9 1973/109 Ex. 4

MATEMATISK STATISTIKK

Forelesninger ved Norges veterinærhøgskole

ved

Ivar Kristianslund

Revidert utgave

ISBN 82-557-0012-9

LANDBRUKSBOKHANDELEN

1432 ÅS-NLH

1973

Norges landbrukshøgskoles bibliotek

q1973/109

ex 4

MATEMATISK STATISTIKK

Forelesninger ved Norges veterinærhøgskole

ved

Ivar Kristianslund

Revidert utgave

ISBN 82-557-0012-9

LANDBRUKSBOKHANDELEN

1432 ÅS-NLH

1973

F o r o r d t i l 1 9 6 9 - u t g a v e n

Dette heftet er en noe utvidet gjengivelse av forelesninger som jeg holdt for veterinærstudenter i vårsemesteret 1969. Forelesningene bygger til dels på erfaringer fra et kurs i biostatistikk som jeg holdt for veterinærer ved Norges veterinærhøgskole høsten 1968. Framstillingen er i stor grad og på mange måter påvirket av prof. dr. Per Ottestads forelesningshefter for studenter ved Norges landbrukshøgskole da jeg selv har undervist etter disse hefter ved Landbrukshøgskolen i flere år. Symbolene som er brukt er stort sett de samme som i prof. Ottestads hefter. Et viktig unntak er at jeg har brukt understrekede symboler for random variable og de tilsvarende symboler uten understrekning for verdier av random variable.

Det er lagt stor vekt på å forsøke å gjøre framstillingen enkel og lettfattelig for lesere med små forkunnskaper i matematikk, men en har forsøkt å unngå å gjøre den upresis og flytende. Gjennomgåelsen av teorien er i stor utstrekning knyttet til eksempler fra veterinærmedisin og husdyrfag. På passende steder i teksten er det skutt inn øvelser som er ment å skulle utdype og festne stoffet etter hvert som studiene går fram.

Heftet ble skrevet etter håndskrevet manuskript direkte på stensil etter hvert som forelesningene ble holdt. Det ble derfor ikke anledning til noen endelig omredigering etter at alt var skrevet.

Da det p.g.a. forskjellige sammentreff ble nødvendig å bruke hele fem forskjellige personer til maskinskrivingen og da skrivingen foregikk under et visst tidspress, så en gjennom fingrene med en del uregelmessigheter ved skrivingen som ville ha blitt rettet ved en strengere korrektur.

Vollebekk, juni 1969

Ivar Kristianslund

F o r o r d t i l 1 9 7 3 - u t g a v e n

I denne utgaven er det føyd til et nytt hovedavsnitt om regresjon og et om kji-kvadrat test. Disse avsnittene er delvis en gjengivelse av et tillegg som jeg skrev i 1971 da timetallet ble noe utvidet.

Den nye utgaven bygger på de gamle stensilene, men noen sider er skrevet helt om, og det er foretatt atskillige forandringer på de gamle stensilene. Siktemålet har spesielt vært å gjøre stoffet så lettfattelig som mulig. For å hjelpe studentene til også å kunne lese andre lærebøker er det bl.a. føyd til en kort forklaring av visse symboler og begreper som brukes i mengdelæren.

Det er føyd til flere nye oppgaver der hvor det har vært ledig plass på slutten av et avsnitt. Disse oppgavene er gitt nummerne b, c osv. eller de er bare gitt betegnelsen "oppgave". Bakerst i heftet er det nå tatt med en facit til oppgavene.

I en ikke alt for fjern framtid håper jeg å få anledning til en mer fullstendig omskrivning av heftet enn det som har vært mulig innen rammen av de gamle stensilene. Jeg vil derfor være meget takknemlig for ethvert forslag til forbedringer.

Ås, april 1973

Ivar Kristianslund

I n n h o l d

I. Innledning

Ordet statistikk brukes ofte om en stor samling tall (jordbruksstatistikk, kriminalstatistikk, importstatistikk osv.). Det vi kaller matematisk statistikk (ofte forkortet til statistikk) har visse tilknytninger til statistikk i den ovenfor nevnte betydningen av ordet, men det er egentlig noe annet. Å gi en noenlunde presis definisjon av begrepet matematisk statistikk er [illegible]. Som en første tilnærmelse kunne vi si at matematisk statistikk er læren om hvorledes vi skal innrette oss når vi omgås tall. Denne definisjonen omfatter imidlertid store deler av matematikken og fagområder som bokholderi o.l. På den annen side er det viktige deler av statistikken som faller utenom en slik definisjon.

For å gi et inntrykk av hva matematisk statistikk er må vi nevne noe av det faget omfatter. En viktig arbeidsmåte i den matematiske statistikken er å ~~sette opp en~~ statistisk modell som viser hvorledes et tallmateriale kan tenkes framkommet. En slik statistisk modell er en tankekonstruksjon som kan uttrykkes eksplisitt ved hjelp av matematiske og statistiske begreper og symboler. Den statistiske modell er et metodisk hjelpemiddel til å oppdage og kvantifisere mer eller mindre faste regler og lovmessigheter i den virkelighet som omgir oss. I statistikken arbeides det med typer av modeller som er så generelle at de kan anvendes på en rekke helt forskjellige områder innen vitenskap og teknikk m.v. Statistisk teori og statistiske metoder kommer til anvendelse både når vi skal skaffe oss viten om vår verden og når vi skal ta praktiske avgjørelser på grunnlag av denne viten. Den matematiske statistikk utgjør således et arsenal av nyttige begreper og metoder som fagmannen kan forsyne seg av. Matematikk er et viktig hjelpemiddel innen statistikken,

men statistikken skiller seg likevel klart fra matematikken. Noe av det som særmerker den matematiske statistikken er at den anvendes på fenomener hvor det inngår et element av usikkerhet og hvor denne usikkerheten blir tatt direkte hensyn til i selve behandlingen av problemet. Noen spredte eksempler på slike fenomener er terningkast, forsikringsvirksomhet, kvalitetskontroll ved stikkprøver, politiske meningsmålinger, representative jordbrukstellinger, og sist, men ikke minst vitenskapelige forsøk av forskjellige slag.

La oss nå forsøke å gi en brukbar definisjon av begrepet matematisk statistikk.

Matematisk statistikk er en vitenskap som har å gjøre med problemer som oppstår og metoder som kan brukes når vi skal samle viten og treffe beslutninger under forhold hvor det hersker usikkerhet. Usikkerheten kommer som regel inn i bildet fordi vi er nødt til å bruke våre kunnskaper om et begrenset antall tilfelle på et større antall tilfelle som vi vet lite om på forhånd. Statistikken kan hjelpe oss til å innrette oss på en formålstjenlig måte i slike situasjoner.

Et eksempel vil hjelpe til å klarlegge vår definisjon. En gruppe veterinærer får i oppdrag å bekjempe en hittil ukjent husdyrsykdom. De undersøker da et begrenset antall dyr for å bli kjent med sykdommen. Videre foretar de en rekke eksperimenter med behandlingsmåter og medisiner, og endelig setter de i gang en praktisk bekjempelse av sykdommen ved hjelp av isolasjon, vaksinasjon, hygieniske tiltak e.l. På alle stadier i dette programmet vil det herske usikkerhet. Etter hvilket prinsipp bør en ta ut dyr til undersøkelse? Hvorledes bør eksperimentene legges opp og konklusjonene utledes av tallmaterialet? I

hvilken utstrekning er konklusjonene gyldige under praktiske forhold? Slike spørsmål blir behandlet systematisk i statistikken på et helt generelt grunnlag uten tilknytning til noe bestemt fagområde som veterinærmedisin, plantekultur, sosialøkonomi osv. Derfor er statistisk teori og statistiske metoder anvendelige i all empirisk forskning, dvs. forskning hvor vi samler erfaring om variable fenomener i den virkelighet som omgir oss.

Matematisk statistikk er i dag et meget stort og variert fagområde med flere spesialdisipliner. Faget er et uhyre viktig hjelpemiddel i naturvitenskapelig og samfunnsvitenskapelig forskning og på en rekke andre høyst forskjellige områder. Den egentlige matematiske statistikk bygger på sannsynlighetsregningen som har røtter helt tilbake til 1500-tallet. Det er imidlertid i vårt århundre at statistikken har fått en så voldsom vekst i omfang og betydning. Utviklingen av faget har i stor utstrekning skjedd i tilknytning til jordbruksforskningen. Fagets relative betydning vil sikkert fortsette å øke betraktelig i tiden framover, bl.a. på grunn av nye muligheter for elektronisk databehandling.

II. Noen statistiske begreper

Den virkelighet vi lever i er uhyre komplisert. De problemer vi får å løse ved hjelp av statistiske metoder er derfor ofte innfløkte og sammensatte. Som regel må vi derfor spalte opp et problemkompleks som det vi støtte på i eksemplet ovenfor i en rekke forholdsvis enkle delproblemer som lar seg beskrive ved hjelp av statistiske modeller. Ved behandlingen av et slikt statistisk problem får vi bruk for en rekke begreper som vi skal definere i det følgende.

Som tidligere nevnt er statistikken et hjelpemiddel til å skaffe oss viten. For å skaffe oss viten må vi foreta observasjoner, dvs. vi må bruke våre sanser og registrere egenskaper ved de objekter vi er interessert i. Som regel vil et viktig ledd i forskningsprosessen være å ta for seg et begrenset antall objekter av et bestemt slag og undersøke disse etter tur. Hvert av disse objekter blir kalt et gjentak eller en telleenhet. Skal vi f.eks. skaffe oss kunnskap om en hittil ukjent sykdom hos sau må vi ta for oss et antall sauer og undersøke hver av disse. Hver sau i dette eksemplet er et gjentak. Et gjentak behøver ikke å være et objekt. Ofte kan det bedre karakteriseres som et fenomen, en situasjon, et eksperiment, e.l. Skal vi undersøke om historien om sjøormen har noe for seg kan f.eks. hver ny rapport om at sjøormen er sett være et gjentak. Andre eksempler på gjentak er grisekull, skåler med en bakteriekultur av et bestemt slag, trekkprøver med en bestemt hingst, besetninger, land, veterinærbesøk, osv. Samlingen av gjentak som vi undersøker i forbindelse med et bestemt problem (f.eks. en samling på 30 sauer) blir kalt et sampel eller utvalg.

Det er hensiktsmessig å oppfatte samplet som et utvalg fra en større samling av gjentak som vi kaller universet eller populasjonen. Et sampel som består av 30 sauer kan f.eks. oppfattes som et sampel fra et univers som består av alle norske sauer. Alternativt kunne vi ha oppfattet universet i dette tilfelle som samlingen av alle nålevende sauer i hele verden, eller som alle sauer i sin alminnelighet. Hva vi vil oppfatte som univers når vi i et konkret tilfelle starter en undersøkelse er til en viss grad gjenstand for valg, men det får konsekvenser for hvorledes samplet bør tas ut. Målet for vår undersøkelse er å komme fram til utsagn, regler eller lovmessigheter som vi kan ha håp om er gyldige for hele universet, selv om de bare bygger på en undersøkelse av samplet. Vi kan altså si at universet er samlingen av alle de gjentak som vi tar sikte på at våre resultater skal gjelde for. Samplet, som alltid er en del av universet, er samlingen av alle de gjentak som vi faktisk undersøker. Som regel er det praktisk eller økonomisk umulig å undersøke hele universet, og det er nettopp dette forhold som gjør at vi får bruk for statistiske metoder.

Skjematisk kan relasjonen mellom gjentak, sampel og univers illustreres som i fig. 1.

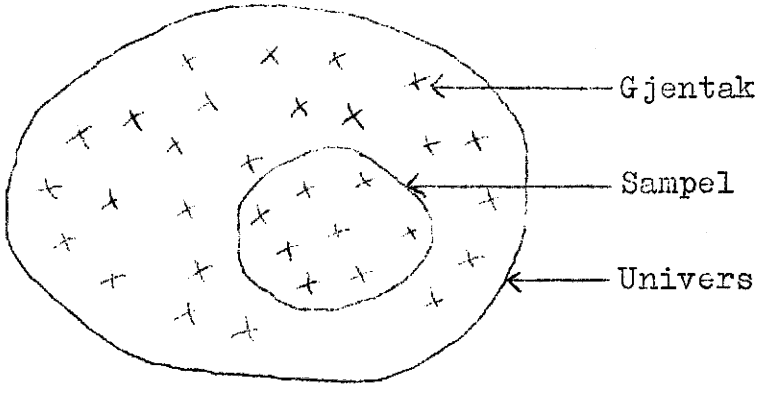


Fig. 1

Det kan stundom være meget vanskelig eller umulig å trekke opp konkrete grenser for et univers. Ofte er universet en abstrakt tankekonstruksjon som ikke har noe reelt motstykke. Likevel er universet et meget nyttig begrep som letter formuleringen og analysen av våre problemer. Hvis f.eks. gjentaket er en trekkprøve med en bestemt hingst, kan universet bestå av alle trekkprøver som kunne tenkes utført med denne hingsten under nærmere spesifiserte forsøksbetingelser (seletøy, sko, veidekke, osv. av en bestemt type).

Universet kan bestå av et begrenset eller et ubegrenset antall gjentak. Av og til blir ordet univers brukt om ubegrensede universer, mens ordet populasjon reserveres for begrensede universer, men språkbruken er noe varierende. Vi vil her holde oss til ordet univers for begge tilfelle. Når ikke noe annet går fram av sammenhengen vil de universene vi får å gjøre med i det følgende alltid være ubegrensede.

Undersøkelsen av hvert gjentak i samplet går ut på å observere og registrere et eller flere kjennetegn. Et kjennetegn er en karakteristikk av en kvantitativ eller kvalitativ egenskap som knytter seg til gjentaket. Kjennetegnet kan enten observeres direkte ved hjelp av syn, hørsel, lukt, smak eller følelse; eller indirekte ved bruk av mer eller mindre kompliserte instrumenter, apparater og metoder. Som eksempler på kjennetegn kan vi nevne "hann" når gjentaket er en forsøkskanin, "bolleformet jur" når gjentaket er en ku, "utemmet" når gjentaket er en hest, "spekktykkelse 25 mm" når gjentaket er en gris, "9 grisunger" når gjentaket er et grisekull, "all import av levende storfe forbudt" når gjentaket er et land, osv. Av og til bruker en betegnelsen begivenhet eller hendelse i stedet for kjennetegn når dette faller naturlig, språklig sett.

Et kjennetegn som naturlig kan uttrykkes ved hjelp av et

enkelt tall blir kalt et kvantitativt kjennetegn. Et kjennetegn som ikke kan uttrykkes på denne måte blir kalt et kvalitativt kjennetegn eller konstant kjennetegn. Kjennetegnet "røde øyne" hos gjentaket "kanin nr. 33" er et kvalitativt kjennetegn. Kjennetegnet "levendevekt 519 kg" hos gjentaket "328 Litago" er et kvantitativt kjennetegn. Kvalitative kjennetegn kan alltid uttrykkes som om de var kvantitative ved at vi velger et tall for hvert av dem. Øyenfarge kan f.eks. defineres ved at vi setter blå =1, brun = 2, rød = 3, osv. Vi skal senere se at et kvantitativt kjennetegn også kan oppfattes som en verdi av en såkalt random variabel.

Som regel knytter det seg en uendelighet av kjennetegn til hvert gjentak, men alle kjennetegn er ikke av samme interesse. Opplegget for en undersøkelse fører naturlig til en gruppering av kjennetegnene.

(1) Noen kjennetegn er felles for alle gjentak av den type som er gjenstand for vår undersøkelse i et gitt tilfelle. De er med andre ord felles for alle gjentak i universet, og universet kan defineres ved å regne opp alle disse kjennetegn. Hvis f.eks. vårt univers består av alle kuer, kan universet i prinsippet defineres ved at vi regner opp alle kjennetegn som er felles for alle kuer.

(2) Visse andre kjennetegn brukes undertiden til en videre klassifisering av gjentakene for analytiske formål. Sammen med de kjennetegn som definerer universet definerer disse kjennetegn et subunivers eller delunivers. I prinsippet kan vi f.eks. definere et subunivers av universet som omfatter alle kuer ved å regne opp alle de kjennetegn som er felles for alle kuer av rasen NRF.

(3) Det finnes også kjennetegn som er av interesse fordi de identifiserer det enkelte gjentak, eksempler er nummeret på

den enkelte ku, eierens navn og adresse, osv.

(4) Endelig har vi de kjennetegn som vår undersøkelse egentlig dreier seg om. Hvert slikt kjennetegn oppfatter vi som et alternativ eller en verdi (hvis det er et kvantitativt kjennetegn) av et sett av alternative kjennetegn. Kjennetegnet "blå øyne" er f.eks. et alternativ blant et sett av alternative kjennetegn som omfatter alle forskjellige øyenfarger som forekommer i vedkommende univers. Kjennetegnet "hankjønn" er et alternativ i et sett som også omfatter kjennetegnet "hunkjønn". Kjennetegnet "8 grisunger" når gjentaket er et grisekull er et alternativ eller en verdi i et sett som omfatter alle naturlige tall fra 0 og opp til, la oss si 20. Kjennetegnet "årlig melkeytelse 4120 kg" er en verdi i et sett som kanskje omfatter hele tall-linjen fra 0 og opp til de ytelser vi finner hos verdens rekordkuer. Vi ser altså at hvert sett av alternative kjennetegn inneholder de mulige alternativer av en egenskap som f.eks. øyenfarge, kjønn, kullstørrelse, melkeytelse, osv.

Kjennetegnene må være definert på en slik måte at hvert gjentak har ett og bare ett kjennetegn fra hvert sett av alternative kjennetegn som vi betrakter. Vi sier at de forskjellige kjennetegn fra samme sett av alternative kjennetegn utelukker hverandre, idet de ikke kan opptre hos ett og samme gjentak. Vi sier også at et gjentak må ha enten det ene eller det andre (eller det tredje, osv.) av dem. Når det finnes bare to alternative kjennetegn i ett sett, kalles disse motsatte kjennetegn. Når gjentaket er en voksen hest, er kjennetegnene "hingst", "hoppe" og "vallakk" alternative kjennetegn. Kjennetegnene "død og levende" er eksempler på motsatte kjennetegn.

Når et sett av alternative kjennetegn inneholder mer enn to alternativer er det alltid mulig å innrette seg slik at vi

kan operere med bare to motsatte kjennetegn ved å gruppere flere kjennetegn under en felles betegnelse. Dette er ofte praktisk og kan lette visse definisjonsproblemer. I stedet for å operere med forskjellige grader av sykdom, kan det f. eks. være hensiktsmessig å bare definere de to motsatte kjennetegnene "syk" og "ikke syk". Man står selvsagt fritt når det gjelder å velge hva man vil mene med "syk".

Når to kjennetegn fra to forskjellige sett av alternative kjennetegn kan opptre samtidig hos ett og samme gjentak, sier vi at de to kjennetegnene ikke utelukker hverandre. Et gjentak kan altså da ha både det ene og det andre kjennetegnet. Terminologien blir tilsvarende når vi har å gjøre med mer enn to kjennetegn fra mer enn to sett av alternative kjennetegn.

Hvis vi f.eks. betrakter et univers av kuer, kan kjennetegnet "NRF" og kjennetegnet "melkemengde 5120 kg siste regnskapsår" tenkes å opptre samtidig hos et gjentak. De utelukker hverandre altså ikke. Som vi ser, tilhører de to kjennetegnene forskjellige sett. Det ene settet omfatter alle raser, og det andre omfatter alle tenkelige melkemengder.

III. Matematisk sannsynlighet

Et helt grunnleggende begrep i statistikken er matematisk sannsynlighet. En sannsynlighet kan defineres på flere måter. Ofte brukes en aksiomatisk framstilling hvor en gjerne også tar i bruk mengdelære. I dette kurset skal vi definere matematisk sannsynlighet på en meget enkel og lettfattelig måte. Dette oppnår vi ved å tenke oss at vi kan ta for oss alle gjentakene i hele universet (selv om dette er ubegrenset) og undersøke hvert enkelt av dem. Når universet er ubegrenset er det selvsagt umulig å undersøke alle gjentak. Vår definisjon er derfor noe kunstig og heller ikke matematisk stringent. Dette spiller imidlertid ingen rolle for vår anvendelse av sannsynlighetsbegrepet. Sannsynligheten for kjennetegnet E_1 i et univers som vi betegner med U skrives som $P(E_1|U)$ og er lik den brøkdelen av alle gjentakene i universet som har kjennetegnet E_1.
Vi har altså

$$(1) \qquad P(E_1|U) = \frac{\text{Antall gjentak med } E_1 \text{ i } U}{\text{Antall gjentak i alt i } U}$$

Sannsynligheten for kjennetegnet "oksekalv" hos en nyfødt kalv kan vi altså, teoretisk sett, finne ved å notere kjønnet for alle kalver og dividere antall oksekalver med antall kalver i alt. Siden en sannsynlighet er en brøkdel av gjentakene i universet innser vi lett at den må være et tall mellom 0 og 1.

Eksemplet ovenfor viser vel umiddelbart at summen av sannsynlighetene for to motsatte kjennetegn er lik 1. Har vi et tilfelle med mer enn to alternative kjennetegn, innser vi på tilsvarende måte at summen av sannsynlighetene for et sett av alternative kjennetegn, når vi tar med alle alternativene, er lik 1. Summen av den brøkdelen av gjentakene som har kjennetegnet "hoppe" og den brøkdelen som har kjennetegnet "hingst" og

den brøkdelen som har kjennetegnet "vallakk" i et univers av voksne hester er selvsagt lik 1.

Tar vi for oss mer enn ett, men ikke alle, kjennetegn fra et sett av alternative kjennetegn, ser vi at sannsynligheten for enten det ene eller det andre eller det tredje, osv. av disse kjennetegn er lik summen av sannsynlighetene for de enkelte kjennetegn. Således er sannsynligheten for enten "hingst" eller "vallakk" i et univers av voksne hester lik summen av sannsynligheten for "hingst" og sannsynligheten for "vallakk". Når vi tenker på sannsynlighetene som brøkdeler av gjentakene i universet er dette innlysende. Den setningen vi nettopp har referert blir kalt enten-eller setningen. Legg merke til at denne enten eller-setningen har å gjøre med kjennetegn som alle tilhører samme sett av alternative kjennetegn, altså for kjennetegn som utelukker hverandre. (Det finnes en mer generell form av enten-eller setningen som gjelder også for kjennetegn som kan opptre samtidig.)

Vi kan også være interessert i sannsynligheten for en kombinasjon av kjennetegn hvor kjennetegnene tilhører forskjellige sett av alternative kjennetegn, f.eks. i sannsynligheten for "hingst" av "vestlandsrase". Kjennetegnet "hingst" tilhører et sett av alternative kjennetegn som også omfatter "hoppe" og "vallakk", mens kjennetegnet "vestlandsrase" tilhører et annet sett, nemlig et sett som omfatter hver enkelt av de aktuelle raser. Definisjonen av sannsynlighet er den samme som før. Følgelig tenker vi oss at vi teller opp alle hingster av vestlandsrase i hele universet og dividerer med antall gjentak i hele universet.

Av eksemplet ser vi også at det kan dannes nye sett av alternative kjennetegn med utgangspunkt i gitte sett. Vårt nye sett omfatter foruten "hingst av vestlandsrase" også "hoppe av vestlandsrase", "hingst av østlandsrase", osv. idet vi kombinerer alle alternativene av de to opprinnelige settene. Summen av sannsynlighetene for alle alternativene i det nye settet vil også være lik 1. Hvis vi skal kombinere mer enn to sett, gjøres dette på tilsvarende måte.

Eksemplet viser at vi må være omhyggelige med å presisere hvilket univers vi tenker på når vi snakker om en sannsynlighet. I eksemplet kan jo universet bestå av alle hester, alle hingster, alle norske hester, osv. Hvis U er universet av alle hester, E_1 = "hingst", E_2 = "norsk" og E_3 = "vallakk", bruker vi følgende skrivemåte.

$P(E_1|U)$ = Sannsynligheten for "hingst" i hele universet av voksne hester.

$P(E_1|UE_2)$ = Sannsynligheten for "hingst" i universet av norske hester.

$P(E_1E_2|U)$ = Sannsynligheten for både "hingst" og "norsk" i hele universet. (Dvs. den brøkdel som norske hingster utgjør av alle hester).

$P(\text{enten } E_1 \text{ eller } E_3|U)$ = Sannsynligheten for enten "hingst" eller "vallakk" i hele universet. (Dvs. den brøkdel som hester av hankjønn utgjør av alle hester).

$P(E_1|U\text{i}E_2)$ = Sannsynligheten for "hingst" i universet av ikke-norske (dvs. utenlandske) hester.

Sannsynligheten $P(E_1|UE_2)$ er en <u>betinget sannsynlighet.</u> Det er sannsynligheten for E_1 betinget av E_2. Dette er det samme som sannsynligheten for E_1 i det subunivers av U hvor alle gjentak har kjennetegnet E_2.

Den såkalte <u>både-og setningen</u> gjelder for kjennetegn (eller kombinasjoner av kjennetegn) som tilhører forskjellige sett av alternative kjennetegn. Det er jo bare kjennetegn fra forskjellige sett som kan opptre samtidig slik at det kan bli snakk om et både-og. Tar vi for oss to slike kjennetegn og gir dem betegnelsene E_1 og E_2 kan både-og setningen skrives på følgende måte:

(2) $$P(E_1E_2|U) = P(E_1|U)\, P(F_2|UE_1)$$

Denne setningen brukes bl.a. når vi kjenner to av de sannsynlighetene som inngår i likningen (2) og ønsker å finne den tredje direkte uten å gå veien om opptelling i universet.

Vi skal ikke gi noe formelt bevis for setningen, men overlater til leseren å illustrere den ved eksemplet i øvelse 1.

Øvelse 1.

La E_1 være kjennetegnet "østlending" og E_2 kjennetegnet "ikke røker". La universet U bestå av alle som for tiden er registrert som studenter ved NLH. Bruk brøkregningens regler og forviss deg om at både-og setningen gjelder for dette eksemplet. Lag om nødvendig et konkret talleksempel.

Øvelse 2.

Kan både-og setningen skrives på følgende måte?

(3) $$P(E_1E_2|U) = P(E_2|U)\cdot P(E_1|UE_2)$$

Hint: Studer symbolene på begge sider av likhetstegnet.

Hvis $P(E_2|UE_1) = P(E_2|U)$ får vi ved innsetting i (2):

(4) $$P(E_1E_2|U) = P(E_1|U)\cdot P(E_2|U)$$

Vi sier da at E_1 og E_2 er uavhengige kjennetegn. Begrepet uavhengighet i statistikken er uhyre viktig. Legg merke til at uavhengighet i dagligtalen kan bety noe annet.

Øvelse 3.

Lag et talleksempel hvor $P(E_2|UE_1) = P(E_2|U)$. Vis at $P(E_2|UE_1) = P(E_2|UiE_1)$. Formuler i ord hva som ligger i at to kjennetegn E_1 og E_2 er uavhengige.

Øvelse 4.

Forsøk å generalisere både-og setningen (2) og uavhengighetskriteriet (4) til å gjelde tre kjennetegn E_1, E_2 og E_3. (Generaliseringen kan føres videre til et vilkårlig antall kjennetegn).

Enten-eller setningen og både-og setningen må ikke blandes sammen. Den enkle enten-eller setningen vi har gjennomgått har å gjøre med kjennetegn (eller kjennetegnkombinasjoner) som tilhører et og samme sett av alternative kjennetegn (eller kjennetegnkombinasjoner). Både-og setningen gjelder kjennetegn (kjennetegnkombinasjoner) som kommer fra hvert sitt sett av alternative kjennetegn (kjennetegnkombinasjoner).

Øvelse 5.

La oss tenke oss at vi har et begrenset univers som består av 100 studenter og at vi betrakter to sett av motsatte kjennetegn hvorav det ene har å gjøre med alder og det andre gjelder ekteskapelig status. Studentene er fordelt på følgende måte:

	E_2 ="fylt 23 år"	iE_2="ikke fylt 23 år"	
E_1="gift"	15	5	20
iE_1="ugift"	40	40	80
	55	45	100

Skriv ned (som brøker) følgende sannsynligheter: $P(E_1|U)$, $P(E_2|U)$, $P(E_1|UE_2)$ $P(E_2|UE_1)$ $P(E_1E_2|U)$ og $P(E_2E_1|U)$. Illustrer riktigheten av både-og setningen skrevet på formen (2) og på formen (3). Er E_1 og E_2 uavhengige kjennetegn? Forsøk å lage en utvidet enten-eller setning som kan brukes til å finne følgende sannsynlighet: P(enten E_1 eller E_2 eller $E_1E_2|U$).

Øvelse 6.

Bytt ut tallene i øvelse 5 med følgende tall:

	E_2="fylt 23 år"	iE_2="ikke fylt 23 år"	
E_1="gift"	15	5	20
iE_1="ugift"	60	20	80
	75	25	100

Undersøk om følgende kjennetegn er uavhengige: (1) E_1 og E_2. (2) E_1 og iE_2. (3) iE_1 og E_2. iE_1 og iE_2. Kommenter resultatet og tallene i tabellen.

IV. Litt om anvendelsen av sannsynlighetsbegrepet.

La oss se på noen av de spørsmålene som melder seg i forbindelse med sannsynlighetsbegrepet. Vi har allerede forklart hva som menes med en sannsynlighet. Et nærliggende spørsmål er da: Hvordan finner vi tall for sannsynligheter? I denne sammenhengen er det viktig å skjelne mellom to forskjellige oppgaver. På den ene siden har vi den oppgaven å finne tall som vi, i hvert fall foreløpig, oppfatter som eksakt riktige. Denne oppgaven lar seg ikke alltid løse, men sannsynlighetsbegrepet er likevel nyttig fordi det gir et konkret uttrykk for hva vi egentlig er ute etter. I noen tilfelle løser vi oppgaven ved å postulere sannsynligheter. Vi bygger da på vår forhåndsviten om universet.

Øvelse 7.

Hvordan ville du postulere sannsynligheten for å få (a) "krone" ved kast med et gitt pengestykke (b) "seks" ved kast med en gitt terning (c) "galte" ved en tilfeldig trekning fra en grisebinge hvor det er 18 galter og 6 purker? Gjør rede for hvorledes du kom fram til postulatene. Prøv å formulere en regel for postulering av sannsynligheter.

