2$$

Likhet mellom disse to kvadratsummene vil inntreffe når $b_{o1} = 0$, dvs. når regresjonslinjen er parallell med x_1-aksen.

Verdien av $\Sigma d_i^2/\Sigma(x_{oi}-\bar{x}_o)^2$ må derfor ligge mellom 0 og 1. Det samme må være tilfelle med størrelsen

$$r_{o1}^2 = 1 - \Sigma d_i^2/\Sigma(x_{oi}-\bar{x}_o)^2$$

Vi ser at $r^2 = 0$ når $b_{o1} = 0$, dvs. når regresjonslinjen er parallell med x_1-aksen. Og vi ser at $r^2 = 1$ når $\Sigma d_i^2 = 0$, dvs. når alle punktene i punktsvermen ligger på regresjonslinjen eller, når det ikke finnes noen spredning av dem omkring regresjonslinjen.

Det kan derfor være naturlig å bruke r_{o1} som uttrykk for styrken av samvariasjonen mellom x_o og x_1. Vi ser at når $\Sigma d_i^2 = \Sigma(x_{oi}-\bar{x}_o)^2$ som svarer til at x_o varierer fritt og uavhengig av

x_1, er $r_{o1} = 0$. Er $\Sigma d_i^2 = 0$, som svarer til at x_o er helt bundet til x_1, er $r_{o1} = 1$. Lar vi så r_{o1} ha samme fortegn som b_{o1}, vil r_{o1} også vise om samvariasjonen er positiv eller negativ. Det er denne r som har fått betegnelsen korrelasjonskoeffisienten.

En må legge merke til, og ikke glemme det når en i et aktuelt tilfelle ønsker å bruke r som uttrykk for styrken av samvariasjonen, at det er forutsatt at regresjonsfunksjonen er lineær. Den er ikke tilfredsstillende uttrykk for graden eller styrken av samvariasjonen i tilfelle hvor regresjonen ikke er lineær. Vi skal imidlertid senere vise at vi etter samme prinsipp kan danne en annen størrelse som kan brukes i slike tilfelle.

Vi bør også legge merke til at på samme måte som $\bar{x}$ er en estimator av $E(x)$ og b_{o1} en estimator av β_{o1} er r_{o1} en estimator av den korrelasjonskoeffisienten vi ville ha funnet dersom samplet av gjentak hadde omfattet hele universet. Bruker vi betegnelsen ρ_{o1} på den siste koeffisienten, er derfor r_{o1} å oppfatte som en estimator av ρ_{o1}. Hva dette betyr for bruken av r_{o1} skal vi komme tilbake til.

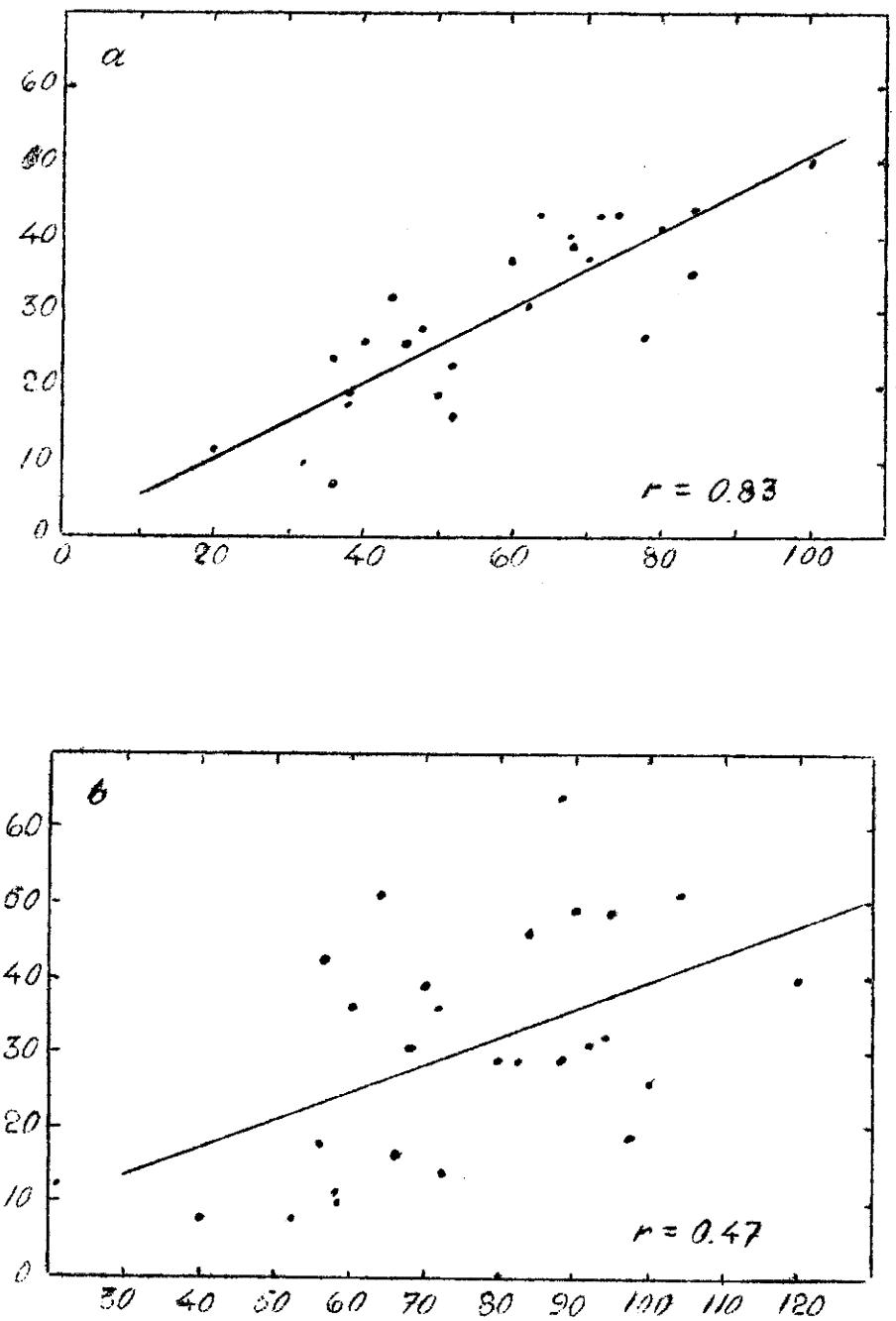
Korrelasjonskoeffisienten er en meget brukt størrelse. I litteraturen finner vi derfor forskjellige uttrykk for den. Hvis vi i formelen for r_{o1}^2 setter inn formelen for Σd_i^2 og for b_{o1}, vil vi lett kunne utlede at

$$r_{o1} \equiv \frac{\Sigma(x_{1i}-\bar{x}_1)(x_{oi}-\bar{x}_o)}{\sqrt{\Sigma(x_{1i}-\bar{x}_1)^2\Sigma(x_{oi}-\bar{x}_o)^2}}$$

Dette er det vanligste uttrykket for r_{o1}. Ved å bruke denne formelen vil en finne at

$$r_{o1} = b_{o1}\frac{s_1}{s_o}$$

I praksis er det naturligvis likegyldig hvilken av disse formlene en benytter seg av.


a
60
50
40
30
20
10
0
0
20
40
60
80
100
r = 0.83
b
60
50
40
30
20
10
0
30
40
50
60
70
80
90
100
110
120
r = 0.47

Figur B.3.

Vi har foran brukt x_1 som betegnelse på den variabel vi oppfatter som den uavhengig variable. Hvilken dette er må det være tatt standpunkt til før en går i gang med beregningene, det er noe som bør være med i planen for den undersøkelsen det gjelder. Vi kan imidlertid bytte om rollene og betrakte den andre, dvs. etter vårt opplegg x_o, som den uavhengig variable i regresjonsfunksjonen. Vi vil da ha en regresjonsfunksjon for x_1 m.h.p. x_o som i det lineære tilfelle skrives slik

$$r(x_o) = \beta_1 + \beta_{1o}\, x_o$$

Estimatoren av β_{1o} betegnes med b_{1o} og uttrykket for den får vi ved i uttrykket for b_{o1} å bytte om fotskriftene 0 og 1. Vi vil da finne at $r_{1o} = r_{o1} = r$, og at

$$r_{1o} = b_{1o} \frac{s_o}{s_1}$$

dvs. at

$$r_{o1} \cdot r_{1o} = r^2 = b_{o1} \cdot b_{1o}$$

eller at

$$r = \pm \sqrt{b_{o1} b_{1o}}$$

De to regresjonskoeffisientene b_{o1} og b_{1o} har samme fortegn, nemlig samme fortegn som summen $\Sigma(x_{1i}-\bar{x}_1)(x_{oi}-\bar{x}_o)$. Korrelasjonskoeffisienten har også samme fortegn som denne summen, altså samme fortegn som de to regresjonskoeffisientene. Til positiv korrelasjons eller samvariasjon svarer m.a.o. en positiv korrelasjonskoeffisient, til negativ samvariasjon en negativ korrelasjonskoeffisient.

B.10. Foreløpig om estimat og informasjon.

Til en foreløpig orientering skal vi nå ta for oss spørsmålet om hvilken informasjon estimatet av en parameter kan gi oss om parameteren, f.eks. hva gjennomsnittet $\bar{x}$ kan fortelle oss om forvent-

ningen E(x). Vi vil da tenke oss at vi har et sampel på n gjentak og n observasjoner av den random variable x. Beregner vi så gjennomsnittet, dvs. at vi finner verdien av estimatoren $\bar{x}$, har vi skaffet oss et estimat av E(x). Vi har nevnt foran at denne verdien må oppfattes som en tilnærmet riktig verdi av E(x). Men spørsmålet er hva som ligger i dette at estimatet gir en tilnærmet riktig verdi.

Vi skal senere forklare hvordan en, når en har verdien av gjennomsnittet og verdien av middelavviket (s), kan avgrense et tallområde eller intervall omkring $\bar{x}$ og så med god grunn påstå at dette intervallet inneholder verdien av E(x). For eksemplet i Tab.B.1. har vi at $\bar{x}$ = 20,14 og s = 3,09, og ved hjelp av disse tallene kan vi da finne at grensene for intervallet er 18,96 og 21,32. Vi har altså her et intervall med en bredde på 21,32-18,96 = 2,36. For eksemplet i avsnitt A.4. har vi bare n = 3 observasjoner (de tre differensene) og for disse har vi at $\bar{x}$ = 24,69 og s = 3,79. Og vi vil da finne at intervallet strekker seg fra 15,25 til 34,09. Intervallet har derfor en bredde på 18,84.

Dette intervallet kalles et $k_n$fidensintervall og grensene for det $k_n$fidensgrensene. Vi skal senere forklare hvordan vi kan beregne disse grensene og hvorfor vi kan si at vi har god grunn til å påstå at intervallet inneholder verdien av E(x).

Tar vi for oss igjen eksemplet fra avsnitt A.4., ser vi at konfidensintervallet har en betydelig bredde (18,84). Dette betyr naturligvis at den informasjon vi har om E(x) er lite tilfredsstillende. At den er lite tilfredsstillende vil vi lett kunne innse ved å tenke oss at vi skal bruke informasjonen til et eller annet, f.eks. som grunnlag for en økonomisk kalkulasjon. Vi skal vise senere at bredden av konfidensintervallet for E(x) er omvendt propor-

sjonal med kvadratroten av antall gjentak (n). Dette betyr naturligvis at størrelsen av samplet har meget å si. At informasjonen om E(x) i vårt eksempel er så utilfredsstillende kommer rimeligvis av at samplet ikke inneholder mer enn n = 3 gjentak.

På tilsvarende måte kan vi beregne konfidensgrenser også for slike parametrer som σ og β_{01} . Tar vi for oss σ_1 for x_1 i eksemplet i Tab.B.8. hvor n = 15, har vi at s_1 = 0,93 og vi vil da finne at konfidensgrensene for σ_1 er 0,69 og 1,47. For samme eksempel er konfidensgrensene for regresjonskoeffisienten β_{01} lik 0,21 og 0,33.

C. Sannsynlighetsregning.

C.1. Matematisk sannsynlighet.

I de foregående avsnitt har vi beskrevet noen metoder vi har bruk for når vi skal ordne og gi en kort karakteristikk av våre observasjoner. Ved hjelp av disse metodene kan vi skaffe oss kunnskap om det sampel av gjentak vi har hentet observasjonene fra. Vi har imidlertid også forklart at samplet må oppfattes som en representant for et univers av gjentak og at det er universet vi ønsker å vite noe om. Viten eller kunnskap om dette universet skaffer vi oss så ved å generalisere. Vi induserer. Som vi nå etter hvert skal få se, har vi da bruk for sannsynlighetsregning.

Sannsynlighetsregningen bygger på begrepet matematisk sannsynlighet, og dette har igjen sammenheng med begrepet univers eller populasjon. Når vi derfor nå skal prøve å introdusere begrepet matematisk sannsynlighet, er det enklest å begynne med universet.

I avsnitt A.3. nevnte vi at et univers består av gjentak, et abstrakt univers av et ubegrenset antall gjentak. La oss nå ta for oss et eksperiment av den enklest tenkelige sort, f.eks. et som går ut på at vi gjør et kast med en vanlig terning. Et enkelt kast . . er da å oppfatte som et gjentak. Mange har vansker med å godta at en bruker betegnelsen gjentak på et enkelt kast fordi, sier de, betegnelsen kan ikke være riktig med mindre det er minst to kast, slik at kast nr.2 er et gjentak av kast nr.1. Når vi likevel sier at et enkelt kast er et gjentak, så er det i den forstand at det er ett av en mengde kast.

Både i dette enkle eksemplet og ellers når det gjelder eksperimenter, er gjentaket en handling, dvs. noe vi foretar oss. Det må bare føyes til at det er noe vi foretar oss etter en bestemt opp-

skrift eller beskrivelse. Vi kan f.eks. oppfatte det å lage risgrynsgrøt som et eksperiment. Dette er jo en handling utført etter en bestemt oppskrift. Slår vi opp i en kokebok, vil vi finne to oppskrifter: risgrynsgrøt av vann og risgrynsgrøt av melk. Grøt av vann og grøt av melk er ikke gjentak i samme univers fordi oppskriften ikke er den samme. For ikke å havne i rotet tankegang må vi holde **strengt fast** på det krav at for at to eksperimenter skal være gjentak i samme univers, er det nødvendig at oppskriften eller beskrivelsen er den samme og at den blir fulgt.

La oss nå fremdeles tenke oss at gjentaket er ét kast med en terning. Ser vi bort fra de muligheter at terningen kan bli stående på en kant eller et hjørne, er det seks mulige utfall. Vi kan også si at gjentaket har seks mulige og alternative kjennetegn. E = "seks" er et av disse. Gjentar vi så kastet n ganger, f.eks. n = 25 ganger, vet vi antallet av gjentak med kjennetegn E. Dette antallet vil vi betegne med z. Det relative antall gjentak med E er da z/n.

I praksis er det naturligvis en grense for hvor mange gjentak vi kan skaffe oss. Men det er ikke vanskelig å forestille seg at en mengde blir større og større. Kan vi forestille oss en mengde gjentak på f.eks. 100, er det ikke særlig vanskelig å forestille seg en mengde pa 1000, ti tusen, hundre tusen o.s.v.

I dette universet vil noen av gjentakene ha kjennetegnet E, f.eks. E = "seks". Med sannsynligheten (P) for E forstår vi den relative frekvens for E i universet. Sannsynligheten har derfor en verdi mellom 0 og 1, grensene regnet med. Er E et kjennetegn som er felles for alle gjentak, dvs. at det nødvendigvis må inntreffe i et gjentak, må det ha sannsynligheten P = 1. Er det et kjennetegn som ingen av gjentakene kan ha, dvs. at det umulig kan

inntreffe i et gjentak, må det ha sannsynligheten P = 0. I et sampel på f.eks. n = 100 gjentak, vil et antall (z) gjentak ha E til kjennetegn. Verdien av z er da naturligvis 0, 1, 2, ...n. Den relative frekvens z/n er estimatoren av sannsynligheten P. At z = 0 betyr derfor ikke at P = 0, og at z = n betyr ikke at P = 1. Er derimot P = 1, må z være lik 0, og er P =1 må z = n.

I noen tilfelle som vi skal komme inn på senere, er universet en konkret mengde gjentak. I en viss sammenheng består universet av f.eks. de gardsbruk (eventuelt de som har mer enn 5 dekar dyrket jord) som finnes i Akershus fylke. Men i regelen er universet en abstraksjon, og det blir da som oftest oppfattet som ubegrenset.

Det at en betrakter universet som ubegrenset, må imidlertid antas til en viss grad å være diktert av bekvemmelighetshensyn. Det er ikke vanskelig å forestille seg antall terningkast økt over alle grenser, men i andre tilfelle kan nok mengden av mulige gjentak være begrenset. En kan imidlertid også da gå ut fra at antallet er meget stort, og det er av liten betydning om det er millioner eller hundre millioner det er tale om. Det som har betydning for utviklingen av sannsynlighetsregningen og bruken av den, er at universet omfatter et så stort antall gjentak at det ikke blir merkbart forandret av at det blir redusert med et sampel av vanlig størrelse.

Vi har sagt foran at E er et kjennetegn som noen av gjentakene i et univers har, uten å si noe om hva det er for univers vi tenker på. Det er imidlertid klart at det samme kjennetegnet kan være kjennetegn på gjentak i forskjellige universer. Kjennetegnet E = "gutt" er jo et kjennetegn som noen av gjentakene i universet av nyfødte barn har. Men det er også et kjennetegn som noen av gjen-

takene i universet av 10 år gamle barn har. Undersøkelser har vist at sannsynligheten for E = "gutt" ikke er den samme i disse to universene. Det er derfor ikke tilstrekkelig å oppgi et tall for sannsynligheten for et kjennetegn. Vi må alltid føye til hva for univers det er tale om.

Universet kan defineres ved at en nevner opp kjennetegn som er felles for gjentakene. Med universet av nyfødte barn må vi derfor, hvis ikke noe annet er sagt, forstå et som er representert ved alle nyfødte barn i verden. Føyer vi til at foreldrene er norske statsborgere, får vi et annet univers. Andre eksempler på ulike universer er: "norsk statsborger, alder 50 år", "mannlig norsk statsborger, alder 50 år" og "mannlig norsk statsborger, alder 50 år, murer". Vi må naturligvis også være oppmerksom på at gjentakene i f.eks. universet definert ved "mannlig norsk statsborger, alder 50 år" som oftest ikke består bare av de mannlige norske statsborgere, alder 50 år, som lever på et bestemt tidspunkt. Disse er bare et sampel fra det universet vi tenker på.

Det er kanskje lettest å innse hvor viktig det er at vi har klart for oss hva for univers det er tale om og at det er tilstrekkelig nøye definert, når vi tenker oss at vi skal estimere sannsynligheten for et kjennetegn. Ønsker vi f.eks. å estimere sannsynligheten for E = "gutt" i universet av nyfødte barn hvis foreldre er norske statsborgere, vil vi ikke i samplet ta med nyfødte barn hvis foreldre er indonesiske statsborgere. Og ønsker vi å estimere dødssannsynligheten i universet av "50-årige mannlige norske statsborgere med murer som yrkestittel", vil vi i samplet ikke ta med kvinner, ikke britiske statsborgere og ikke snekkere. Definisjonen av universet tillater nemlig ikke det.

Estimeringen av sannsynligheten for et kjennetegn forutsetter imidlertid ikke bare at universet er klart nok definert, slik at en kan vite hva slags gjentak en skal ta med i samplet. Estimeringen forutsetter også at samplet er et random sampel eller kan brukes som et random sampel. Spørsmålet blir da hva vi skal forstå med et random sampel.

Vi har vært inne på dette spørsmålet i avsnitt A.3. og forklarte da at gjentakene i et random sampel må være valt ved hjelp av en eller annen teknikk for loddtrekning. Men vi må nå komme noe nærmere inn på saken.

La oss først tenke oss at universet består av et endelig antall gjentak og at antallet er så lite at vi i praksis kan sette merkelapp på hvert gjentak, merkelapp med f.eks. nummer. Hvis vi da under uttaking av et sampel på n gjentak bruker et apparat for loddtrekning som er slik konstruert at alle gjentak i universet stilles likt, vil samplet bli et random sampel. Etter at vi nå har innført begrepet matematisk sannsynlighet, kan vi si at vi med å stille alle gjentak likt mener at sannsynligheten for et bestemt gjentak å komme med i samplet skal være den samme for alle gjentak. Består universet av N gjentak, nummerert fra 1 til N, skal sannsynligheten for at f.eks. gjentak nr.25 skal komme med i samplet være lik 1/N.

Er antall gjentak i universet meget stort, er det f.eks. et ubegrenset antall gjentak, kan vi bare forestille oss at en slik teknikk blir benyttet. Den skisserte fremgangsmåten forteller oss da hva vi skal forstå med et random sampel. Det er bare i de få tilfelle hvor gjentakene i universet er kjent eller registrert og antallet er beskjedent, slik at vi kan sette merkelapp på hvert

gjentak, at vi kan skaffe oss et random sampel ved loddtrekning. Fremgangsmåten kan brukes til å ta ut random sampler for de såkalte representative jordbrukstellingene fordi universet da består av f.eks. alle gardsbrukene i Akershus.

All bruk av sannsynlighetsregningen og statistisk metodikk forutsetter som vi skal se, at det eller de sampler av gjentak vi gjør bruk av er random sampler. Hvis da gjentakene ikke kan tas ut ved hjelp av en eller annen random prosess, blir vi stilt overfor spørsmålet om hva vi så skal forstå med universet. De nyfødte barn, født i 1970, hvis foreldre er norske statsborgere, er naturligvis også et sampel som representerer et abstrakt univers. Men hva skal vi så forstå med universet? Det kan ikke være noe annet enn det universet som samplet representerer i egenskap av et random sampel. Vi er m.a.o. nødt til, kan vi si, å operere med et univers som er en ren tankekonstruksjon og så å betrakte det samplet vi har, som et random sampel tatt fra dette universet. Dette er imidlertid også tilstrekkelig. Ved hjelp av samplet av nyfødte barn, født i 1970, hvis foreldre er norske statsborgere, kan vi estimere sannsynligheten for f.eks. E = "blå øyne". Estimatoren er det relative antall blåøyde barn i samplet. Har vi da samtidig et sampel av nyfødte barn, født i 1970, hvis foreldre er italienske statsborgere, kan vi på samme måte forestille oss et annet univers som altså er det universet dette samplet representerer i egenskap av et random sampel. Vi kan så på samme måte estimere sannsynligheten for E = "blå øyne" i dette universet. Og et av våre problemer blir å undersøke om det kan sies at sannsynligheten for E er forskjellig i disse to universene.

C.2. Deluniverser eller subuniverser.

Som vi nå har sett, er sannsynligheten for et kjennetegn knyttet til et bestemt univers. La oss betegne det med U. Gjentakene i dette universet har visse felles kjennetegn, og det er da disse som gir oss beskrivelsen av universet. For at dette skal komme med i den betegnelsen vi bruker, skal vi for sannsynligheten for E i universet U bruke symbolet P(E;U).

Et univers kan alltid deles opp i mindre omfattende universer, deluniverser eller subuniverser, ved hjelp av kjennetegn som ikke er felles for alle gjentak i universet. Universet av nyfødte barn, født i 1970, hvis foreldre er norske statsborgere, kan f.eks. deles i to universer ved hjelp av kjennetegnene "jente" og "gutt". Setter vi E_1 = "jente" og/fordi det er bare ett alternativ til dette, kan vi sette "gutt" = ikke-E_1 eller kort iE_1. Betegner vi så det uoppdelte universet med U, kan vi betegne de to deluniversene med UE_1 og U iE_1. Sett at vi så er interesserte i sannsynligheten for E_2 = "blå øyne". Vi har da tre sannsynligheter for E_2, nemlig sannsynligheten i det uoppdelte univers og sannsynligheten i hvert av de to deluniversene, altså sannsynlighetene $P(E_2;U)$, $P(E_2;UE_1)$ og $P(E_2;Ui E_1)$. De to siste er det en kaller betingede sannsynligheter. $P(E_2;UE_1)$ er sannsynligheten for E_2 betinget av E_1. Dette er altså det samme som sannsynligheten for E_2 i det delunivers hvor alle gjentak har kjennetegnet E_1.

Hvis det ikke betyr noe for sannsynligheten for E_2 om gjentakene har E_1 eller iE_1, dvs. hvis $P(E_2;UE_1) = P(E_2;Ui E_1)$, sier vi at de to kjennetegnene E_1 og E_2 opptrer uavhengig av hverandre eller at de er uavhengige kjennetegn. Vi skal se senere at det kan ha stor interesse å undersøke om to kjennetegn er avhengige av hverandre. Foreløpig skal vi ta for oss et eksempel hvor antall gjen-

tak (n) er så stort at det relative antall gjentak med et kjennetegn er en nokså nær riktig verdi for sannsynligheten for kjennetegnet.

Ved en bestemt krysning av bananfluer finner en hos avkommet E_1 = "normale børster" og E_2 ="normale øyne". De to alternative kjennetegn er iE_1 = "reduserte børster" og iE_2 = "reduserte øyne". Ved en bestemt krysning forstår vi en krysning mellom en hun og en han som har nærmere spesifiserte kjennetegn, og ved et gjentak forstår vi et avkom.

I et krysningsforsøk fikk en n = 2835 avkom og blant disse fant en

1705 med E_1 og E_2 (normale børster og normale øyne)
506 med E_1 og iE_2 (normale børster og reduserte øyne)
489 med iE_1 og E_2 (reduserte børster og normale øyne)
135 med iE_1 og iE_2 (reduserte børster og reduserte øyne)

Betegnes estimatoren av en sannsynlighet P med $\hat{P}$ finner vi at

$$\hat{P}(E_2;UE_1) = \frac{1705}{1705 + 506} = 0,772$$

$$\hat{P}(E_2;UiE_1) = \frac{489}{489 + 135} = 0,784$$

Vi ser at det er liten forskjell mellom de to estimatene, noe som tyder på at de to kjennetegn er uavhengige. Vi skal vise senere at det iallfall ikke kan påvises at de er avhengige av hverandre.

C.3. Enten-eller setningen og både-og setningen.

La oss tenke oss at et gjentak kan ha ett av m alternative kjennetegn E_1, E_2, E_3, ...E_m. Vi forutsetter altså at disse kjennetegn utelukker hverandre, dvs. at et gjentak ikke kan ha to eller flere av dem. Vi vil også forutsette at det ikke finnes andre

alternativer, slik at et gjentak nødvendigvis må ha ett av dem. Gjentaket nyfødt barn kan ha ett av kjennetegnene E_1 = "jente" og E_2 = "gutt" og må nødvendigvis ha ett av dem. Gjentaket "tvilling" kan ha ett av kjennetegnene E_1 = "2 jenter", E_2 ="2 gutter" og E_3 = "2 barn av ulike kjønn" og må ha ett av dem.

La oss tenke oss at det i et random sampel på n gjentak er z_1 gjentak med E_1, z_2 gjentak med E_2, ... og z_m gjentak med E_m. Forutsetter vi at det ikke finnes andre alternative kjennetegn enn disse m og at de utelukker hverandre, er det klart at $z_1 + z_2 + z_3 + \ldots + z_m = n$. Følgelig er summen av de relative frekvensene lik enheten:

$$\frac{z_1}{n} + \frac{z_2}{n} + \ldots + \frac{z_m}{n} = 1$$

Denne ligningen må imidlertid gjelde også om vi har med alle gjentakene i universet, men da må vi erstatte de relative frekvensene med sannsynligheten for E_1, E_2, ... E_m, altså med $P(E_1;U)$, $P(E_2;U)$, ... $P(E_m;U)$. Følgelig har vi at

$$P(E_1;U) + P(E_2;U) + \ldots + P(E_m;U) = 1$$

Forutsetningen er da at E_1, E_2, ...E_m er kjennetegn som utelukker hverandre og at det ikke finnes andre alternative kjennetegn.

Vi skal så videre tenke oss at g ($g < m$) av de m alternative kjennetegn, f.eks. de g første E_1, E_2, ... E_g, kan innordnes under en felles betegnelse E. Som eksempler kan nevnes at E_1 = "2 jenter" og E_2 = "2 gutter" kan innordnes under kjennetegnet E = "2 barn av samme kjønn", og at E_1 = "spar 2", E_2 = "spar 3" ..... E_{13} = "spar ess" kan innordnes under kjennetegnet E = "spar".

Enten-eller setningen går da ut på at sannsynligheten for E er lik summen av sannsynlighetene for E_1, E_2, ...E_g, eller

$$P(E;U) = P(E_1;U) + P(E_2;U) + ... + P(E_g;U)$$

Forutsetningen er som før at kjennetegnene utelukker hverandre i et gjentak i universet U.