Den andre oppgaven vi kan bli stilt ovenfor er å finne tall som vi ikke i noe tilfelle oppfatter som annet enn tilnærmingsverdier for de tilsvarende sannsynligheter. Det å finne slike tall kaller vi å estimere sannsynligheter. De tilnærmingsverdiene vi kommer fram til kaller vi estimater. Estimeringen bygger normalt på en undersøkelse av et sampel fra det universet vi er interessert i.

Øvelse 8

Hvordan ville du estimere de sannsynlighetene som er nevnt i øvelse 7? Prøv å formulere en regel for hvorledes samplet bør tas ut når vi skal estimere en sannsynlighet.

I mange tilfelle hvor vi kjenner visse sannsynligheter eksakt, blir vi stilt overfor den oppgaven å finne nye beslektede sannsynligheter. Enten-eller setningen og både-og setningen er da ofte til stor hjelp. En tredje meget nyttig formel er den såkalte binomialloven som vi skal gjennomgå senere.

Det er viktig å merke seg at en sannsynlighet er et utsagn om universet. Den sier hvor stor brøkdel av gjentakene i universet det er som har et bestemt kjennetegn eller en bestemt kjennetegnkombinasjon. Hvis det er universet som interesserer oss (og det er det som regel i forskningen), er kjennskapet til sannsynligheter eller estimater av sannsynligheter av umiddelbar nytte. Vi skal imidlertid merke oss at en sannsynlighet også har visse implikasjoner for et sampel, ja endog for et enkelt gjentak (som er et spesialtilfelle av et sampel). Vi kan altså trekke slutninger både fra et sampel til universet (estimering) og den andre veien fra universet til et nytt sampel (prediksjon).

La oss ta et eksempel. Sett at sannsynligheten for at en veterinærstudent er oppvokst på landet er 2/3. Denne sannsynligheten er et utsagn om universet av veterinærstudenter. Jeg står overfor et sampel av veterinærstudenter (f.eks. en klasse på 30 studenter) og får i oppdrag å forutsi (lage en prediksjon om) hvor mange av disse det er som er oppvokst på landet. Under visse forutsetninger,(nemlig at studentene utgjør et random sampel som er definert på neste side), vil jeg da gjette på 20. Tilsvarende, hvis jeg står overfor en enkelt student vil jeg kanskje uten videre gjette på at han er oppvokst på landet. Under samme forutsetning vil jeg da i det lange løp kunne regne med å gjette riktig i 2 av 3 tilfelle.

Øvelse 9.

Ville det være en bedre strategi om jeg i det siste eksemplet, hvor jeg sto overfor en enkelt student, gjettet på at han var fra landet i 2 av 3 tilfelle? (Vi forutsetter altså at jeg blir stilt overfor en enkelt student gjentatte ganger.)

Et viktig spørsmål er hvilke betingelser et sampel må oppfylle forat en sannsynlighet skal ha visse klare implikasjoner for dette samplet slik som vist i eksemplet ovenfor. Svaret er at samplet må være et random sampel eller tilfeldig utvalg (random = tilfeldig). I prinsippet er det lett å definere et random sampel. Derimot er det ikke alltid så lett i praksis å skaffe seg et slikt sampel. Et random sampel er et sampel som tas ut på en slik måte at alle gjentakene i hele universet har like stor sannsynlighet for å komme med i samplet. Vi skal se på et eksempel. Hvis vi skal ta ut et random sampel på n gjentak (studenter) fra et univers som består av N gjentak, kunne vi i prinsippet gå fram på følgende måte: Gjentakene nummereres fra 1 til N. Dessuten nummererer vi N papirlapper fra 1 til N. Lappene legges i en beholder og blandes godt. Deretter trekker vi ut tilfeldig n lapper. Gjentakene med de tilsvarende nummer utgjør samplet.

I det hele tatt skaffer vi oss ofte et random sampel ved en eller annen form for loddtrekning (randomisering). Populært uttrykt kan vi si at randomiseringen er et middel vi bruker til å kunne stille oss helt "upartiske" ved uttakingen av samplet. Alle gjentak i hele universet får samme "sjanse" til å bli med i vårt sampel. Derved kan vi ha best mulig håp om at samplet i en viss forstand "representerer" eller "likner" universet. Dette gir oss igjen et grunnlag for å våge å anvende det vi vet om

universet på samplet eller omvendt, alt etter hva oppgaven er i det enkelte tilfelle.

La oss illustrere betydningen av å arbeide med random sampler ved igjen å ta for oss eksemplet med veterinærstudenter. Hvis jeg står overfor en klasse veterinærstudenter, er det mulig at disse kan oppfattes, i hvert fall tilnærmet, som et random sampel fra hele universet av veterinærstudenter. Det kan imidlertid også hende at dette samplet ikke er et random sampel fra dette universet. Hvis f.eks. faget som det undervises i er valgfritt, kan det tenkes at dette faget ikke øver samme tiltrekning på alle kategorier av studenter. Samplet må da kanskje oppfattes som et sampel fra et subunivers av veterinærstudenter, f.eks. et subunivers som består av studenter som kan tenke seg å velge dette faget. Hvis faget øver helt ulik tiltrekning på landsungdom og byungdom, vil samplet ikke kunne brukes til å estimere sannsynligheten for kjennetegnet "oppvokst på landet" i hele universet av veterinærstudenter. Heller ikke ville en kunne løse den motsatte oppgaven, nemlig å forutsi hvor mange av studentene i klassen som hadde kjennetegnet "oppvokst på landet" i en situasjon hvor sannsynligheten for dette kjennetegnet i hele universet av veterinærstudenter var kjent. Vanskeligheter av denne art må en alltid søke å ta hensyn til i forskningsarbeidet og i den praktiske anvendelsen av statistikken.

Hvis et random sampel består av bare et eneste gjentak, snakker vi om et tilfeldig gjentak. Som nevnt kan en sannsynlighet ha visse implikasjoner også for det enkelte gjentak. Riktignok kan vi si om et enkelt gjentak at enten har det et bestemt kjennetegn, eller så har det ikke dette kjennetegn. Hvorfor skal vi da trekke inn en sannsynlighet? Svaret er at vi kan være interessert i å lage en prediksjon. Det kan hende at vi ikke har

undersøkt gjentaket enda eller at kjennetegnet hittil ikke har gitt seg observerbare utslag. Forat en sannsynlighet skal gi grunnlag for en slik prediksjon, bør gjentaket være et tilfeldig gjentak.

La oss se på et eksempel. Sett at det var kjent at 1/3 av alle norske menn dør av hjertesykdommer. Sannsynligheten på 1/3 er et utsagn om hele universet av norske menn og den sier ingen ting om hva jeg, som er et gjentak i dette universet, kommer til å dø av. Hvis jeg kan oppfatte meg selv om et tilfeldig gjentak fra dette universet, har imidlertid sannsynligheten visse implikasjoner også for meg, selv om den ikke sier noe som helst sikkert om hvorledes det vil gå meg.

Sett at jeg hele mitt liv har sørget for riktig kosthold og passende mosjon. Da kan jeg neppe oppfatte meg selv som et tilfeldig gjentak. I stedet må jeg trolig oppfatte meg selv som et gjentak fra et subunivers av norske menn. I dette subuniverset kan den nevnte sannsynligheten være mindre enn 1/3.

Vi har tidligere vært inne på vanskelighetene ved å avgrense et univers. La oss til slutt nevne at det i mange tilfelle kan være hensiktsmessig uten videre å oppfatte det samplet en har som et random sampel og å definere universet ved hjelp av samplet. En sier da at universet er det universet som samplet representerer i egenskap av et random sampel. En slik definisjon har vært brukt av professor Ottestad. (P.Ottestad, Matematisk Statistikk. Forelesninger ved Norges Landbrukshøgskole. Oslo-Vollebekk 1962, s. 32.) Da universet bare er et tankemessig hjelpemiddel er det ofte både vanskelig og lite påkrevet å konkretisere universet ytterligere.

Hvis gjentaket er et eksperiment, består universet av et ubegrenset antall eksperimenter utført etter samme oppskrift. En rekke eksperimenter utført etter oppskriften vil da være et random sampel.

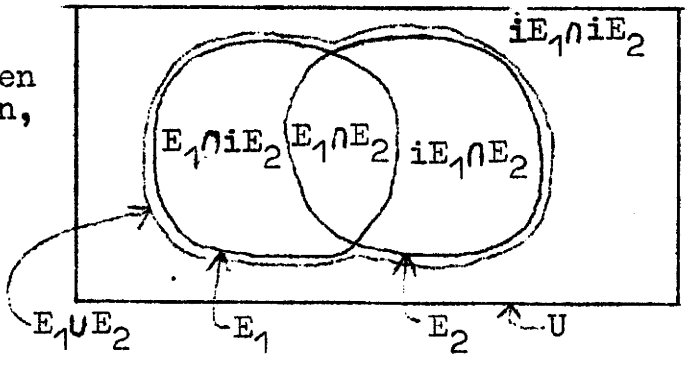
Vi vil nå vise sammenhengen mellom vår framstilling og visse symboler og begreper i mengdelæren. (Resten av dette hovedavsnittet kan overspringes uten at det går ut over sammenhengen.) Selv om vi ikke har definert en sannsynlighet som en mengdefunksjon, er det ikke noe i veien for at vi kan bruke mengdelærens symboler. Sannsynlighetsregningens regler blir de samme i alle tilfelle.

Til universet, U svarer mengdelærens begrep, <u>den universale mengden</u> som vi vil gi samme betegnelse, U. Til hvert kjennetegn som forekommer i universet svarer det en <u>delmengde</u> av den universale mengden, nemlig mengden av alle gjentakene i universet som har kjennetegnet. Vi vil her bruke samme symbol for et kjennetegn og den delmengden som svarer til kjennetegnet.

Sannsynligheten $P(E_1E_2|U)$ kan med mengdelærens symboler skrives som $P(E_1 \cap E_2)$. Her står $E_1 \cap E_2$ for <u>snittet</u> av mengdene E_1 og E_2. Et annet nyttig begrep fra mengdelæren er <u>unionen</u> av to mengder E_1 og E_2. Sannsynligheten for denne skrives $P(E_1 \cup E_2)$ som betyr P(enten $E_1 iE_2$ eller iE_1E_2 eller $E_1E_2|U$).

Nedenfor er vist et såkalt <u>Venn diagram</u> hvor den universale mengden er representert ved et rektangel og mengdene E_1 og E_2 ved sirkler. Slike diagrammer er spesielt nyttige når vi betrakter kjennetegn som framkommer ved å kombinere mer enn to sett av motsatte kjennetegn.

<u>Oppgave</u>: Prøv å illustrere den utvidede enten-eller setningen, $P(E_1 \cup E_2)=P(E_1)+P(E_2)-P(E_1 \cap E_2)$ ved hjelp av Venn diagrammet (Euler diagrammet) på denne siden.



V. Binomialloven

Vi skal ta for oss en viktig formel i sannsynlighetsregningen som kan utledes ved hjelp av både-og setningen og enten-eller setningen.

Sett at sannsynligheten for kjennetegnet E i universet U er lik P(E|U)=p. Sannsynligheten for det motsatte kjennetegnet, iE vil vi betegne med q. Altså er P(iE|U)=q=1-p.

La oss tenke oss at vi tar ut et random sampel på n gjentak fra universet U. Hva er da sannsynligheten for at X av disse gjentakene har kjennetegnet E og at altså de øvrige n-X gjentak har kjennetegnet iE? Binomialloven gir oss svaret på dette spørsmålet direkte uten at vi behøver å gå veien om både-og setningen og enten-eller setningen. Formelen som uttrykker binomialloven er viktig fordi vi ofte støter på problemer av denne type. Den kan også brukes som en fordelingsfunksjon, noe vi kommer tilbake til senere.

Vi skal gjennomgå et eksempel for å antyde hvorledes formelen kan bevises.

La oss betrakte et univers, U som består av alle storfefødsler. La E stå for kjennetegnet "oksekalv" og iE for "kvigekalv". La sannsynligheten for oksekalv være p=0,52. Sannsynligheten for kvigekalv blir da q=0,48. Anta at vi kan skaffe oss et random sampel som består av n=4 storfefødsler. Hva er da sannsynligheten for at X=3 av de 4 kalvene (fødslene) skal ha kjennetegnet "oksekalv"?

Den søkte sannsynligheten kunne vi skrive som P(X=3|Un=4). Vi skal først forklare hva denne sannsynligheten betyr. Fra universet U kan det tas ut en uendelighet av forskjellige

sampler på n=4 gjentak (fødsler). La oss nå definere et nytt univers Un=4 hvor hvert gjentak er et sampel på 4 fødsler fra universet U. Universet Un=4 består altså av alle mulige forskjellige sampler på 4 gjentak som kan tas ut fra universet U. Sannsynligheten P(X=3|Un=4) refererer seg til universet Un=4. Den kan tolkes som den brøkdel av alle sampler på n=4 gjentak fra universet U som har kjennetegnet "X=3".

Det vil lette framstillingen om vi tenker oss at et random sampel på 4 fødsler fra U tas ut på følgende måte: Først tar vi for oss fødsel nr 1 (som altså er en tilfeldig fødsel) og undersøker om resultatet er oksekalv eller kvigekalv. Deretter tar vi for oss fødsel nr. 2 (som også er en helt tilfeldig fødsel fra U), osv. Vi vil forutsette at sannsynligheten for at fødsel nr. 2 har kjennetegnet "oksekalv" (hva slags brøkdel uttrykker denne sannsynligheten?) er den samme, nemlig p=0,52, uansett om fødsel nr. 1 hadde kjennetegnet "oksekalv" eller ikke. Av det vi har lært i øvelse 3, s. 13 går det da fram at kjennetegnene "oksekalv ved fødsel nr. 1" og "oksekalv ved fødsel nr. 2" er uavhengige kjennetegn. Av dette følger det også at f.eks. kjennetegnet "kvigekalv ved fødsel nr. 2" og "oksekalv ved fødsel nr.1" er uavhengige. (Se øvelse 6, s. 14.) Vi summerer de forskjellige uavhengigheter ved å si kort at fødsel nr. 1 og fødsel nr. 2 er uavhengige. Tilsvarende uavhengighet kan vises å gjelde for alle par av fødsler blant de fire. Vi kan uttrykke dette kort ved å si at de enkelte fødslene er uavhengige. (Uavhengigheten er altså en følge av forutsetningen om at sannsynligheten p er den samme ved alle fødsler. Hvis vi hadde startet med å <u>forutsette</u> uavhengighet, måtte vi derimot i tillegg også ha forutsatt konstant p, da uavhengighet ikke impliserer samme p ved hver ny fødsel.)

At vår forutsetning kan ventes å være oppfylt følger av det faktum at vi opererer med et <u>ubegrenset</u> univers og et <u>begrenset random</u> sampel. "Beholdningen" av f. eks. oksekalver minker ikke selv om vi skulle slumpe til å få mange oksekalver etter hverandre.

Øvelse 10.

Tenk deg et endelig univers, f.eks. en kalvebinge som inneholder 13 oksekalver og 12 kvigekalver og vis at forutsetningen ovenfor ikke er oppfylt når vi trekker ut et random sampel av kalver fra denne bingen. (Den er oppfylt hvis vi slipper hver kalv tilbake i bingen igjen før vi foretar neste trekning).

Vi kan si at binomialloven gjelder for uttak av et random sampel fra et uendelig univers. For et endelig univers (øvelse 10, første del) har vi en tilsvarende lov, den hypergeometriske lov som vi ikke skal komme inn på. Hvis vi hele tiden tenker på et endelig univers (øvelse 10) kan vi si at den hypergeometriske lov gjelder for trekning uten tilbakelegging og binomialloven gjelder for trekning med tilbakelegging.

La oss nå vende tilbake til vårt eksempel. Det resultatet vi var interessert i, nemlig 3 oksekalver og 1 kvigekalv kan vi få på 4 forskjellige måter, alt etter som om det er den 1.eller 2. eller 3. eller 4. kalven vi trekker ut som er kvigekalv. Lar vi "o" stå for oksekalv og "k" stå for kvigekalv, kan de mulige resultatene skrives som kooo okoo ooko og oook. Bruker vi nå vår forutsetning samt en utvidet form av både-og setningen (øvelse 4) ser vi at sannsynlighetene for hvert av resultatene kan skrives som henholdsvis qppp pqpp ppqp og pppq. Her er f.eks. qppp sannsynligheten for å få både kvigekalv ved den første fødselen og oksekalv ved den andre og oksekalv ved den tredje og oksekalv ved den fjerde fødselen. Vi ser at sannsynligheten er den samme, nemlig p^3q for alle de fire typer av sampler som inneholder 3 oksekalver og en kvigekalv. Vi er interessert i å få 3 okser og 1 kvige på enten den

første måten eller den andre måten eller den tredje måten eller den fjerde måten. I følge enten-eller setningen blir dette $p^3q+p^3q+p^3q+p^3q=4p^3q$ og dette er svaret på vår oppgave.

Vi ser at resultatet, $4p^3q$ også kan skrives som $4p^Xq^{n-X}$. Vi merker oss også at tallet 4 står for antall måter et sampel på 3 okser og 1 kvige kan bli trukket ut på.

Hvis vi nå tar for oss det generelle tilfelle hvor n og X er hvilke som helst tall som er forenlige med problemstillingen, kan det vises at sannsynligheten, P_X for å få X gjentak med E og n-X gjentak med iE i et random sampel på n gjentak kan skrives på følgende måte:

$$(5) \quad P_X = \binom{n}{X}p^Xq^{n-X} = \frac{n\,!}{X!(n-X)\,!}\,p^Xq^{n-X}$$

Vi ser at leddet p^Xq^{n-X} er det samme som i vårt spesielle tilfelle. Leddet $\binom{n}{X}$ som leses "n over X" er et matematisk symbol som også kan skrives som $\frac{n!}{X!(n-X)!}$. Her er igjen n! et matematisk symbol som står for produktet av alle naturlige tall fra 1 til n. (Dvs. n! = 1.2.3....n.)

$\binom{n}{X}$ er det generelle uttrykk for antall måter vi kan få X gjentak med E og n-X gjentak med iE på i et random sampel på n gjentak.

Øvelse 11

(a) Bruk tallene i vårt eksempel (p=0,52, q=0,48, n=4 og X=3) og regn ut P_X etter formelen (5).

(b) Gjenta utregningen for hver av de øvrige verdier av X som kan forekomme når n=4 (X=0, X=1, X=2 og X=4).

(c) Summer verdiene av P_X for alle de 5 verdiene av X og kommenter resultatet.

Øvelse 12.

En meieristudent påstår at han kan skille melk fra meieri A fra melk fra meieri B på smaken. Vi gir ham i tre dager på rad et melkeglass som han ikke vet hvor er fra. Hver dag avgjør vi på tilfeldig måte om melken skal anskaffes fra meieri A eller fra meieri B. Sett at studenten i alle tre tilfelle greier å bestemme hvilket av de to meieriene melken er fra etter å ha smakt på den. Hva er sannsynligheten for et slikt resultat hvis vi forutsetter at han driver med ren gjetning?

Øvelse 13.

En pølsefabrikant har grunn til å anta at 90% av de middagspølser fabrikken produserer inneholder mindre enn 6% stivelse. For å undersøke om denne antakelse holder stikk ble det tatt ut tilfeldig 5 pølser til analyse, og av disse var det 2 som inneholdt minst 6% stivelse.

Hva er sannsynligheten for å få et slikt observasjonsresultat hvis pølsefabrikantens antakelse er riktig?
Forutsetninger: Vi ser bort fra alle praktiske problemer som er forbundet med samplingen og den kjemiske analysen.

Øvelse 14.

Sannsynligheten for at en hund som ikke vaksineres mot hvalpesyke skal få denne sykdommen (før eller senere) er 1/4 mens sannsynligheten for at en hund som er vaksinert skal få sykdommen er 1/9. Sannsynligheten for at en hund som får hvalpesyke skal dø er 1/3 (uansett om den er vaksinert eller ikke).

En mann eier to hunder som er stasjonert på to forskjellige steder slik at faren for smitte mellom dem ikke er annerledes enn mellom hunder i sin alminnelighet. Den ene er vaksinert, den andre ikke.

a. Hva er sannsynligheten for at begge hunder skal dø av hvalpesyke?

b. Hva er sannsynligheten for at både den hunden somer vaksinert skal dø av hvalpesyke og at den som ikke er vaksinert ikke skal få hvalpesyke?

c. Hva er sannsynligheten for at minst en av hundene skal dø hvalpesyke?

VI. Random variable.

Vi skal nå forklare hva som menes med en random variabel. Videre skal vi gjennomgå flere beslektede begreper og se på noen eksempler.

En random variabel er i mange lærebøker definert på en noe abstrakt måte ved hjelp av mengdelære. Vi skal her forsøke å gi en enklere definisjon.

La oss først undersøke hva som er forskjellen på en variabel i matematikken og en random variabel i statistikken.

En variabel i matematikken kunne vi f.eks. definere som en størrelse som kan anta visse verdier innenfor et sett av verdier. En variabel kan f.eks. anta alle verdier fra 0 til + ∞ . I matematikken sier vi imidlertid ikke noe om hvor sannsynlige de forskjellige verdiene av den variable er.

En random variabel i statistikken skiller seg fra en variabel i matematikken ved at vi ikke bare spesifiserer hvilke verdier den variable kan anta (et sett av alternative verdier), men vi har også spesielt i tankene at de ulike verdiene opptrer med forskjellig sannsynlighet. Disse sannsynlighetene behøver ikke alltid å være kjente og spesifiserte, men de inngår i alle tilfelle som en viktig del av begrepet en random variabel.

Vi skjelner mellom to slags random variable. En diskret random variabel kan bare anta visse atskilte verdier på tall-linjen. Et eksempel på en diskret random variabel er resultatet (utkommet) ved et terningkast. Denne random variable kan anta verdiene 1, 2, 3, 4, 5, og 6, altså seks atskilte verdier på tall-linjen. (Til hver verdi knytter det seg en sannsynlighet.) Antall oksekalver i et random sampel på fire storfefødsler er et

annet eksempel på en diskret random variabel. Denne kan anta verdiene 0, 1, 2, 3 og 4. (Til hver verdi knytter det seg en sannsynlighet som under visse forutsetninger kan regnes ut etter binomialloven som antydet i øvelse 11).

En kontinuerlig random variabel kan (i prinsippet om ikke i praksis) anta alle verdier på tall-linjen eller på et intervall av tall-linjen. Vekten av en gris som er et tilfeldig gjentak fra et uendelig univers av griser kan oppfattes (tilnærmet) som en kontinuerlig random variabel. Denne random variable kan i prinsippet anta alle verdier på tall-linjen fra 0 og opp til vekten av de største griser i universet. Et annet eksempel på en kontinuerlig random variabel er melkeytelsen til en tilfeldig ku fra et uendelig univers av kuer.

En random variabel er forbundet med et bestemt univers av gjentak (f.eks. universet av alle terningkast med en bestemt terning). Hvert gjentak har et bestemt kvantitativt kjennetegn (f.eks. kjennetegnet "6") fra et bestemt sett av alternative kvantitative kjennetegn (f.eks. settet 1, 2, 3, 4, 5 og 6). Hvert gjentak har altså ett og bare ett kjennetegn fra dette settet. (Vi er i denne sammenheng ikke interessert i kjennetegn fra andre sett.) Foreløpig vil vi tenke oss at vårt sett består av et begrenset (endelig) antall kjennetegn. Dette svarer til at vi bare betrakter diskrete random variable. Senere skal vi utvide vår definisjon til også å gjelde kontinuerlige random variable. Til hvert kjennetegn i vårt sett av alternative kjennetegn knytter det seg en bestemt sannsynlighet i det universet vi betrakter slik at summen av sannsynlighetene for alle kjennetegnene i settet er lik 1. (Vi kunne f.eks. tenke oss at det til settet 1, 2, 3, 4, 5 og 6 svarer sannsynlighetene 1/6, 1/6, 1/6, 1/6, 1/6, og 1/6).

Med en diskret random variabel mener vi et slikt sett av alternative kvantitative kjennetegn som det knytter seg bestemte sannsynligheter til. Hvert enkelt kvantitativt kjennetegn i settet av alternative kvantitative kjennetegn blir kalt en verdi av den random variable.

Det er uhyre viktig å få helt klart for seg hva en random variabel er og å kunne skjelne den fra en variabel i matematikken. Når vi snakker om en random variabel tenker vi på alle dens verdier under ett idet vi samtidig erkjenner at hver verdi har sin bestemte (men ofte ukjente) sannsynlighet.

Rent praktisk kan vi si at verdien av en random variabel gjerne er et resultat av en telling, veiing eller måling av et eller annet slag. Hvert fenomen som tellingen, veiingen eller målingen knytter seg til oppfattes som et (tilfeldig) gjentak i et univers.

Vi skal se på noen eksempler på random variable.

Eks. 1. Som nevnt er resultatet av et terningkast en random variabel. Hvis terningen er riktig avbalansert, kan vi lage følgende oppstilling:

Verdier av den random variable (d.v.s. sett av alternative kvantitative kjennetegn):	1	2	3	4	5	6
Tilhørende sannsynligheter:	1/6	1/6	1/6	1/6	1/6	1/6

Universet kan her være et univers av alle mulige kast med denne terningen. (Det kunne også være universet av alle mulige kast med alle riktig avbalanserte terninger.) Gjentaket er det enkelte terningkast med denne terningen (eller med en riktig avbalansert terning).

Den random variable er antall øyne ~~som kommer opp~~ idet vi underforstår at det til hvert antall øyne knytter seg en sannsynlighet. Vi kan kanskje tillate oss å si at ordet variabel har forbindelse med tallene 1, 2, 3, 4, 5 og 6 mens ordet random er en følge av at vi opererer med de seks sannsynlighetene på 1/6.

Begrepet en random variabel er noe helt nytt i forhold til det vi er vant til fra matematikken. Det er ikke lett å få en klar forestilling om hva det er uten å se for seg en hel tabell (eller figur) som i eksemplet ovenfor.

I litteraturen brukes ofte samme symbol for en random variabel som for verdien av en random variabel. Det blir overlatt til leseren å avgjøre av sammenhengen hva forfatteren har i tankene. For den som skal lære faget er det absolutt nødvendig å forstå forskjellen på de to begreper. I nyere lærebokslitteratur er det vanlig å bruke f.eks. fete typer for random variable og de tilsvarende vanlige bokstaver for verdier av random variable (hvis verdiene ikke er spesifisert som tall). Det er litt omstendelig å bruke denne skrivemåten, men den gir klarhet i begrepene. Vi skal bruke den i alle sammenhenger hvor vi ønsker å oppnå stor presisjon i framstillingen. I stedet for fete bokstaver skal vi da streke under symbolene.

Vi skal illustrere denne skrivemåten ved hjelp av eksemplet ovenfor. $\underline{X}$ står for den random variable, altså for tallsettet 1, 2, 3, 4, 5 og 6 idet vi underforstår at det til hvert tall knytter seg en sannsynlighet. X står for ett av tallene 1, 2, 3, 4, 5 og 6, men vi har enda ikke spesifisert hvilket.

Når vi tar for oss et enkelt gjentak fra universet, finner vi at den random variable har en bestemt verdi, X.

Vi bruker da skrivemåten $\underline{X} = X$, men merk at strengt tatt er det ikke noen likhet her idet $\underline{X}$ står for et sett av kjennetegn, mens X står for et enkelt kjennetegn.

Den random variable i eksemplet ovenfor kunne defineres på følgende måte:

$$P(\underline{X} = 1 \mid U) = \frac{1}{6} \qquad P(\underline{X} = 2 \mid U) = \frac{1}{6} \qquad P(\underline{X} = 3 \mid U) = \frac{1}{6}$$

$$P(\underline{X} = 4 \mid U) = \frac{1}{6} \qquad P(\underline{X} = 5 \mid U) = \frac{1}{6} \qquad P(\underline{X} = 6 \mid U) = \frac{1}{6}$$

Vi kunne også føye til:

$P(\underline{X} = X \mid U) = 0$ for alle andre X enn X=1, X=2, X=3, X=4, X=5 og X=6.

I mer sammentrengt form kunne vi skrive:

$P(\underline{X} = X) = \frac{1}{6}$ $X = 1, 2, 3, 4, 5$ og 6

$P(\underline{X} = X) = 0$ for alle andre X.

En annen skrivemåte er følgende:

$f(X) = \frac{1}{6}$ $X = 1, 2, 3, 4, 5$ og 6

$f(X) = 0$ for alle andre X.

I dette siste tilfelle har vi skrevet sannsynligheten for de forskjellige verdier av $\underline{X}$ som en funksjon av X (ikke av $\underline{X}$ for det er ikke mulig). I dette spesielle eksemplet er imidlertid funksjonen uhyre enkel, nemlig en konstant funksjon, altså egentlig ikke en funksjon av X.

Funksjonen f(X) blir kalt fordelingsfunksjonen for $\underline{X}$. Den kan selvsagt også framstilles grafisk, f.eks. som i fig. 2.