Setningen følger umiddelbart av at hvis det i et sampel på n gjentak er z_1 med E_1, z_2 med E_2, ... og z_g gjentak med E_g, er antall gjentak med E - som jo er enten E_1 eller E_2, eller .... eller E_g - være lik $z = z_1 + z_2 + ... + z_g$. Den relative frekvens for E er derfor lik

$$\frac{z}{n} = \frac{z_1 + z_2 + ... + z_g}{n} = \frac{z_1}{n} + \frac{z_2}{n} + ... + \frac{z_g}{n}$$

en ligning som også må gjelde om vi har med alle gjentak i universet.

La oss anta at vi vet at sannsynligheten for E_1 = "2 jenter" og sannsynligheten for E_2 = "2 gutter" i et univers av tvillingfødsler er 0,31 og 0,34. Da er sannsynligheten for E = "2 barn av samme kjønn" lik $P(E;U) = 0,31 + 0,34 = 0,65$.

Kan vi innordne g av m kjennetegn som utelukker hverandre under kjennetegnet E, kan vi innordne resten av dem under kjennetegnet ikke-E eller iE. Det følger da uten videre at

$$P(E;U) + P(iE;U) = 1$$

Kjennetegnene E og iE kalles motsatte. Summen av sannsynligheten for to motsatte kjennetegn er altså lik enheten.

Kjennetegnene "2 barn av samme kjønn" og "2 barn av ulike kjønn" er motsatte ved en tvillingfødsel. Følgelig er

P(2 barn av ulike kjønn;U) = 1-P(2 barn av samme kjønn;U)

Er sannsynligheten for "2 barn av samme kjønn" lik 0,65, er sannsynligheten for "2 barn av ulike kjønn) lik 1-0,65 = 0,35.

La oss tenke oss at E_1 og E_2 ikke utelukker hverandre, dvs. at noen av gjentakene kan ha begge kjennetegn. Et barn kan ha både kjennetegnet E_1 = "gutt" og E_2 = "blå øyne". Har vi et sampel på n gjentak, kan vi dele det i to sampler slik at det ene omfatter alle gjentak med E_1, det andre alle gjentak med iE_1. Både E_1 og E_2 kan da naturligvis forekomme bare i det første samplet, og la oss si at antallet av gjentak med E_2 i dette samplet er z_2'. Dette antallet er da antallet av gjentak i hele samplet som har både E_1 og E_2, eller med en kort betegnelse, det sammensatte kjennetegn E_1E_2. Det relative antall gjentak med dette sammensatte kjennetegn er da z_2'/n. Sett nå at det første delsamplet hvor alle gjentak har E_1, består av z_1 gjentak. Da har vi at

$$\frac{z_2'}{n} = \frac{z_1}{n} \cdot \frac{z_2'}{z_1}$$

Her er da z_1/n det relative antall gjentak i hele samplet som har kjennetegnet E_1. Og z_2'/z_1 er det relative antall gjentak i delsamplet på z_1 gjentak som har kjennetegnet E_2. Disse to relative antall er estimatorer av sannsynlighetene $P(E_1;U)$ og $P(E_2;UE_1)$.

Denne operasjonen som går ut på deling av samplet i to delsampler, kan vi lett tenke oss gjennomført også om samplet omfatter hele universet. Resultatet er et delunivers hvor alle gjentak har E_1 og et annet delunivers hvor alle gjentak har iE_1. Og vi vil da komme fram til samme resultat, men nå som uttrykk for sannsynligheten for det sammensatte kjennetegn E_1E_2. Vi vil finne at sannsynligheten for både E_1 og E_2, dvs. $P(E_1E_2;U)$ er lik sannsynligheten for E_1 i hele universet multiplisert med sannsynligheten for E_2 i det deluniverset hvor alle gjentak har E_1, dvs. at

$$P(E_1E_2;U) = P(E_1;U) \cdot P(E_2;UE_1)$$

Er E_1 og E_2 uavhengige kjennetegn, er $P(E_2;UE_1) = P(E_2;U)$ og derfor

$$P(E_1E_2;U) = P(E_1;U) \cdot P(E_2;U)$$

Antar vi f.eks. at E_1 = "normale børster" og E_2 = "normale øyne" i vårt eksempel i avnsitt C.2. er uavhengige og at hvert har sannsynligheten 0,75, er sannsynligheten et gjentak med begge kjennetegn lik $P(E_1E_2;U) = 0,75^2 = 0,5625$. Som vi skal se senere, stemmer dette ikke med de data vi har. Vi ser da også at estimatet av denne sannsynligheten er 1705/2835 = 0,6014.

Både-og setningen kan utvides til å omfatte flere enn to kjennetegn. For tre kjennetegn har vi at

$$P(E_1E_2E_3;U) = P(E_1;U) \cdot P(E_2;UE_1) \cdot P(E_3;UE_1E_2)$$

Her er $P(E_3;UE_1E_2)$ sannsynligheten for E_3 i deluniverset UE_1E_2, dvs. sannsynligheten for E_3 i den del av universet U hvor alle gjentak har både E_1 og E_2. $P(E_3;UE_1E_2)$ er også sannsynligheten for E_3 betinget av E_1 og E_2.

Er E_2 uavhengig av E_1 og E_3 uavhengig av det sammensatte kjennetegn E_1E_2, har vi at

$$P(E_1E_2E_3;U) = P(E_1;U) \cdot P(E_2;U) \cdot P(E_3;U)$$

dvs. at sannsynligheten for at et gjentak skal ha alle tre kjennetegn er lik produktet av sannsynlighetene for hvert enkelt av dem.

Vi kan nå også finne sannsynligheten for et kjennetegn E = "enten E_1 eller E_2 eller både E_1 og E_2", dvs. sannsynligheten for i det minste ett av kjennetegnene E_1 og E_2. Hvis de to kjennetegn ikke utelukker hverandre, er det i alt fire mulige sammensatte kjennetegn, nemlig

$$E_1E_2,\ E_1iE_2,\ iE_1E_2 \text{ og } iE_1iE_2$$

Disse fire sammensatte kjennetegn utelukker hverandre. Svaret på

vårt spørsmål blir derfor etter enten-eller setningen:

$$P(E;U) = P(E_1E_2;U) + P(E_1iE_2;U) + P(iE_1E_2;U)$$

som kan vises å være lik

$$P(E;U) = P(E_1;U) + P(E_2;U) - P(E_1E_2;U)$$

Hvis E_1 og E_2 utelukker hverandre, kan de ikke forekomme sammen. Da er naturligvis $P(E_1E_2;U) = 0$, og $P(E;U)$ blir redusert til den enkle enten-eller setningen.

C.4. Binomialfunksjonen.

La oss tenke oss at E er et felles kjennetegn for noen av gjentakene i et univers U. For sannsynligheten for E skal vi bruke betegnelsen $P(E;U) = p$. Sannsynligheten for det motsatte kjennetegn er da $P(iE;U) = 1-p = q$.

Det vi er interesserte i er antall gjentak med E i et random sampel på n gjentak. La oss betegne dette antall med z. Da er n-z antall gjentak med iE. De verdier z kan ha er $z = 0,1,2,\ldots n$, og vi er interesserte i sannsynligheten for hver av disse verdier.

Forutsettes det at universet er ubegrenset eller i hvert fall så stort at uttaket av et sampel ikke medfører noen merkbar forandring i det, kan en vise at sannsynligheten for z gjentak med E og n-z gjentak med iE er

$$P_z = \binom{n}{z} p^z q^{n-z} \qquad (z = 0,1,2,\ldots n)$$

Her er

$$\binom{n}{z} = \frac{n!}{z!\,(n-z)!}$$

hvor $n! = 1.2.3.\ldots.n$, $z! = 1.2.3\ldots z$ og $(n-z)! = 1.2.3\ldots(n-z)$.

La oss anta at sannsynligheten for E = "gutt" i et univers av nyfødte barn er p = 0,52. Sannsynligheten for iE = "jente" er da q = 1-p = 0,48. Da er sannsynligheten for z gutter og 4-z jenter i et random sampel på n = 4 lik

$$P_z = \binom{4}{z}\ 0{,}52^z\ 0{,}48^{4-z}$$

I denne formelen kan vi nå sette z = 0,1,2,3 og 4 og regne ut og finner da sannsynlighetene for 0 gutter, 1 gutt, 2 gutter, 3 gutter og 4 gutter. Utregningen er vist i Tab.C.1.

Tabell C.1.

z	$\binom{4}{z}$	$0{,}52^z$	$0{,}48^{4-z}$	P_z
0	1	1,0000	0,0531	0,0531
1	4	0,5200	0,1106	0,2300
2	6	0,2704	0,2304	0,3738
3	4	0,1406	0,4800	0,2700
4	1	0,0731	1,0000	0,0731
				1,0000

Vi ser at $\Sigma P_z = 1$. Dette må være riktig for alle verdier av n og p. Etter Newtons binomialformel er nemlig

$$\Sigma P_z = \Sigma\binom{n}{z}\ p^z\ q^{n-z} = (p+q)^n$$

og her er jo $(p+q) = 1$.

At summen er lik enheten stemmer også med det vi gjennomgikk i foregående avsnitt. Vi kan nemlig sette $E_1 = (z=0)$, $E_2 = (z=1)$ .....$E_5 = (z=4)$. Vi ser at $E_1 \ldots E_5$ utelukker hverandre og at det ikke finnes andre alternative muligheter. Følgelig må summen av sannsynlighetene for $E_1 \ldots E_5$ være lik enheten.

Sannsynlighetene P_z kan også beregnes ved hjelp av den såkalte differensligningen. Det kan lett utledes at

$$P_{z+1} = \frac{p(n-z)}{q(z+1)}\ P_z$$

En kan da først beregne $P_0 = q^n$ og så trinnvis P_1, $P_2 \ldots P_n$ ved hjelp av ligningen, slik som vist i Tab. C.2.

Tabell C.2.

z	n-z	z+1	p(n-z)	q(z+1)	$\frac{p(n-z)}{q(z+1)}$	P_z
0	4	1	2,08	0,48	4,33333	0,05308
1	3	2	1,56	0,96	1,62500	0,23001
2	2	3	1,04	1,44	0,72222	0,37376
3	1	4	0,52	1,92	0,27083	0,26993
4	0	5	0	2,40	0	0,07311

Vi ser av vårt eksempel at de forskjellige verdier av z er ulike sannsynlige, noe som kommer av at P_z er en funksjon av z. I mange, eller kanskje de fleste, tilfelle er en ikke interessert i å kjenne sannsynlighetene P_z for alle verdier av z. En er bare interessert i å bestemme den sannsynligste verdien av z og sannsynligheten for denne. I vårt eksempel er z = 2 den sannsynligste verdien og sannsynligheten for denne er, som vi ser av Tab. C.1., lik 0,3738.

La oss betegne den sannsynligste verdien av z med a. Det kan vises at P_z som funksjon av z har bare ett maksimum, og går vi her ut fra det, er det klart at P_a må være større enn både P_{a-1} og P_{a+1}. Av den differenslikningen vi gjengav foran, fremgår det at

$$P_{a+1} = \frac{p(n-a)}{q(a+1)} \; P_a$$

Da det også kan vises at

$$P_{a-1} = \frac{qa}{p(n-a+1)} \; P_a$$

fører ulikhetene

$$P_{a-1} < P_a > P_{a+1}$$

til

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

Løsningen av disse to ulikhetene er

$$np-q < a < np+p$$

Differensen mellom de to grensene for a er

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

Er (np-q) og (np+p) brudne tall, er det ett og bare ett tall mellom dem som er et helt tall. Dette tallet er da den sannsynligste verdi (a) av z. Det hender naturligvis at (np-q) og (np+p) er hele tall. I så fall er z = np-q og z = np+p like sannsynlige verdier og sannsynligere enn alle andre verdier av z.

I vårt eksempel er n = 4 og p = 0,52. Vi finner da at np-q = 1,6 og np+p = 2,6. Følgelig er a = 2. Sannsynligheten for denne verdien av z finner vi da naturligvis ved innsetting av z = 2 i formelen for P_z.

Er n et meget stort tall, blir beregningen av n!, a! og (n-a)! meget arbeidskrevende. Det er utarbeidet en tabell over log(r!) for alle hele tall fra r = 1 til r = 1000 som vi da kan ta i bruk. Det er imidlertid oftest nøyaktig nok å beregne en tilnærmingsverdi for P_a. Tilnærmingsformelen er

$$P_a \approx \frac{1}{\sqrt{2\pi npq}}$$

Er f.eks. n = 1000 og p = 0,25, finner vi at np-q = 249,25 og np+p = 250,25, dvs. at den sannsynligste verdi av z er a = 250. Sannsynligheten for denne verdien er da tilnærmet lik

$$P_a \approx \frac{1}{\sqrt{2\pi 1000.0,25.0,75}} = 0,029$$

Direkte beregnet finner vi at $P_a = 0,0304$.

Et interessant og kanskje også nyttig grensetilfelle av binomialfunksjonen får vi ved å la $n \to \infty$ $p \to 0$ på en slik måte at $np \to m$ hvor m er et endelig tall. En finner da at binomialfunksjonen går over til Poissonfunksjonen som er

$$P_z = \frac{e^{-m} m^z}{z!} \qquad (z = 0,1,2,.... \infty)$$

hvor e er grunntallet i det naturlige logoritmesystem, e = 2,718..

I dette tilfelle vil en finne at den sannsynligste verdi av z er det ene hele tallet mellom grensene

$$m-1 < a < m$$

Er m et helt tall, vil en lett finne at $P_{m-1} = P_m$, dvs. at da er z = m-1 og z = m like sannsynlige og sannsynligere enn alle andre verdier av z.

C.5. Den hypergeometriske funksjon. *se til s. 76*

La oss tenke oss at universet omfatter N gjentak, hvor N er et så lite antall at uttak av et sampel på n gjentak forandrer universet merkbart. La oss videre tenke oss at H av disse gjentak har kjennetegnet E. Resten av gjentakene, altså N-H gjentak, har det motsatte kjennetegn iE. Sannsynligheten for E i dette universet er da P(E;U) = p = H/N.

Tar vi nå et random sampel på n gjentak fra dette universet, vil en finne at sannsynligheten for z gjentak med E i samplet er

$$P_z = \frac{\binom{H}{z}\binom{N-H}{n-z}}{\binom{N}{n}} \qquad (z=0,1,2,....n)$$

Dette uttrykket for P_z er viktig som grunnlag for de såkalte representative tellinger som vi skal komme inn på senere. Vi skal da gå noe nærmere inn på funksjonen. Her skal vi nøye oss med å foreta en sammenligning med den binomiale funksjonen.

La oss tenke oss at N = 20 og H = 10, dvs. at P(U) = p = $\frac{1}{2}$. La størrelsen av samplet være n = 3. Vi har da at

$$P_z = \frac{\binom{10}{z}\binom{10}{3-z}}{\binom{20}{3}}$$

Verdiene av P_z for z = 0,1,2 og 3 er vist i Tab. C.3. Til sammenligning har vi tatt med verdiene av P_z for den binomiale funksjon med n = 3 og p = $\frac{1}{2}$.

Tabell C.3.

z	P_z hypergeom	P_z binomial
0	0,1052	0,125
1	0,3948	0,375
2	0,3948	0,375
3	0,1052	0,125
	1,0000	1,0000

Vi ser her at for de ekstreme verdiene, z = 0 og z = 3, er P_z mindre for den hypergeometriske enn for den binomiale funksjon.

Det er vist at for økende verdier av N og H blir forskjellen mellom de to funksjonene mindre og mindre. Binomialfunksjonen er derfor et grensetilfelle av den hypergeometriske.

D. Fordelingsfunksjoner.

D.1. Diskrete random variable.

La oss tenke oss at vi har et random sampel på n gjentak som er representant for et univers U, og at vi for hvert av disse gjentak har en observasjon av en diskret random variabel x. Som forklart i avsnitt B.2. kan vi da ordne observasjonene i en frekvensfordeling. La frekvensen til verdien x_i være z_i. For denne verdien av x vil det da til den relative frekvens z_i/n svare en sannsynlighet $P(x_i;U)$ i universet. Som oftest er denne sannsynligheten avhengig av verdien av x, slik at vi kan oppfatte den som en funksjon av x og sette

$$P(x;U) = f(x)$$

Denne funksjonen kalles fordelingsfunksjonen for den random variable x i universet U. Siden x er en diskret random variabel, har funksjonen f(x) gyldighet for bare bestemte atskilte verdier av x.

De verdier den random variable kan ha, er alternativer som utelukker hverandre. Etter enten-eller setningen er derfor

$$\Sigma P(x;U) = \Sigma f(x) = 1$$

når en under summeringen tar med alle de verdier av x som kan forekomme.

Vi har vist tidligere hvordan vi kan bruke gjennomsnittet og middelavviket til karakteristikk av observasjonene og dermed også av frekvensfordelingen. Til karakteristikk av en fordelingsfunksjon brukes tilsvarende størrelser. De to viktigste er forventningen og standardavviket. Forventningen svarer, kan vi si til gjennomsnittet og standardavviket til middelavviket. For gjennomsnittet og middelavviket har vi brukt betegnelsene $\bar{x}$ og s. For å unngå forvekslinger er det vanlig, som nevnt i avsnitt B.8., å betegne forventningen med μ og standardavviket med σ. Forventningen

betegnes også ofte med E(x), og kvadratet på standardavviket med var(x).

Forventningen er summen av produktene av x og sannsynligheten P(x;U) = f(x), eller:

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

Standardavvikets kvadrat (variansen, som den ofte kalles) er lik produktet av $(x-\mu)^2$ og f(x), dvs.

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

Under begge disse summeringene skal en ta med alle de verdiene av x som kan forekomme i et gjentak.

Vi kan skrive om formlene for gjennomsnittet og middelavviket slik:

$$\bar{x} = \frac{1}{n} \Sigma z_i x_i = \Sigma \frac{z_i}{n} x_i$$

$$s^2 = \frac{1}{n-1} \Sigma z_i (x_i - \bar{x})^2 = \frac{n}{n-1} \Sigma \frac{z_i}{n} (x_i - \bar{x})^2$$

Her er nå den relative frekvensen z_i/n estimator for sannsynligheten $P(x_i;U) = f(x_i)$. Vi ser da at $\bar{x}$ og forventningen μ svarer til hverandre. Ser vi bort fra faktoren $\frac{n}{n-1}$ som vi skal komme tilbake til, ser vi at også s^2 og σ^2 svarer til hverandre. De er dannet på samme måte; $\bar{x}$ og s ved frekvensfordelingen, μ og σ ved fordelingsfunksjonen.

I en aktuell situasjon har vi observasjoner av x i et sampel og kan ordne disse i en frekvensfordeling. Noen ganger kan det da være av interesse å undersøke hvilken fordelingsfunksjon den random variable har. Vi har imidlertid ikke noe middel til å utlede fordelingsfunksjonen fra frekvensfordelingen. En har derfor vært nødt til å ta utgangspunkt i visse nærmere spesifiserte forutsetninger og ut fra disse utledet typer av fordelingsfunksjoner. I et aktuelt tilfelle kan en så undersøke om en av disse typene gir

en tilfredsstillende beskrivelse av frekvensfordelingen. En oppfatter da fordelingsfunksjonen som en modell, og en er interessert i om modellen passer for det tilfelle en har for seg.

En av disse modellene er den som går under navn av den binomiale fordelingsfunksjon. Den er

$$f(x) = \binom{k}{x} p^x q^{k-x} \qquad x = 0,1,2,\ldots k$$

hvor k, p og q = 1-p, er parametrer. Tallverdien av p, og da også av q, er positiv < 1.

Erstatter vi k med n og x med z, gir denne funksjonen som vi har forklart i avsnitt C.4., sannsynligheten for z gjentak med et kjennetegn E og (n-z) gjentak med det motsatte kjennetegn iE i et random sampel på n gjentak. Vi kan derfor si at binomialfunksjonen er fordelingsfunksjonen for den random variable z i dette tilfelle. Det er imidlertid meget som tyder på at denne funksjonen er en tilfredsstillende modell også i andre tilfelle.

La oss, for å ta for oss et eksempel, anta som hypotese at den er en tilfredsstillende modell for beskrivelse av frekvensfordelingen i Tab.B.3. hvor x er antall arrstråler i arret hos en valmueart. For å kunne foreta en sammenligning er det naturligvis nødvendig å kjenne verdiene av parametrene k og p. Som oftest har vi ikke hypotetiske verdier for disse parametrene. Vi må derfor estimere dem før vi kan beregne funksjonsverdiene og foreta sammenligning.

Det har vært og er fremdeles ulike meninger om hvilken metode en skal bruke for parameterestimeringen. Vi kan her ikke komme inn på dette meget vanskelige emne og må derfor nøye oss med å beskrive en av de metodene som er foreslått og som er i bruk.

En kan vise at forventningen og variansen (σ^2) for x i dette

tilfelle er

$$E(x) = \mu = kp$$

og

$$var(x) = \sigma^2 = kpq$$

Den metoden vi nå skal bruke til estimering av k og p, går ut på at vi setter likhetstegn mellom forventningen og gjennomsnittet på den ene side og mellom standardavviket og middelavviket på den annen side. For vårt eksempel har vi at $\bar{x} = 12,76$ og $s^2 = 5,0018$. Vi setter derfor

$$kp = 12,76$$

og

$$kpq = 5,0018$$

Løser vi disse to ligningene, finner vi at $k = 20,99$ og $p = 0,608$. Runder vi så av for k til det nærmeste hele tall og setter $k = 21$, finner vi av første ligning at $\hat{p} = 0,61$.*

Innsettes så disse verdiene for k og p i funksjonen, får vi de funksjonsverdiene som er gitt i Tab. D.1. Til sammenligning er verdiene av de relative frekvensene også tatt med. Vi ser at det er god overensstemmelse mellom funksjonsverdiene og de relative frekvensene, og vi må her nøye oss med å konstatere det.

Tabell D.1.

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

* For å markere at de verdier av k og p vi har her, er estimater, har vi satt hatt over k og p. Sml. avsnitt B.9. Det samme har vi gjort med f(x) i tabellen fordi estimatene av k og p er brukt.

For k = 21 skal naturligvis alle hele verdier av x fra x = 0 til x = 21 forekomme, men verdiene for f(x) for x = 0,1,2,3, og 21 er så små at de er ikke tatt med i tabellen.

Lar en $k \to \infty$ og $p \to 0$, samtidig med at $kp \to m$, hvor m er et endelig tall, går binomialfunksjonen som nevnt i avsnitt C.4 over til

$$f(x) = \frac{e^{-m} m^x}{x!} \qquad x=0,1,2,3,\ldots\infty$$

Denne funksjonen kalles Poissons fordelingsfunksjon. Den har bare en parameter, nemlig m, og en kan vise at

$$\mu = \sigma^2 = m$$

Gjennomsnittet $\bar{x}$ er således en estimator av m.

La oss som eksempel på anvendelsen av denne funksjonen anta som hypotese at den er en treffende modell for eksemplet i Tab.B.7. Den observerte random variable er her antall eksemplarer av Sandlilje innen ruter på 0,25 kvadratmeter på et felt. Vi finner at $\bar{x} = 1,73$ og bruker dette tallet som estimator av m. I Tab.D.2. er så gitt verdiene av $\hat{f}(x)$ for x = 0 til x = 8. For større verdier av x er $\hat{f}(x)$ så små tall at de er ikke tatt med i tabellen. De relative frekvensene z/n er også oppgitt. Vi ser at også i dette tilfelle er det god overensstemmelse mellom verdiene av $\hat{f}(x)$ og de relative frekvensene.

Tabell D.2.

x	$\hat{f}(x)$	z/n
0	0,1773	0,18
1	0,3067	0,31
2	0,2653	0,27
3	0,1530	0,13
4	0,0662	0,08
5	0,0229	0,01
6	0,0066	0,02
7	0,0016	
8	0,0004	
	1,0000	1,00

Etter Poissons fordelingsfunksjon har den random variable et variasjonsområde som strekker seg fra $x = 0$ til $x \to \infty$ Den observerte random variable har imidlertid alltid et variasjonsområde med endelige grenser. Vi kan kanskje derfor si at Poissons funksjon ikke kan være en realistisk modell. Det kan imidlertid vises at denne fordelingsfunksjonen er et grensetilfelle for andre mer kompliserte funksjoner med begrenset variasjonsområde for den random variable. For vårt eksempel kan det være en av disse funksjonene som er den realistiske modellen.

Etter hvert er det blitt utviklet et stort antall fordelingsfunksjoner for diskrete random variable. Men her har vi ikke anledning til å ta med flere modeller enn de to vi har nevnt.

D.2. Kontinuerlige random variable.

I avsnitt B.1 ble det sagt at en kontinuerlig random variabel er en random variabel som kan ha en hvilken som helst reell tallverdi mellom en nedre og en øvre grense. Dette har vi neppe full dekning for. Det vi vet er at hvis vi har et meget stort antall observasjoner og avsetter disse som punkter på en rett linje, vil det se ut som punktene dekker linjen i hele dens utstrekning. Ser vi imidlertid nærmere etter, vil vi kanskje likevel oppdage åpninger mellom punktene. Våre observasjoner er nemlig alltid avrundede verdier fordi vi bruker graderte måleinstrumenter.

En observasjon inneholder også en observasjonsfeil. Vi kan si at en observert verdi er sammensatt additivt av den egentlige verdi av den random variable og en positiv eller negativ observasjonsfeil. Det er lite fruktbart å spekulere på hvilke verdier den

egentlige variable kan ha. Det er rimelig å tro at en i et bestemt tilfelle ville komme til at det bare kan være tale om et visst sett av rasjonale tall, dvs. at vi ville måtte oppfatte variasjonen som diskret. Men vi ville nok komme til at antallet av mulige verdier innen et intervall er meget stort og sprangene fra en verdi til den neste meget små. Dessuten måtte vi regne med observasjonsfeil og avrundingsfeil. Det ville derfor bli nokså umulig i praksis å operere med diskrete verdier. Vi bør imidlertid være oppmerksomme på at når vi forutsetter kontinuerlig variasjon, må dette oppfattes som en approksimasjon.

Ved fordelingsfunksjonen for den random variable x forstår vi så en **kontinuerlig** funksjon av x, f(x), som bl.a. tilfredsstiller det krav at for alle x innen variasjonsområdet er f(x) >0. Hvis variasjonsområdet for x strekker seg fra x = a til x = b, krever vi dessuten at arealet av den flaten som i et rettvinklet koordinatsystem avgrenses av grafen for funksjonen, X-aksen og ordinatene til x = a og x = b, er lik enheten, dvs. at

$$\int_a^b f(x)\,dx = 1$$

Avgrenser vi innen variasjonsområdet et intervall fra x = c til x = d (se Fig. D.1), er arealet av den flaten som avgrenses av grafen for funksjonen, X-aksen og ordinatene til x=c og x=d, lik sannsynligheten for en verdi av x innen dette intervallet. Dette vil da si at

$$\int_c^d f(x)\,dx = P(c \leq x \leq d;\ U)$$

Legg merke til her at f(x) ikke er en sannsynlighet. Men f(x) dx er en sannsynlighet, nemlig sannsynligheten for en x-verdi innen differensialelementet dx.

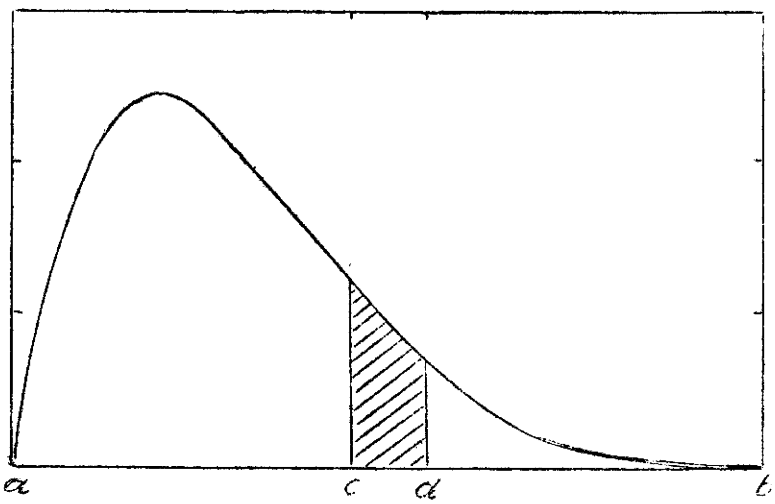


Fig. D.1

Også kontinuerlige random variable har forventning og standardavvik. Disse størrelsene defineres på samme måte som for diskrete random variable. Forskjellen er bare at vi må uttrykke dem ved integraler i stedet for ved summer. Strekker variasjonsområdet seg fra x = a til x = b, er forventningen definert ved integralet

$$E(x) = \mu = \int_a^b f(x)\ x\ dx$$

og kvadratet på standardavviket ved integralet

$$var(x) = \sigma^2 = \int_a^b f(x)\ (x-\mu)^2 dx$$

Også for kontinuerlige random variable er det utviklet mange fordelingsfunksjoner, altså modeller. Den viktigste av dem er den såkalte normale fordelingsfunksjon. Den er så viktig for praktisk anvendelse av statistiske metoder at vi må studere den nokså nøye. Vi har ikke høve til å beskjeftige oss med de forutsetninger som er lagt til grunn for utledningen av den. Vi må nøye oss med å presentere den ved formlen

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

hvor μ og σ er forventningen og standardavviket, og e er grunntallet i det naturlige logaritmesystem. Vi ser at μ og σ er de eneste parametrene i funksjonen. Den har dessuten noen meget enkle egenskaper, og det er kanskje dette som har gjort at det er blitt brukt så meget. Variasjonsområdet for x strekker seg fra $-\infty$ til $+\infty$ og blant annet av den grunn er funksjonen neppe en realistisk modell. Vi skal imidlertid se etter hvert at flere praktisk uunnværlige metoder som er utviklet ved at en har bygget på denne fordelingsfunksjonen, er gyldige innen et meget vidt spektrum av situasjoner og er ikke stemplet av den normale fordelingsfunksjon som fundament. Vi kaller gjerne slike metoder robuste.