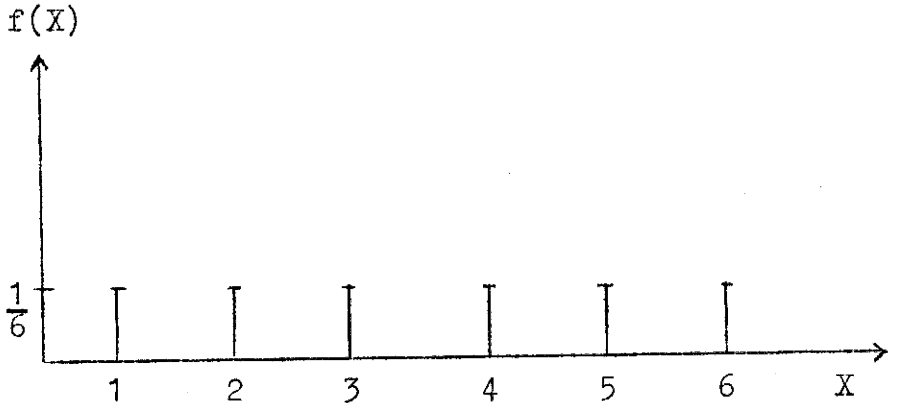


Fig. 2

Eks. 2. Antall oksekalver i et random sampel på 4 storfefødsler kan oppfattes som en random variabel. Vi kan f.eks. tenke oss et univers som består av alle tenkelige sampler på 4 gjentak fra et univers av storfefødsler (se s. 21-22). Hvert slikt sampel er et gjentak. Hvis sannsynligheten for "oksekalv" ved en enkeltfødsel er konstant lik 0,52 er det lett å regne ut sannsynligheten for alle mulige antall oksekalver i følge binomialloven. Vi kan da lage følgende oppstilling:

Verdier av den random variable (dvs. sett av alternative kvantitative kjennetegn):	0	1	2	3	4	Sum
Tilhørende sannsynligheter:	0,0531	0,2300	0,3738	0,2700	0,0731	1,0000

Denne oppstillingen kan vi si representerer den random variable $\underline{X}$ som er antall oksekalver i et random sampel på 4 storfefødsler. Den random variable kunne også ha blitt representert på følgende måte:

$P(\underline{X} = 0 | Un=4) = 0,0531$ $P(\underline{X} = 1 | Un=4) = 0,2300$

$P(\underline{X} = 2 | Un=4) = 0,3738$ $P(\underline{X} = 3 | Un=4) = 0,2700$

$P(\underline{X} = 4 | Un=4) = 0,0731$

$P(\underline{X} = X | Un=4) = 0$ for alle andre X enn X=0, X=1, X=2, X=3 og X=4.

Vi har tidligere lært å regne ut sannsynligheten for X oksekalver i et random sampel på n storfefødsler ved hjelp av formelen $P_X = \binom{n}{X} p^X q^{n-X}$. Det er denne formelen som ble brukt ved utregning av sannsynlighetene øverst på denne siden.

Da vi brukte formelen i sannsynlighetsregningen tidligere, oppfattet vi X som et kvantitativt kjennetegn, og vi var bare interessert i en eller noen få forskjellige X. Som vi nettopp har sett, er det imidlertid ikke noe i veien for at vi kan ta for oss alle mulige X og oppfatte disse som verdier av en random variabel $\underline{X}$.

Vi kan da skrive:

(6) $P(\underline{X} = X) = \binom{n}{X} p^X q^{n-X}$ for X = 0, 1, 2, 3, ...,n

$P(\underline{X} = X) = 0$ for alle andre X.

Siden sannsynligheten på venstre side er en funksjon av X, kan vi også skrive:

(7) $f(X) = \binom{n}{X} p^X q^{n-X}$ for X = 0, 1, 2, 3, ...,n

$f(X) = 0$ for alle andre X.

Funksjonen f(X) er her fordelingsfunksjonen for $\underline{X}$ (ikke for X). Den er en funksjon av X. n og p er parametre i fordelingsfunksjonen. (Når p betraktes som parameter er ikke q parameter da q = 1-p.) En parameter er en størrelse som er konstant i et bestemt problem, men som kan ha forskjellige verdier i forskjellige oppgaver. I eksempel 2 ovenfor har parameteren n verdien 4 og p har verdien 0,52. Fordelingsfunksjonen for antall oksekalver i et random sampel på 4 storfefødsler er derfor

(8) $f(X) = \binom{4}{X} 0{,}52^X \ 0{,}48^{4-X}$ for X = 0, 1, 2, 3 og 4

$f(X) = 0$ for alle andre X.

Hvis vi setter inn verdiene 0, 1, 2, 3 og 4 etter tur i denne formelen, får vi de sannsynlighetene som er gjengitt i oppstillingen ovenfor.

Merk at når en fordelingsfunksjon presenteres blir ofte tilføyelsen "f(X) = 0 for alle andre X" utelatt, da den blir ansett som underforstått.

Begrepet fordelingsfunksjon er like viktig som begrepet random variabel og er knyttet sammen med det. Vi tenker oss gjerne at hver random variabel har sin fordelingsfunksjon, uansett om denne funksjonen er kjent eller ukjent, matematisk handterbar eller uhandterbar.

Vi har hittil bare definert en diskret random variabel og fordelingsfunksjonen for denne. Når vi får å gjøre med en kontinuerlig random variabel er det vanskelig å bruke nøyaktig samme definisjon fordi en slik random variabel kan anta en uendelighet av verdier. Siden settet av alternative kvantitative kjennetegn inneholder en uendelighet av alternativer, blir den brøkdelen av gjentakene i universet som har et bestemt alternativ forsvinnende liten. Sannsynligheten for et bestemt kvantitativt kjennetegn (dvs. for en bestemt verdi av den random variable) må derfor settes lik 0.

Vi skal belyse disse problemene nærmere ved et par eksempler og vise hvorledes vanskelighetene kan løses.

Eks. 3. Den årlige melkeytelse hos kuer i et bestemt uendelig univers kan oppfattes som en random variabel. De forskjellige ytelser fra 0 kg og opp til ytelsene hos rekordkuene i universet utgjør et sett av alternative kvantitative kjennetegn. Settet inneholder imidlertid en uendelighet av kjennetegn slik at sannsynligheten for et bestemt kjennetegn (melkeytelse med alle tenkelige desimalers nøyaktighet) må settes lik 0.

En mulig måte å løse problemet på kunne være å se på intervaller av melkeytelser og tilsvarende sannsynligheter slik som nedenfor:

Intervaller for verdier av den random variable (dvs. sett av alternative kvantitative kjennetegn):	0	2000	4000	6000	8000	10000	12000
Tilhørende sannsynligheter:	0,06	0,22	0,29	0,26	0,13	0,03	0,01

Dette er imidlertid ikke noen god måte å presentere en kontinuerlig random variabel på. Den måten statistikerne har valgt er mye bedre. Den går ut på å la sannsynlighetene som knytter seg til hvert intervall være reprosentert ved et areal. Dette er illustrert i fig. 3.

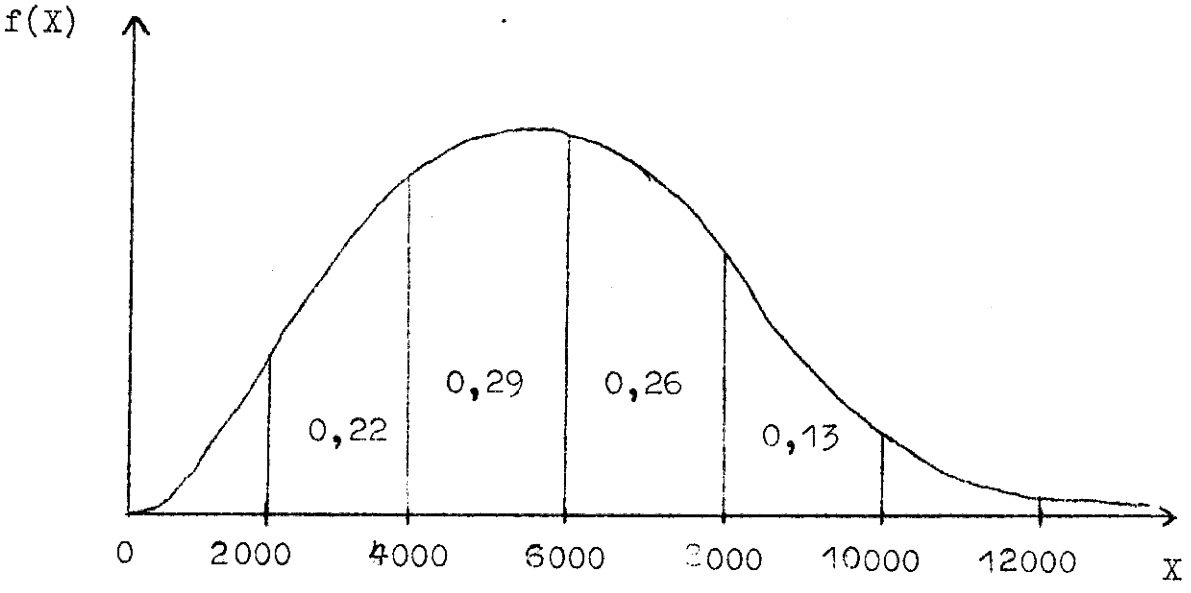


Fig. 3

Kurven er tenkt trukket på en slik måte at sannsynligheten for en melkeytelse i et visst intervall, likegyldig hvilket, er lik det arealet som begrenses av X-aksen, kurven og ordinatene i intervallets endepunkter. Således er $P(2000 \leq X \leq 4000 \mid U) = 0,22$.

Hele arealet under kurven, men over X-aksen er lik 1. Vi kan dele opp X-aksen i vilkårlige intervaller. De tilhørende sannsynlighetene finnes alltid som arealer over intervallet, men under kurven. Hvordan vi kan komme fram til slike kurver er et spørsmål for seg som vi skal la ligge foreløpig.

Vi ser at med litt tilpassing kan vår definisjon av en diskret random variabel også anvendes på kontinuerlige random variable. Imidlertid må vi da operere med intervaller og arealer som vist ovenfor. Intervallene kan gjøres så små vi ønsker, men de må ha en lengde som er større enn 0. Har et intervall lengden 0, settes den tilsvarende sannsynligheten lik 0.

Når vi har å gjøre med en kontinuerlig random variabel (som i fig. 3) framstiller funksjonen f(X) ordinatene til kurven som vi har trukket opp. Også i slike tilfelle vil vi kalle funksjonen f(X) fordelingsfunksjonen for X.

Det er en meget viktig forskjell på fordelingsfunksjonen for en diskret og en kontinuerlig random variabel. Til hver verdi X av en diskret random variabel X hører det en bestemt sannsynlighet f(X). En fordelingsfunksjon f(X) for en diskret random variabel er altså en funksjon som har de forskjellige verdiene X av den random variable som argumenter og de tilsvarende sannsynlighetene f(X) som funksjonsverdier.

Når vi har å gjøre med en kontinuerlig random variabel er argumentene i fordelingsfunksjonen fremdeles verdier av den random variable. Funksjonsverdiene er imidlertid ikke lenger sannsynligheter. De er simpelthen ordinater i diagrammet for fordelingsfunksjonen. Sannsynligheter kan i det kontinuerlige tilfelle bare knytte seg til intervaller, og sannsynligheter kan også bare finnes ved arealberegninger. Matematisk kan slike arealberegninger utføres ved å integrere fordelingsfunksjonen over et visst intervall.

Svært mange av de random variable vi får å gjøre med i praksis er (tilnærmet) kontinuerlige random variable. Som eksempler kan nevnes kroppstemperaturen hos dyr av et bestemt slag, spekktykkelsen hos griser, tilveksten hos slaktedyr, omsatt mengde kjøtt pr. år i Norge, osv.

Til slutt skal vi si litt om terminologien i forbindelse med de begrepene vi har gjennomgått i dette hovedavsnittet (kan overspringes). Denne er nokså varierende både på norsk og engelsk. I dette heftet har vi benyttet den terminologien som er i bruk ved Norges landbrukshøgskole, men vi skal kort angi noen andre nokså vanlige varianter.

I stedet for random variabel brukes ofte betegnelsen stokastisk variabel eller tilfeldig variabel.

Uttrykket "fordelingsfunksjon" brukes ofte om en såkalt kumulativ fordelingsfunksjon (et begrep som ikke er gjennomgått her). Hvis det er fare for misforståelser kan vi bruke betegnelsen elementær fordelingsfunksjon om de fordelingsfunksjonene vi har gjennomgått.

Vil vi markere hva slags random variabel vi har med å gjøre, kan vi bruke betegnelsen sannsynlighetsfunksjon (probability function) for fordelingsfunksjonen for en diskret random variabel (siden funksjonsverdiene er sannsynligheter) og betegnelsen tetthetsfunksjon (density function) for fordelingsfunksjonen for en kontinuerlig random variabel.

Oppgave: Et medisinsk preparat, M kan inneholde varierende mengder av et stoff, A, men på grunn av en spesiell framstillingsmåte vil mengden av A i M alltid ligge mellom 2% og 8%.

La X være mengden av A i prosent i en tilfeldig prøve av M. Det har vist seg at X varierer fra prøve til prøve på denne måten:

$f(X) = a(X+3) \qquad (2 \leq X \leq 8)$ hvor a er en konstant.

a) Bestem konstanten a. b) Framstill funksjonen grafisk. c) Finn sannsynligheten for at $3 < X < 5$. d) Finn sannsynligheten for at $X \geq 4$.

VII. Karakteristikker av fordelingsfunksjonen for en random variabel

Hvis vi kjenner fordelingsfunksjonen for en random variabel, vet vi implisitt alt av statistisk interesse om denne random variable. Likevel ønsker vi vanligvis å få greie på formelen for og gjerne også verdien av visse spesielle konstanter (momenter) som karakteriserer fordelingsfunksjonen. De konstantene vi er mest interessert i kalles forventningen og standardavviket. Det er ofte hensiktsmessig å operere med kvadratet av standardavviket som kalles (den teoretiske) variansen.

Også hvis fordelingsfunksjonen er ukjent vil det være svært nyttig å vite noe om forventningen og variansen.

A. Forventningen

Populært uttrykt er forventningen gjennomsnittet av verdiene av den random variable for alle gjentak i hele universet. Hvis den random variable er antall griseunger i et univers av grisekull, kan altså forventningen betraktes som det gjennomsnittlige antall griseunger pr. kull i hele universet. Legg merke til at det i universet selvsagt kan være mange grisekull som består av f.eks. 8 griseunger. Hvis forventningen skulle beregnes etter den vanlige formelen for et aritmetisk gjennomsnitt (summen av verdiene dividert med antallet) måtte derfor visse verdier, som f.eks. tallet 8 i vårt eksempel, komme med mer enn én gang under summeringen.

Siden universet i mange tilfelle er ubegrenset eller abstrakt, er det naturligvis ofte umulig å regne ut

forventningen etter formelen for det aritmetiske gjennomsnittet. Den egentlige definisjonen av forventningen framstiller da heller ikke denne som et aritmetisk gjennomsnitt, slik som vi har gjort ovenfor, men bygger på fordelingsfunksjonen for den random variable. Det er nemlig en direkte korrespondanse mellom universet og fordelingsfunksjonen. Fordelingsfunksjonen har å gjøre med en bestemt random variabel og viser hvorledes verdiene av denne random variable er fordelt, relativt sett, blandt gjentakene i universet. (I ett og samme univers kan vi tenke oss flere random variable og således flere fordelingsfunksjoner, men dette skal vi ikke komme inn på nå.) Hvis fordelingsfunksjonen er kjent, eller kan forutsettes kjent, er forsåvidt også universet kjent. Det er jo da tilstrekkelig godt beskrevet med hensyn til den random variable. Vi er nemlig vanligvis ikke interessert i det absolutte, men i det relative antall gjentak som har en gitt verdi av en random variabel eller som har en verdi som faller i et gitt intervall. Dette gir fordelingsfunksjonen opplysning om.

La oss her skyte inn noen betraktninger som ligger litt på siden av det vi nå behandler. Det faktum at fordelingsfunksjonen så å si avspeiler universet gjør vi ofte bruk av i den praktiske anvendelsen av statistikken. På grunnlag av a priori viten og teoretiske overveielser kan vi ofte driste oss til å postulere at en random variabel som vi er interessert i (f.eks. vekten av marsvin) har en bestemt kjent fordelingsfunksjon (f.eks. den såkalte normale fordelingsfunksjon), dvs. en fordelingsfunksjon som vi kjenner formelen for selv om vi ikke kjenner verdiene av dens parametre i det konkrete tilfelle.

På denne måten kvitter vi oss til en viss grad med de vanskelighetene som er forbundet med univers-begrepet. Oppgaven i neste omgang blir da ofte å undersøke ved hjelp av et sampel om postuleringen kan sies å være brukbar, samt å estimere parameterne i fordelingsfunksjonen.

Siden fordelingsfunksjonen rommer alle opplysninger av interesse om universet, er det klart at forventningen kan beregnes med utgangspunkt i fordelingsfunksjonen hvis denne er kjent. Forventningen kan faktisk betraktes som en karakteristikk både av universet og av den fordelingsfunksjonen vi betrakter.

Da vi ovenfor forklarte hva forventningen er tok vi utgangspunkt i universet. I vår egentlige definisjon vil vi ta utgangspunkt i fordelingsfunksjonen. Igjen kan vi definere forventningen som et gjennomsnitt, men denne gangen som et veid gjennomsnitt. Før vi går videre vil vi derfor ved et eksempel minne om hva et veid gjennomsnitt er. Eksemplet kan samtidig tjene til å repetere bruken av indekser og summetegn.

Sett at det til en årseksamen gis karakterer i 3 fag, F_i (i=1,2,3). Karakterene, som er tallkarakterer, betegnes med X_i (i=1,2,3). Karakterene i de enkelte fag har vekttallene (vektene) V_i (i=1,2,3). Gjennomsnittskarakteren når vi bruker vekttall er egentlig det vi kaller et veid gjennomsnitt. Formelen for det veide gjennomsnittet i dette tilfelle er:

$$(9)\quad \text{Veid gjennomsnitt} = \bar{X}_v = \frac{\sum_{i=1}^{3} V_i X_i}{\sum_{i=1}^{3} V_i} = \frac{V_1X_1+V_2X_2+V_3X_3}{V_1+V_2+V_3}$$

I ord kan vi si at et veid gjennomsnitt av en rekke tall er summen av produktene (dvs. produktsummen) av tallene og vektene, dividert med summen av vektene. Hvis summen av vektene er lik 1 kan vi selvsagt se bort fra divideringen.

Nedenfor er det gjengitt et talleksempel.

F_i	F_1	F_2	F_3	
X_i	1,5	3,0	1,2	$\sum_{i=1}^{3} X_i = 5,7 \quad \bar{X} = \frac{\sum_{i=1}^{3} X_i}{3} = 1,9$
V_i	3	1	2	$\sum_{i=1}^{3} V_i = 6$
V_iX_i	4,5	3,0	2,4	$\sum_{i=1}^{3} V_iX_i = 9,9$

$$\bar{X}_v = \frac{\sum_{i=1}^{3} V_iX_i}{\sum_{i=1}^{3} V_i} = \frac{9,9}{6} = \underline{\underline{1,65}}$$

La oss vende tilbake til forventningen. Forventingen, μ eller $E(\underline{X})$ for en <u>diskret</u> random variabel $\underline{X}$ kan defineres på følgende måte:

$$(10) \qquad \mu = E(\underline{X}) = \sum_{X} f(X) \cdot X$$

μ er gresk m og minner om middeltall. Symbolet E står for "expectation" som betyr forventning. Når vi skriver $\sum_{X}$ mener vi at summeringen skal skje over alle <u>ulike</u> verdier den random variable kan anta. Legg merke til at antall ledd i summen vanligvis er mindre enn antall gjentak i universet. Hvis, som i vårt tidligere eksempel, hvert gjentak er et grisekull,

vil det i universet vanligvis være mer enn ett kull som består av f.eks. 8 griseunger. Likevel skal altså tallet X=8 tas med bare eń gang under summeringen i (10),siden vi bare skal ta med ulike verdier. Tallet 8 blir nemlig multiplisert med tallet f(8) som er det relative antall kull i hele universet som består av 8 grisunger.

f(X) er sannsynligheten for at den random variable X skal ha verdien X. Funksjonsverdiene f(X) for de forskjellige X ka oppfattes som vekter, og summen av vektene, når vi summerer over alle ulike X som kan forekomme,er lik 1.

(11) $$\sum_X f(X)=1$$

Vi ser derfor at forventningen er et veid gjennomsnitt av alle verdiene den random variable kan ha. De tilhørende sannsynlighetene er vekter.

Forventningen kan altså oppfattes som et aritmetisk gjennomsnitt når utgangspunktet er alle gjentakene i universet og som et veid gjennomsnitt når utgangspunktet er fordelingsfunksjonen. Vår dobbelte omtale av eksemplet hvor universet består av grisekull og hvor den random variable er antall griseunger skulle gjøre det klart hvorfor det må være slik.

Forventningen er et uttrykk for "tyngdepunktet" i X-verdiene i universet. I uttrykket for forventningen blir det både tatt hensyn til hvilke X-verdier som forekommer i universet, og til hvor ofte hver X-verdi forekommer, relativt sett. La oss f.eks. se på den random variable, antall øyne ved terningkast. Situasjonen er denne:

Verdier av den random variable:	1	2	3	4	5	6
Tilhørende sannsynligheter:	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Forventningen for antall øyne blir:

$$(12)\quad E(\underline{X}) = \sum_X f(X)\cdot X = \frac{1}{6}\cdot 1+\frac{1}{6}\cdot 2+\frac{1}{6}\cdot 3+\frac{1}{6}\cdot 4+\frac{1}{6}\cdot 5+\frac{1}{6}\cdot 6 = \frac{21}{6} = \underline{3,5}$$

Øvelse 15.

Finn forventningen for antall oksekalver i et random sampel på 4 storfefødsler (eksemplet ovenfor). Bruk formelen (10).

Legg merke til at forventningen ikke er en sannsynlighet. Den er heller å sammenlikne med en X-verdi og har samme benevning. Forventningen er ikke det samme som den sannsynligste verdien av den random variable. Forventningen kan tvert imot være en verdi av X som ikke forekommer i universet (se eksemplet med terningkast).

Forventningen for en kontinuerlig random variabel $\underline{X}$ defineres og tolkes på en tilsvarende måte, men vi bruker integral i stedet for summetegn i (10):

$$(13)\qquad \mu = E(\underline{X}) = \int_X f(X)X\,dX$$

B. Variansen

Vi så ovenfor at forventningen er et uttrykk for "tyngdepunktet" i verdiene av den random variable. Hvis alle verdier av $\underline{X}$ var lik μ, ville vi vite alt om den random variable når μ var kjent, idet $\underline{X}$ faktisk ville være en konstant. I almminnelighet finnes verdiene av $\underline{X}$ spredt omkring μ i større eller mindre avstand fra μ, og vi er interessert i å få

et slags gjennomsnittsmål for avstanden fra μ idet vi både tar hensyn til hvilke ulike verdier av $\underline{X}$ som forekommer i universet og hos hvor mange gjentak hver verdi forekommer, relativt sett. En kunne i og for seg tenke seg å bruke flere slags mål av denne typen. Det en vanligvis bruker, fordi det matematisk sett er gunstig, er variansen (den teoretiske variansen) som for en diskret random variabel $\underline{X}$ er definert på følgende måte:

$$(14) \qquad \sigma^2 = \text{var}(\underline{X}) = \sum_X f(X)(X-\mu)^2$$

Kvadratroten av variansen betegnes med σ og kalles standardavviket. Symbolet σ er en gresk s og bringer tanken hen til et spredningsmål. Ordet "varians" gir liknende assosiasjoner. Var er å oppfatte som et operasjonssymbol.

La oss se hvorledes vi rent praktisk kan tolke variansen. Det er naturlig å basere et spredningsmål på differensene $X-\mu$ for alle gjentakene i universet. Vi kunne f.eks. tenke oss å ta gjennomsnittet av alle $X-\mu$ i hele universet. Dette gjennomsnittet ville imidlertid bli 0. Forventningen er nemlig å oppfatte som et gjennomsnitt av alle X i universet. Følgelig blir summen av $X-\mu$ summen av en rekke talls avvik fra deres gjennomsnitt. Denne summen blir 0 fordi de positive og negative avvik opphever hverandre (øvelse 16). Ved å bruke $(X-\mu)^2$ i stedet for $(X-\mu)$ blir vi kvitt denne vanskeligheten.

Hvis vi studerer variansen nærmere, ser vi at den populært kan betraktes som gjennomsnittet av $(X-\mu)^2$ for alle gjentak i universet. Hvis vi tar utgangspunkt i fordelingsfunksjonen, får vi med bare de X som er ulike, dvs. at gjentak

som har samme X kommer med bare én gang. Vi ser da av (14) at variansen er et veid gjennomsnitt av $(X-\mu)^2$ for alle ulike verdier X den random variable $\underline{X}$ kan anta. Vektene er de tilhørende sannsynlighetene.

Ved å bruke standardavviket σ i stedet for variansen σ^2 opphever vi på en måte virkningen av at vi har kvadrert hver $(X-\mu)$. Standardavviket σ har samme benevning som X. Både σ og σ^2 er i grunnen mål for det samme, men i hver sin skala.

Variansen for en kontinuerlig random variabel $\underline{X}$ defineres og tolkes på en liknende måte som i det diskrete tilfelle, men vi bruker integraler i stedet for summetegn i (14).

$$(15) \qquad \sigma^2 = \mathrm{var}(\underline{X}) = \int_X f(X)(X-\mu)^2 dX$$

Øvelse 16.

Det aritmetiske gjennomsnitt $\bar{X}$ (leses Xbar) av en rekke tall $X_1, X_2, X_3, \ldots\ldots, X_n$ defineres som

$$(16) \qquad \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Vis at summen av avvikene fra gjennomsnittet, $\sum_{i=1}^{n}(X_i-\bar{X})$ er lik 0 og kommenter dette.

Øvelse 17.

Vis at standardavviket for antall øyne ved terningskast med en ideell terning er lik 1,71 (variansen er lik 2,92).

Øvelse 18.

I et bestemt univers er sannsynlighetene, (f(X) for at en drektig søye skal få X levende fødte lam følgende (tallene er valgt noe urealistiske for å gjøre regningen enkel):

X	f(X)
0	0,1
1	0,3
2	0,5
3	0,1

a) Beregn forventningen, μ og variansen, σ^2 for antall levende fødte lam (X).

b) Vi har et random sampel på 3 drektige søyer fra det nevnte universet. Hva er sannsynligheten for at 2 av de tre søyene får bare dødfødte lam ?

Øvelse 19.

Beregn variansen for antall oksekalver i et random sampel på 4 storfefødsler etter formelen (14). Bruk samme forutsetninger som i øvelse 15.

Øvelse 19 b.

La hvert gjentak være to tvillingkalver, og la oss betrakte en random variabel, $\underline{X}$ = antall oksekalver (X=0,1,2). Vi vil tenke oss tre forskjellige universer hvor fordelingsfunksjonen for $\underline{X}$ er henholdsvis

$$f(X) = \frac{1}{3}, \qquad f(X) = \frac{1}{4}\binom{2}{X} \qquad \text{og} \qquad f(X) = 2X - X^2.$$

Sett opp tabeller av det slaget som er oppgitt i øvelse 18 og beregn med utgangspunkt i disse forventningen og standardavviket for $\underline{X}$ i hvert av de tre universene. Hvorfor er standardavviket forskjellig i de tre universene?

Øvelse 19 c.

La $\underline{X}$ være antall dyr som dør pr. dag i en naturlig koloni av et dyreslag på en øy, og anta at fordelingsfunksjone er

$$f(X) = \frac{0,75}{\binom{4+X}{4}} \qquad (X=0,1,2,3, \ldots).$$

Hva er sannsynligheten for at minst to dyr dør en bestemt dag?

C. Tchebycheffs ulikhet.

La $\underline{X}$ være en random variabel (diskret eller kontinuerlig) med forventning μ og varians σ^2. La oss merke av μ på tall-linjen slik som i fig. 4.

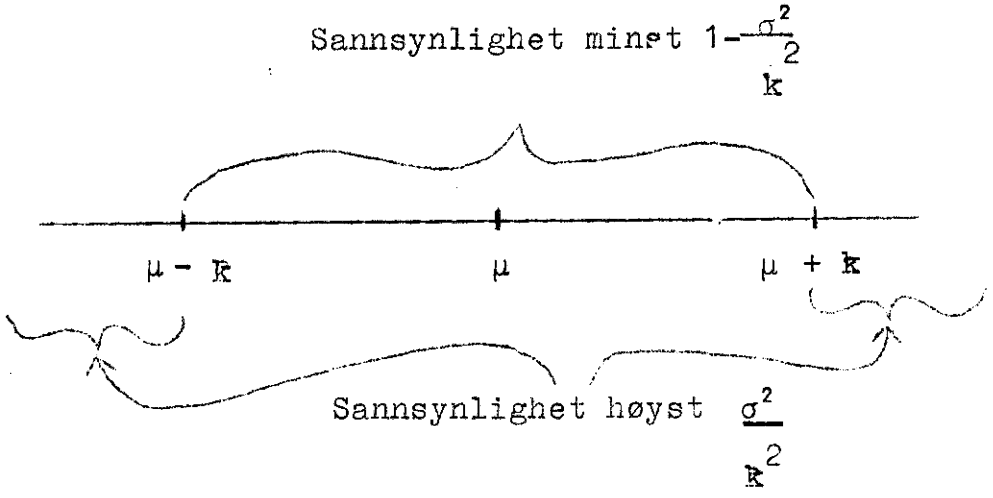


Fig. 4

Vi merker også av to punkter på hver sin side av μ i en avstand k fra μ. Tallet k må være større enn σ men kan for øvrig velges helt fritt. Tchebycheffs ulikhet er en setning som gjelder for en hvilken som helst random variabel uansett hvorledes dens fordelingsfunksjon ser ut. Ulikheten kan skrives på følgende måte:

$$(17) \qquad P(\mu - k \leq \underline{X} \leq \mu + k) \geq 1 - \frac{\sigma^2}{k^2}$$

eller

$$(18) \qquad P(\underline{X} \leq \mu - k \text{ eller } \underline{X} \geq \mu + k) \leq \frac{\sigma^2}{k^2}$$

Merk at vi har utelatt universsbetegnelsen U i (17) og (18). Dette gjøres ofte når det ikke er fare for misforståelser.

Av (18) ser vi at den brøkdelen av alle gjentakene i universet som har verdier av $\underline{X}$ utenfor intervallet fra $\mu - k$ til $\mu + k$ i høyden er lik $\frac{\sigma^2}{k^2}$.

Selv om fordelingsfunksjonen for $\underline{X}$ er ukjent kan vi altså, ved hjelp av Tchebycheffs ulikhet, komme med nyttige sannsynlighetsutsagn hvis μ og σ er kjent. Hvis også fordelingsfunksjonen er kjent, kan vi imidlertid lage skarpere (mer innholdsrike) utsagn.

Øvelse 20.