Vi ser at funksjonen har sitt maksimum for $x = \mu$ og at den er symmetrisk omkring dette maksimum. Hvordan funksjonen tar seg ut geometrisk er demonstrert ved grafen i Fig.D.2.

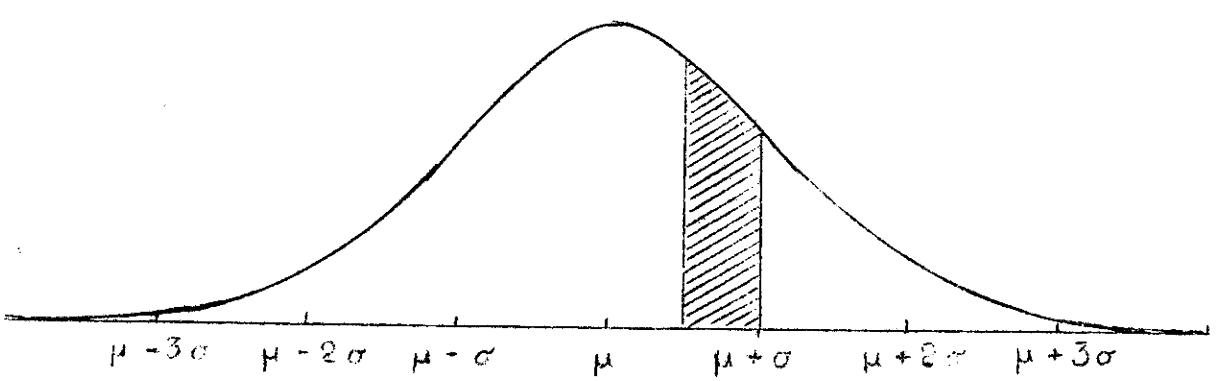


Fig. D.2

La oss nå tenke oss at vi stykker opp variasjonsområdet i et antall like store stykker, altså at vi lager klasser på samme måte som når vi skal ordne observasjonene av en kontinuerlig random variabel i en frekvensfordeling. La oss velge lengden av stykkene, eller klassevidden, lik $\frac{1}{2}\sigma$ og $x = \mu$ som et av delingspunktene. Som vi har forklart foran, er da arealet av den flaten som er avgrenset av grafen, X-aksen og ordinatene til klassens eller stykkets endepunkter, lik sannsynligheten (P) for en verdi av x i vedkommende klasse. Arealet av den skraverte flaten i Fig. D.2 er lik 0,14988 og er altså sannsynligheten for en x-verdi i den klassen som har grensene $\mu + \frac{1}{2}\sigma$ og $\mu + \sigma$ Det ubestemte integral av funksjonen er ikke kjent, og derfor er beregningen av arealene utført ved hjelp av numeriske integrasjonsmetoder. Resultatet er gitt i Tab. D.3 hvor P er sannsynligheten for en x-verdi innen de oppførte klassene. På grunn av symmetrien er bare høyre halvdel (dvs. for $x > \mu$) av funksjonen tatt med. Når unntas den ytterste klassen er klassevidden overalt lik $\frac{1}{2}\sigma$.

Tabell D.3.

Klassegrenser nedre	Klassegrenser øvre	P
μ	$\mu + \frac{1}{2}\sigma$	0,19146
$\mu + \frac{1}{2}\sigma$	$\mu + \sigma$	0,14988
$\mu + \sigma$	$\mu + 1,5\,\sigma$	0,09185
$\mu + 1,5\,\sigma$	$\mu + 2\,\sigma$	0,04406
$\mu + 2\,\sigma$	$\mu + 2,5\,\sigma$	0,01654
$\mu + 2,5\,\sigma$	$\mu + 3\,\sigma$	0,00486
$\mu + 3\,\sigma$	$\mu + 3,5\,\sigma$	0,00112
$\mu + 3,5\,\sigma$	$\mu + 4\,\sigma$	0,00020
$\mu + 4\,\sigma$	∞	0,00003
		0,50000

Vi ser at sannsynligheten for en x-verdi mellom μ og $\mu + \frac{1}{2}\sigma$ er lik P= 0,19146. På grunn av symmetrien er dette da også sann-

synligheten for en x-verdi mellom $\mu-\frac{1}{2}\sigma$ og μ. Følgelig er sannsynligheten for en x-verdi mellom $\mu-\frac{1}{2}\sigma$ og $\mu+\frac{1}{2}\sigma$ lik det dobbelte, eller 2.0,19146 = 0,38292. På samme måte ser vi at sannsynligheten for en x-verdi mellom $\mu-\sigma$ og $\mu+\sigma$ er lik 2(0,19146 + 0,14988) = 0,68268. At x-verdien faller mellom disse to grensene er imidlertid ensbetydende med at $(x-\mu)$ faller mellom grensene $-\sigma$ og $+\sigma$. Og dette er igjen ensbetydende med at tallverdien av $(x-\mu)$ er mindre eller i høyden lik σ, dvs. at sannsynligheten for $|x-\mu| \leq \sigma$ er 0,68268.

Ved å fortsette denne oppsummeringen i Tab.D.3 kan vi lett beregne sannsynligheten for $|x-\mu| \leq a.\sigma$ for a = 1,5, a= 2, a = 2,5 osv. Disse sannsynlighetene er gitt i Tab. D.4 under betegnelsen Q. I denne tabellen er også oppført P = 1-Q som da naturligvis er sannsynligheten for $|x-\mu| \geq a.\sigma$. Vi ser av denne tabellen at sannsynligheten for $|x-\mu| \geq 3.\sigma$ er bare P = 0,0027, dvs. at verdier av x som avviker fra forventningen med mer enn $3.\sigma$ har en meget liten sannsynlighet. Det kan være nyttig å sammenholde dette med den påstand som ble fremsatt i avsnitt B.6, at observasjoner som avviker fra gjennomsnittet med mer enn tre ganger middelavviket, forekommer sjelden.

Tabell D.4.

a	Q	P
0,5	0,38292	0,61708
1,0	0,68268	0,31732
1,5	0,86638	0,13362
2,0	0,95450	0,04550
2,5	0,98758	0,01242
3,0	0,99730	0,00270
3,5	0,99954	0,00046
4,0	0,99994	0,00006

Tabellene D.3 og D.4 gir en god beskrivelse av den normale fordelingsfunksjon. Det er imidlertid en annen tabell som er nyttigere for praktiske anvendelser. Tab. D.4 er en tabell over sannsynlighetene P for valte verdier av a. For praktiske formål er Tab. D.5 av større nytte fordi den er en tabell over a for valte verdier av P. Vi skal senere komme inn på hvordan den kan brukes.

Tabell D.5

P	a
0,05	1,960
0,02	2,326
0,01	2,576
0,001	3,291

Som nevnt er den normale fordelingsfunksjonen meget viktig, kanskje særlig fordi den har vært brukt som basis for utvikling av praktisk-statistiske metoder, f.eks. de metodene vi bruker til analyse av data fra forsøk. Både fordi variasjonsområdet strekker seg fra $-\infty$ til $+\infty$ og fordi funksjonen er helt symmetrisk, er den lite realistisk som modell. Når vi i praksis kan bruke metoder som er utviklet under forutsetning av at den observerte random variable har normal fordelingsfunksjon, er det fordi det er vist at metodene er robuste.

Den normale fordelingsfunksjon ble funnet opp av Gauss og Laplace som var opptatt av variasjonen i observasjonsfeil. Senere har en prøvd den som modell for andre random variable. Vi skal her ta for oss ett eksempel. I Tab. D.6 er gitt frekvensfordelingen, for x = hodeskallens største bredde hos voksne menn. Gjennomsnittet og middelavviket for de n = 2000 observasjonene er $\bar{x} = 156,16$ og $s = 5,73$.

For å kunne foreta en sammenligning må naturligvis de to parametrene i fordelingsfunksjonen estimeres. Det kan være en viss

tvil om hvordan dette skal gjøres. Vi har her nøyet oss med å sette $\hat{\mu} = \bar{x}$ og $\hat{\sigma} = s$. En skulle så beregne de estimerte sannsynlighetene for x-verdier innen de oppgitte klasser ved numerisk integrasjon av funksjonen, f.eks. ved hjelp av Simpsons formel. Når klassevidden ikke er for stor vil en imidlertid også få approksimativt riktige estimater av disse sannsynligheter ved beregning av arealet av rektangulære flater. Høyden av rektanglene settes da lik funksjonsverdien for klassens midtverdi og bredden lik klassevidden. I vårt eksempel er observasjonene gitt i hele millimeter, dvs. at en har brukt et måleredskap med millimetergradering og har så under hver måling avlest til nærmeste hele millimeter. Dette vil da si at f.eks. den klassen som er oppført som (155-159), har grensene 154,5... og 159,4999... Klassevidden er derfor her 5 mm.

De estimerte sannsynligheter er oppført i tabellen under betegnelsen $\hat{p}$ og vi kan da foreta en sammenligning med de relative frekvensene z/n. Vi kan konstatere at overensstemmelsen er tilfredsstillende.

Tabell D.6.

Klasser	Midtverdi	z		z/n	$\hat{p}$
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

Det er funnet opp andre modeller for beskrivelsen av frekvensfordelingen for kontinuerlige random variable, som er langt mer realistiske. Vi skal her kort nevne den modellen som synes å være den mest realistiske. Det er den såkalte Beta-fordelingsfunksjonen som er gitt ved formelen

$$f(x) = K.\ (x-a)^m\ (b-x)^k \qquad a \leq x \leq b$$

hvor m og k er parametrer med positive verdier. Variasjonsområdet for den random variable strekker seg fra x = a til x = b, og konstanten K har da en slik verdi at integralet av funksjonen fra x=a til x=b er lik enheten. Er m og k hele positive tall, har vi at

$$k = \frac{(k+m+1)!}{k!\ m!}\ (b-a)^{-(k+m+1)}$$

Det er grafen for denne funksjonen (a=o, b=10, k=3 og m=1) som er vist i Fig. D.1.

Setter vi m = k, får vi en funksjon som er symmetrisk omkring et maksimum for $x = \frac{a+b}{2}$ som også da er forventningen for x. Når $m \lesseqgtr k$, er funksjonen usymmetrisk.

Denne fordelingsfunksjonen som tilfredsstiller de viktigste krav til en realistisk modell, er dessverre meget vanskelig å håndtere matematisk. Men vi skal senere se hvordan den er blitt benyttet til å undersøke om metoder som er basert på den normale fordelingsfunksjonen, er tilstrekkelig robuste.

Vi må også her ta med enda en fordelingsfunksjon, den såkalte Gamma fordelingsfunksjon. Variasjonsområdet for den random variable strekker seg fra x = 0 til $x \to \infty$ og funksjonen er

$$f(x) = K\ x^m\ e^{-kx}$$

hvor m og k er parametrer. Hvis m er et helt positivt tall, er

$$K = \frac{k^{m+1}}{m!}$$

Det kan vises at funksjonen har sitt maksimum for $x = m/k$ og at

$$E(x) = \mu = (m+1)/k \quad \text{og} \quad \text{var}(x) = \sigma^2 = (m+1)/k^2$$

Et viktig tilfelle av denne fordelingsfunksjonen har vi når vi setter $x = \chi^2$ hvor χ er den greske bokstav kji. Da er $k = \frac{1}{2}$ og $m = \frac{1}{2}(f-2)$. Bokstaven f står her for antall frihetsgrader som vi skal forklare betydningen av senere. Ved å benytte formlene for $E(x)$ og $\text{var}(x)$ finner vi at

$$E(\chi^2) = f \quad \text{og} \quad \text{Var}(\chi^2) = 2f$$

D.3. Standardavviket som målestokk for størrelsen av variasjonen.

I avsnitt B.6 viste vi ved noen eksempler at det alt overveiende antall observasjoner faller innenfor et område som strekker seg fra $\bar{x} - 3s$ til $\bar{x} + 3s$. I de eksemplene som ble benyttet, var det iallfall et relativt meget lite antall observasjoner utenfor dette området. Dette er imidlertid et altfor spinkelt grunnlag for generalisering, og dessuten vil vi gjerne vite om en ved hjelp av forventningen og standardavviket kan si noe om størrelsen av variasjonen. Har nå den random variable normal fordelingsfunksjon, viser Tab. D.4 at sannsynligheten for en observasjon utenfor området fra $\mu-3\sigma$ til $\mu+3\sigma$ er meget liten, nemlig bare 0,0027. Sannsynligheten for en observasjon utenfor området fra $\mu-2\sigma$ til $\mu+2\sigma$ er også liten, nemlig 0,0455.

Vi vet imidlertid praktisk talt aldri hvilken fordelingsfunksjon den random variable har. Det ville derfor være av en viss interesse å ha en tabell svarende til Tab. D.4 eller D.5 som hadde gyldighet for alle random variable uansett hvilken fordelingsfunksjon de har.

Det er vist at for $a > 1$ er sannsynligheten for

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

mindre eller i høyden lik $1/a^2$. Denne setningen går under navn av den Bienaymé-Tchebycheffske ulikhet etter de to matematikere som oppdaget den. I Tab. D.7 er vist noen sammenhørende verdier av a og P. Vi ser at P avtar for voksende verdi av a på samme måte som for den normale fordelingsfunksjonen (Tab. D.4). Og vi ser at sannsynligheten for en x-verdi utenfor området fra $\mu-4\sigma$ til $\mu+4\sigma$ er liten, nemlig mindre eller i høyden lik 0,062.