I et bestemt univers av voksne menn er legemshøyden $\underline{X}$ en random variabel med forventning $\mu = 176$ cm og varians $\sigma^2 = 64$ cm^2. (Konstruert eksempel.) Hva kan sies om sannsynligheten for legemshøyder mindre enn 160 cm eller større enn 192 cm i dette universet?

Øvelse 20 b.

Sett $k = a \cdot \sigma$ i Tchebycheffs ulikhet (a er en konstant). Hvorledes kan Tchebycheffs ulikhet skrives da? Tegn figur og forklar hva sannsynlighetsutsagnet betyr. Hvilket krav må stilles til konstanten a?

VIII. Noen begreper som brukes i forbindelse med et sampel

Hittil har vi for det meste hatt å gjøre med hele universet. Vi skal nå gjennomgå noen få av de viktigste begreper som anvendes for å karakterisere et sampel. Slike begreper er nyttige for rent beskrivende formål, men viktigere er det at de inngår som sentrale hjelpemidler i mange praktisk-statistiske metoder.

A. Gjennomsnitt, empirisk varians og frekvensfordeling[1]

For å gjøre framstillingen så lettfattelig som mulig skal vi knytte den til eksempler.

En bonde har et år 12 drektige søyer. I sine notater har han nummerert søyene vilkårlig fra 1 til 12. Resultatet av lammingen ble følgende:

Tabell 1

Søye nr.:	1	2	3	4	5	6	7	8	9	10	11	12
Antall levende fødte lam:	2	0	2	1	2	0	1	2	2	2	3	1

Disse 12 lammingene kan betraktes som et sampel på 12 gjentak fra et univers av lamminger. Antall levende fødte lam er en random variabel $\underline{x}$ og vi har altså 12 observasjoner av $\underline{x}$. Vi skal bruke eksemplet til å illustrere noen viktige begreper og formler som anvendes i forbindelse med sampler.

Det aritmetiske gjennomsnitt har vi stiftet bekjentskap med tidligere. Verdiene av en random variabel i et sampel på n gjentak betegnes gjerne med $x_1, x_2, x_3, \ldots, x_n$. Nummereringen kan velges helt vilkårlig. Når vi vil referere til en av verdiene uten å spesifisere hvilken av de n verdiene vi har i tankene, bruker vi symbolet x_i. Her kan i oppfattes som en indeks som "løper" fra 1 til n. Vi indikerer dette ved å føye til $(i = 1,2,3,\ldots,n)$.

---

[1] I dette avnsitt er det p.g.a. skrivefeil brukt små x-er i stedet for store som ellers i heftet.

Det aritmetiske gjennomsnitt (eller bare gjennomsnittet) $\bar{x}$ av de n verdiene defineres på følgende måte:

$$(19) \qquad \bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}$$

eller

$$(20) \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (i = 1,2,3,\ldots,n)$$

I vårt eksempel får vi:

$$(21) \qquad \bar{x} = \frac{2+2+1+2+1+2+2+2+3+1}{12} = \frac{18}{12} = 1,5$$

(Hvorfor må også søye nr. 2 og 6 telles med ved utregning av gjennomsnittet?)

Det aritmetiske gjennomsnittet (i samplet) svarer på en måte til forventningen (i universet) og oppfattes ofte som et estimat (en tilnærmingsverdi) av denne (se s. 38).

Den empiriske variansen V (i samplet) svarer på tilsvarende måte til den teoretiske variansen σ^2 (i universet). Ordene empirisk og teoretisk sløyfes ofte, men er nyttige å ta med når det er fare for forvekslinger.

Den empiriske variansen defineres på følgende måte:

$$(22) \qquad V = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \qquad (i = 1,2,3,\ldots,n)$$

Hvis vi sløyfer et-tallet i nevneren, blir vi stående tilbake med gjennomsnittet av $(x - \bar{x})^2$ for alle gjentak i samplet. Denne størrelsen kan sees å svare til den teoretiske variansen (se s. 44 nederst). Grunnen til den vesle "uoverensstemmelsen" mellom σ^2 og V som introduseres ved at vi trekker fra 1 i nevneren for V skal vi komme tilbake til siden.

Variansen i vårt eksempel blir:

$$(23) \qquad V = \frac{(2-1,5)^2+(0-1,5)^2+(2-1,5)^2+\ldots+(1-1,5)^2}{12-1} = 0,82$$

(Merk at vi her må ta med søye nr. 2 og 6 også i nevneren.)

V kan oppfattes som et estimat av σ^2 eller også som et mål for hvor mye verdiene av $\underline{x}$ varierer i samplet. Kvadratroten av V betegnes med s og kalles middelavviket. Middelavviket s (i samplet) svarer altså til standardavviket σ (i universet).

Vi kunne spørre om det i samplet finnes noe som tilsvarer fordelingsfunksjonen, som jo gjelder universet. Svaret er bekreftende. La oss belyse dette ved vårt eksempel.

Lammeresultatet for de 12 søyene kunne stilles opp på en mer oversiktlig måte enn i tabell 1. Dette er gjort i tabell 2. Vi ser av tabell 1 at det finnes bare 4 ulike verdier av $\underline{x}$ i samplet. Vi vil gi disse verdiene nummer fra 1 til 4 i rekkefølge etter størrelsen og betegne en vilkårlig av dem med x_j ($j = 1,2,3,4$). Ved denne nummereringen får vi altså at $x_1 = 0$, $x_2 = 1$, $x_3 = 2$ og $x_4 = 3$. Merk at f.eks. x_1 nå ikke betyr lammeresultatet for søye nr. 1 slik som tidligere. For å markere forandringen bruker vi nå fotindeksen j til erstatning for i. I sin alminnelighet vil vi la j løpe fra 1 til m ($m = 4$ i vårt eksempel). Tallet m er alltid mindre enn (eller i høyden lik) n.

Vi vil la Z_j stå for antall ganger verdien x_j forekommer i samplet. I tabell 2 har en stilt opp de ulike verdier x_j av $\underline{x}$ som forekommer i samplet idet en samtidig har angitt hvor mange ganger hver verdi forekommer, absolutt og relativt.

Tabell 2

x_j	Z_j	$\frac{Z_j}{n}$
0	2	2/12
1	3	3/12
2	6	6/12
3	1	1/12
$\sum_{j=1}^{4} Z_j = n =$	12	$\sum_{j=1}^{4} \frac{Z_j}{n} = 1$

De to første kolonner i tabell 2 utgjør tilsammen et eksempel på en frekvensfordeling. En frekvensfordeling for en diskret random variabel er en tabell over de ulike verdier av den random variable som forekommer i et sampel og antall gjentak som har hver verdi. Tallene Z_j kalles absolutte frekvenser, mens tallene Z_j/n blir kalt relative frekvenser. Summen av de absolutte frekvensene er alltid lik sampelstørrelsen.

$$(24) \qquad \sum_{j=1}^{m} Z_j = n$$

Summen av de relative frekvensene er alltid lik 1.

$$(25) \qquad \sum_{j=1}^{m} \frac{Z_j}{n} = \frac{1}{n} \sum_{j=1}^{m} Z_j = 1$$

Vi har tidligere vist at fordelingsfunksjonen for en diskret random variabel kan skrives i form av en tabelloppstilling (se f.eks. s. 32-33). Kolonne 1 og 3 i tabell 2 er et eksempel på en tilsvarende oppstilling for et sampel. Motstykket i samplet til fordelingsfunksjonen som jo gjelder universet er altså en oppstilling av alle x_j med tilhørende relative frekvenser Z_j/n.

Hvis vi vil undersøke ved hjelp av et sampel om en diskret random variabel x følger en gitt fordelingsfunksjon $f(x)$, må vi altså sammenlikne $f(x)$ med Z_j/n for alle x som kan forekomme i universet. Nå hender det rett som det er at det finnes verdier av den random variable i universet som ikke forekommer i et gitt sampel. I vårt eksempel kunne det f.eks. tenkes at det forekommer søyer med firlinger i universet, selv om det største antall lam i vårt sampel er 3. Vi setter da selvsagt frekvensen for 4 lam lik 0 i vårt sampel. Derimot blir $f(4)$ forskjellig fra 0. Når vi summerer over alle verdier den random variable kan ha blir såvel summen av $f(x)$ som summen av Z/n lik 1. Så langt er altså

de to rekker sammenliknbare. La oss til slutt tilføye at en sammenlikning av de relative frekvenser i samplet med de tilsvarende sannsynligheter som antas å gjelde for universet selvsagt har liten interesse hvis samplet ikke er et random sampel.

Gjennomsnittet og variansen kan naturligvis beregnes med utgangspunkt i en frekvensfordeling. I vårt eksempel kan vi altså bruke tallene i tabell 2 til å beregne de nevnte størrelser. Formlene er følgende:

$$(26) \qquad \bar{x} = \frac{\sum_{j=1}^{m} Z_j x_j}{n} = \sum_{j=1}^{m} \frac{Z_j}{n} x_j$$

$$(27) \qquad V = \frac{\sum_{j=1}^{m} Z_j (x_j - \bar{x})^2}{n-1} = \sum_{j=1}^{m} \frac{Z_j}{n-1} (x_j - \bar{x})^2$$

Riktigheten av formlene skulle være innlysende (og kan demonstreres ved vårt eksempel).

Vi ser at det aritmetiske gjennomsnittet av X_i $(i = 1,2,3,..,n)$ kan betraktes som et veid gjennomsnitt av x_j $(j = 1,2,3,...,m)$ med de relative frekvensene Z_j/n som vekter. Hvis vi ser bort fra et-tallet i nevneren for variansen kan variansen på tilsvarende måte betraktes som et veid gjennomsnitt av $(x_j - \bar{x})^2$ med de samme vekter. Siden vektene Z_j/n i samplet svarer til verdiene $f(x)$ av fordelingsfunksjonen, ser vi også her at det aritmetiske gjennomsnittet svarer til forventningen (se formel 10, s. 41) og at den empiriske variansen svarer til den teoretiske variansen (se formel 14, s. 44).

Den random variable vi har brukt som eksempel i hele dette avsnittet er en diskret random variabel. Når vi har å gjøre med en kontinuerlig random variabel kommer det visse forandringer inn i bildet. Definisjonene (20) og (22) av gjennomsnittet og den

empiriske variansen blir de samme som før. Det er imidlertid lite hensiktsmessig å stille opp observasjonene av en kontinuerlig random variabel $\underline{x}$ i en frekvensfordeling hvor alle ulike verdier av $\underline{x}$ i samplet er tatt med. Som regel finnes det nemlig relativt få like verdier av en kontinuerlig random variabel i et sampel. Derimot finnes det ofte mange ulike verdier spredt med ujamn avstand utover tall-linjen.

Når vi skal ordne observasjonene av en kontinuerlig random variabel i en frekvensfordeling, deler vi opp variasjonsområdet for den random variable i intervaller. Gjentak som har verdier i et gitt intervall sies å tilhøre den tilsvarende klasse. En frekvensfordeling for observasjoner av en kontinuerlig random variabel er altså en tabell over de forskjellige intervaller eller klasser (gjerne også klassenes midtverdier) og antall gjentak som tilhører hver klasse.

Vi skal se på et eksempel. I tabell 3 har vi gjengitt vekten av hver gris i et random sampel på 60 7-ukers grisunger.

Tabell 3

17,1	22,5	17,7	18,0	19,4	15,7	21,1	12,8	12,8	12,6
22,2	25,6	20,8	19,9	19,6	22,2	16,2	15,0	16,0	16,5
26,6	25,9	19,3	16,2	16,7	16,5	12,5	12,7	15,0	16,5
23,3	23,2	19,3	15,4	16,9	15,1	13,5	15,4	17,6	15,6
19,4	20,5	17,4	20,6	15,7	13,9	13,9	13,0	16,6	14,0
19,6	19,0	17,6	19,4	18,7	15,9	17,6	16,9	14,4	16,1

I tabell 4 har vi ordnet disse observasjonene i en frekvensfordeling.

Tabell 4

Klasser	x_j	Z_j	Klasser	x_j	Z_j
12,0-12,9	12,45	5	20,0-20,9	20,45	3
13,0-13,9	13,45	4	21,0-21,9	21,45	1
14,0-14,9	14,45	2	22,0-22,9	22,45	3
15,0-15,9	15,45	9	23,0-23,9	23,45	2
16,0-16,9	16,45	11	24,0-24,9	24,45	0
17,0-17,9	17,45	6	25,0-25,9	25,45	2
18,0-18,9	18,45	2	26,0-26,9	26,45	1
19,0-19,9	19,45	9			n=60

Den midterste verdi i intervall nr. j har vi betegnet x_j og antall gjentak i klasse nr. j har vi kalt Z_j. Med slike betegnelser er formlene (26) og (27) for gjennomsnittet og variansen tilnærmet riktige også for observasjoner av en kontinuerlig random variabel. De er eksakt riktige hvis midtverdien i hvert intervall også er lik gjennomsnittet av de observasjonene som faller i vedkommende intervall. Denne betingelsen er imidlertid neppe noen gang oppfylt i praksis. Når vi skal regne ut gjennomsnittet og den empiriske variansen for observasjonene av en kontinuerlig random variabel bør vi derfor ta utgangspunkt i de opprinnelige observasjonene (tabell 3 i vårt eksempel) og ikke i frekvensfordelingen (tabell 4) hvis vi vil ha nøyaktige resultater.

En klasseinndeling bør alltid foretas på en slik måte at hver observasjon kan henføres til en og bare en klasse. For øvrig er det mye av et praktisk spørsmål hvorledes en vil innrette seg. Klassevidden er gjerne den samme for alle klasser bortsett kanskje fra den ene eller begge ytterklassene. Da hver klasse ved beregninger, grafiske framstillinger o.l. gjerne blir representert ved sin midtverdi, er det en fordel at midtverdien har få desimaler, men at den likevel virkelig er i midten av intervallet. (Hvis en skal være helt nøyaktig vil spørsmålet om hva som blir

å oppfatte som de eksakte klassegrensene og den egentlige midtverdien bl.a. avhenge av hvor mange desimalers nøyaktighet observasjonene er notert med.)

Hvis vi har et random sampel av observasjoner av en kontinuerlig random variabel $\underline{x}$ og vi vil undersøke om fordelingsfunksjonen for $\underline{x}$ er en gitt funksjon f(x), må vi som i det diskrete tilfelle sammenlikne Z_j/n med sannsynligheter i universet. Disse sannsynlighetene regnes imidlertid nå ut på en annen måte enn før. Hvis vi ikke har anledning til å integrere funksjonen kan sannsynligheten som svarer til klasse j regnes ut tilnærmet som produktet av $f(x_j)$ og klassevidden v. Framgangsmåten er illustrert i fig. 6.

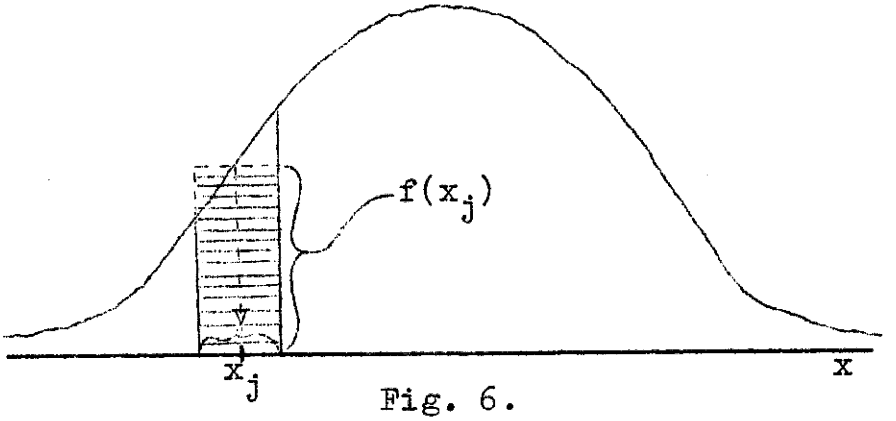


Fig. 6.

B. Praktiske regneformler til beregning av gjennomsnittet og den empiriske variansen.

Når vi skal regne ut gjennomsnittet er det av og til hensiktsmessig å gjøre dette på følgende måte: Først trekker vi et passende konstant tall c fra alle observasjonene. Hvis f.eks. alle tall har første siffer felles, kan vi altså ganske enkelt sløyfe dette sifferet. Deretter regner vi ut gjennomsnittet av det som blir stående igjen etter fratrekningen. Til det gjennomsnittet vi da får adderer vi tallet c. Svaret er da det søkte gjennomsnittet. Skal vi regne ut gjennomsnittshøyden for en rekke personer kan vi f.eks. la c være lik 1 meter. Riktigheten av framgangsmåten skulle være innlysende.

Den empiriske variansen for en rekke observasjoner blir uforandret om vi trekker et konstant tall c fra alle observasjonene. Variansen er nemlig et spredningsmål, og spredningen blir den samme om observasjonsmassen som helhet blir forskjøvet f.eks. til venstre (hvis c er positiv) på tall-linjen. Ved å velge en passende c kan vi av og til lette regnearbeidet betraktelig.

Definisjonsformlene (22) og (27) for den empiriske variansen er som regel tungvindte å arbeide med. Regningen går i de aller fleste tilfelle raskere og mer feilfritt om vi omskriver formlene på følgende måte:

$$(28) \qquad V = s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}}{n-1}$$

$$(29) \qquad V = s^2 = \frac{\sum_{j=1}^{m} Z_j(X_j - \bar{X})^2}{n-1} = \frac{\sum_{j=1}^{m} Z_j X_j^2 - \frac{(\sum_{j=1}^{m} Z_j X_j)^2}{n}}{n-1}$$

Den første omskrivingen kan bevises på følgende måte:

$$(30)\quad \Sigma (X-\bar{X})^2 = \Sigma (X^2 - 2\bar{X}X+\bar{Y}^2) = \Sigma X^2 - 2\bar{X}\Sigma X + n\bar{X}^2$$
$$= \Sigma X^2 - 2\bar{X}n\bar{X} + n\bar{X}^2 = \Sigma X^2 - n\bar{X}^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

C. Gjennomsnittet og den empiriske variansen oppfattet som random variable.

La oss tenke oss et univers U som består av alle voksne menn. Hver mann er altså et gjentak og vi vil forestille oss at universet er ubegrenset. Vi er interessert i en random variabel, nemlig legemshøyden, $\underline{X}$. Fordelingsfunksjonen for $\underline{X}$ er f(X) med forventning $E(\underline{X}) = \mu_{\underline{X}}$ og varians $var(\underline{X}) = \sigma^2_{\underline{X}}$. Fordelingsfunksjonen, forventningen og variansen kan være kjente eller ukjente.

Vi tar et random sampel på n gjentak (menn) fra dette universet og observerer verdien av den random variable (høyden) for hvert gjentak. I slike tilfelle sier vi ofte kort at vi har et random sampel på n observasjoner av $\underline{X}$.

La oss nummerere de n menn i en vilkårlig rekkefølge fra 1 til n. De tilhørende legemshøydene vil vi betegne med X_{11}, X_{21} ..., X_{n1} eller også X_{i1},(i=1,2 ..., n). Disse X-ene er da n verdier av den random variable $\underline{X}$. Gjennomsnittshøyden for de n menn vil vi betegne med $\bar{X}_1$. Vi har da

$$(31)\qquad \bar{X}_1 = \frac{\sum_{i=1}^{n} X_{i1}}{n}$$

Vi vil nå tenke oss at vi tar et nytt random sampel på n menn fra det samme universet. Også denne gangen nummererer vi gjentakene vilkårlig fra 1 til n. Legemshøydene til disse n nummererte menn betegner vi med henholdsvis X_{12}, X_{22}, X_{32}, ..., X_{n2}. Dette kan også skrives som X_{i2} (i=1,2,3,...,n). Disse n X-ene er n nye verdier av den random variable $\underline{X}$. Vi får et nytt gjennomsnitt som vi betegner med $\bar{X}_2$.

$$(32) \qquad \bar{X}_2 = \frac{\sum_{i=1}^{n} X_{i2}}{n}$$

Vi kunne fortsette på samme måte og ta nye random sampler på n gjentak og beregne nye gjennomsnitter $\bar{X}_3$, $\bar{X}_4$, $\bar{X}_5$, osv. Hvert gjennomsnitt er tenkt beregnet på grunnlag av et sett av n verdier av den random variable $\underline{X}$. Disse n verdiene vil vanligvis variere fra sampel til sampel. Derfor vil også gjennomsnittet variere fra sampel til sampel. Vi skal nå innføre en meget viktig og nyttig betraktningsmåte. Hvert gjennomsnitt, $\bar{X}_1$, $\bar{X}_2$, $\bar{X}_3$, osv. kan betraktes som en verdi av en random variabel som vi betegner med $\bar{\underline{X}}$. Mens vi tidligere opererte med et univers U hvor hvert gjentak var en mann, har vi nå å gjøre med et nytt univers hvor hvert gjentak er en samling av n menn. Det nye universet som er avledet av det gamle universet U er altså et univers av sampler. Hvert sampel er et gjentak.

Tidligere har vi betraktet gjennomsnittet $\bar{X}$ som en størrelse som karakteriserer et enkelt sampel. Nå har vi vist at vi også kan se på gjennomsnittet som en random variabel $\bar{\underline{X}}$. Det gjennomsnittet vi finner for et enkelt sampel blir å betrakte som en verdi av denne random variable.

Det er ofte nyttig å vite noe om fordelingsfunksjonen for $\underline{\bar{X}}$. Vi skal ikke gå i detalj her, men bare presentere to uhyre viktige resultater.

La den random variable $\underline{X}$, som nevnt, ha forventning $\mu_{\underline{X}}$ og varians $\sigma^2_{\underline{X}}$. Det kan da vises at forventningen $E(\underline{\bar{X}}) = \mu_{\underline{\bar{X}}}$ og variansen $\text{var}(\underline{\bar{X}}) = \sigma^2_{\underline{\bar{X}}}$, for $\underline{\bar{X}}$ er gitt ved følgende formler:

(33) $$\mu_{\underline{\bar{X}}} = \mu_{\underline{X}}$$

(34) $$\sigma^2_{\underline{\bar{X}}} = \frac{\sigma^2_{\underline{X}}}{n} \quad \text{eller} \quad \sigma_{\underline{\bar{X}}} = \frac{\sigma_{\underline{X}}}{\sqrt{n}}$$

Disse to formlene er helt generelle. De gjelder uansett hvilken fordelingsfunksjon $\underline{X}$ har, og de gjelder for alle endelige sampelstørrelser (verdier av n). I ord kan (33) og (34) kort uttrykkes slik:

Forventningen for gjennomsnittet er lik forventningen for den opprinnelige random variable. Variansen for gjennomsnittet er lik variansen for den opprinnelige random variable dividert med n.

Øvelse 21.

Forklar på en praktisk måte hvorfor formelen (33) synes å være riktig. Forklar også hvorfor en n i nevneren i formelen (34) ikke synes å være urimelig.

På samme måte som vi nå har vist at $\bar{X}$ er en verdi av en random variabel $\underline{\bar{X}}$ kunne vi også vise at s^2 for et gitt sampel er en verdi av en random variabel $\underline{s}^2$. Det skulle være unødvendig å gjenta hele resonnementet. Også fordelingsfunksjonen for $\underline{s}^2$ er av atskillig interesse. Her vil vi imidlertid nøye oss med å legge fram følgende resultat: Uansett hvilken fordelingsfunksjon

$\underline{X}$ har og uansett hvilken verdi n har så kan det vises at forventningen for s^2 er gitt ved følgende formel:

(35) $$E(\underline{s}^2) = \sigma^2$$

I ord kan dette uttrykkes slik: Forventningen for den empiriske variansen er lik den teoretiske variansen. Hvis vi hadde brukt n i stedet for n-1 i nevneren da vi definerte den empiriske variansen ville vi ikke ha fått resultatet (35). Her har vi altså begrunnelsen for den tilsynelatende noe merkelige nevneren.

Oppgave:
Nedenfor er gjengitt antall celler av et bestemt slag i en liten prøve av en spesiell vevsveske. Det er i alt 100 prøver.

4	4	3	4	0	1	3	3	3	3
4	0	1	2	1	2	5	2	4	4
4	4	1	0	0	4	4	3	6	1
6	7	4	2	3	4	3	3	2	5
3	2	0	2	2	2	1	4	1	2
5	4	5	3	1	4	2	1	2	2
2	3	1	4	2	4	2	1	4	4
4	1	2	5	4	3	3	3	4	2
3	1	1	3	1	3	2	3	2	2
3	3	4	2	0	1	5	3	2	2

a) Regn ut gjennomsnittet og middelavviket for antall celler pr. prøve.
b) Sett opp en frekvensfordeling for antall celler pr. prøve.
c) Regn ut gjennomsnittet og middelavviket på grunnlag av frekvensfordelingen.

IX. Noen spesielle fordelingsfunksjoner.

A. Generell orientering

Matematisk sett er en fordelingsfunksjon en vanlig funksjon eller formel. De størrelsene som kan inngå i en slik formel kan deles i tre grupper. For det første kan det inngå visse konstanter som f.eks. tallet 2, tallet π (= 3,14) eller tallet e (= 2,72). Disse konstantene er selvsagt alltid de samme i en og samme fordelingsfunksjon.

For det andre kan det inngå en eller flere parametre (s. 33). En parameter i en bestemt fordelingsfunksjon er ofte pålagt visse begrensninger som f.eks. at den bare kan anta verdier som ligger mellom 0 og 1 eller at den bare kan anta verdier som er hele positive tall. For øvrig kan verdien av en parameter velges fritt slik at den passer til det konkrete fenomen fordelingsfunksjonen skal beskrive i et gitt tilfelle. En parameter i en fordelingsfunksjon kan på mange måter sammenliknes med giret i en bil (fullfør analogien selv). Det er parametrene i en fordelingsfunksjon som gjør at en og samme fordelingsfunksjon (f.eks. den normale fordelingsfunksjon) ofte kan brukes til tilnærmet å beskrive vidt forskjellige fenomener som f.eks. melkeytelsen hos kuer, "intelligensen" hos mennesker, feilen ved astronomiske målinger, osv.

Når vi har for oss en ganske bestemt fordelingsfunksjon og størrelsen av fordelingsfunksjonens parametre er fastlagt, er det bare en størrelse som kan variere, nemlig verdien X av den random variable. Vi er da i den situasjon at vi, om vi vil, kan sette inn tall for X og finne verdien av f(X) som altså er en sannsynlighet (i det diskrete tilfelle) eller bare en ordinat i diagrammet (i det

kontinuerlige tilfelle). Når en får å gjøre med en bestemt fordelingsfunksjon bør en alltid merke seg hvilke verdier den random variable kan ha. Videre bør en aldri glemme at summen eller integralet av fordelingsfunksjonen over alle disse verdiene må være lik 1.

Statistisk sett kan en fordelingsfunksjon oppfattes som en modell (eller en del av en modell) som beskriver et fenomen vi er interessert i. Det finnes et utvalg av slike fordelingsfunksjoner som har vist seg å kunne beskrive, med god tilnærmelse, en hel rekke vidt forskjellige fenomener. Forsåvidt kan vi plukke ut en fordelingsfunksjon som passer til vårt problem på samme måte som vi i konfeksjonsmagasinet plukker ut en dress som passer til vår figur. Dressen har, grovt sett, to "parametre", nemlig armlengden og benlengden. Selv om vi tilpasser disse "parametre", er det imidlertid ikke sikkert at resultatet blir tilfredsstillende. Slik er det også i statistikken. En "skreddersydd" fordelingsfunksjon er oftest å foretrekke, men det er langt vanskeligere å "skreddersy" en fordelingsfunksjon enn det er å få skreddersydd en dress.

Formålet med å ta i bruk en fordelingsfunksjon kan være rent beskrivende. Det kunne nevnes utallige eksempler på dette. Størrelsen av et bestemt organ hos et dyr av et bestemt slag varierer fra individ til individ. Denne variasjonen kunne vi tenke oss å beskrive for hele universet under ett ved hjelp av en fordelingsfunksjon.

Det rent beskrivende er imidlertid ikke hovedsaken. Viktigere er det at fordelingsfunksjoner inngår på forskjellige måter i alle de metoder vi bruker i den praktiske anvendelsen av statistikken.

Vi har allerede nevnt at en bestemt fordelingsfunksjon som

f.eks. den normale fordelingsfunksjon egentlig ikke bare er en enkelt fordelingsfunksjon, men utgjør en hel "familie" av fordelingsfunksjoner. Hvert "medlem" av "familien" framkommer ved at parametrene gis et bestemt sett av verdier. Vi snakker således om en 2-parameter familie av normale fordelingsfunksjoner.

La oss nevne at de fleste familiene vi opererer med også er "i slekt" med hverandre. Ved at visse parametre i en fordelingsfunksjon gjøres tilstrekkelig små, tilstrekkelig store, e.l. vil denne fordelingsfunksjonen ofte bli identisk med en annen kjent fordelingsfunksjon.

Med utgangspunkt i en random variabel som har en bestemt kjent fordelingsfunksjon kan vi også forme en ny random variabel etter en bestemt regel og vite at denne nye random variable da har en annen kjent fordelingsfunksjon. Denslags manipulasjoner er svært viktige ved den praktiske anvendelsen av statistiske metoder.

Vi har her bare anledning til å nevne litt om noen av de viktigste fordelingsfunksjonene.

B. Fordelingsfunksjoner for diskrete random variable

En slik fordelingsfunksjon kan framstilles grafisk ved et stolpediagram. Se f.eks. fig. 2, s. 31.

1. Den binomiale fordelingsfunksjon.

Denne har vi stiftet bekjentskap med tidligere (s. 33). Den kan skrives på følgende måte:

$$(36)\quad f(X) = \binom{k}{X} p^X q^{k-X} \qquad \text{for } X = 0, 1, 2, \ldots, k$$

Parameteren p må være et tall mellom 0 og 1. Hvis $p = \frac{1}{2}$ er stolpediagrammet symmetrisk (hvorfor?). Hvis $p \neq 0{,}5$ er funksjonen skjev den ene eller andre veien (forklar hvordan). Parameteren k kan i mange anvendelser oppfattes som størrelsen av et sampel, men den behøver ikke være det. k må alltid være et helt positivt tall.

Forventningen for en diskret random variabel kan alltid finnes etter formelen (10), s. 41. Setter vi inn (36) for f(X) i (10), er det mulig å vise at forventningen μ for en binomialt fordelt random variabel $\underline{X}$ blir:

$$(37) \qquad \mu = E(\underline{X}) = \sum_{X} f(X) \cdot X = \sum_{X=0}^{k} \binom{k}{X} p^X q^{k-X} X = kp$$

Riktigheten av det siste likhetstegnet er slett ikke innlysende matematisk og må bare aksepteres i dette kurset.

Når vi vil regne ut forventningen for en binomialt fordelt random variabel kan vi altså bruke den enkle formelen $\mu = kp$ som gjelder for denne spesielle fordelingsfunksjonen i stedet for å gå den tunge omveien om den generelle formelen (10).

Øvelse 22.

Løs oppgaven i øvelse 15 ved hjelp av formelen $\mu = kp$. Svaret skal bli det samme som før. Forklar ved hjelp av et praktisk resonnement hvorfor formelen $\mu = kp$ må være riktig.

Variansen for en diskret random variabel kan generelt finnes med utgangspunkt i formelen (14), s. 44. Setter vi inn (36) for f(X) i (14) kan det vises at resultatet blir følgende:

$$(38)\qquad \sigma^2 = \mathrm{var}(\underline{X}) = \sum_X f(X)(X-\mu)^2 = \sum_{X=0}^{k} \binom{k}{X} p^X q^{k-X} (X-\mu)^2 = kpq.$$

Det siste likhetstegnet er heller ikke her innlysende, matematisk sett.

Formelen $\sigma^2 = kpq$ brukes alltid når vi skal regne ut variansen for en binomialt fordelt random variabel. Det er unødvendig å bruke den generelle definisjonsformelen (14).

Øvelse 23.

Løs oppgaven i øvelse 19 etter formelen $\sigma^2 = kpq$.

Verdiene av koeffisienten $\binom{k}{X}$ for en gitt k og for de forskjellige verdier av X fra X = 0 til X = k kan beregnes på vanlig måte, men kan ofte finnes raskere ved hjelp av Pascals trekant.

1 1

1 2 1

1 3 3 1

1 4 6 4 1

1 5 10 10 5 1

osv.

Trekanten formes ved at hvert tall gjøres lik summen av de to som det står midt under. Dessuten fyller vi ut med tallet 1 ytterst.

For en gitt k bruker vi bare en linje i trekanten, nemlig den linjen som har k ved siden av 1-tallet. I vårt tidligere eksempel var k lik 4. Vi bruker da linjen 1 4 6 4 1. Disse 5 tallene er da lik $\binom{4}{X}$ for henholdsvis X=0, X=1, X=2, X=3 og X=4 (kontroller!).

2. Poissons fordelingsfunksjon.

Et annet eksempel på en fordelingsfunksjon for en diskret random variabel $\underline{X}$ er Poissons fordelingsfunksjon. Den kan skrives på følgende måte:

$$(39) \qquad f(X) = \frac{e^{-m} m^{X}}{X!} \qquad \text{for } X = 0, 1, 2, \ldots$$

Verdiene som den random variable kan ha er alle naturlige tall 0, 1, 2, 3 osv. e står for grunntallet i det naturlige logaritmesystemet (e = 2,72). Fordelingsfunksjonen har én parameter, m. Hvis vi setter inn (39) for f(X) i de generelle formlene (10) og (14), finner vi for denne spesielle fordelingsfunksjonen at $\mu = E(\underline{X}) = m$ og $\sigma^2 = \text{var}(\underline{X}) = m$. For denne fordelingsfunksjonen er altså forventningen og den teoretiske variansen alltid like store og lik m.

$$(40) \qquad \mu = E(\underline{X}) = \sigma^2 = \text{var}(\underline{X}) = m$$

Poissons fordelingsfunksjon er "i slekt" med den binomiale fordelingsfunksjon. La oss tenke oss at vi fra den binomiale familien av fordelingsfunksjoner tar for oss en rekke av medlemmer som er slik at hvert nytt medlem i rekken har mindre p men større k enn det foregående medlem. Samtidig skal kp være samme tall,

nemlig m, for alle medlemmer. Hvis vi på denne måten gjør p liten nok og k stor nok kan det vises matematisk at den binomiale fordelingsfunksjonen vi ender opp med er lik Poissons fordelingsfunksjon (39). I matematisk mer presise ordelag kan vi si at Poissons fordelingsfunksjon framkommer av den binomiale ved en grenseovergang hvor vi lar $p \to 0$ og $k \to \infty$ samtidig som $kp=m$ (eller $kp \to m$).

Mange biologiske fenomener kan antas i følge tilnærmet Poissons fordelingsfunksjon. Til en viss grad kan vi resonnere oss til dette ut fra vårt kjennskap til hvorledes fordelingsfunksjonen framkommer. Antall tilfelle pr. år i Norge av en relativt sjelden ikke smittsom sykdom hos et bestemt husdyrslag kan tenkes å følge Poissons fordelingsfunksjon. Resonnementet er følgende: Alle dyr i Norge av vedkommende slag i et enkelt år kan oppfattes som et stort ($k \to \infty$) random sampel av dyr fra universet av alle dyr av vedkommende slag. Sykdommen er relativt sjelden ($p \to 0$) og ikke smittsom (uavhengighet, se s. 22). I det lange løp vil gjennomsnittlig $k \cdot p$ dyr pr. år få sykdommen.

Slike resonnementer er gjerne noe omtrentlige og må ikke tillegges for stor vekt. Fordelingsfunksjonen har for øvrig også mange anvendelser hvor et tilsvarende resonnement ikke kan gjennomføres, og hvor den binomiale fordelingsfunksjonen ikke kommer inn i bildet. Antall celler av et bestemt slag i en liten prøve av en viss størrelse av en bestemt vevsvæske er en annen type fenomener som ofte kan antas å følge Poissons fordelingsfunksjon.

Oppgave: Undersøk om Poissons fordelingsfunksjon synes å passe som modell for observasjonene i oppgaven på s. 61.

C. Fordelingsfunksjoner for kontinuerlige random variable.

Disse fordelingsfunksjonene kan framstilles grafisk som kurver. Se f.eks. fig. 3, s. 35.

1. Den normale fordelingsfunksjon

Den normale eller Gauss-Laplaceske fordelingsfunksjonen er en uhyre viktig fordelingsfunksjon. Den kan skrives på følgende måte:

$$(41) \qquad f(X) = \frac{1}{\tau\sqrt{2\pi}} e^{\frac{-(X-\nu)^2}{2\tau^2}} \qquad \text{for} \qquad -\infty < X < \infty$$

Fordelingsfunksjonen har to parametre som vi har betegnet med ν (gresk n, leses ny) og τ (gresk t, leses tau).

Forventningen og den teoretiske variansen for en random variabel som følger den normale fordelingsfunksjonen kan i prinsippet finnes på vanlig måte ved å sette inn (41) i (13) og (15). Utledningen er ikke så helt enkel, så vi setter her bare opp resultatene.

$$(42) \qquad \mu = E(\underline{X}) = \int_{-\infty}^{\infty} f(X)X\,dX = \int_{-\infty}^{\infty} \frac{1}{\tau\sqrt{2\pi}} e^{\frac{-(X-\nu)^2}{2\tau^2}} X\,dX = \nu$$

$$(43) \qquad \sigma^2 = \mathrm{var}(\underline{X}) = \int_{-\infty}^{\infty} f(X)(X-\nu)^2 dX = \int_{-\infty}^{\infty} \frac{1}{\tau\sqrt{2\pi}} e^{\frac{-(X-\nu)^2}{2\tau^2}} (X-\nu)^2 dX =$$

Resultatene er meget enkle. Vi ser at forventningen μ er lik parameteren ν mens standardavviket σ er lik parameteren τ. Vi kan derfor like gjerne bruke symbolene μ og σ i stedet for ν og τ og skrive fordelingsfunksjonen på følgende måte:

$$(44) \qquad f(X) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(X-\mu)^2}{2\sigma^2}} \qquad \text{for} \quad -\infty < X < \infty$$

Dette er den vanlige skrivemåten, og vi vil bruke denne heretter. Når en random variabel $\underline{X}$ er normalt fordelt er altså forventningen $\mu = E(X)$ en parameter i selve fordelingsfunksjonen. Videre er da også standardavviket σ en parameter i fordelingsfunksjonen. Når vi vil angi kort at en random variabel $\underline{X}$ har fordelingsfunksjonen (44) gjør vi ofte dette ved å skrive $\underline{X}$: $N(\mu, \sigma)$.

Parameteren μ kan ha en hvilken som helst verdi i intervallet $-\infty < \mu < \infty$. Parameteren σ kan ha en hvilken som helst verdi i intervallet $0 < \sigma < \infty$.

Kurven som framstiller denne fordelingsfunksjonen har én topp, nemlig der hvor $X = \mu$, og den er symmetrisk om dette punktet. Kurven synker stadig når vi beveger oss bort fra μ. Den nærmer seg asymptotisk til X-aksen når $X \to -\infty$ eller $X \to \infty$. Verdien av μ bestemmer plasseringen av toppen (og dermed av hele kurven) på tall-linjen. Verdien av parameteren σ bestemmer formen på kurven. Når σ er liten blir arealet under kurven konsentrert omkring μ (kurven blir høy og relativt spiss). Når σ er stor blir arealet under kurven spredt mer utover tall-linjen (kurven blir lavere og flatere). σ kan finnes igjen på kurven som avstanden fra μ til vendepunktet for kurven på den ene eller andre siden av μ. Kurver med samme σ har nøyaktig

samme form selv om μ er forskjellig. De er bare flyttet langs X-aksen i forhold til hverandre.

Svært mange av de statistiske metodene vi skal gjennomgå bygger strengt tatt på den forutsetningen at den random variable vi har å gjøre med er normalt fordelt. Dette kan synes å være en meget dristig forutsetning, men det er to forhold som gjør at forutsetningen tross alt ikke er så hasardiøs. For det første kan det vises på teoretisk grunnlag at den normale fordelingsfunksjon er en forholdsvis generell modell. For det andre har empiriske undersøkelser vist at de metodene det her er snakk om ofte leder til tilnærmet riktig konklusjon selv om fordelingsfunksjonen avviker nokså sterkt fra den normale. Vi sier om slike metoder at de er robuste overfor avvik fra forutsetningen om normal fordeling.

Det finnes mange tabeller av forskjellig slag over den normale fordelingsfunksjonen, men vi skal ikke gå så mye inn på dette her. Vi skal bare nevne noen tall i forbindelse med en ganske spesiell normal fordelingsfunksjon, nemlig den funksjonen som har forventning 0 og standardavvik 1. Denne funksjonen kalles den standardiserte (eller normaliserte) normale fordelingsfunksjonen. Den er gjengitt i fig. 7. Denne fordelingsfunksjonen

får en bruk for i visse praktiske anvendelser av statistikken. Problemstillingen, som vi skal begrunne nærmere senere, er da følgende: Vi ønsker å finne et tall a som er slik at sannsynligheten for at den random variable $\underline{X}$ skal ha en verdi som faller i intervallet fra -a til a skal være lik et oppgitt tall Q. Mer sammentrengt kan vi skrive den betingelsen som a skal oppfylle på følgende måte:

$$(45) \qquad P(-a \leq \underline{X} \leq a) = Q$$

Q er gjenstand for valg. Som regel velges Q enten lik 0,95 eller lik 0,99. Q er en sannsynlighet og derfor et areal i diagrammet for fordelingsfunksjonen. Ofte opererer en også med en størrelse P som er lik 1-Q. I stedet for å velge Q kan en like gjerne velge P, og denne velges da gjerne lik 0,05 eller 0,01.

Hele problemstillingen er illustrert i fig. 7 hvor en har valgt Q = 0,95 og dermed P = 0,05. Det skraverte arealet i hver "hale" av fordelingen er lik $\frac{P}{2} = 0,025$ mens det uskraverte området i midten er lik Q = 0,95. Hele arealet under kurven, men over X-aksen er selvsagt lik 1. I dette tilfellet kan det vises at a = 1,96. Dette tallet er funnet en gang for alle og gjengitt i tabeller. Hvis vi hadde valgt Q lik 0,99 ville a selvsagt blitt større, nemlig 2,576.

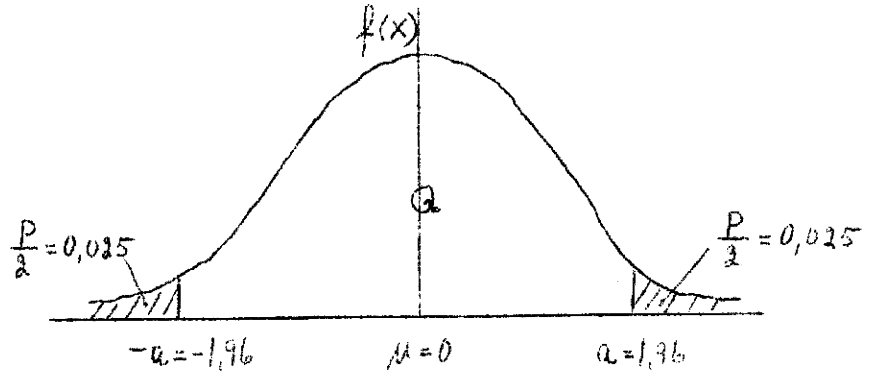


Fig. 7.

2. Students t-fordeling.

Vi viste i avsnitt VIII, C at gjennomsnittet og den empiriske variansen også kan oppfattes som random variable. Det samme gjelder middelavviket som er kvadratroten av variansen og altså en funksjon av denne. En størrelse som er en funksjon av en random variabel er i alminnelighet selv en random variabel. Det samme gjelder hvis den er en funksjon av flere random variable.

Vi vil nå betrakte en random variabel $\underline{X}$ som er normalt fordelt med forventning μ og standardavvik σ. Vi tenker oss at vi skaffer oss et random sampel på n observasjoner av $\underline{X}$. På grunnlag av dette samplet kan vi regne ut $\overline{X}$ og s. La oss definere en størrelse t som er en funksjon av $\overline{X}$ og s.

$$(46) \qquad t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

Hvis vi på grunnlag av vårt sampel regner ut $\overline{X}$ og s, kan vi, hvis μ er kjent, også regne ut t.

Hvis vi tenker oss en uendelighet av random sampler, hvert på n gjentak, kan vi som vist i avsnitt VIII, C gå over til å betrakte gjennomsnittet og middelavviket som random variable. Til hvert sampel svarer det en verdi av t. Det er derfor klart at vi også kan betrakte denne størrelsen som en random variabel. Denne random variable vil vi betegne med $\underline{t}$. Den er en funksjon av de random variable $\underline{\overline{X}}$ og $\underline{s}$.

$$(47) \qquad \underline{t} = \frac{\underline{\overline{X}} - \mu}{\underline{s}/\sqrt{n}}$$

Under våre forutsetninger kan det vises at $\underline{t}$ har en ganske spesiell fordelingsfunksjon som kalles t-fordelingen eller Students t-fordeling. (Navnet kommer av at W. S. Gosset, som først utledet fordelingen skrev under pseudonymet "Student".) Fordelingsfunksjonen f(t) har en komplisert matematisk form, så det er liten grunn til å gjengi formelen her. I formelen inngår det en parameter f som kalles antall frihetsgrader. f må være et helt positivt tall. I de sammenhenger hvor vi skal bruke fordelingen er antall frihetsgrader fastlagt av problemets natur, og vi finner f etter ganske spesielle regler eller formler. Når f.eks. den random variable $\underline{t}$ fremkommer som vist i (47), finner vi f av (48)

(48) $\qquad f = n - 1$

Kurven som framstiller t-fordelingen er noe flatere enn kurven for den standardiserte normale fordelingsfunksjon (s. 71), men likner for øvrig denne. Den er således symmetrisk om $E(\underline{t})$ som er lik 0 (når $f > 1$). Variasjonsområdet for $\underline{t}$ går fra $-\infty$ til ∞.

Hvis vi tenker oss en rekke t-fordelinger med voksende verdier av f vil fordelingene nærme seg den standardiserte normale fordelingsfunksjon. t-fordelingen faller helt sammen med denne når $f \to \infty$.

I våre praktiske anvendelser av t-fordelingen er vi på samme måte som beskrevet på s. 71 interessert i å finne en a som svarer til en valgt P eller Q og som tilfredsstiller sannsynlighets-utsagnet $P(-a \leq \underline{t} \leq a) = Q$. Siden t-fordelingen er flatere enn normalfordelingen vil disse a-verdiene være større

enn for normalfordelingen så lenge $f < \infty$. Til hver f svarer det en bestemt t-kurve og derfor et bestemt sett av a-verdier. I hovedtabell I bak i heftet har en gjengitt inne i tabellen de a-verdier som svarer til ulike f-verdier for P = 0,05 (Q = 0,95) og for P = 0,01 (Q = 0,99). I nedre linje av tabellen (som altså også kan oppfattes som en tabell for den standardiserte normalfordelingen) finner en de tallene som er gjengitt på s. 71.

Vi skal ta med et annet viktig eksempel på en størrelse som er fordelt etter Students t-fordeling. La oss betrakte to forskjellige universer U_1 og U_2. Til gjentakene i U_1 knytter det seg en random variabel $\underline{X}_1$: N (μ_1, σ) og til gjentakene i U_2 knytter det seg en random variabel $\underline{X}_2$: N (μ_2, σ). De to random variable har altså begge normal fordelingsfunksjon med standardavvik σ, men forventningene er forskjellige.

Vi tenker oss at vi tar et random sampel på n_1 observasjoner fra U_1 og et random sampel på n_2 observasjoner fra U_2. Gjennomsnittet og middelavviket for de n_1 observasjonene gir vi betegnelsene $\overline{X}_1$ og s_1, mens vi for de tilsvarende størrelsene i det andre samplet bruker betegnelsene $\overline{X}_2$ og s_2. (Legg merke til at fotindeksene 1 og 2 først og fremst refererer seg til universene U_1 og U_2, selv om de også refererer seg til samplene fra de to universene. $\overline{X}_1$ og $\overline{X}_2$ er altså gjennomsnitter for sampler fra <u>to forskjellige</u> universer, mens vi i avsnitt VIII, C brukte symbolene $\overline{X}_1$ og $\overline{X}_2$ for gjennomsnitter for sampler fra <u>et enkelt</u> univers.)

Hvis vi nå for hvert av de to universene gjennomfører samme resonnement som i avsnitt VIII, C, dvs. at vi ser for oss en uendelighet av tenkte sampler, er det klart at gjennomsnittene og middelavvikene kan betraktes som random variable.

I overensstemmelse med vår vanlige skrivemåte gir vi disse random variable betegnelsene $\underline{\bar{X}}_1$, $\underline{\bar{X}}_2$, $\underline{s}_1$ og $\underline{s}_2$. Når vi på en ganske spesiell måte former en ny random variabel som en funksjon av disse fire random variable, kan det vises at den nye random variable under våre forutsetninger er fordelt etter t-fordelingen. Vi betegner den derfor med $\underline{t}$ og skriver den på følgende måte:

$$(49) \qquad \underline{t} = \frac{\underline{\bar{X}}_1 - \underline{\bar{X}}_2 - (\mu_1 - \mu_2)}{\underline{s}_p \sqrt{\frac{n_1+n_2}{n_1 \cdot n_2}}}$$

Her er den random variable $\underline{s}_p$ et veid gjennomsnitt av de to random variable $\underline{s}_1$ og $\underline{s}_2$:

$$(50) \qquad \underline{s}_p^2 = \frac{(n_1-1)\,\underline{s}_1^2 + (n_2-1)\,\underline{s}_2^2}{(n_1-1) + (n_2-1)}$$

Når den random variable $\underline{t}$ framkommer slik som vist i (49) og (50) er parameteren f gitt ved følgende regel:

$$(51) \qquad f = (n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$$

Hvis vi har bare et eneste random sampel fra hvert av de to universene U_1 og U_2, finner vi ved utregning en verdi av hver av de random variable $\underline{\bar{X}}_1$, $\underline{\bar{X}}_2$, $\underline{s}_1$, $\underline{s}_2$ og $\underline{s}_p$, nemlig verdiene $\bar{X}_1$, $\bar{X}_2$, s_1, s_2 og s_p. Disse verdiene er i et konkret tilfelle tall som vi regner ut. s_p finnes av (50) idet vi sløyfer understrekningen under s-ene. Formlene (50) og (49) gjelder nemlig like fullt om vi sløyfer understrekningene under de random variable. Vi får da formler til å finne <u>verdiene</u> s_p^2 og t av $\underline{s}_p^2$ og $\underline{t}$ i konkrete tilfelle. For å finne t må vi imidlertid også kjenne differensen mellom μ_1 og μ_2.

Ved praktiske beregninger (som vi kommer til senere) lønner det seg ofte å skrive om telleren på høyre side i formelen (50). En eventuell omskrivning kan foretas ved hjelp av formelen (28) eller formelen (29) på s. 57.

X. Litt om forsøksplaner i biologien.

I biologisk forskning gjøres det utstrakt bruk av forsøk (eksperimenter). I veterinærmedisinen er det f.eks. ofte aktuelt å gjøre forsøk med forskjellige behandlingsmåter mot en bestemt sykdom hos et bestemt dyreslag. For at forsøkene skal gi oss den informasjonen vi ønsker må de planlegges omhyggelig. Forsøksplanlegging og analyse av forsøksdata er et stort og viktig felt innen statistikken. Vi skal her bare kort omtale grunntrekkene i to av de viktigste forsøksplanene. Framstillingen vil bli knyttet til eksempler.

A. Fri randomisering.

Anta at vi ønsker å sammenlikne 4 forskjellige måter å behandle voksne hester på etter en bestemt operasjon (f.eks. etter kastrering). Vi vil tenke oss at tiden som går fra operasjonen fram til full helbredelse (definert på en bestemt måte) er et brukbart mål for hvor vellykket en bestemt behandlingsmåte har vært. De 4 behandlingsmåtene vil vi betegne med T_1, T_2, T_3 og T_4 eller generelt med T_j ($j = 1,2, \ldots,k$). (T står for treatment = behandlingsmåte eller forsøksledd. Symbolene j og k har vi brukt tidligere, men her står de for noe nytt.)

Vi tenker oss at vi har et random sampel på f.eks. N = 36 hester til rådighet for våre forsøk. Vi kan kanskje tillate oss å oppfatte de 36 første hester som kommer inn til behandling som et random sampel fra universet U av alle hester som det kan bli aktuelt å operere.

La oss tenke oss at vi bestemmer oss for å prøve hver behandlingsmåte på 9 hester. Spørsmålet blir da hvilket prinsipp vi skal følge når vi skal fordele de 36 hestene på de 4 behandlingsmåtene. Hvis vi velger prinsippet som kalles fri randomisering, må vi fordele de 36 hestene helt tilfeldig på de 4 behandlingsmåtene. Vi kan f.eks. ha en beholder med 36 lapper: 9 som det står skrevet T_1 på, 9 som det står T_2 på, osv. Hver gang vi får en hest til behandling trekker vi tilfeldig en lapp for å bestemme hvilken behandlingsmåte vi skal bruke.

Ved denne forsøksplanen oppnår vi at alle de 4 samplene på 9 hester før operasjonen kan oppfattes som random sampler fra et og samme univers, U. Hvis vi i stedet hadde fordelt hestene etter et eller annet kjennetegn, f.eks. slik at vi fortrinnsvis lot de yngste hestene få behandlingsmåten T_1 og de eldste behandlingsmåten T_4, ville samplet som ble brukt til T_1 være et sampel fra et subunivers av U, nemlig et subunivers av unge hester. I et slikt tilfelle ville det nærmest være umulig å avgjøre om eventuelle forskjeller i helbredelsestiden skulle tilskrives behandlingsmåten eller alderen. Når vi bruker fri randomisering kan de 4 samplene etter operasjonen oppfattes som random sampler fra 4 subuniverser av U. Disse 4 subuniversene atskiller seg fra hverandre kun på en måte, nemlig ved at behandlingsmåtene er forskjellige.

Når forsøket er gjennomført, oppfatter vi observasjonene av helbredelsestiden som observasjoner av 4 random variable, $\underline{X}_1$, $\underline{X}_2$, $\underline{X}_3$ og $\underline{X}_4$, en for hvert subunivers. Forventningene for $\underline{X}$-ene betegner vi med μ_1, μ_2, μ_3 og μ_4 og standardavvikene med σ_1, σ_2, σ_3 og σ_4.

Observasjon nr. i av random variabel nr. j betegner vi med X_{ij}

$(i = 1,2,\ldots, n;\ j = 1,2,\ldots,k)$. I vårt eksempel er $n = 9$ og $k = 4$.

Hvis vi har den spesielle situasjonen at alle $\underline{X}$-ene har identisk samme fordelingsfunksjon, vil dette si at alle behandlingsmåtene har identisk samme virkning. Alle $\underline{X}$-ene har da samme forventning som vi vil betegne med μ og samme standardavvik som vi vil betegne med σ. I en slik situasjon kan vi like gjerne si at vi har et random sampel på $nk = 36$ observasjoner av en enkelt random variabel $\underline{X}$ som å si at vi har 4 random sampler på $n = 9$ gjentak, et sampel for hver random variabel $\underline{X}_j$.

I vårt eksempel brukte vi like mange hester til hver behandlingsmåte. Det kan vises at dette er den beste måten å utnytte et gitt antall hester på. Det kan imidlertid hende at en hest eller to må slaktes eller faller ut av forsøket på annen måte. I slike tilfelle vil vi betegne antall hester som får behandlingsmåte nr. j med n_j $(j = 1,2,\ldots,k)$.

Vi skal i senere hovedavsnitt se eksempler på hvorledes observasjonene fra forsøk etter forskjellige forsøksplaner kan analyseres statistisk.

B. Blokkplanen.

Sett at vi ønsker å sammenlikne vekten av kyllinger som er oppvokst på $k = 3$ forskjellige måter, nemlig T_1 ="inne på bur", T_2 = "inne i binge" og T_3 = "ute". Det er ingen ting i veien for at vi kunne bruke den samme forsøksplanen her som i foregående avsnitt, men vi skal nå illustrere en annen forsøksplan som i dette tilfelle antagelig er bedre.

Først skaffer vi oss et random sampel av "kyllingkull" (kyllinger som er helsøsken). Hvert "kull" må være på minst 3 kyllinger. Kullene bør være et random sampel av kull fra det universet U av kull som vi ønsker å uttale oss om. Fra hvert kull plukker vi ut ved loddtrekning 3 kyllinger som fordeles tilfeldig på de tre oppalsmåtene.

Denne forsøksplanen kalles blokkplanen. Hvert kull eller foreldrepar kalles en blokk. Som nevnt kan vi tenke oss et univers U av blokker. Siden kyllingene som tildeles de forskjellige oppalsstedene tilhører de samme blokkene og er blitt fordelt på oppalsstedene ved loddtrekning, kan de 3 gruppene av kyllinger oppfattes som sampler fra et og samme univers før forsøket settes i gang, nemlig et univers av enkeltkyllinger. De 3 samplene av kyllinger er imidlertid ikke uavhengige av hverandre da kyllingene i de forskjellige samplene jo er søsken. Ved analysen av et slikt forsøk må en derfor bruke en annen teknikk enn ved analysen av forsøk etter prinsippet fri randomisering.

De tre samplene kan ved forsøkets slutt oppfattes som sampler fra tre subuniverser av kyllinger. Disse subuniversene atskiller seg fra hverandre kun ved at oppalsmåten har vært forskjellig.

Også ved bruk av denne forsøksplanen kan observasjonene, når forsøket er gjennomført, oppfattes som observasjoner av k random variable og betegnes med X_{ij} ($i = 1,2,\ldots,n$; $j = 1,2..,k$). Forventningen for den random variable X_j betegnes som før med μ_j, og standardavviket med σ_j.

Hvis de k random variable X_j har identisk samme fordelings-

funksjon slik at altså oppalsmåtene er helt likeverdige kan vi igjen si at vi har nk observasjoner av en enkelt random variabel $\underline{X}$ med forventning μ og standardavvik σ.

(Modellen for et blokkforsøk er videre behandlet på s. 94.)

C. Sammenlikning av planene

Vi skal her bare kort påpeke at en forutsetning for at det skal være noen fordel å bruke blokkplanen er at variasjonen i observasjonene fra blokk til blokk er stor sammenliknet med variasjonen innenfor en og samme blokk. I vårt eksempel må altså vektforskjellen mellom kyllingene etter forskjellige foreldrepar være betydelig sammenliknet med vektforskjellen mellom kyllinger som er søsken.

I de fleste situasjoner lar det seg gjennomføre å bruke en hvilken som helst av de to planene. I vårt eksempel med hester kunne vi f.eks. latt forskjellige raser eller alternativt forskjellige aldersgrupper utgjøre blokkene. I eksemplet med kyllinger kunne vi også lett ha brukt planen fri randomisering. Som regel er imidlertid en av planene å foretrekke for et gitt problem.

Hvilken plan en vil bruke i et gitt tilfelle er et spørsmål som en kan avgjøre fritt etter å ha vurdert situasjonen på beste måte. Når en har valgt en bestemt plan og har gjennomført forsøket, må en imidlertid bruke de statistiske analysemetoder som er forenlige med vedkommende plan. De observasjonene en får vil jo selvsagt være forskjellige alt etter hvilken plan en har brukt. Å analysere et blokkforsøk som om det skulle være et forsøk etter planen fri randomisering, eller omvendt, er helt utillatelig.

XI. Estimering

Vi har tidligere (s. 15) nevnt litt om estimering av sannsynligheter. Vi skal nå ta med mer om estimering i sin alminnelighet. Estimeringsteorien er et av de sentrale felter innen statistikken.

A. Punktestimering

Vi skal først gjennomgå et eksempel for å belyse en av de mange typer av problemstillinger vi møter i estimeringsteorien.