Tabell D.7

a	P
2	0,2500
3	0,1111
4	0,0625
5	0,0400

Denne setningen har gyldighet for alle mulige fordelingsfunksjoner. Den gjelder også for alle usedvanlige typer, slike som en sjelden eller aldri kommer over i forskningsarbeid. De fordelingsfunksjoner en oftest har å gjøre med, avviker ikke så meget fra den normale. For å gi en liten illustrasjon til dette har vi i Tab. D.8 gitt sannsynligheten for P for $|x-\mu| \geq a.\sigma$ for et enkelt tilfelle av en symmetrisk Beta fordelingsfunksjon, nemlig

$$f(x) = K\, x^2(b-x)^2$$

Ved å sammenligne med Tab. D.4 ser vi at for a=2 og a=2,5 er verdiene av P mindre enn for den normale fordelingsfunksjonen.

Tabell D.8.

a	P
1,0	0,3598
1,5	0,1530
2,0	0,0300
2,5	0,0004

D.4. Funksjoner av en random variabel.

I flere sammenhenger har vi bruk for funksjoner av den observerte random variable. Som eksempel kan vi tenke oss at temperaturen er observert i et sampel av gjentak og at målingene er utført med et Fahrenheittermometer. Vi har f.eks. følgende n=12 observasjoner

40,2	41,3	41,0	41,2
41,4	40,6	40,9	40,3
40,7	40,9	40,9	40,6

Gjennomsnittet og middelavviket for disse observasjoner er

$\bar{x} = 40,83$ og $s = 0,37$.

La oss tenke oss at vi av en eller annen grunn ville ha ønsket at temperaturene hadde vært målt med et Celsiustermometer. Siden vi kjenner den funksjonelle sammenhengen mellom Fahrenheitgrader og Celsiusgrader, kan vi naturligvis regne om hver av de 12 observasjonene til Celsiusgrader. Er x antall Fahrenheitgrader og y antall Celsiusgrader, har vi at

$$y = \frac{5}{9}(x-32) = \frac{5}{9}x - \frac{160}{9}$$

En slik ligning kaller vi en transformasjonslikning eller en transformasjon. Vi ser at den er lineær og er et bestemt eksempel på en alminnelig lineær transformasjon $y = a + bx$.

I noen tilfelle kan en naturligvis være interessert i å transformere x til y for hvert av de n gjentak. Men er vi interessert i bare gjennomsnittet og middelavviket for de transformerte observasjonene (y), er det lett å vise at en kan sette

$$\bar{y} = a + b.\bar{x} \quad \text{og} \quad s_y = b.s_x$$

Bruker vi disse to formlene på vårt eksempel, vil vi finne at

$$\bar{y} = \frac{5}{9}\,40,83 - \frac{160}{9} = 4,90$$

og

$$s_y = \frac{5}{9}\,.0,37 = 0,21$$

Det er enkelt å vise at dette må være riktig. Vi har nemlig at

$$\bar{y} = \frac{1}{n}\Sigma y_i = \frac{1}{n}\Sigma(a+bx_i) = \frac{1}{n}(n.a+b.n.\bar{x}) = a + b.\bar{x}$$

og

$$s_y^2 = \frac{1}{n-1}\Sigma(y_i-\bar{y})^2 = \frac{1}{n-1}\Sigma[(a+bx_i)-(a+b\bar{x})]^2$$
$$= \frac{1}{n-1} b^2 \Sigma(x_i-\bar{x})^2 = b^2 s_x^2$$

Transformasjonsligningen kan også skrives slik:

$$y = b(x + \frac{a}{b})$$

som klarere viser at bruken av den betyr både en forandring av nullpunktet og av måleenheten. Er a = 0, betyr det at vi forandrer bare måleenheten. Er b = 1, betyr det at vi endrer bare nullpunktet.

Det er vel umiddelbart innlysende at disse omregningsformlene for gjennomsnittet og middelavviket også gjelder forventningen og standardavviket. La forventningen og standardavviket for x være μ_x og σ_x. For y = a + bx har vi da at

$$\mu_y = a + b\,\mu_x \quad \text{og} \quad \sigma_y = b.\sigma_x$$

La p = P(E;U) være sannsynligheten for E i universet U. Forventningen og standardavviket for antall (z) gjentak med E i et random sampel på n gjentak er da (sml. C.4 og D.1)

$$\mu_x = np \quad \text{og} \quad \sigma_x = \sqrt{np(1-p)}$$

Her er nå z det absolutte antall gjentak med E. Sett at vi så er interesserte i det relative antall

$$y = \frac{z}{n} = \frac{1}{n} z$$

Dette er en enkel linær transformasjonsligning (a=0 og b= $\frac{1}{n}$) og forventningen og standardavviket for y blir derfor lik

$$\mu_y = \frac{1}{n}\mu_x = \frac{1}{n}.np = p$$

og

$$\sigma_y = \frac{1}{n}\sigma_x = \frac{1}{n}\sqrt{np(1-p)} = \sqrt{\frac{p(1-p)}{n}}$$

Er sannsynligheten for E = "gutt" i universet av nyfødte barn

lik p = 0,52, er forventning og standardavvik for det relative antall gutter i et random sampel på n = 4 lik

$$\mu_y = p = 0,52 \quad \text{og} \quad \sigma_y = \sqrt{\frac{0,52 \cdot 0,48}{4}} = 0,2498$$

Det finnes mange eksempler på transformasjoner som er i bruk i praksis og hvor ligningen ikke er lineær. Vi har da at y = g(x) hvor funksjonen ikke er lineær. Eksempler er $y = \sqrt{x}$ og y = log x. I slike tilfelle er det meget vanskeligere å foreta omregning av forventning og standardavvik eller gjennomsnitt og middelavvik. Den fremgangsmåten som en har slått inn på, er å erstatte funksjonen g(x) med dens rekkeutvikling (Taylor-rekken) og så kaste bort alle ledd unntatt konstantleddet og det lineære ledd. Det er vel umiddelbart innlysende at en da ikke kan regne med annet enn i beste fall tilnærmet riktige resultater. Og det er vel også klart at nøyaktigheten vil avhenge av hvor fort rekkeutviklingen konvergerer. Det er grunn til å være på vakt overfor resultater som er oppnådd på denne måten.

D.5. Fordelingsfunksjonen for flere random variabler.

I mange tilfelle skaffer en seg observasjoner av flere random variabler for hvert gjentak. Et eksempel er vist i Tab.B.9. La oss nå ta for oss et annet eksempel. Vi vil tenke oss at et forsøksfelt er delt i n ruter og at det i hver rute er plantet et valt konstant antall kålrotplanter. La oss si 50. Under høstingen av feltet vil en da kanskje finne at noen av plantene er gått ut, slik at antall planter (x_1) varierer rutene imellom. Vi vil så tenke oss at vi for hvert gjentak også observerer antall planter (x_0) som er angrepne av en viss sykdom. Vi har da n samtidige observasjoner av to random variabler.

I dette eksemplet er begge random variabler av den diskrete typen. Ved fordelingsfunksjonen for to slike random variabler forstår vi en funksjon av dem begge, $f(x_1,x_o)$, som gir sannsynligheten for en bestemt verdi av x_1 og en bestemt verdi av x_o i et gjentak. Dette vil si at hvis $x_1=a$ og $x_o=b$ kan forekomme i samme gjentak i universet U, er

$$P(x_1=a \text{ og } x_o=b;U) = f(a,b)$$

Ved universet U forstår vi her som ellers det universet som er representert av de n gjentakene i egenskap av et random sampel. Til funksjonen $f(x_1,x_o)$ må vi sette som krav at

$$\Sigma\Sigma\, f(x_1,x_o) = 1$$

når vi under summeringen tar med alle de verdier av de to variabler som kan forekomme i et gjentak.

La oss så tenke oss at samplet består av n=100 gjentak eller ruter. Vil vi da skaffe oss en oversikt over observasjonene, kan vi begynne med å ordne gjentakene etter verdiene av x_1, dvs. at vi ordner observasjonene av x_1 i en frekvensfordeling. La oss tenke oss at denne fordeling ser slik ut:

x_{1i}	z_{1i}	z_{1i}/n
47	5	0,05
48	10	0,10
49	60	0,60
50	25	0,25
	100	1,00

Til de relative frekvensene (z_{1i}/n) svarer da i universet en fordelingsfunksjon for x_1, $f(x_1)$, slik at sannsynligheten for $x_1=a$ i et gjentak i universet er

$$P(x_1=a;U) = f(a)$$

Vi kan så ta for oss de gjentakene hvor x_1 har samme verdi, f.eks. de 10 gjentakene som har $x_1=48$. Observasjonene av x_o for denne

delen av samplet kan så ordnes i en frekvensfordeling. La oss tenke oss at denne blir:

x_{oi}	z_{oi}	z_{oi}/z_{1i}
0	6	0,6
1	2	0,2
2	1	0,1
3	1	0,1
	10	1,0

Denne delen av samplet må vi oppfatte som en random representant for et delunivers av U, og til de relative frekvensene (z_{oi}/z_{1i}) vil det i dette deluniverset svare en fordelingsfunksjon for x_o. Sannsynligheten for $x_o=b$ er da $P(x_o=b;U,x_1=48)$ eller i alminnelighet for $x_1=a$, $P(x_o=b;U,x_1=a)$. Dette er en betinget sannsynlighet, og i samsvar med det må vi operere med en betinget fordelingsfunksjon som vi skal betegne med $f(x_o;x_1)$.

Etter både-og setningen (se C.3) er sannsynligheten for $x_1=a$ og $x_o=b$ i et gjentak lik

$$P(x_1=a \text{ og } x_o=b;U) = P(x_1=a;U).P(x_o=b;U,x_1=a)$$

Brukes symbolene for fordelingsfunksjonene, må dette skrives slik:

$$f(a,b) = f(a).f(b;a)$$

eller i sin alminnelighet:

$$f(x_1,x_o) = f(x_1).f(x_o;x_1)$$

Til den betingede fordelingsfunksjonen må vi stille det krav (se D.1) at

$$\Sigma f(x_o;x_1) = 1$$

når vi under summeringen tar med alle de verdier x_o kan ha i det deluniverset hvor x_1 har en fast verdi. At vi sier at fordelingsfunksjonen er betinget, kommer av at x_1 opptrer som en parameter i funksjonen.

Vi skal ta for oss et nytt eksempel. La oss tenke oss at E_2 er en egenskap eller et kjennetegn som nedarves til bare hunlig avkom. La sannsynligheten for at et gjentak (dvs. et avkom) har kjennetegnet E_1 = "hun" er $P(E_1; U) = P$ og at sannsynligheten for at en hun har kjennetegnet E_2er $P(E_2;U,E_1) = p$. Sannsynligheten for x_1 hunner i et sampel på n avkom er da

$$f(x_1) = \binom{n}{x_1} P^{x_1} (1-P)^{n-x_1}$$

Sannsynligheten for x_o avkom med E i et sampel på x_1 hunner er

$$f(x_o;x_1) = \binom{x_1}{x_o} P^{x_o} (1-p)^{x_1-x_o}$$

I den siste funksjonen som er den betingede, opptrer x_1 som en parameter. Den vil derfor også finnes i uttrykkene for den betingede forventning og det betingede standardavvik. Brukes formlene som er gitt i avsnitt D.1, finner vi at den betingede forventning er px_1 og at det betingede standardavvik er $\sqrt{p(1-p)x_1}$

Den betingede fordelingsfunksjonen spiller stor rolle i praksis. Og det en da er særlig interessert i, er den betingede forventningen som vi vil betegne med $E(x_o;x_1)$. Etter definisjonen som er gitt i avsnitt D.1 er

$$E(x_o;x_1) = \Sigma f(x_o;x_1).x_o$$

Under summeringen må da tas med alle de verdier av x_o som kan forekomme i et gjentak i deluniverset (U,x_1).

Hvis det finnes avhengighet mellom de to random variable, vil praktisk talt alltid denne forventningen være en funksjon av x_1. Denne funksjonen kalles regresjonsfunksjonen for x_o m.h.p. x_1. I avsnitt B.9 ble den betegnet med $r(x_1)$. I det lineære tilfelle er

$$r(x_1) = \beta_o + \beta_{o1}x_1$$

En krever imidlertid at funksjonen skal tilfredsstilles av $x_1=\mu_1$ og $x_o=\mu_o$, hvor μ_o er forventningen for x_o. Dette fører til at

$$r(x_1) = \mu_o + \beta_{o1}(x_1-\mu_1)$$

Hvis vi lar x_1 og x_o bytte rolle, kan vi sette

$$f(x_1,x_o) = f(x_o).f(x_1;x_o)$$

hvor da $f(x_1;x_o)$ er fordelingsfunksjonen for x_1 i deluniverset (U,x_o) eller den betingede fordelingsfunksjon for x_1. Forventningen for x_1 i dette deluniverset vil som oftest være en funksjon av x_o, og denne funksjonen er da regresjonsfunksjonen for x_1 m.h.p. x_o. I det lineære tilfelle har vi (se B.9) at

$$r(x_o) = \mu_1 + \beta_{1o}(x_o-\mu_o)$$

I et aktuelt tilfelle er det bare en av de to regresjonsfunksjonene vi har bruk for. Vi står imidlertid fritt i valget av betegnelse på de to random variable, og vi vil da her og senere, under gjennomgåelsen av mer kompliserte tilfelle, bruke x_o som betegnelsen på den avhengige random variable.

For universet innfører vi en sum som går under navn av kovariansen, $\text{cov}(x_1,x_o)$, og som svarer, kan vi si, til den kovarianssum vi benyttet i avsnitt B.9. Denne summen er

$$\text{cov}(x_1,x_o) = \Sigma\Sigma\, f(x_1,x_o).(x_1-\mu_1)(x_o-\mu_o)$$

Settes

$$S = \Sigma\Sigma\, f(x_1,x_o)\ [x_o - r(x_1)]^2$$

vil vi finne ved minimalisering av S at

$$\beta_{o1} = \frac{\text{cov}(x_1,x_o)}{\text{var}(x_1)}$$

Dette resultat svarer, ser vi, til den estimator (b_{o1}) vi kom fram til i avsnitt B.9.

Vi har også en korrelasjonskoeffisient (ρ) som svarer til den korrelasjonskoeffisient (r) for samplet som vi innførte i avsnitt B.9. Den er

$$\rho = \frac{\mathrm{cov}(x_1,x_o)}{\sigma_1\sigma_o}$$

hvor σ_1 og σ_o er standardavvikene for x_1 og x_o.