Sett at vi ønsker å bestemme innholdet i gram pr. liter av et kjemikalium A i en oppløsning O som vi har fått tilsendt en viss mengde av i en beholder. I forbindelse med den kjemiske analysen foretar vi flere veiinger, målinger, avlesninger, osv. og hver gang er det muligheter for å gjøre større eller mindre feil. Vi foretar derfor flere gjentatte analyser. I slike tilfelle vil analyseresultatene gjerne variere fra analyse til analyse selv om de egentlig skulle bli identiske. (Vi går her ut fra at uttakingen av prøver til analyse ikke er forbundet med feil og at det ikke foregår fordampning e.l. fra beholderen. Vi kan da si at innholdet av A i beholderen i gram pr. liter er konstant og at variasjonene i analyseresultatene bare skyldes feil.)

Vi vil forutsette at den personen som foretar analysene ikke har noen systematisk tendens til å arbeide på en slik måte at feilene går i en bestemt retning. En tilsvarende forutsetning vil vi gjøre om de metoder og apparater, m.v. som brukes. Vi sier da at analysene er fri for systematiske feil. De feilene som forekommer er da tilfeldige feil som ved gjentatte analyser i det

lange løp opphever hverandre. Vi vil også forutsette at feilene ved gjentatte analyser er uavhengige av hverandre.

La oss skyte inn her at systematiske feil som det ikke er mulig å eliminere ved justering av instrumenter o.l. ofte kan nøytraliseres på den måten at de gjøres om til å bli tilfeldige feil ved hjelp av randomisering. Hvis en person har en tendens til å "avlese på skrå" den ene veien, vil det gjerne finnes personer som vanligvis "avleser på skrå" den andre veien. Hvis vi bruker et random sampel av personer i stedet for en enkelt person til å utføre analysene, vil slike feil kunne gå over til å bli tilfeldige feil som snart går i en retning, snart i en annen. På tilsvarende måte kunne vi også tenke oss å bruke et random sampel av instrumenter e.l. for et bestemt formål i stedet for et enkelt instrument. Vanskeligheten i praksis er selvsagt at det gjerne er et meget begrenset antall personer og metoder osv. som står til vår disposisjon.

I vårt eksempel vil vi tenke oss et univers hvor hver analyse er et gjentak. Dette universet vil vi forestille oss er uendelig.

Resultatet av en analyse kan betraktes som en random variabel $\underline{X}$ mens det sanne innholdet av A i gram pr. liter er en konstant, μ . Under våre forutsetninger er

(52) $$E(\underline{X}) = \mu = \text{det sanne innholdet av stoff A.}$$

La oss nå innføre en del viktige begreper som vi så vidt har vært borte i tidligere. Definisjonene vil bli gitt i tilknytning til vårt eksempel, men brukes generelt i en rekke andre sammenhenger.

Oppgaven med å finne et tilnærmet riktig tall for det sanne innholdet av stoffet A er et estimeringsproblem. Siden oppgaven er å finne et enkelt tall, dvs. et punkt på tall-linjen, snakker vi om punktestimering i motsetning til intervallestimering. Ved intervallestimering finner vi et intervall på tall-linjen hvor vi påstår at den ukjente størrelsen (i vårt tilfelle μ) befinner seg. Den ukjente størrelsen, μ som vi dypest sett egentlig er interessert i kan vi kalle en estimand. I mange estimeringsproblemer er estimanden en ukjent parameter i en fordelingsfunksjon. I vårt tilfelle vil således μ være en parameter i den normale fordelingsfunksjon hvis $\underline{X}$ er normalt fordelt. Derfor brukes ofte betegnelsen parameter i stedet for estimand som er lite brukt. Estimanden behøver selvsagt ikke å være en forventning. I den generelle estimeringsteorien bruker en derfor ofte θ i stedet for μ som betegnelse på estimanden.

Utgangspunktet for estimeringen (det å finne en tilnærmingsverdi for estimanden) er et random sampel på n gjentak (n analyser i vårt eksempel). Vi skaffer oss altså n verdier, X_1, X_2, ..., X_n av en random variabel $\underline{X}$.

For å komme fram til en tilnærmingsverdi for estimanden tar vi vanligvis utgangspunkt i en funksjon av disse n observasjonene (f.eks. gjennomsnittet av dem). Siden vi kan tenke oss en uendelighet av sampler på n gjentak, kan vi også tenke oss en uendelighet av verdier for denne funksjonen (se avsnitt VIII C). Funksjonen kan derfor oppfattes som en random variabel. Funksjonen blir kalt en estimator. Når vi ser bort fra degenererte tilfelle er en estimator alltid en random variabel.

Enhver verdi av denne random variable er en funksjon av observasjonene i et sampel. I den generelle estimeringsteorien blir en estimator ofte betegnet med $\underline{\hat{\theta}}$. I vårt eksempel kunne vi tenke oss å bruke en rekke forskjellige estimatorer (gi eksempler). Av grunner som vi skal komme tilbake til vil vi imidlertid bruke gjennomsnittet $\underline{\bar{X}}$.

I et gitt estimeringsproblem tar vi som regel ut bare et eneste sampel. Vi får da n verdier $X_1, X_2, X_3, \ldots, X_n$ av den random variable $\underline{X}$ og disse kan vi bruke til å regne ut tallet $\bar{X}$ som er en verdi av den random variable $\underline{\bar{X}}$. Tallet $\bar{X}$ kaller vi et estimat.

For å summere opp i tilknytning til eksemplet ovenfor, kan vi altså si at det sanne innholdet av kjemikaliet A, f.eks. θ = 16 gram pr. liter er en estimand (en ukjent konstant), gjennomsnittet $\underline{\bar{X}} = \underline{\hat{\theta}}$ er en estimator (en random variabel) og tallet $\bar{X} = \hat{\theta}$ = 18,83 gram pr. liter er et estimat (en verdi av den random variable $\underline{\hat{\theta}} = \underline{\bar{X}}$)

Vi kan som regel velge mellom flere estimatorer, og det gjelder da å velge en god estimator. De egenskapene som avgjør om en estimator er god eller ikke er knyttet til estimatorens fordelingsfunksjon.

I vårt eksempel valgte vi å bruke estimatoren $\underline{\bar{X}}$. Selv om vi ikke kjenner fordelingsfunksjonen for $\underline{\bar{X}}$, vet vi (se (33), s. 60 og (52), s. 83) at forventningen for $\underline{\bar{X}}$ er lik μ. Forventningen for estimatoren $\underline{\bar{X}}$ er altså lik estimanden μ. Denne viktige egenskapen ved estimatoren har vi et bestemt navn på. Vi sier generelt at en estimator $\underline{\hat{\theta}}$ er forventningsrett (unbiased) hvis forventningen for estimatoren er lik estimanden θ, altså hvis

(53) $$E(\underline{\hat{\theta}}) = \theta$$

Hvis $E(\underline{\hat{\theta}}) - \theta \neq 0$ sier vi at $\underline{\hat{\theta}}$ er <u>forventningsskjev</u> (biased) og $E(\underline{\hat{\theta}}) - \theta$ blir kalt forventningsskjevheten (the bias).

Populært uttrykt kan betydningen av å bruke forventningsrette estimatorer illustreres på følgende måte: Hvis vi alltid bruker forventningsrette estimatorer, vil vi i det lange løp få riktige resultater i gjennomsnitt.

Vi har tidligere ((34), s. 60) vist at $var(\underline{\bar{X}}) = \sigma^2/n$. Det er naturligvis en fordel at en estimator har liten varians. Variansen for $\underline{\bar{X}}$ er, som vi kunne vente, mindre når n er stor enn når n er liten. (Gjennomsnittshøyden for 100 mann varierer mindre fra sampel til sampel enn gjennomsnittshøyden for 2 mann). Hvor stor vi skal gjøre n i et gitt tilfelle vil bl.a. avhenge av hvilke økonomiske ressurser som står til vår disposisjon.

Hvis vi har valget mellom en rekke estimatorer som alle er forventningsrette, velger vi gjerne den som har <u>minst varians</u> for den sampelstørrelsen vi vil bruke.

Det finnes også en rekke andre ønskelige egenskaper ved estimatorer, men vi har ikke tid til å komme inn på disse her.

Vi har tidligere (s. 52-53 og s. 56) vært inne på at det av og til er aktuelt å sammenlikne de relative frekvensene i et sampel med sannsynligheter i universet for å undersøke om en random variabel kan antas å følge en gitt fordelingsfunksjon. For å kunne regne ut sannsynlighetene må vi kjenne fordelingsfunksjonens parametre. Hvis disse er ukjente må de estimeres. Dette kan bl.a. gjøres ved at vi setter forventningen lik gjennomsnittet og den teoretiske variansen lik den empiriske. Når vi f.eks. har å gjøre med den normale fordelingsfunksjonen, får vi da de estimatene vi trenger direkte. Har vi derimot å gjøre med

den binomiale fordelingsfunksjonen, får vi to likninger til bestemmelse av estimatene, nemlig $kp=\bar{X}$ og $kp(1-p)=s^2$ (se (37) s. 65 og (38) s. 66). Hvis vi kjenner den største verdien $\underline{X}$ kan anta kan k settes lik denne. Vi slipper da å estimere k og estimerer da bare p ved hjelp av likningen $kp=\bar{X}$. Parameteren m i Poissons fordelingsfunksjon kan estimeres ved $\bar{X}$.

B. Intervallestimering

Intervallestimeringen bygger på de samme grunnleggende prinsipper som punktestimeringen og blir gjerne foretatt i tilknytning til og som en videreføring av denne. Problemstillingen er følgende: Vi ønsker å konstruere et intervall på tall-linjen som har en slik bredde og plassering at det er en bestemt, på forhånd valgt sannsynlighet Q for at intervallet skal falle slik at det inneholder estimanden ϑ. Et slikt intervall blir kalt et konfidensintervall, og grensene for intervallet betegnes som konfidensgrensene. Sannsynligheten Q blir kalt konfidenssannsynligheten.

Det faktum at det er mulig å lage konfidensintervaller for helt ukjente størrelser er et slående eksempel på hva en kan oppnå ved bruk av statistiske metoder.

1. Konfidensintervall for forventningen for en normalt fordelt random variabel.

Sett at vi er interessert i å beregne konfidensgrenser for forventningen μ for en normalt fordelt random variabel $\underline{X}$. Vi skaffer oss da et random sampel på n observasjoner av $\underline{X}$. Hvis

også standardavviket σ er ukjent, tar vi utgangspunkt i den random variable $\underline{t}$ som er definert ved (47), s. 72. Denne er fordelt etter Students fordelingsfunksjon med n-1 frihetsgrader. I hovedtabell I bak i heftet er det gjengitt sammenhørende verdier av a (inne i tabellen) og P som er lik 1-Q. Den definisjonsmessige sammenhengen mellom Q og a er gitt ved følgende sannsynlighetsutsagn om $\underline{t}$:

$$P(-a \leq \underline{t} \leq a)=Q \tag{54}$$

Dette er også illustrert i fig. 8 hvor det skraverte arealet er lik Q.

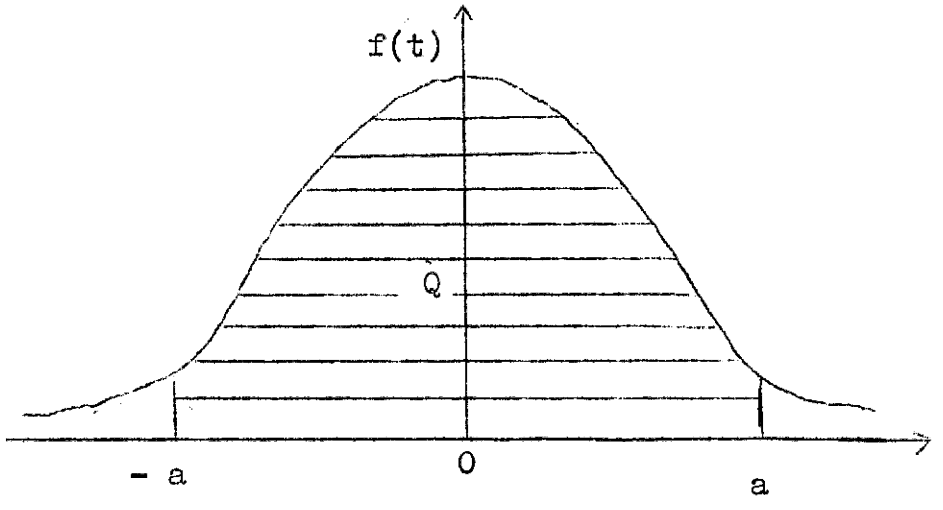


Fig. 8

Ved å sette (47) inn i (54) får vi da

$$P\left(-a \leq \frac{\overline{\underline{X}}-\mu}{\underline{s}/\sqrt{n}} \leq a\right)=Q \tag{55}$$

Dette er fremdeles et sannsynlighetsutsagn om $\underline{t}$, men det kan i denne form også oppfattes som et sannsynlighetsutsagn om $\overline{\underline{X}}$ og $\underline{s}$.

Det som står inne i parentesen i (55) er ulikheter eller likheter som selvsagt kan omformes etter matematikkens regler uten at dette forstyrrer riktigheten av sannsynlighetsutsagnet. Hvis vi inne i parentesen først multipliserer alle tre ledd med $\underline{s}/\sqrt{n}$, deretter trekker $\bar{\underline{X}}$ fra alle ledd og til sist multipliserer alle ledd med -1 samtidig som vi snur ulikhetstegnene og skriver ulikhetene i motsatte rekkefølge, får vi:

$$(56) \qquad P(\bar{\underline{X}} - a\underline{s}/\sqrt{n} \leq \mu \leq \bar{\underline{X}} + a\underline{s}/\sqrt{n}) = Q$$

Sannsynligheten for at et intervall som strekker seg fra $\bar{\underline{X}} - a\underline{s}/\sqrt{n}$ til $\bar{\underline{X}} + a\underline{s}/\sqrt{n}$ skal falle slik at det inneholder μ er altså lik Q. De to grensene for intervallet er random variable siden de er funksjoner av de to random variable $\bar{\underline{X}}$ og $\underline{s}$. For hvert sampel på n observasjoner av $\underline{X}$ kan vi regne ut en bestemt verdi $\bar{X}$ av $\bar{\underline{X}}$ og en bestemt verdi s av $\underline{s}$. Dermed kan vi også finne et bestemt intervall siden a kan finnes av hovedtabell I for f=n-1 og n er et kjent tall.

Til hvert sampel svarer det altså et ganske bestemt intervall. Intervallene vil variere fra sampel til sampel både i bredde og beliggenhet, men i følge (56) vil en brøkdel Q av alle tenkelige intervaller konstruert på denne måten falle slik at de dekker μ som altså hele tiden ligger fast på tall-linjen.

Når vi i et konkret tilfelle skal finne et konfidensintervall tar vi ut et random sampel av en valgt størrelse n. Idet vi bruker den a som svarer til den Q vi har valgt og til f=n-1, finner vi konfidensgrensene ved hjelp av observasjonene X_1, X_2, ..., X_n av $\underline{X}$ på følgende måte:

(57) Nedre konfidensgrense: $\bar{X} - as/\sqrt{n}$

Øvre konfidensgrense : $\bar{X} + as/\sqrt{n}$

I følge (56) vet vi at en brøkdel Q av alle slike intervaller dekker μ. Sannsynligheten for at det intervallet vi finner ved hjelp at et enkelt vilkårlig sampel skal dekke μ kan derfor sies å være lik Q.

Vi legger merke til at punktestimatet $\bar{X}$ danner midtpunktet i konfidensintervallet. Bredden B av konfidensintervallet er lik øvre konfidensgrense minus nedre grense:

(58) $$B = 2as/\sqrt{n}$$

Vi skal illustrere metodikken ved et par eksempler:

Eks. 1. Sett at vi foretar 6 uavhengige analyser for å bestemme innholdet av et kjemikalium A i en oppløsning O. (Vårt tidligere eksempel.) Vi ønsker å finne et konfidensintervall for det sanne innholdet av A målt i gram pr. liter idet vi bruker konfidenssannsynligheten Q = 0,95. Vi forutsetter at analyseresultatene er fri for systematiske feil og at de tilfeldige feil er normalt fordelt. Analyseresultatene og utregningen av konfidensintervallet er vist nedenfor:

Analyse nr.	Innhold av A, gram pr. liter X
1	23
2	19
3	32
4	14
5	16
6	9
	$\Sigma X = 113$

$\bar{X} = 18,8333$

$\Sigma X^2 = 2447$

$(\Sigma X)^2 = 113^2 = 12769$

$\frac{(\Sigma X)^2}{n} = 2128,166$

$$s^2 = \frac{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}{n-1}$$

$$= \frac{2447 - 2128,166}{5} = 63,7668$$

$s = \sqrt{63,7668} = 7,9854$ $\qquad f = n-1=5$

Nedre grense: $\bar{X} - as/\sqrt{n} = 18,8333-2,571 \cdot 7,9854/\sqrt{6} = \underline{\underline{10,45}}$

Øvre grense : $\bar{X} + as/\sqrt{n} = \underline{\underline{27,21}}$

Hvis vi påstår at det sanne innholdet av kjemikaliet A ligger mellom ca. 10 og ca. 27 gram pr. liter er altså sannsynligheten for at vårt utsagn er riktig lik 0,95. Hvis vi skal oppgi et enkelt tall for innholdet av A er ca. 19 gram pr. liter det beste tallet vi kan bruke i dette tilfelle siden det er funnet ved å bruke en forventningsrett estimator.

Vi støter ofte på den oppgaven å estimere den sanne verdien av en størrelse på grunnlag av et random sampel av målinger, veiinger, e.l. Hvis variasjonen i observasjonene bare skyldes tilfeldige feil som kan antas å være tilnærmet normalt fordelt kan en gå fram som i eksempel 1 ovenfor. Både teoretiske overveielser og praktisk erfaring viser at slike målefeil ofte kan antas å være tilnærmet normalt fordelt.

I eksempel 1 var det kun <u>feil</u> av forskjellige slag som førte til at observasjonene varierte fra gjentak til gjentak. I andre tilfelle kan variasjonen være en del av selve det fenomenet som observeres. Også i slike tilfelle kan en ofte bruke den teknikken som er gjennomgått ovenfor. Et aktuelt eksempel er gitt i øvelse 24 nedenfor.

Øvelse 24.

For å beregne den normale størrelsen av hjertet hos voksne hanner av et bestemt dyreslag, ble det tatt ut et random sampel på 10 voksne handyr. Størrelsen av hjertet hos disse dyrene er gjengitt nedenfor i gram.

48 40 52 32 37 26 24 46 41 37

Vi vil forutsette at vekten av hjertet er en random variabel med normal fordelingsfunksjon. Finn konfidensgrensene for forventningen for vekten av hjertet. Bruk Q = 0,95.

Øvelse 25.

Studer formelen (58), s. 90 for bredden av konfidensintervallet og kontroller at valget av Q og n påvirker bredden i den retning en skulle vente. Forklar også hvorfor det synes rimelig at s inngår i formelen på den måten den gjør.

Vi skal til slutt se hvorledes vi kan anvende de samme formlene på oppgaver av en litt annen type. Anta at vi har et forsøk etter blokkplanen med bare 2 forsøksledd. I vårt eksempel på s. 79 kan vi f.eks. tenke oss at vi bare har med forsøksleddene T_2 = "inne i binge" og T_3 = "ute".

Det vi først og fremst ønsker å finne ut ved vårt forsøk er hvor mye forventningen μ_2 for vekten $\underline{X}_2$ av kyllinger som vokser opp inne i binge avviker fra forventningen μ_3 for vekten $\underline{X}_3$ av kyllinger som vokser opp ute. Vi tenker oss altså to universer og to random variable. (I stedet for universer kunne vi her i stedet snakke om subuniverser (se s. 80), men et subunivers er jo også et univers, så vi sløyfer forstavelsen sub.)

La oss også formulere problemet på en annen måte. Vi husker at to og to kyllinger hadde samme foreldre. Det kunne derfor være naturlig å se på differensen mellom vekten for den kyllingen som har vokst opp inne og den som har vokst opp ute. Når vi betrakter et univers av foreldrepar (blokker) kan denne differensen oppfattes

som en random variabel som vi vil betegne med $\underline{X}$. Forventningen for $\underline{X}$ vil vi betegne med μ. Formulert på denne måte kan vi si at problemet egentlig er å estimere μ. Det kan for øvrig lett vises at $\mu=\mu_2-\mu_3$.

Hvis vi nå kan tillate oss å forutsette at $\underline{X}$ er tilnærmet normalt fordelt, ser vi at problemet formelt sett svarer helt ut til det problemet vi løste i eksempel 1 ovenfor. Forskjellen er bare at det nå er de observerte differensene som oppfattes som verdier av en random variabel $\underline{X}$. Vektene for de to sett av kyllinger kan vi så å si glemme helt etter at vi har regnet ut rekken av differenser. Vi skal belyse metoden nærmere ved et talleksempel med 10 blokker. Vektenheten er 10 gram.

Eks. 2.

Foreldrepar nr.	X_{i2}	X_{i3}	$X_i=X_{i2}-X_{i3}$
1	27	24	3
2	51	45	6
3	42	33	9
4	39	33	6
5	45	27	18
6	30	32	-2
7	33	33	0
8	39	30	9
9	39	42	-3
10	45	42	3
			$X=49$
			$\bar{X}=4,9$

$$s^2=\frac{\Sigma X^2-\frac{(\Sigma X)^2}{n}}{n-1}=\frac{589-\frac{49^2}{10}}{9}=38,7667 \qquad s=6,226$$

$f=n-1=9$ Q velges lik 0,95 $a=2,262$

Ved å sette inn i (57), s. 90 får vi:

Nedre grense = $\underline{0,45}$ Øvre grense = $\underline{9,35}$

Legg merke til at n i det siste eksemplet var lik antall differenser eller foreldrepar eller blokker og altså lik halvparten av antall kyllinger som inngår i forsøket.

De 20 individuelle kyllingvektene i eksempel 2 ovenfor varierer av flere grunner. Vi skal nevne litt mer om den statistiske modellen for et blokkforsøk i tilknytning til dette eksemplet, selv om dette egentlig hører hjemme i et annet avsnitt. De random variable $\underline{X}_2$ og $\underline{X}_3$ kan skrives som en sum av ledd på følgende måte (symbolene betyr det samme som tidligere i dette avsnittet):

$$(59) \qquad \underline{X}_2 = \mu_2 + \underline{Z} + \underline{W}_2 + \underline{\varepsilon}$$

$$(60) \qquad \underline{X}_3 = \mu_3 + \underline{Z} + \underline{W}_3 + \underline{\varepsilon}$$

Uttrykkene (59) og (60) kan også skrives under ett:

$$(61) \qquad \underline{X}_j = \mu_j + \underline{Z} + \underline{W}_j + \underline{\varepsilon} \qquad (j=2,3)$$

Her er $\underline{Z}$ en random variable som har en verdi Z for hver blokk i universet av blokker. (Symbolet Z ble brukt for noe helt annet i hovedavsnitt VIII.) Z gir uttrykk for det vi kaller blokkeffekten eller i vårt tilfelle foreldreeffekten. $\underline{W}_j$ står i vårt tilfelle for to random variable som knytter seg til blokkene på samme måte som Z. Den random variable $\underline{W}_2$ kan nærmest sies å "tre i kraft" for kyllinger som vokser opp inne i binge mens $\underline{W}_3$ "trer i kraft" for kyllinger som vokser opp ute. De random variable $\underline{W}_j$ (j=2,3) gir uttrykk for det vi kaller samspillet mellom forsøksledd og blokk, i vårt tilfelle mellom oppalsmåte og foreldrepar. $\underline{\varepsilon}$ er en random variabel som har en verdi for hver kylling i hele universet av kyllinger. Forventningen for hver av de random variable $\underline{Z}$, $\underline{W}_1$, $\underline{W}_2$ og $\underline{\varepsilon}$ kan settes lik 0 da disse random variable på hver sin måte gir uttrykk for avvik fra μ_2 og μ_3 i (59) og (60).

Hvis vi tar for oss en enkelt av de 20 kyllingene foreligger det en bestemt verdi av $\underline{X}_j$, og dermed også av $\underline{Z}$, $\underline{W}_j$ og $\underline{\varepsilon}$ i (61). Hvis vi bruker verdiene av de random variable i stedet for de random variable selv kan modellen (61) for blokkforsøk skrives på følgende måte:

$$X_{ij} = \mu_j + Z_i + W_{ij} + \varepsilon_{ij} \qquad (i=1,2,..,n;\ j=2,3) \tag{62}$$

For å belyse symbolene nærmere, la oss tenke oss at vi betrakter det tilfelle da i = 4 og j = 3. Vi har da å gjøre med den kyllingen etter foreldrepar 4 (i = 4) som vokste opp ute (j = 3). Vekten av denne kyllingen er X_{43}. I følge (62) kan X_{43} skrives på følgende måte:

$$X_{43} = \mu_3 + Z_4 + W_{43} + \varepsilon_{43} \tag{63}$$

Av tabellen ovenfor ser vi at X_{43} = 33 vektenheter. La oss tenke oss at vi kjente leddene på høyre side i (63) og at disse hadde følgende verdier: μ_3 = 36, Z_4 = -1, W_{43} = -5 og ε_{43} = 3. Da ville vi kunne sette inn disse tallene i (63):

$$33 = 36 - 1 - 5 + 3 \tag{64}$$

Vi skal forklare hva disse tallene kan tenkes å gi uttrykk for. Tallet -1 for Z_4 viser at kyllinger etter foreldrepar nr. 4 under ellers like vilkår er 1 vektenhet lettere enn gjennomsnittet for kyllinger etter alle foreldrepar i universet. Dette kan bl.a. skyldes visse genetiske forhold som er felles for kyllinger etter dette foreldrepar og som gir seg samme utslag enten kyllingene vokser opp ute eller inne. Tallet for W_{43} viser at kyllinger etter foreldrepar nr. 4, når de vokser opp ute, under ellers like vilkår er 5 vektenheter lettere enn gjennomsnittet for kyllinger etter alle foreldrepar i universet. Dette kan f.eks. komme av at kyl-

linger etter dette foreldrepar har en dårlig fjørdrakt som fører til at de fryser og blir lette når de vokser opp ute. Når de vokser opp inne behøver ikke denne fjørdrakten å ha noen spesiell virkning. Tallet $\varepsilon_{43} = 3$ kan bl.a. ha sammenheng med at den spesielle kyllingen det her gjelder er noe mer "grådig i matfatet" enn kyllinger flest slik at den under ellers like vilkår er 3 enheter tyngre enn gjennomsnittet for alle kyllinger i det universet det er snakk om.

De eksempler på årsaksforhold vi har gitt for leddene på høyre side i (64) er selvsagt sterkt forenklede, men illustrerer prinsippene. Uttrykket (62) viser at alle 20 observasjoner kan skrives på liknende måte. (61) og (62) kan også uten videre generaliseres til å gjelde mer enn to forsøksledd. Leddene på høyre side av (62) er aldri kjent i praksis. Vi er imidlertid interessert i å estimere visse funksjoner av μ_j (som f.eks. $\mu_1 - \mu_2$ i vårt eksempel). Disse funksjonene kalles kontraster. Videre er vi ofte interessert i å vite om leddet W_{ij} virkelig eksisterer eller om det kan settes lik 0, da dette ofte avgjør hvilken analyseteknikk vi bør bruke.

Grunnen til at vi tok med disse betraktningene om blokkmodellen i dette avsnittet er at vi ville vise hva som skjer når vi former differensene $X_i = X_{i2} - X_{i3}$ slik som i eksempel 2 ovenfor. Bruker vi (62), ser vi at differensene kan skrives på følgende måte:

$$(65) \qquad \begin{aligned} X_{i2}-X_{i3} &= \mu_2+Z_i+W_{i2}+\varepsilon_{i2}-(\mu_3+Z_i+W_{i3}+\varepsilon_{i3}) \\ &= \mu_2-\mu_3+W_{i2}-W_{i3}+\varepsilon_{i2}-\varepsilon_{i3} \\ &= \mu+(W_{i2}-W_{i3})+(\varepsilon_{i2}-\varepsilon_{i3}) \qquad (i=1,2,\ldots,n) \end{aligned}$$

Vi ser at leddet Z_i som gir uttrykk for foreldreeffekten har falt bort. I siste kolonne i tabellen i eksempel 2 ovenfor har vi altså fjernet en kilde til variasjon, nemlig leddet Z_i. Det er dette vi generelt sett oppnår ved et blokkforsøk. Vi elimi-nerer en kilde til variasjon, nemlig blokkeffekten slik at denne ikke forstyrrer vår sammenlikning av forsøksleddene.

Observasjonene i eksempel 2 er et tilfelle av det vi ofte kaller parobservasjoner. Data fra et blokkforsøk med to forsøks-ledd er et viktig eksempel på parobservasjoner, men det finnes også andre eksempler (se øvelse 26). Hvis vi kan tillate oss å forutsette at differensene mellom parobservasjonene kan oppfattes som observasjoner av en random variabel $\underline{X}$ som er tilnærmet normalt fordelt, kan vi bruke den teknikken som er beskrevet i eksempel 2 til å beregne konfidensgrenser for $E(\underline{X})$.

Hvis vi i eksempel 2 hadde brukt differensene $X_{i3} - X_{i2}$ i stedet for $X_{i2} - X_{i3}$ ville konfidensgrensene blitt -9,35 og -0,45 i stedet for 0,45 og 9,35. En kan derfor komme fram til samme konklusjon uansett hvilken vei en tar differensene i slike tilfelle.

Øvelse 26.

Hos 7 dyr av et bestemt slag ble størrelsen av høyre og venstre nyre registrert. Vektene i gram er gjen-gitt nedenfor.

Dyr nr.	1	2	3	4	5	6	7
Høyre nyre	11,2	13,7	10,3	11,0	15,2	7,2	8,1
Venstre nyre	12,5	13,5	11,3	12,1	16,1	9,1	11,4

Estimer forskjellen mellom størrelsen av nyrene hos vedkommende dyreslag. Presiser problemstillingen og gjør rede for de forutsetninger som må gjøres.

Øvelse 27.

Forklar hvorfor en praktisk bonde kunne være interessert i å vite om standardavviket for $\underline{X}_2$ er forskjellig fra standardavviket for $\underline{X}_3$ i vårt eksempel med kyllinger ovenfor.