Vi har da at

$$\mathrm{cov}(x_1,x_o) = \rho\,\sigma_1\sigma_o$$

og derfor at

$$\beta_{o1} = \rho\,\frac{\sigma_o}{\sigma_1}$$

Hvis $f(x_o;x_1)$ ikke har x_1 som parameter, dvs. at den ikke er en betinget fordelingsfunksjon, vil vi ha at $f(x_o;x_1) = f(x_o)$ Følgelig er

$$f(x_1,x_o) = f(x_1).f(x_o)$$

Vi sier da at x_1 og x_o er uavhengige random variabler. I dette tilfelle vil vi finne at

$$\mathrm{cov}(x_1,x_o) = \Sigma\, f(x_1).(x_1-\mu_1)\quad \Sigma\, f(x_o).(x_o-\mu_o)$$

En finner lett at begge disse summene er lik null og at derfor $\mathrm{cov}(x_1,x_o) = 0$. Dette betyr da naturligvis at $\rho = 0$ og $\beta_{o1} = 0$. Er regresjonsfunksjonen $r(x_1)$ lineær, betyr omvendt $\rho = 0$ at den betingede forventning for x_o er uavhengig av x_1. Da er naturligvis grafen av regresjonsfunksjonen en rett linje parallell med x_1-aksen i avstanden μ_o fra denne.

Også når de to random variablene er kontinuerlige eller en av dem er det, er fordelingsfunksjonen en funksjon av begge, altså $f(x_1,x_o)$. Men i disse tilfelle er ikke funksjonsverdiene sannsynligheter. Er både x_1 og x_o kontinuerlige, er sannsynligheten for $a \leq x_1 \leq b$ og $c \leq x_o \leq d$ i et gjentak lik

$$P(a \leq x_1 \leq b \text{ og } c \leq x_o \leq d\ ;\ U) = \int_a^b \int_c^d f(x_1,x_o)\,dx_1dx_o$$

Til $f(x_1,x_o)$ må vi her stille det krav at det bestemte integral av funksjonen er lik enheten når integrasjonsområdene inneholder alle de verdier av x_1 og x_o som kan forekomme i et gjentak.

Også i dette tilfelle kan vi sette

$$f(x_1,x_o) = f(x_1).f(x_o;x_1)$$

hvor den siste funksjonen er den betingede fordelingsfunksjonen for x_o. Den betingede forventningen for x_o blir da

$$E(x_o;x_1) = \int_c f(x_o;x_1)\, x_o\, dx_o$$

hvor integrasjonsområdet omfatter alle de verdier av x_o som kan forekomme når x_1 er gitt. Er det samvariasjon mellom de to random variable, er den betingede forventningen for x_o som oftest en funksjon av x_1, og denne funksjonen er da regresjonsfunksjonen for x_o m.h.p. x_1.

Den normale fordelingsfunksjon for to random variable spiller en viss rolle. Vi skal nøye oss med å presentere den. Settes som foran $f(x_i,x_o) = f(x_o;x_1)$, er $f(x_1)$ en vanlig normal fordelingsfunksjon med parametrene μ_1 og σ_1. Den betingede fordelingsfunksjonen for x_o er også normal og kan skrives slik:

$$f(x_o;x_1) = \frac{1}{\sigma_o\sqrt{2\pi(1-\rho^2)}}\, e^{-\frac{[x_o-r(x_1)]^2}{2\sigma_o^2(1-\rho^2)}}$$

hvor $r(x_1)$ er regresjonsfunksjonen

$$r(x_1) = \mu_o + \rho\frac{\sigma_o}{\sigma_1}(x_1-\mu_1)$$

Vi ser at i dette tilfelle er regresjonsfunksjonen med i formelen for fordelingsfunksjonen. Det er også nyttig å merke seg at stan-

dardavviket for x_o er uavhengig av x_1 og er $\sigma_o\sqrt{1-\rho^2}$. Dette viser at jo større korrelasjonskoeffisienten er , jo mindre er variasjonen i x_o omkring regresjonsfunksjonen. Når $\rho \to 1$,vil de to variabler bli identiske når en ser bort fra at de to forventningene kan være forskjellige.

Er $\rho = 0$, vil $f(x_o;x_1)$ bli en vanlig normal fordelingsfunksjon med parametrene μ_o og σ_o. Vi ser da at

$$f(x_1,x_o) = f(x_1).f(x_o)$$

dvs. at de to random variablene er uavhengige. En må imidlertid ikke gå ut fra at det er slik i andre tilfelle. At korrelasjonskoeffisienten er lik null betyr ikke alltid at de to random variablene er uavhengige.

I praksis kan vi bare sjelden gå ut fra at regresjonsfunksjonen for x_o m.h.p. x_1 er lineær. I aktuelle tilfelle vil en derfor bli stilt overfor den oppgaven å undersøke hvilken form en skal bruke for regresjonsfunksjonen. Dette er imidlertid et emne vi skal komme inn på senere.

I avsnitt B.9 er nevnt at en kan ha observasjoner av flere enn to random variabler for hvert gjentak i et sampel. I samsvar med dette har vi også fordelingsfunksjoner for tre eller flere random variabler. Fordelingsfunksjonen er da en funksjon av alle variable. Er det f.eks. tre variabler, (x_o,x_1 og x_2), kan vi sette

$$f(x_o,x_1,x_2) = f(x_1,x_2).f(x_o;x_1,x_2)$$

hvor vi også kan sette at

$$f(x_1,x_2) = f(x_1).f(x_2;x_1)$$

Funksjonen $f(x_o;x_1,x_2)$ er da her den betingede fordelingsfunksjon for x_o. I denne er da vanligvis x_1 og x_2 parametrer. Den betingede

forventningen for x_o er derfor oftest en funksjon av både x_1 og x_2 og er regresjonsfunksjonen for x_o m.h.p. x_1 og x_2. I det lineære tilfelle er

$$r(x_1,x_2) = \mu_o + \beta_{o1.2}(x_1-\mu_1) + \beta_{o2.1}(x_2-\mu_2)$$

hvor μ_o, μ_1 og μ_2 er forventningene for de tre random variablene. Koeffisientene $\beta_{o1.2}$ og $\beta_{o2.1}$ kalles regresjonskoeffisientene. Vi skal senere komme inn på hvordan de skal estimeres. Som vi da skal se eksempler på, kan det være flere enn tre variabler i en slik regresjonsfunksjon. Er det flere enn to , taler en om multippel regresjon og korrelasjon.

D.6. Funksjoner av flere random variable.

Funksjoner av to eller flere random variable spiller en viss rolle i statistikken. La oss tenke oss at vi har observasjoner av to random variable, x_1 og x_2. Vi kan da være interesserte i summen (x_1+x_2), i differansen (x_1-x_2) eller i forholdet x_1/x_2. Det en først og fremst er interessert i, er forventningen og standardavviket for funksjonen.

Vi skal her nøye oss med lineære funksjoner, f.eks. funksjonen

$$y = a + b_1x_1 + b_2x_2$$

hvor a, b_1 og b_2 er kjente størrelser. Her er naturligvis da y en random variabel. Forventningen for den vil vi betegne med $E(y)$ eller μ_y og standardavviket med σ_y. De tilsvarende karakteristikker for x_1 og x_2 betegnes med μ_1, μ_2, σ_1 og σ_2. Det kan da bevises at

$$\mu_y = a + b_1\mu_1 + b_2\mu_2$$

og

$$\sigma_y^2 = b_1^2\sigma_1^2 + b_2^2\sigma_2^2 + 2b_1b_2\,\rho\,\sigma_1\sigma_2$$

hvor ρ er korrelasjonskoeffisienten mellom x_1 og x_2. Det er enkelt å utlede disse formlene. Vi vil likevel her nøye oss med den utledningen vi får ved å tenke oss at vi har observasjoner av x_1 og x_2 i hvert gjentak i et sampel. Vi betegner disse observasjonene med x_{1i} og x_{2i} $(i=1,2,...n)$. Forutsatt at verdiene av a, b_1 og b_2 er kjent, kan vi for hvert gjentak beregne

$$y_i = a + b_1 x_{1i} + b_2 x_{2i}$$

For summen har vi da at

$$\Sigma y_i = na + b_1 \Sigma x_{1i} + b_2 \Sigma x_{2i}$$

og ved divisjon med n at

$$\bar{y} = a + b_1 \bar{x}_1 + b_2 \bar{x}_2$$

Videre finner vi at

$$y_i - \bar{y} = b_1(x_{1i} - \bar{x}_1) + b_2(x_{2i} - \bar{x}_2)$$

og ved kvadrering og summering at

$$\Sigma(y_i - \bar{y})^2 = b_1^2 \Sigma(x_{1i} - \bar{x}_1)^2 + b_2^2 \Sigma(x_{2i} - \bar{x}_2)^2 + 2 b_1 b_2 \Sigma(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$$

Divideres så med n-1 finnes

$$s_y^2 = b_1^2 s_1^2 + b_2^2 s_2^2 + 2 b_1 b_2 r s_1 s_2$$

hvor s_y, s_1 og s_2 er middelavvikene for y, x_1 og x_2, og hvor r er korrelasjonskoeffisienten mellom x_1 og x_2. Vi ser at disse formlene for $\bar{y}$ og s_y^2 svarer helt til de formlene vi har gjengitt foran for μ_y og σ_y. Går vi nå tilbake til disse formlene og forutsetter at x_1 og x_2 er ikke-korrelerte, dvs. at $\rho = 0$, ser vi at

$$\sigma_y^2 = b_1^2 \sigma_1^2 + b_2^2 \sigma_2^2$$

Setter vi så f.eks. a = 0, $b_1=1$ og $b_2=-1$, dvs. at $y = x_1 - x_2$, vil vi finne at

$$\mu_y = \mu_1 - \mu_2 \qquad \text{og at} \qquad \sigma_y^2 = \sigma_1^2 + \sigma_2^2$$

Setter vi a = 0, $b_1=b_2=1$, dvs. at $y = x_1 + x_2$, finner vi at

$$\mu_y = \mu_1 + \mu_2 \qquad \text{og at} \qquad \sigma_y^2 = \sigma_1^2 + \sigma_2^2$$

Dette viser da at hvis x_1 og x_2 er ukorrelerte, har summen (x_1+x_2) og differansen (x_1-x_2) samme standardavvik.

Den situasjon vi tok utgangspunkt i foran for utledningen av $\bar{y}$ og s_y^2, var at x_{1i} og x_{2i} var observasjoner av to random variable i samme gjentak. I slike situasjoner må vi jo alltid regne med muligheten for at det eksisterer korrelasjon mellom de to random variable. I andre tilfelle kan vi trygt regne med at det ikke finnes slik korrelasjon. En vanlig situasjon er at x_1 og x_2 er den samme random variable, og at fotskriftene 1 og 2 i x_{1i} og x_{2i} betyr at den første (x_{1i}) stammer fra et tilfeldig gjentak i et univers U_1, den andre (x_{2i}) fra et tilfeldig gjentak i et annet univers U_2. Som eksempel kan vi tenke oss at x_{1i} er observasjon av den dobbelte barktykkelse fra et tilfeldig valt tre i en skog i Telemark og x_{2i} observasjon av den samme random variable fra et tilfeldig tre i en skog i Østfold.

La oss tenke oss at x_{1i} og x_{2i} er observasjoner av samme random variabel (eller det kan være to forskjellige random variabler) som vi har stilt sammen til et par. Den ene (x_{1i}) skriverseg fra et tilfeldig gjentak i universet U_1 og den andre (x_{2i}) fra et tilfeldig gjentak i universet U_2. Vi kan så ta et nytt tilfeldig gjentak fra U_1 og et nytt tilfeldig gjentak fra U_2 og få to nye observasjoner, x_{12} og x_{22}. Dette kan vi så tenke oss fortsatt til vi har n gjentak fra U_1 og n gjentak fra U_2. Vi danner så, la oss si, differensene $y_i = x_{1i}-x_{2i}$, altså differensene mellom de to observasjonene vi selv har paret sammen. Summerer vi disse differensene, finner vi at

$$\Sigma y_i = \Sigma x_{1i} - \Sigma x_{2i}$$

og dividerer med antallet (n), finner vi at $\bar{y} = \bar{x}_1 - \bar{x}_2$.

Danner vi så med sikte på å finne en formel for s_y^2, differensene

$y_i-\bar{y}$, vil vi få følgende differenser:

$$y_1-\bar{y} = (x_{11}-x_{21})-(\bar{x}_1-\bar{x}_2) = (x_{11}-\bar{x}_1)-(x_{21}-\bar{x}_2)$$
$$y_2-\bar{y} = (x_{12}-x_{22})-(\bar{x}_1-\bar{x}_2) = (x_{12}-\bar{x}_1)-(x_{22}-\bar{x}_2)$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$y_n-\bar{y} = (x_{1n}-x_{2n})-(\bar{x}_1-\bar{x}_2) = (x_{1n}-\bar{x}_1)-(x_{2n}-\bar{x}_2)$$

Kvadrerer vi så disse differensene og summerer kvadratene, får vi at

$$\Sigma(y_i-\bar{y})^2 = \Sigma(x_{1i}-\bar{x}_1)^2 + \Sigma(x_{2i}-\bar{x}_2)^2 - 2\,\Sigma(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)$$

Divideres så med (n-1), finner vi at

$$s_y^2 = s_1^2 + s_2^2 - 2s_1s_2r$$

hvor r er korrelasjonskoeffisienten.

Hvis vi utførte i praksis en slik datainnsamling som den vi har beskrevet - noe ingen forhåpentlig vil finne på - ville vi finne at verdien av korrelasjonskoeffisienten ikke er lik null. Men dette ville da være en rent tilfeldig eller random effekt. Er nemlig x_1 og x_2 uavhengige, som vi har forutsatt, vil fordelingsfunksjonen for de to random variable kunne skrives som et produkt av fordelingsfunksjonene for hver av dem (se avsnitt D.5), dvs. at

$$f(x_1,x_2) = f(x_1).f(x_2)$$

For funksjonen $y = a + b_1x_1 + b_2x_2$ vil vi da lett kunne utlede at

$$\mu_y = a + b_1\mu_1 + b_2\mu_2$$

og

$$\sigma_y^2 = b_1^2\sigma_1^2 + b_2^2\sigma_2^2$$

Vi skal senere se at disse formlene har stor betydning i praktiske anvendelser, f.eks. når $a = 0$, $b_1 = 1$ og $b_2 = -1$.

Disse formlene for forventning og standardavvik for en lineær funksjon av to random variabler kan lett generaliseres til flere

variabler. Sett at vi har at

$$y = a + b_1x_1 + b_2x_2 + b_3x_3$$

For denne funksjonen kan det vises at

$$\mu_y = a + b_1\mu_1 + b_2\mu_2 + b_3\mu_3$$

og

$$\sigma_y^2 = b_1^2\sigma_1^2 + b_2^2\sigma_2^2 + b_3^2\sigma_3^2 + 2b_1b_2\rho_{12}\sigma_1\sigma_2 + 2b_1b_3\rho_{13}\sigma_1\sigma_3 + 2b_2b_3\rho_{23}\sigma_2\sigma_3$$

Her er ρ_{12}, ρ_{13} og ρ_{23} korrelasjonskoeffisientene mellom x_1 og x_2, mellom x_1 og x_3 og mellom x_2 og x_3.