2. Konfidensgrenser for differensen mellom forventningene for to random variable som begge har normal fordelingsfunksjon

Sett at vi er interessert i å beregne konfidensgrenser for differensen $\mu_d = \mu_1 - \mu_2$ mellom forventningene μ_1 og μ_2 for to random variable $\underline{X}_1$ og $\underline{X}_2$ som begge har normal fordelingsfunksjon og som knytter seg til hvert sitt univers (eller subunivers). La oss anta at vi har et random sampel på n_1 observasjoner av $\underline{X}_1$ og et uavhengig random sampel på n_2 observasjoner av $\underline{X}_2$. Hvis standardavvikene σ_1 og σ_2 er ukjente, men kan forutsettes å være tilnærmet like ($\sigma_1 = \sigma_2 = \sigma$) kan vi ta vårt utgangspunkt i en random variabel $\underline{t}$ som er definert ved (49) og (50) s. 75. Denne følger Students t-fordeling med n_1+n_2-2 frihetsgrader.

Vi vet da (sammenlikn s. 88) at følgende sannsynlighetsutsagn må være riktig for alle samhørende verdier av Q, n_1+n_2 og a:

$$(73) \qquad P\left(-a \leq \frac{\overline{\underline{X}}_1 - \overline{\underline{X}}_2 - (\mu_1 - \mu_2)}{\underline{s}_p} \sqrt{\frac{n_1 n_2}{n_1+n_2}} \leq a\right) = Q$$

Etter litt matematisk omforming av det som står inne i parentesen får vi da:

$$(74) \qquad P\left(\overline{\underline{X}}_1 - \overline{\underline{X}}_2 - a\underline{s}_p \sqrt{\frac{n_1+n_2}{n_1 \cdot n_2}} \leq \mu_1 - \mu_2 \leq \overline{\underline{X}}_1 - \overline{\underline{X}}_2 + a\underline{s}_p \sqrt{\frac{n_1+n_2}{n_1 \cdot n_2}}\right) = Q$$

Vi kan altså regne ut konfidensgrenser for $\mu_1 - \mu_2$ på følgende måte:

(75)

Nedre konfidensgrense: $\bar{X}_1-\bar{X}_2-as_p\sqrt{\frac{n_1+n_2}{n_1\cdot n_2}}$

Øvre konfidensgrense : $\bar{X}_1-\bar{X}_2+as_p\sqrt{\frac{n_1+n_2}{n_1\cdot n_2}}$

I dette tilfelle er det punktestimatet $\bar{X}_1-\bar{X}_2$ som danner midtpunktet i intervallet og intervallets bredde er

(76) $$2as_p\sqrt{\frac{n_1+n_2}{n_1\cdot n_2}}$$

La oss se på et eksempel.

Eks. 3. Vi vil ta for oss det eksemplet vi beskrev da vi gjennomgikk forsøksplanen etter prinsippet fri randomisering (s.77-78). Vi vil anta at vi har med bare to behandlingsmåter T_1 og T_2 i forsøket. Videre vil vi tenke oss at 9 hester fikk behandlingsmåten T_1 og at 9 fikk T_2, men at en av de 9 sistnevnte hestene måtte slaktes. Hvis vi nå kan forutsette at $\underline{X}_1$ og $\underline{X}_2$ (definert s.78) begge har normal fordelingsfunksjon og at $\sigma_1=\sigma_2$, ser vi at vilkårene er til stede for å beregne konfidensgrensene for $\mu_1-\mu_2$ ved hjelp av (75). Dette er gjort nedenfor.

T_1	T_2
25	34
37	26
27	26
32	27
38	32
30	35
26	36
34	26
41	

$\Sigma X_1 = 290$ $\Sigma X_2 = 242$

$n_1 = 9$ $n_2 = 8$

$\bar{X}_1 = 32,22$ $\bar{X}_2 = 30,25$

$\Sigma X_1^2 = 9604$ $\Sigma X_2^2 = 7458$

$(\Sigma X_1)^2 = 84100$ $(\Sigma X_2)^2 = 58564$

$\frac{(\Sigma X_1)^2}{n_1} = 9344,4$ $\frac{(\Sigma X_2)^2}{n_2} = 7320,5$

$\bar{X}_1 - \bar{X}_2 = 32,22-30,25 = 1,97$ $Q=0,95$. $f=n_1+n_2-2=15$ $a=2,131$

$$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)(n_2-1)} = \frac{\Sigma(X_1-\bar{X}_1)^2+\Sigma(X_2-\bar{X}_2)^2}{n_1+n_2-2}$$

$$= \frac{\Sigma X_1^2 - \frac{(\Sigma X_1)^2}{n_1} + \Sigma X_2^2 - \frac{(\Sigma X_2)^2}{n_2}}{n_1+n_2-2} = \frac{9604-9344,4+7458-7320,5}{9+8-2}$$

$$= \frac{397,1}{15} = 26,4733 \qquad s_p = \sqrt{26,4733} = 5,145$$

$$as_p\sqrt{\frac{n_1+n_2}{n_1 \cdot n_2}} = 2,131 \cdot 5,145 \cdot \sqrt{\frac{9+8}{9 \cdot 8}} = 5,33$$

Nedre konfidensgrense: $\bar{X}_1 - \bar{X}_2 - as_p\sqrt{\frac{n_1+n_2}{n_1 \cdot n_2}} = 1,97-5,33 = \underline{\underline{-3,36}}$

Øvre konfidensgrense: $\bar{X}_1 - \bar{X}_2 + as_p\sqrt{\frac{n_1+n_2}{n_1 \cdot n_2}} = 1,97+5,33 = \underline{\underline{7,30}}$

Den metoden vi har gjennomgått i eksempel 3, s. 99 er vanlig brukt til analysere data fra forsøk etter planen fri randomisering, mens altså framgangsmåten i eksempel 2, s.93 brukes for blokkplanen. Eksempel 3 er et tilfelle av data som ikke er parobservasjoner, men vi kan også støte på slike ikke-parobservasjoner som er framkommet på annen måte enn ved forsøk (se øvelse 28 nedenfor). Hvis de forutsetninger metoden bygger på kan antas å være tilnærmet oppfylt, kan en da bruke samme analysemetode som i eksempel 3.

At tallene i eksempel 3 ikke er parobservasjoner er for det første klart av det faktum at $n_1 \neq n_2$. Men selv om n_1 og n_2 hadde vært like, ville vi fremdeles ikke hatt parobservasjoner. Den vertikale arrangeringen av observasjonene i eksempel 3 er nemlig helt vilkårlig. Vi kan foreta en vertikal omplassering av tallene i en av kolonnene uten at dette forandrer noe som helst på realitetene. Det er altså svært mange måter å gruppere sammen en observasjon fra T_1 og en observasjon fra T_2 på. I eksempel 2, derimot, kan vi ikke bytte om to tall i den ene kolonnen uten av vi foretar en tilsvarende ombytting av tallene i den andre kolonnen. Tallene på en og samme linje hører sammen, og de knytter seg også til et felles gjentak, i dette tilfelle til et felles foreldrepar.

De metodene til å beregne konfidensintervall som vi hittil har gjennomgått er basert på t-fordelingen. Disse metodene har heldigvis vist seg å være robuste både overfor avvik fra forutsetningen om normalitet og overfor avvik fra den tilleggsforutsetningen vi måtte gjøre i avsnitt IX B 2 om at $\sigma_1 = \sigma_2$. Den størrelsen vi i hvert tilfelle har forutsatt er t-fordelt er altså tilnærmet t-fordelt selv om de nevnte forutsetningene ikke er oppfylt. Følgelig er altså den konfidenssannsynligheten vi opererer med tilnærmet riktig selv om de nevnte forutsetningene svikter en del. Vi ser da bort fra helt ekstreme tilfelle. Merk også at n_1 og n_2 helst bør være noenlunde like hvis forutsetningen om like standardavvik ikke er oppfylt.

Øvelse 28.

Det prosentiske innholdet av et stoff S i et organ O ble bestemt hos et random sampel på 5 dyr av dyreslag A og hos 4 dyr av dyreslag B. Observasjonene ble følgende:

Dyreslag A	24,3	20,8	23,7	17,4	21,3	Sum: 107,5
Dyreslag B	18,2	16,9	20,2	16,7		Sum: 72,0

Finn konfidensgrensene for differensen mellom forventningene for innholdet av stoff S hos de to dyreslagene. Bruk Q=0,95.

Øvelse 29.

Analyser formelen (76), s.99 for bredden av konfidensintervallet på liknende måte som nevnt for formelen (58) i øvelse 25, s.92.

3. Konfidensgrenser for en sannsynlighet

La sannsynligheten for et kjennetegn E (f.eks. "kommer seg ikke") i et univers U (f.eks. universet av de dyr som får en bestemt sykdom) være $P(E|U) = p$. La videre Z være antall gjentak med E i et random sampel på n gjentak fra U. $\underline{Z}$ kan da oppfattes som en random variabel som vil ha forskjellige verdier fra sampel til sampel. Det samme kan sies om den relative frekvensen $\frac{\underline{Z}}{n}$.

Det er tidligere vist at $\underline{Z}$ er binomialt fordelt med forventningen np og standardavviket $\sqrt{npq}$. Det kan også vises at $\frac{\underline{Z}}{n}$ har forventningen p og standardavviket $\sqrt{\frac{pq}{n}}$. Siden forventningen for $\frac{\underline{Z}}{n} = p$, er $\frac{\underline{Z}}{n}$ en forventningsrett estimator for p.

Spørsmålet er nå hva vi kan si om fordelingsfunksjonen for $\frac{\underline{Z}}{n}$ i tillegg til det som allerede er sagt om forventningen og standardavviket.

Når n ikke er for liten, kan det vises at $\underline{Z}' = \frac{\underline{Z}}{n}$ er tilnærmet normalt fordelt. Forutsetter vi nå at $\underline{Z}'$ er normalt fordelt med forventning p og standardavvik $\sqrt{\frac{pq}{n}}$, kan vi ved å trekke fra forventningen og dividere med standardavviket, dvs. ved å standardisere, forme en ny random variabel $\underline{X}$

som følger den spesielle fordelingsfunksjonen som er beskrevet nederst på s. 70 og på s. 71.

$$(77) \qquad \underline{X} = \frac{\underline{Z}'-p}{\sqrt{\frac{pq}{n}}}$$

Av fig. 7, s.71 ser vi da at følgende sannsynlighetsutsagn er oppfylt for alle samhørende verdier av a og Q:

$$(78) \qquad P(-a \leq \frac{\underline{Z}'-p}{\sqrt{\frac{pq}{n}}} \leq a) = Q$$

På grunnlag av (78) kan en lett ved litt regning komme fram til følgende formel for konfidensgrensene for p:

$$(79) \quad \text{Konfidensgrenser: } \frac{n}{n+a^2}\left[Z'+\frac{a^2}{2n} \mp a\sqrt{\frac{Z'(1-Z')}{n}+\frac{a^2}{4n^2}}\right]$$

Når vi bruker - foran rottegnet, får vi nedre grense, når vi bruker + får vi øvre grense. Til Q=0,95 svarer a=1,96. Til Q=0,99 svarer a=2,576.

Når n er stor kan en bruke følgende forenklede formel:

$$(80) \quad \text{Konfidensgrenser: } Z' \mp a\sqrt{\frac{Z'(1-Z')}{n}}$$

En undersøkelse ga som resultat at 57 av 300 dyr ikke kom seg etter en bestemt sykdom. Konfidensgrensene for dødelighetsraten for denne sykdommen finnes da på følgende måte når vi bruker Q=0,95:

$$\frac{300}{300+1,96^2}\left[0,19+\frac{1,96^2}{600} \mp 1,96\sqrt{\frac{0,19\cdot 0,81}{300}+\frac{1,96^2}{4\cdot 300^2}}\right]$$

Ved utregning finner en at nedre grense blir 0,15 og øvre grense blir 0,24.

Øvelse 30.

Av et stort antall personer vaksinert mot kopper, ble (etter 25 år eller mer siden vaksinasjonen) 932 angrepne av sykdommen og av disse 483 hardt angrepet. La E_1 = "angrepet" og E_2 = "hardt angrepet". Beregn konfidensgrensene for $P(E_2 \mid UE_1)$. Bruk Q = 0,95.

Øvelse 31.

Studer bredden av konfidensintervallet (80) på liknende måte som i øvelse 25, s. 92.

Øvelse 31 II.

Når NRF-kuer insemineres med sed fra okser av kjøttrasen Charolais, er det en tendens til at det oppstår fødselsvansker på grunn av at kalvene er relativt store. For å undersøke hvor alvorlige disse problemene er, ble det samlet inn opplysninger om i alt 366 fødsler etter slike inseminasjoner. Av disse 366 fødslene var det 11 vanskelige fødsler.

a) Estimer sannsynligheten, p for fødselsvansker ved slike fødsler.
b) Beregn et konfidensintervall for p. Bruk konfidenssannsynligheten 0,95.

Øvelse 31 III.

Et foringsforsøk med kalkuner ble utført etter planen fri randomisering. 8 kalkuner ble foret etter forplan A og 8 etter blanen B. I løpet av en bestemt tid ble det registrert følgende tilvekster i kg levende vekt:

Plan A:	3,5	2,0	2,3	3,5	2,0	2,0	1,8	1,3
Plan B:	1,8	2,5	2,5	2,8	2,0	1,5	1,8	2,5

Beregn konfidensgrensene for differensen mellom forventningene for tilveksten ved bruk av de to forplanene. Bruk Q=0,95.

Øvelse 31 IV.

Et blokkforsøk med 5 blokker (grisekull) og i alt 10 griser ble utført for å sammenlikne to forplaner for slaktegriser et sted i Afrik Kullene ble tatt ut tilfeldig blant dyrematerialet i vedkommende områd Tilveksten i kg levende vekt i løpet av forsøksperioden ble følgende:

Blokk nr.	1	2	3	4	5
Forplan A:	67	57	74	61	70
Forplan B:	62	51	75	58	65

Beregn konfidensgrensene for differensen mellom forventningene for tilveksten ved bruk av de to planene. Bruk Q=0,95.

XII. Hypotesetesting

A Generelt

Vi skal nå gå over til et annet viktig felt innen statistikken, nemlig testing av statistiske hypoteser. Oppgavene vi får å løse er nær beslektet med de problemene vi arbeidet med i estimeringsteorien, men vi skal betrakte dem fra en litt annen synsvinkel.

En hypotese er et utsagn om virkeligheten, og det er et utsagn av foreløpig karakter, et utsagn som vi ønsker å undersøke nærmere riktigheten av. Statistiske hypoteser er hypoteser om sannsynligheter eller fordelingsfunksjoner. De fleste statistiske hypoteser vi får å gjøre med er hypoteser om parametre i fordelingsfunksjoner. Slike parametre har vi også arbeidet med tidligere. I eksempel 1, s.90 var vi f.eks. interessert i det sanne innholdet av kjemikalium A, dvs. i forventningen for en random variabel. I eksempel 2, s. 92-93 var vi interessert i differensen mellom to forventninger, nemlig forventningene for vekten av kyllinger som var oppvokst (1) inne i binge og (2) ute.

En hypotese representerer et metodisk hjelpemiddel som vi bruker til å skaffe oss kjennskap til virkeligheten. Forskning består i stor utstrekning i en prøve- og-feile·prosess. En hypotese er en slags gjetning om virkeligheten, og vi er interessert i å undersøke om det er en god gjetning. Vi foretar derfor observasjoner og undersøker om disse er i overensstemmelse med det vi skulle vente hvis hypotesen var riktig. Er de ikke det, forkaster vi hypotesen og prøver oss kanskje med en ny hypotese som vi da bør undersøke holdbarheten av ved hjelp av nye data. (Hvorfor kan vi ikke like gjerne bruke de observasjonene vi har ?).

Prøvingen eller testingen av en hypotese består i å avlede visse konsekvenser av den og undersøke om disse harmonerer med de observerte data. En hypotese som vi har testet mange ganger ved hjelp av forskjellige sett av data og som vi aldri har kunnet forkaste vil vi etter hvert fatte stor tiltro til, og kanskje oppfatte den som en lovmessighet.

Ofte støter vi på uttrykket null-hypotese. En null-hypotese er en hypotese som går ut på at to eller flere parametre er like, dvs. at forskjellen mellom dem er lik null. (Eller at en enkelt parameter er lik 0.)

En null-hypotese er særmerket ved at det ikke behøver å ligge noen observasjoner eller tidligere erfaringer til grunn for den. Alle og enhver som vet hva hypotesetesting er kan uten vanskelighet sette fram en nullhypotese. Andre statistiske hypoteser, derimot, blir vanligvis satt fram på grunnlag av kjennskap til det fenomen hypotesen gjelder. Når vi framsetter en null-hypotese, er hensikten ganske enkelt å undersøke om den kan forkastes. I såfall må vi jo godta dens alternativ, nemlig at det er en forskjell på parametrene, og det var kanskje det vi ønsket å påvise da vi framsatte null-hypotesen.

En null-hypotese kommer også i en særstilling fordi vi vanligvis ikke godtar denne. Det skal jo mye til at f.eks. to behandlingsmåter skal være nøyaktig like med hensyn til forventet helbredelsestid. For null-hypotesen er situasjonen følgende: Enten forkaster vi den, eller så unnlater vi å forkaste den, og vi sier da at det ikke er noen påviselig forskjell.

For å få en felles terminologi for hypoteser i sin alminnelighet (som det kan bli snakk om å godta) og null-hypoteser, skal vi i det følgende ikke bruke uttrykket godta om en hypotese. I stedet skal vi snakke om å unnlate å forkaste. Hypotesetestingen kan etter dette lede til to tenkelige resultater: (1) Vi forkaster hypotesen, eller (2) Vi unnlater å forkaste hypotesen.

Det er to menn ved navn Neyman og Pearson som spesielt har æren for den moderne hypotesetestingsteorien. Vi kan ikke ta med stort av denne teorien her, men vi skal nevne noen trekk av teorien.

For å konstruere et kriterium for å teste en statistisk hypotese er det alltid nødvendig å formulere en alternativ hypotese. Vi skal heretter bruke symbolet H_o for den hypotesen vi tester og H_A for dens alternativ eller alternativer. Det er nemlig som regel flere alternative hypoteser.

Når vi tester statistiske hypoteser kan vi komme i skade for å gjøre to slags feil: (1) Vi kan forkaste H_o selv om H_o er riktig. Dette blir kalt feil av 1.slag (forkastningsfeil). (2) Vi kan unnlate å forkaste H_o selv om H_o er feil. Dette blir kalt feil av 2.slag (forkastningsunnlatelsesfeil).

Situasjonen kan skisseres i en enkel tabell. Når det gjelder virkeligheten, er vårt problem spesifisert slik at enten er H_o riktig eller så er H_A riktig. Noen tredje mulighet finnes ikke. Når det gjelder vår handling så går den ut på enten å forkaste H_o eller å unnlate å forkaste H_o. Resultatet av hypotesetestingen i et gitt tilfelle kan derfor settes opp i følgende tabell:

		Virkeligheten	
		H_o er riktig	H_A er riktig (H_o feil)
Vår handling	Forkaster H_o	Forkastningsfeil	Riktig handling
	Unnl. å fork. H_o	Riktig handling	Forkastningsunnlatelsesfeil

Vår oppgave er å innrette oss slik at vi best mulig unngår begge typer av feil. Dette er imidlertid et meget kinkig problem. Når vi minsker sannsynligheten for å gjøre en type av feil er det i alminnelighet nemlig ikke til å unngå at vi samtidig øker sannsynligheten for å gjøre den andre typen av feil. Vi må derfor foreta en viss avveining mellom ulempene ved å gjøre de to typer av feil.

Vi skal belyse situasjonen ved et eksempel. Når vi skal teste en hypotese, skaffer vi oss alltid et sett av data som oppfattes som observasjoner av en eller flere random variable. Disse data stammer ofte fra forsøk, men kan også være skaffet til veie på annen måte. Som eksempel på et sett av data kan vi ta tallene som er oppstilt i begynnelsen av eksempel 4 s.114. De to første kolonnene er de samme som i eksempel 3, s. 99. Imidlertid har vi føyd til 2 nye kolonner, idet vi nå tenker oss at hele 4 behandlingsmåter

ble sammenliknet ved forsøket, slik som opprinnelig beskrevet på s. 77-79. Disse data eller observasjoner er et produkt av "virkeligheten" og er et resultat av en mangfoldighet av ukjente og kjente årsaksforhold. Blandt disse årsaksforhold må vi også innbefatte selve forsøksplanen når det som her er snakk om data fra et forsøk.

Av våre data former vi et tall F etter en bestemt formel som vi skal komme tilbake til i detalj i neste avsnitt. Under den egentlige hypotesetestingen er dette ene tallet det eneste vi er interessert i når det gjelder observasjonene.

Dette var litt om observasjonene. La oss så se på den andre siden av hypotesetestingsproblemet, nemlig selve hypotesen. Foruten hypotesen har vi gjerne et sett av á priori forutsetninger eller opplysninger om den "mekanismen" som har frambrakt våre data. I eksemplet ovenfor er våre forutsetninger følgende: (1) At hvert av de 4 samplene er et random sampel. (Denne forutsetningen kunne vi la være å nevne da den nærmest er automatisk oppfylt når forsøket er riktig utført etter planen som beskrevet på s.78.) (2) At hver av de 4 random variable $\underline{X}_1$, $\underline{X}_2$, $\underline{X}_3$ og $\underline{X}_4$ har normal fordelingsfunksjon. (3) At $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4 = \sigma$. Selve hypotesen eller om vi vil den hypotetiske forutsetningen, går i vårt tilfelle ut på at

$$\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu$$

Hvis vi hadde gjentatt vårt forsøk under tilsvarende forhold (dvs. etter samme forsøksplan, med de samme forsøksledd, det samme antall gjentak, osv.) er det klart at vi måtte regne med å få nye data, og dermed et nytt tall F. Tenker vi oss en uendelighet av gjentatte forsøk, får vi altså også en uendelighet av tall, F. Noen av disse er store, andre er små. Vi kan altså snakke om en random variabel F.

De tre á priori forutsetningene og den ene hypotetiske forutsetningen som er nevnt ovenfor går tilsammen ut på at alle observasjonene fra et og samme forsøk er observasjoner av en enkelt random variabel $\underline{X}$: $N(\mu,\sigma)$ (se s.79). På grunnlag av disse fire forutsetningene er det mulig å bevise at $\underline{F}$ følger en bestemt fordelingsfunksjon som gjerne kalles F-fordelingen til ære for statistikeren R.A. Fischer. Vi skal ikke gjengi formelen for fordelingsfunksjonen for $\underline{F}$ her, da den er forholdsvis komplisert og har liten interesse for oss. $\underline{F}$ er en kontinuerlig random variabel som kan anta alle positive verdier. Fordelingsfunksjonen f(F) har to parametre f_1 og f_2 som er hele positive tall og som vi i våre anvendelser vil kalle henholdsvis antall frihetsgrader for telleren og antall frihetsgrader for nevneren. Når $f_2 > 2$ (som den vanligvis er i våre anvendelser) er forventningen for $\underline{F}$ lik

$$E(\underline{F}) = \frac{f_2}{f_2-2} \tag{81}$$

Forventningen er altså lik 3 for $f_2=3$ og avtar mot 1 hvis vi betrakter nye medlemmer av "F-familien" med voksende f_2. Funksjonen har i grove trekk et forløp som skissert i fig. 9.

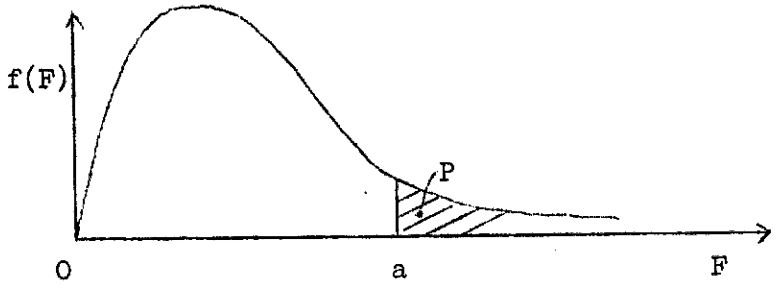


Fig. 9

Selv om formen på F-kurven er forskjellig for hvert sett av parametre, har alle F-kurver det til felles at de har en lang "hale" til høyre.

I våre anvendelser av funksjonen er vi interessert i å finne et tall a som er slik at sannsynligheten for at $\underline{F}$ skal være større enn eller lik a er lik et oppgitt tall P.

(82) $$P(\underline{F} \geq a) = P$$

(Symbolet P er brukt i to forskjellige betydninger i (82)). Sammenhengen mellom P og a er også illustrert i fig. 9. Som regel velges P = 0,05 eller P = 0,01. Sammenhørende verdier av P, f_1, f_2 og a er gjengitt i hovedtabellene II og III bak i heftet.

Hvert enkelt forsøk gir oss en enkelt verdi F av $\underline{F}$, og denne F kan selvsagt være et hvilket som helst tall i variasjonsområdet for $\underline{F}$, dvs. et hvilket som helst positivt tall. Alle F er derfor forenlige med den antagelse at hypotesen er riktig. (Fordelingsfunksjonen er utledet under den forutsetning at hypotesen er riktig). Det vi gjerne burde ha er en serie av uavhengige forsøk etter samme plan. Hvis da de F vi fikk fordelte seg langs tall-linjen på en måte som samsvarte med den tilsvarende F-kurve, kunne vi ta dette som et indisium på at hypotesen var riktig. Hvis de derimot fordelte seg på en måte som tilsvarte en helt annen kurve, ville vi få meget sterk mistanke om at hypotesen var feilaktig.

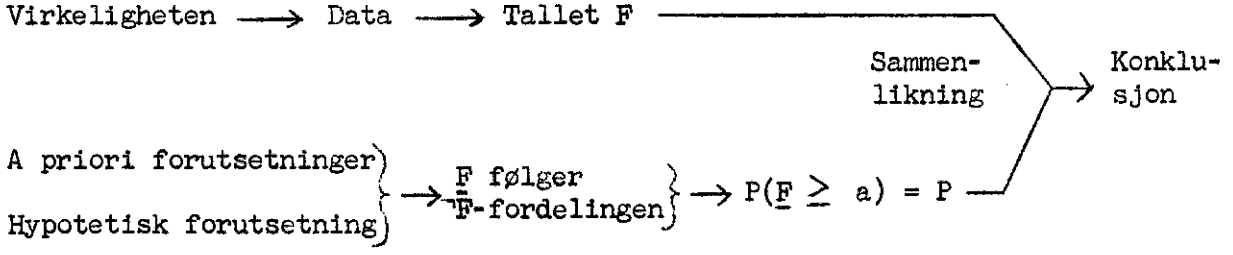
Nå er det imidlertid nokså vanlig å trekke konklusjoner (i hvert fall foreløpige konklusjoner) på grunnlag av et eneste forsøk, og det er teknikken for å gjøre dette på beste måte vi skal lære her. (Vi forstår likevel at verdien av et enkelt forsøk er nokså begrenset hvis resultatene ikke blir etterprøvet ved nye forsøk.)

Det eneste spor "virkeligheten" har etterlatt seg er tabellen over våre data, og den informasjonen tabellen inneholder kan for vårt formål best summeres i form av et enkelt tall F. Hvis denne F ligger i "halen" til høyre, vil dette forhold svekke vår tiltro til hypotesen, selv om en stor F som nevnt aldri er direkte uforenlig med hypotesen. Skal vi i det hele tatt komme til noen konklusjon, er vi imidlertid nødt til å ta en viss risiko for å gjøre feil. Det vi da gjør når vi tester en hypotese er følgende: Vi bestemmer oss

på forhånd for et tall P som er lik sannsynligheten for å forkaste hypotesen selv om den er riktig. Hvis den F vi finner på grunnlag av tallene fra vårt forsøk er større enn den a som svarer til P (se fig. 9.) forkaster vi hypotesen. P blir kalt sannsynlighetsnivået eller signifikansnivået. Hvis vi bruker P = 0,05 og resultatet blir at vi må forkaste hypotesen, sier vi gjerne at vi forkaster den på 5 % nivået. Vi sier også da at F er signifikant på 5 % nivået. Det området av F-aksen som fører til forkastning av H_o, altså området til høyre for a, blir kalt et forkastningsområde for H_o eller et kritisk område for F. Til hver P svarer det et forkastningsområde.

Når F er større enn a, tar vi altså det faktum at vi har funnet et tall F som er "uvanlig stort" under vår hypotetiske forutsetning som et tilstrekkelig "bevis" på at det må være noe i veien med hypotesen.

Hypotesetestingen går altså ut på at vi på den ene siden regner ut et tall F ved hjelp av våre data og at vi på den andre siden finner fram til et tall a. Hvis F er større enn eller lik a, forkaster vi hypotesen. Det hele kan summeres i følgende skjema hvor en pil kan leses "resulterer i" eller "leder til":



Det er bare når hypotesen (den hypotetiske forutsetningen) er riktig at F følger F-fordelingen. Hvis vår hypotese: H_o er feilaktig, følger F en annen fordelingsfunksjon. Sett at det er et bestemt alternativ H_A til H_o som er den riktige hypotesen. H_A går ut på at forventningene $\mu_1 \ldots \mu_4$ avviker fra hverandre med visse nærmere angitte tall. I dette tilfelle følger

F en fordelingsfunksjon som vi vil betegne med $f(F \mid H_A)$ i motsetning til F-fordelingen som vi nå vil betegne med $f(F \mid H_o)$ Situasjonen er illustrert i fig. 10 hvor vi har tenkt oss at vi kjenner $f(F \mid H_A)$.