E. Sampel random variabler.

E.1 Innledning.

La oss tenke oss at vi fra et større skogområde med noenlunde ensaldret gran, tar ut et sampel på f.eks. n = 25 grantrær. Vi tenker oss videre at vi måler brysthøydediameteren på hvert enkelt tre og beregner gjennomsnitt og middelavvik.

Tar vi et nytt sampel på n = 25 trær, er det meget lite sannsynlig at vi skal få de samme observasjonene om igjen. Vi vil få et nytt sett observasjoner og dermed et gjennomsnitt og et middelavvik som er forskjellig fra dem vi fikk i vårt første sampel. Både gjennomsnittet og middelavviket vil altså variere samplene imellom, og vi kaller dem derfor sampel random variabler. Har vi r sampler og har beregnet gjennomsnittet for de n observasjonene i hvert sampel, har vi r observasjoner av en ny random variabel, nemlig $\bar{x}$. Gjentakene er da her de enkelte samplene. På samme måte må vi også oppfatte middelavviket s som en sampel random variabel.

I Tab. E.1 er gjengitt gjennomsnitt ($\bar{x}_j$) og middelavvik (s_j) for r = 10 sampler, hvert på n = 25 gjentak. Samplene er tatt fra et univers hvor fordelingsfunksjonen for den observerte random variable er normal med E(x) = 10 og σ = 1.

Tabell E.1.

Sampel nr. j	$\bar{x}_j$	s_j
1	10,13	0,96
2	9,90	1,13
3	9,90	0,99
4	10,26	1,01
5	9,93	1,04
6	9,76	0,69
7	10,02	0,89
8	9,61	0,92
9	9,95	0,90
10	10,22	1,03
	99,78	9,56

De to gjennomsnittene er $\bar{\bar{x}} = 9,98$ og $\bar{s} = 0,96$. Vi kan også beregne middelavvikene for de $r = 10$ gjennomsnittene og de 10 middelavvik på vanlig måte. Vi finner da at

$$s^2_{\bar{x}} = \frac{1}{r-1} \Sigma(\bar{x}_j - \bar{\bar{x}})^2 = \frac{0,3656}{9} = 0,0406$$

og

$$s^2_s = \frac{1}{r-1} \Sigma(s_j - \bar{s})^2 = \frac{0,1264}{9} = 0,0140$$

I neste avsnitt skal vi vise at disse resultatene er omtrent de vi vil vente.

Vi kan også beregne korrelasjonskoeffisienten mellom $\bar{x}$ og s (se avsnitt B.9) og finner at $r = 0,12$. Vi ser at verdien av denne korrelasjonskoeffisienten er meget liten. Også det er et resultat vi skal vente oss i dette tilfelle.

Vi skal vise senere at vi i praksis har bruk for slike størrelser som $E(\bar{x})$ og $var(\bar{x})$, og at vi har bruk for estimatorer som f.eks. middelavviket $s_{\bar{x}}$. I praksis må vi nøye oss med det ene samplet vi har. Og konsekvensen av dette er at vi må ha en formel for f.eks. $var(\bar{x})$ som viser hvordan vi skal gå fram for å beregne $s_{\bar{x}}$ når vi har bare det ene samplet.

Vi skal vise at

$$var(\bar{x}) = \frac{var(x)}{n} = \frac{\sigma^2}{n}$$

hvor n er antall gjentak i det aktuelle samplet. I vårt eksempel i Tab. E.1 er $n = 25$ og $var(x) = \sigma^2 = 1$. Etter formelen er da $var(\bar{x}) = 1/25 = 0,04$. For de 10 observasjonene av $\bar{x}$ fant vi at $s^2_{\bar{x}} = 0,0406$, dvs. praktisk talt det vi skulle vente etter formelen.

E.2. Fordelingsfunksjonene for gjennomsnittet og variansen.

Vi vil nå tenke oss at vi har et random sampel på n gjentak og observasjonene x_i (i=1,2,3...n) av den random variable x.

Gjennomsnittet $\bar{x}$ og middelavvikets kvadrat $V = s^2$ må vi da som forklart foran, oppfatte som random variable, og vi er naturligvis interessert i hvilke fordelingsfunksjoner disse har.

Forutsetter vi nå at den observerte random variable x har normal fordelingsfunksjon med forventningen $E(x) = \mu$ og variansen $\text{var}(x) = \sigma^2$, kan det bevises* at $\bar{x}$ og s^2 er uavhengige random variable. Hva dette betyr skal vi komme nærmere inn på senere. Her må vi nøye oss med å si at det bl.a. betyr at korrelasjonskoeffisienten (ρ) mellom de to random variable er lik null.

Det kan videre bevises* at fordelingsfunksjonen for $\bar{x}$ er normal med

$$E(\bar{x}) = \mu \quad \text{og} \quad \text{var}(\bar{x}) = \frac{\text{var}(x)}{n} = \frac{\sigma^2}{n}$$

dvs. at

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

Når det gjelder middelavviket, kan det bevises at fordelingsfunksjonen for $V = s^2$ er en Gamma fordelingsfunksjon (se avsnitt D.2) nemlig

$$f(V) = K \, V^{\frac{1}{2}(n-3)} \, e^{-\frac{n-1}{2\sigma^2}V} \qquad (V \geq 0)$$

Sammenligner vi med formelen på side 87, ser vi at $m = \frac{1}{2}(n-3)$ og at $k = (n-1)/2\sigma^2$. Bruker vi så formlene for $E(x)$ og $\text{var}(x)$, finner vi at

$$E(V) = \sigma^2 \quad \text{og} \quad \text{var}(V) = \frac{2}{n-1}\sigma^4$$

Skriver vi $(n-1)V/\sigma^2 = \chi^2$ og $n-1 = f$, ser vi at fordelingsfunksjonen blir identisk med den fordelingsfunksjon for χ^2 som ble nevnt i avsnitt D.2. Det sies derfor ofte at $(n-1)V/\sigma^2$ er et kjikvadrat med $f = n-1$ frihetsgrader.

* Se Tillegg III.

La oss så se noe nærmere på fordelingsfunksjonen for $\bar{x}$. Siden den er normal, vil Q i Tab. D.4 (avsnitt D.2) være sannsynligheten for $\bar{x}$ mellom grensene

$$E(\bar{x}) - a\sqrt{var(\bar{x})} \quad \text{og} \quad E(\bar{x}) + a\sqrt{var(\bar{x})}$$

dvs. mellom grensene $\mu - a.\sigma/\sqrt{n}$ og $\mu + a.\sigma/\sqrt{n}$. Differensen mellom disse to grensene er lik $2a.\sigma/\sqrt{n}$. Settes a =3, er sannsynligheten for $\bar{x}$ mellom de to grensene lik Q = 0,9973, og differensen mellom grensene lik $6\ \sigma/\sqrt{n}$. Vi ser at differensen mellom de to grensene avtar med voksende n, dvs. at jo større samplet er, jo mindre avstand mellom grensene for $\bar{x}$. Er n et meget stort tall, vil de to grensene falle sammen omtrent. Samtidig er det praktisk talt sikkert (Q=0,9973) at $\bar{x}$ faller mellom disse grensene. Og det vil jo da si at $\bar{x}$ er praktisk talt lik forventningen $E(x) = \mu$.

Vi kunne ha gjennomført et tilsvarende resonnement for variansen V. Men fordelingsfunksjonen er i dette tilfelle ikke normal, den er bl.a. skjev. Vi måtte derfor bruke en forskjellig koeffisient a for den nedre og den øvre grensen. Resultatet ville imidlerid blitt det samme, nemlig at for voksende n vil grensene for V komme nærmere hverandre og tilslutt smelte sammen. For store sampler er derfor $V = s^2$ praktisk talt lik σ^2.

Det er her forutsatt at fordelingsfunksjonen for den observerte random variable (x) er normal. Er fordelingsfunksjonen ikke normal, vil både $\bar{x}$ og V ha andre fordelingsfunksjoner enn dem vi har referert foran. Det kan imidlertid bevises at uansett hvilken fordelingsfunksjon den observerte random variable har, er $E(\bar{x}) = \mu$ og $var(\bar{x}) = \sigma^2/n$. For variansen V har vi alltid at $E(V) = \sigma^2$, men formelen for var(V) er mer komplisert enn den vi har gjengitt foran.

Tillegg til side 108.

I avsnitt D.6 har vi funnet at hvis x_1 og x_2 er uavhengige og $y = x_1 - x_2$, er $\text{var}(y) = \sigma_1^2 + \sigma_2^2$ hvor $\sigma_1^2 = \text{var}(x_1)$ og $\sigma_2^2 = \text{var}(x_2)$.

La oss nå tenke oss at $\bar{x}_1$ er gjennomsnittet av n_1 observasjoner av en random variabel x i et sampel på n_1 gjentak i universet U_1. La tilsvarende $\bar{x}_2$ være gjennomsnittet av n_2 observasjoner av den samme random variable i et sampel på n_2 gjentak i universet U_2. Sett så

$$d = \bar{x}_1 - \bar{x}_2$$

Vi har at $\text{var}(\bar{x}_1) = \sigma_1^2/n_1$ og $\text{var}(\bar{x}_2) = \sigma_2^2/n_2$. Formelen for var(y) anvendt på dette tilfelle gir da at

$$\text{var}(d) = \text{var}(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Dette er en meget viktig formel som vi vil få bruk for senere.

E.3. Fordelingsfunksjonen for t.

Som i foregående avsnitt vil vi forutsette at den observerte random variable (x) har normal fordelingsfunksjon med forventningen μ og standardavviket σ. Vi vil så tenke oss at vi har observasjoner av x i et sampel på n gjentak. Gjennomsnittet og middelavviket betegnes som før med $\bar{x}$ og s.

La oss sette

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{\bar{x} - \mu}{s}\sqrt{n}$$

Siden både $\bar{x}$ og s er random variable, er også t en random variabel. Fordelingsfunksjonen for t kan lett utledes ved hjelp av fordelingsfunksjonene for $\bar{x}$ og V fra foregående avsnitt. Den er kjent under navnet Students fordelingsfunksjon* og er

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

hvor K er en konstant. Parameteren f har fått betegnelsen antall frihetsgrader. I det tilfelle vi har for oss her, er $f = n-1$. I andre tilfelle som vi skal ta for oss senere, bestemmes f på annen måte.

Variasjonsområdet for t strekker seg fra $-\infty$ til $+\infty$. Fordelingsfunksjonen er symmetrisk omkring $t = 0$. Den nærmer seg mer og mer til den normale fordelingsfunksjon når f vokser og faller sammen med denne når $f \to \infty$.

Oftest når en skal bruke denne funksjonen, har en ikke interesse av fortegnet for t, en er bare interessert i tallverdien

$$|t| = \frac{|\bar{x} - \mu|}{s}\sqrt{n}$$

På grunn av symmetrien omkring $t = 0$ vil fordelingsfunksjonen for $|t|$ være identisk med høyre halvdel av fordelingsfunksjonen

* Se Tillegg IV.

for t når alle ordinatene fordobles. I Fig.E.1 er tegnet inn grafen av funksjonen for det tilfelle at f =4. Arealet av den skraverte flaten i denne figuren er da lik sannsynligheten for $|t| \geq a$, altså sannsynligheten $P(|t| \geq a)$. For praktiske formål er verdien av a beregnet for valte verdier av P. Siden antall frihetsgrader (f) er en parameter i funksjonen, må verdien av a også avhenge av verdien av f. I Tabell I bak i denne boka er gjengitt en tabell over a for f = 1,2,3.... og $P(|t| \geq a)$ = 0,05, 0,025 og 0,01. Er f.eks. f = 10, og P = 0,05, er a = 2,228. Vi ser også at verdien av a synker for voksende verdi av f.

Fordelingsfunksjonen for t er utledet under den forutsetning at den observerte random variable har normal fordelingsfunksjon. Vi har sagt foran (avsnitt D.2) at den normale fordelingsfunksjon neppe kan betraktes som en realistisk modell. Hvis fordelingsfunksjonen for t hadde vært sterkt avhengig av den normale fordelingsfunksjon som forutsetning, ville konsekvensen bli at vi også måtte betrakte fordelingsfunksjonen for t som lite realistisk og dermed ubrukelig i de aller fleste sammenhenger. Det har imidlertid vist seg at fordelingsfunksjonen for t er meget robust i den forstand at fordelingsfunksjonen for den observerte random variable har lite å si*. Derfor har t og dens fordelingsfunksjon etter hvert fått en meget stor betydning i mange sammenhenger som vi skal komme inn på senere.

* Se Tillegg IV.

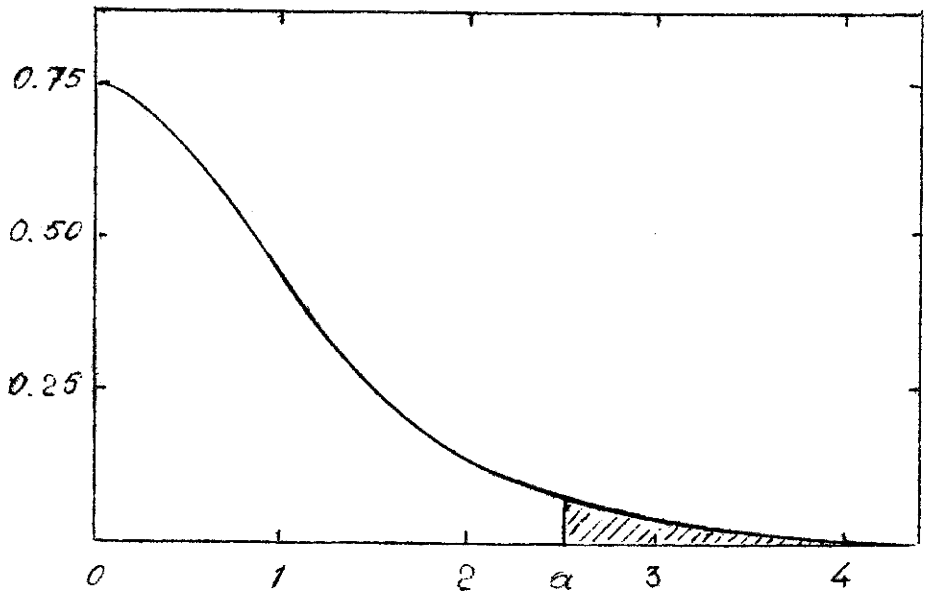


Fig. E.1