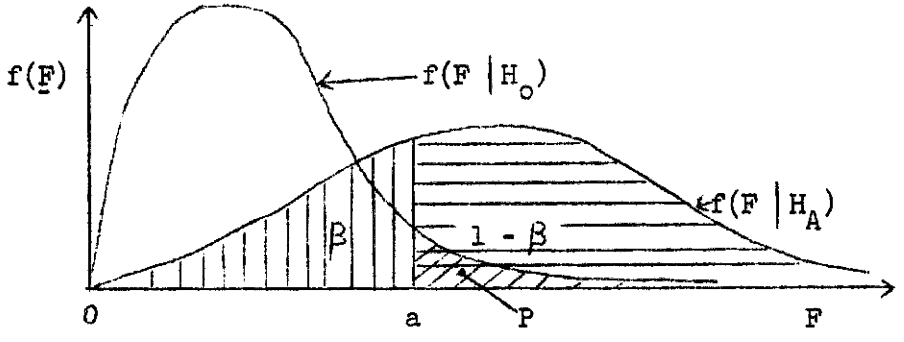


Fig. 10

Vi ser at det er en nokså stor sannsynlighet for å få en F større enn a hvis H_A er riktig. I figuren er denne sannsynligheten representert ved det horisontalt skraverte arealet. Størrelsen av dette arealet er $1 - \beta$ som blir kalt teststyrken for alternativet H_A.

Uansett hvilken hypotese som er den riktige så unnlater vi å forkaste H_o når våre data resulterer i en F som er mindre enn a og forkaster H_o når vi får en F større enn a. Vi ser derfor at sannsynligheten for forkastningsunnlatelsesfeil er lik β (det vertikalt skraverte arealet i figuren). Slike feil gjør vi jo bare når H_A er riktig, og følgelig må sannsynligheten for å gjøre slike feil regnes ut på grunnlag av den kurven som gjelder når H_A er riktig. På tilsvarende måte ser vi at sannsynligheten for forkastningsfeil er lik P. Det er de to arealene β og P vi gjerne skulle gjøre så små som mulig. Vi ser imidlertid at hvis vi minsker P så øker vi β og omvendt. Det er derfor vanlig i praksis å bestemme seg for en P og la β bli det den blir. Svært ofte velges P = 0,05 eller P = 0,01. Valget beror nærmest på en konvensjon, og på det faktum at vi har tabeller for nettopp disse P-verdiene.

Egentlig er vårt eksempel enda mer komplisert fordi det ikke bare finnes ett alternativ til H_o, men en hel serie av alternativer, H_A. Hvert alternativ

resulterer i en fordelingsfunksjon for $\underline{F}$ og altså i en kurve $f(F \mid H_A)$ i vår figur. For å finne ut noe om hvilket forkastningsområde en bør bruke, kan en i slike tilfelle studere teststyrken for de aktuelle alternativer. Den ideelle situasjon ville vi ha om teststyrken $1 - \beta$ var lik 1 for alle alternativer og at samtidig sannsynlighetsnivået P var lik 0. Å finne et forkastningsområde som tilfredsstiller disse krav er umulig i vårt eksempel og forøvrig også i alle andre praktiske tilfelle.

Vi skal senere begrunne ytterligere hvorfor vi har området til høyre for a være forkastningsområde for H_o. I andre hypotesetestingssituasjoner kan vi ha både et forkastningsområde til høyre og et til venstre.

Vi har nå gjennomgått noen grunntrekk av hypotesetestingsteorien i tilknytning til et konkret eksempel. Teorien er imidlertid helt generell. I stedet for den random variable $\underline{F}$ opererer en i andre tilfelle med andre random variable som f.eks. $\underline{t}$ som vi har gjennomgått tidligere eller $\underline{\chi}^2$ som vi har tabell over bakerst i heftet.

B. Variansanalyse og F-test

1. En-veis gruppering.

Vi skal nå gjennomgå i detalj det eksemplet vi brukte i foregående avsnitt da vi illustrerte hypotesetestingsteorien. Da vi allerede har forklart problemstillingen og hovedtrekkene i analysen (s. 77-79 og s. 107-108), skal vi her bare vise hvorledes vi regner ut tallet F. Dessuten skal vi si litt om anvendbarheten av denne testen og om tolkingen av resultatene.

Hele analysen av dette eksemplet (eks. 4) finner vi på s. 114-115.

Eks. 4. Fire måter å behandle voksne hester på etter en operasjon (se s. 77-79)

	T_1	T_2	T_3	T_4	
	25	34	32	30	
	37	26	29	29	
	27	26	35	25	
	32	27	31	26	
	38	32	31	32	
	30	35	32	22	
	26	36	33	32	
	34	26	38	26	
	41		36	35	
S_j	290	242	297	257	$\sum_{j=1}^{k} S_j = S = 1086$
S_j^2	84100	58564	88209	66049	
n_j	9	8	9	9	$\sum_{j=1}^{k} n_j = N = 35$
$\frac{S_j^2}{n_j}$	9344,44	7320,50	9801,00	7338,78	$\sum_{j=1}^{k} \frac{S_j^2}{n_j} = 33805$
$\sum_{i=1}^{n_j} X_{ij}^2$	9604	7458	9865	7475	$\sum_{j=1}^{k} \sum_{i=1}^{n_j} X_{ij}^2 = 34402$
$\sum_{i=1}^{n_j} X_{ij}^2 - \frac{S_j^2}{n_j}$	259,56	137,50	64,00	136,22	
$V_j = \frac{\sum_{i=1}^{n_j} X_{ij}^2 - \frac{S_j^2}{n_j}}{n_j - 1}$	32,45	19,64	8,00	17,03	$\frac{S^2}{N} = \frac{1086^2}{35} = 33697$
$\bar{X}_j$	32,22	30,25	33,00	28,56	

Variasjons-årsak	Frihetsgrader (DF)	Kvadratsum (SS)	Varians ($\frac{SS}{DF}$)	$F=\frac{V_T}{V_R}$
Total	N-1=34	$\Sigma\Sigma(X_{ij}-\bar{X})^2=\Sigma\Sigma X_{ij}^2-\frac{S^2}{N}=34402-33697=705$		
Mellom-grupper	k-1=3	$\Sigma n_j(\bar{X}_j-\bar{X})^2=\Sigma\frac{S_j^2}{n_j}-\frac{S^2}{N}=33805-33697=108$	$V_T=36$	1,87
Innen grupper	N-k=31	$\Sigma\Sigma(X_{ij}-\bar{X}_j)^2$ $=597$	$V_R=19,3$	

$P=0,05 \quad a_{0,05}(3,31)=$ca. 2,92

F er ikke signifikant. Null-hypotesen kan ikke forkastes. Det er ingen påviselig forskjell på forventningen for helbredelsestiden for de fire behandlingsmåtene.

Vi vil nå forklare symbolene vi vil bruke generelt og oppgi i parantes noe om hva symbolene står for i eksempel 4. Antall forsøksledd (behandlingsmåter av hester etter en operasjon) betegner vi med k (k=4). Et vilkårlig av disse k forsøksledd betegnes med T_j. Observasjon nr. i av den random variable $\underline{X}_j$ (helbredelsestiden etter behandlingsmåte nr. j) betegner vi med X_{ij}. I tabelloppstillingen i begynnelsen av eksempel 4 er X_{ij} det tallet som står i linje nr. i og kolonne nr. j. Altså står X_{ij} for et vilkårlig tall i tabellen. ($X_{42} = 27$). Antall observasjoner av den random variable X_j betegner vi med n_j. ($n_2 = 8$). Antall observasjoner i alt betegner vi med $N = \sum_{j=1}^{k} n_j$. Summen av alle observasjonene for T_j betegner vi med $S_j = \sum_{i=1}^{n_j} X_{ij}$. ($S_3 = \sum_{i=1}^{9} X_{i3} = 297$) Summen av kvadratene av alle observasjonene i gruppe j skrives $\sum_{i=1}^{n_j} X_{ij}$. ($\sum_{i=1}^{9} X_{i4}^2 = 7475$) Den empiriske variansen V_j for observasjonene i gruppe j er gitt ved følgende formel:

$$V_j = \frac{\sum_{i=1}^{n_j} X_{ij}^2 - \frac{S_j^2}{n_j}}{n_j - 1} \qquad (83)$$

Merk at dette er nøyaktig samme formel som formelen (28), s. 57. Vi har nå bare brukt andre symboler. Gjennomsnittet av alle observasjonene i gruppe j betegner vi med $\bar{X}_j$. Gjennomsnittet av alle N observasjonene, altså $\frac{S}{N}$ betegner vi med $\bar{X}$.

På s. 114 har en under selve observasjonsmaterialet satt opp på linjer en del av de størrelsene som hittil er nevnt. Deretter er noen av disse linjene summert horisontalt. Dette blir da en summering fra j=1 til j=k. Noen av de størrelsene som er tatt med på s.114 er ikke nødvendig for utregningen av F, men vi kommer tilbake til dette senere.

På s. 115 har en satt opp i tabellform selve utregningen av F som vi nå skal forklare. F er lik forholdet mellom to empiriske varianser. (At

disse virkelig er variansер av den typen vi er vant til skal vi forklare senere da det ikke uten videre er klart av formelen). Den første variansen betegnes med V_T og kalles mellom-gruppe variansen. Vi skal senere vise at denne har sammenheng med variasjonen i gjennomsnittene $\bar{X}_j$ fra gruppe til gruppe.

$$(84) \qquad V_T = \frac{\sum_{j=1}^{k} n_j(\bar{X}_j-\bar{X})^2}{k-1} = \frac{\sum_{j=1}^{k} \frac{S_j^2}{n_j} - \frac{S^2}{N}}{k-1}$$

Den andre variansen kalles innen-gruppe variansen og betegnes V_R. Vi skal vise senere at denne er et veid gjennomsnitt av gruppevariansene V_j.

$$(85) \qquad V_R = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)^2}{N-k}$$

For å spare regnearbeid er det vanlig å trekke inn i analysen telleren i den empiriske variansen vi får hvis vi oppfatter alle N observasjonene som observasjoner av en enkelt random variabel $\underline{X}$ (se s.79). La oss betegne denne variansen V_{Tot}. Formelen er gitt ved (28), s.57, men i våre nåværende symboler blir formelen følgende:

$$(86) \qquad V_{Tot} = \frac{\sum\sum(X_{ij}-\bar{X})^2}{N-1} = \frac{\sum\sum X_{ij}^2 - \frac{S^2}{N}}{N-1}$$

Telleren i hver av disse tre variansene er en kvadratsum. Vi snakker således om kvadratsummen for V_T, kvadratsummen for V_R og den totale kvadratsummen. Det er lett å vise at den totale kvadratsummen er lik summen av de to andre. Kvadratsummen for V_R kan derfor beregnes av følgende likning:

$$(87) \qquad \sum\sum(X_{ij}-\bar{X}_j)^2 = \sum\sum X_{ij}^2 - \frac{S^2}{N} - \left(\sum \frac{S_j^2}{n_j} - \frac{S^2}{N}\right)$$

Nevneren i hver av de tre variansene (84)-(86) blir kalt

antall frihetsgrader for kvadratsummen i telleren. Størrelsen F finnes som nevnt på følgende måte:

$$F = \frac{V_T}{V_R} \tag{88}$$

Når vi skal slå opp i hovedtabell II eller III for å finne a må vi kjenne parameterne i fordelingsfunksjonen for F. Disse finnes på følgende måte: Antall frihetsgrader for teller er lik nevneren i V_T (V_T er jo telleren i F). Antall frihetsgrader for nevner er lik nevneren i V_R (V_R er jo nevneren i F). Hele framgangsmåten ved analysen er illustrert i eksempel 4. Det er vanlig å sette opp resultatene i tabellform slik som på s. 115, men denne oppstillingen er gjerne mer sammentrengt, uten formler.

At V_R er et veid gjennomsnitt av V_j med n_j-1 som vekter kan vises på følgende måte:

$$V_R = \frac{\sum_{j=1}^{k}\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)^2}{N-k} = \frac{\sum_{j=1}^{k}(n_j-1)\dfrac{\sum_{i=1}^{n_j}(X_{ij}-\bar{X}_j)^2}{(n_j-1)}}{\sum_{j=1}^{k}n_j-k} = \frac{\sum_{j=1}^{k}(n_j-1)V_j}{\sum_{j=1}^{k}(n_j-1)} \tag{89}$$

Vi skal nå forsøke å begrunne praktisk hvorfor det synes rimelig å forkaste H_0 når F er stor. Vi skal da tenke oss et tilfelle hvor alle n_j er like og lik n. Formlene blir da noe enklere. Setter vi $n_j=n$ blir formelen for V_T følgende:

$$V_T = \frac{\sum_{j=1}^{k} n(\bar{X}_j-\bar{X})^2}{k-1} = n\,\frac{\sum_{j=1}^{k}(\bar{X}_j-\bar{X})^2}{k-1} \tag{90}$$

Når vi ser bort fra faktoren n er V_T en empirisk varians for de k gruppegjennomsnittene. Dette ser vi lett ved å sammenlikne (90) med (28), s. 57. V_T blir stor når det er stor variasjon

i gruppegjennomsnittene $\bar{X}_j$. Siden gruppegjennomsnittene $\bar{X}_j$ er estimater av μ_j, er en stor V_T et tegn på at det kanskje er noe i veien med hypotesen om at alle μ_j er like. Faktoren n kan vi praktisk oppfatte på følgende måte: Vi har tidligere (34), s. 60 presentert formelen $\sigma^2_{\underline{X}} = n\sigma^2_{\underline{\bar{X}}}$. Hvis en tilsvarende formel hadde gyldighet også for empiriske variansar (dette er ikke tilfelle eksakt), måtte vi multiplisere variansen beregnet for k gjennomsnitter med n (antall observasjoner bak hvert gjennomsnitt) for å bringe den opp på samme nivå som variansen for en rekke enkeltobservasjoner. Uten å gjøre krav på at dette er noe eksakt resonnement, kan vi derfor si at faktoren n i (90) bringer V_T opp på samme nivå som en vanlig empirisk varians.

La oss vise at også V_R er å sammenlikne med en vanlig empirisk varians. Vi setter $n_j = n$ i (89) og får da følgende:

$$(91) \qquad V_R = \frac{\sum(n-1)V_j}{N-k} = \frac{(n-1)\sum V_j}{k(n-1)} = \frac{\sum V_j}{k}$$

Når $n_j = n$ er altså innen-gruppe variansen lik gjennomsnittet av gruppevariansene. Gruppevariansene gir i vårt eksempel uttrykk for variasjonen i helbredelsestiden for hester som har fått samme behandlingsmåte. Den variasjonen som skyldes at hestene har fått forskjellig behandling kommer forsåvidt ikke til uttrykk i V_R, men derimot i V_T. Når V_T er stor, tyder dette på at behandlingsmåtene gir forskjellig resultat. Når vi regner ut F kan vi si at vi måler V_T med V_R som målestokk. Det kan for øvrig vises at V_T og V_R har samme forventning hvis H_0 er riktig. Tidligere har vi nevnt at forventningen for V_T/V_R er noe større enn 1 under samme forutsetning. Hvis F er stor

synes det etter dette rimelig å forkaste H_0. Tabellene bak i heftet gir oss et middel til å anslå hvor stor F må være.

Hittil har vi bare sett på selve variansanalysen (dvs. beregninen av kvadratsummer og variansер) og F-testen. Det er imidlertid visse problemer som kan komme til i forbindelse med tolkningen av resultatet og den eventuelle videre analyse. Vi kan ikke ta med mer enn hovedkonklusjonene her, men viser for øvrig til referater fra professor Ottestads forelesninger ved Norges landbrukshøgskole.

Sett at vi finner en signifikant F. Det kan da tenkes at dette ikke skyldes at det er noe i veien med H_0, men at det er noe i veien med våre a priori forutsetninger om normalitet og like standardavvik. Ved å regne ut gruppevariansene V_j slik som vi har gjort i eksempel 4 kan vi få en viss indikasjon på om de teoretiske variansene er ulike for de forskjellige forsøksledd. Det finnes også testmetoder som kan brukes til å undersøke dette. Forutsetningen om normalitet skaper sjelden vesentlige problemer, da det har vist seg at F-testen er robust overfor avvik fra denne forutsetningen.

Hvis vi har kommet til den konklusjon at forventningene μ_j er forskjellige, kan det ha atskillig interesse å estimere forskjellen mellom to og to forventninger. Sett at vi ønsker å finne et konfidensintervall for $\mu_1 - \mu_2$ i eksempel 4. Vi kan da gå fram slik som vist i eksempel 3, s. 99 hvor vi har de samme data. Eksempel 3 kan altså oppfattes enten som et selvstendig forsøk med bare to forsøksledd eller som en videre analyse av eksempel 4. På tilsvarende måte kunne vi også finne et konfidensintervall for $\mu_3 - \mu_4$. Hvis vi i tillegg også ville beregne et

konfidensintervall for $\mu_2 - \mu_3$, kan det reises visse teoretiske innvendinger mot dette fordi vi da bruker gruppegjennomsnittene $\bar{X}_3$ og $\bar{X}_4$ og gruppevariansene V_3 og V_4 mer enn en gang. Empiriske undersøkelser ved professor Ottestad tyder imidlertid på at dette likevel ikke er så farlig. Stort sett kan vi derfor estimere de kontrastene vi er interessert i etter metoden i eksempel 3 og likevel være nokså trygge på at konfidenssannsynligheten ikke er en annen enn den vi opererer med i følge våre tabeller.

Observasjonene i eksempel 4 er et tilfelle av det vi kaller en-veis grupperte data. Observasjonene fra et forsøk etter planen fri randomisering er et viktig eksempel på en-veis grupperte data, men vi har også eksempler på at slike data framkommer på annen måte enn ved forsøk (se øvelse 32). Også i slike tilfelle kan vi bruke den analysemetoden vi nå har gjennomgått hvis forutsetningene er til stede. Når vi har en-veis grupperte data med bare to forsøksledd e.l. har vi det vi på s. 101 kalte ikke-parobservasjoner.

Øvelse 32.

En type ormer blir klassifisert i tre grupper etter en strukturell karakter.

Det ble tatt et random sampel på 11 ormer fra hver gruppe og lengden av hver orm ble målt. Lengdene er gjengitt nedenfor.

Undersøk om det er signifikant forskjell på lengden av ormer i de tre gruppene. Forklar også kort hvilken hypotese det er som testes og forutsetningene for testmetoden.

	Ormegruppe Nr. 1	Nr. 2	Nr. 3
	8,9	12,2	9,5
	9,7	12,0	8,0
	11,5	11,5	8,3
	8,2	8,7	10,0
	10,5	10,5	9,5
	10,8	9,0	10,0
	11,0	10,5	11,3
	8,0	13,0	10,5
	9,9	13,0	8,0
	11,0	11,0	8,0
	11,0	11,1	9,2
Sum	110,5	122,5	102,3
Sum av kvadrater	1124,69	1384,49	963,97

2. To-veis gruppering.

To-veis grupperte data fremkommer bl.a. som resultat av forsøk etter blokkplanen (s. 79). Et eksempel på to-veis grupperte data som ikke stammer fra forsøk er gitt i øvelse 33, s. 126. Når vi har bare to behandlingsmåter e.l. svarer to-veis grupperte data til det vi kalte par-observasjoner på s. 97.

Vi vil ta for oss det eksemplet vi brukte da vi gjennomgikk blokkplanen (s. 79). Dette forsøket gikk ut på å sammenlikne k=3 forskjellige oppalsmåter T_j for kyllinger. Vi brukte n=10 blokker (foreldrepar), B_i, idet tre og tre kyllinger hadde samme foreldre. Detaljer med hensyn til symbolbruk, m. v. finnes for øvrig på s. 79-81.

Som i tilfellet med en-veis gruppering er vi også nå først og fremst interessert i å teste hypotesen $H_0: \mu_j = \mu \quad (j=1,2,..k)$. Alternativet er som før at minst en μ_j er forskjellig fra μ. Også testingen kan foretas på liknende måte som da vi hadde en-veis grupperte data (eksempel 4, s. 114). Symbolene er de samme, bortsett fra at alle n_j nå er like og lik n som altså her står for antall blokker. Hele analysen er vist på neste side for vårt eksempel (eksempel 5).

Eks. 5. Tre oppalsmåter for kyllinger (se s. 79-81)

	T_1	T_2	T_3	S_i
	36	27	24	81
	35	51	45	131
	27	42	33	102
	29	39	33	101
	30	45	27	102
	31	30	32	93
	25	33	33	91
	28	39	30	97
	33	39	42	114
	36	45	42	123
S_j	304	390	341	1035= S.
$\sum_{i=1}^{n} X_{ij}^2$	9350	15696	12049	$\sum_{i=1}^{n}\sum_{j=1}^{k} X_{ij}^2 = 37095$

N=nk=10·3=30

$\sum_{i=1}^{n} S_i^2 = 109195$ $\qquad \frac{\sum_{i=1}^{n} S_i^2}{k} = 36398$

$\sum_{j=1}^{k} S_j^2 = 360797$ $\qquad \frac{\sum_{j=1}^{k} S_j^2}{n} = 36080$ $\qquad \frac{S^2}{N} = \frac{1035^2}{30} = 35708$

Variasjonsårsak	Frihetsgrader (DF)	Kvadratsum (SS)		Varians ($\frac{SS}{DF}$)	$F=\frac{V_T}{V_R}$
Total	N-1=29	$\sum\sum(X_{ij}-\bar{X})^2=\sum\sum X_{ij}^2-\frac{S^2}{N}$ =37095-35708	= 1387		
Blokk (Foreldre)	n-1=9	$k\sum(\bar{X}_i-\bar{X})^2=\frac{\sum S_i^2}{k}-\frac{S^2}{N}$ = 36398-35708	= 690		
Forsøksfaktor (Oppalsmåte)	k-1=2	$n\sum(\bar{X}_j-\bar{X})^2=\frac{\sum S_j^2}{n}-\frac{S^2}{N}$ =36080-35708=372	1062	186	10,28
Rest	(n-1)·(k-1)= 18	$\sum\sum(X_{ij}-\bar{X}_i-\bar{X}_j+\bar{X})^2$ =	325	18,1	

P=0,05 $\qquad a_{0,05}\,(2,18) = 3,55$

F er signifikant. Null-hypotesen må forkastes. Forventningen for vekten er forskjellig for de tre oppalsmåtene.

Vi skal forklare analysen nærmere i det følgende. Våre a priori forutsetninger når vi har et slikt blokkforsøk er følgende: (1) De n blokkene er et <u>random samvel av blokker</u> fra det universet av blokker vi ønsker å uttale oss om. Dessuten er forsøksleddene fordelt tilfeldig innen hver blokk. (2) Hver av de k random variable X_j har <u>normal fordelingsfunksjon</u>. (3) $\sigma_j = \sigma$ for alle j. Under disse forutsetninger, som svarer til de forutsetningene vi gjorde for en-veis gruppering, kan vi teste H_0 ved hjelp av F-testen.

De nye symbolene som kommer til sammenlikning med en-veis gruppering er følgende: Summen av alle observasjonene for blokk (eller linje) nr. i betegner vi med $S_i = \sum_{j=1}^{k} X_{ij}$. Gjennomsnittet av alle observasjonene i blokk nr. i betegner vi med $\bar{X}_i$.

Hvis vi for en-veis gruppering kunne trekke inn i analysen tre forskjellige variansar eller kvadratsummer, får vi nå en i tillegg for blokk. Vi vil skrive ned formlene for fire variansar, selv om vi strengt tatt bare trenger to av dem til å regne ut F. Den første vil vi betegne med V_{Tot}. Den svarer helt ut til (83), s. 117.

$$(92) \qquad V_{Tot} = \frac{\sum\sum (X_{ij} - \bar{X})^2}{N-1} = \frac{\sum\sum X_{ij}^2 - \frac{S^2}{N}}{N-1}$$

Den andre, V_B vil vi kalle <u>blokkvariansen</u>. Den er lik produktet av en vanlig empirisk varians for de n blokkgjennomsnittene $\bar{X}_i$ og antall observasjoner k som ligger bak hvert av disse gjennomsnittene.

$$(93) \qquad V_B = \frac{k \sum_{i=1}^{n} (\bar{X}_i - \bar{X})^2}{n-1} = \frac{\sum \frac{S_i^2}{k} - \frac{S^2}{N}}{n-1}$$

Den neste variansen V_T svarer til (90), s. 118. Denne kan vi kalle <u>variansen for forsøksledd</u>.

$$(94) \qquad V_T = \frac{n\sum_{j=1}^{k}(\bar{X}_j-\bar{\bar{X}})^2}{k-1} = \frac{\sum\frac{S_j^2}{n} - \frac{S^2}{N}}{k-1}$$

Den siste variansen, V_R kan vi kalle restvariansen.

$$(95) \qquad V_R = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k}(X_{ij}-\bar{X}_i-\bar{X}_j+\bar{\bar{X}})^2}{(n-1)(k-1)}$$

Summen av kvadratsummene (tellerne) for de tre siste variansene er lik kvadratsummen for den første, altså lik den totale kvadratsummen. (En tilsvarende sammenheng gjelder også for frihetsgradene (nevnerne) for de fire variansene.) Kvadratsummen for V_R kan derfor mest praktisk regnes ut etter følgende formel:

$$(96) \quad \sum\sum(X_{ij}-\bar{X}_i-\bar{X}_j+\bar{\bar{X}}) = \sum\sum X_{ij}^2 - \frac{S^2}{N} - \left[\left(\sum\frac{S_i^2}{k} - \frac{S^2}{N}\right) + \left(\sum\frac{S_j^2}{n} - \frac{S^2}{k}\right)\right]$$

Hvis vi tenker oss at vi måler variasjonen i observasjonene med kvadratsummen, kan vi altså si at vi har spaltet opp den totale variasjonen i en variasjon som skyldes blokkinndelingen, en variasjon som skyldes de forskjellige forsøksledd, og en restvariasjon.

F finnes som forholdet mellom V_T og V_R.

$$(97) \qquad F = \frac{V_T}{V_R}$$

Antall frihetsgrader for teller er lik k-1 og antall frihetsgrader for nevner er lik (n-1)(k-1).

Vi har ikke tid til å diskutere i detalj tolkningen av en signifikant F. Det som er sagt om dette for en-veis gruppering på s. 180 gjelder stort sett også for to-veis gruppering. Vi kan føye til her at vår analyse av to-veis grupperte data bygger på den forutsetningen at det ikke finnes samspill mellom

forsøksledd og blokk. Undersøkelser utført av professor Ottestad tyder på at når det finnes slikt samspill er F-testen for følsom, dvs. at H_0 blir forkastet i visse tilfelle da den ikke skulle ha vært forkastet. Hvis en mener at det finnes slikt samspill kan det derfor være grunn til å bruke et lavere signifikansnivå enn det en ellers ville ha gjort, f.eks. 0,025 i stedet for 0,05.

Konfidensgrenser for differenser mellom forventninger, f. eks. for $\mu_2 - \mu_3$ kan for to-veis gruppering finnes etter metoden som er beskrevet i eksempel 2, s. 93. Teoretisk sett er det visse problemer forbundet med å bruke denne metoden flere ganger på et og samme sett av data. Undersøkelser av professor Ottestad synes imidlertid å tyde på at disse vanskelighetene er uten betydning i praksis.

Øvelse 33.

Følgende tall er observert av første blomstringsdag i antall dager fra 1. januar for fire arter (T_j) på fem steder (stasjoner) i et bestemt år (1932). Test nullhypotesen $\mu_j = \mu$. Kommenter resultatet.

Sted	T_1	T_2	T_3	T_4	S_i
A_1	67	95	102	103	367
A_2	61	78	99	113	351
A_3	66	85	117	134	394
A_4	74	105	103	120	402
A_5	57	91	102	112	362
S_j	327	454	523	592	1893

C. Andre testmetoder

Det er nokså vanlig å teste hypoteser om forventninger eller differenser mellom forventninger ved hjelp av en random variabel t som fremkommer slik som vist på s. 74-75. Hvis vi f. eks. setter inn en hypotetisk verdi for μ i (43), s. 78

og finner et tall t som er mindre enn -a eller større enn a (se fig. 8, s. 88) må H_0 forkastes hvis alternativet er at μ er forskjellig fra den hypotetiske verdien (to-sidig alternativ).

Når vi har funnet et konfidensintervall for μ er det imidlertid ikke nødvendig å foreta denne beregningen av t. Alt vi har å gjøre er å se etter om den hypotetiske verdien ligger innenfor eller utenfor konfidensintervallet. Hvis den hypotetiske verdien ligger utenfor konfidensintervallet, må hypotesen forkastes på 100(1-Q)% nivået. Denne måten å teste hypotesen på gir alltid samme resultat som når vi regner ut t. En tilsvarende framgangsmåte kan også brukes for andre typer av konfidensintervaller.

Når vi har funnet et konfidensintervall, bør vi alltid undersøke om intervallet inneholder null. Hvis konfidensintervallet inneholder null, kan den tilsvarende null-hypotesen ikke forkastes. (Vi forutsetter som før at alternativet er to-sidig.) I eksempel 3, s. 99-100 fant vi konfidensgrensene -3,36 og 7,30 for differensen $\mu_1-\mu_2$ mellom forventningene for helbredelsestiden ved bruk av to behandlingsmåter etter en operasjon. Siden konfidensintervallet inneholder null, kan en null-hypotese om at de to behandlingsmåtene er like gode ikke forkastes på 5% nivået. (Q=0,95, derfor blir P=0,0

Vi har hittil for det meste vært opptatt med hypoteser om forventninger og differenser mellom forventninger. Når en ønsker å teste hypoteser om sannsynligheter, kan en ofte gjøre bruk av en fordelingsfunksjon som kalles kji-kvadrat fordelingen (χ^2-fordelingen I hovedavsnitt XIII skal vi se hvordan dette kan gjøres. Der skal vi også vise hvordan χ^2-fordelingen kan brukes til å teste om det er uavhengighet mellom kjennetegn og til å teste en hypotese om at en frekvensfordeling som er funnet i et sampel tilsvarer en kjent fordelingsfunksjon. I hovedavsnitt XII skal vi bruke students t til å teste hypoteser i forbindelse med regresjonsanalysen.

Det finnes i det hele tatt et stort antall testmetoder som passer til hver sine typer av problemer. Vi har bare anledning til å gjennomgå et lite utvalg.

XII. Regresjon

A. Innledning

Både i forskningsarbeid og i det praktiske liv er vi ofte interessert i sammenhengen mellom to eller flere størrelser. En statistisk teknikk som da ofte kommer til anvendelse går under navn av regresjonsanalyse. Denne metodikken er meget fleksibel og derfor uhyre viktig. Det kan f.eks. nevnes at variansanalysen kan fremstilles på en slik måte at den blir et spesialtilfelle av regresjonsanalysen.

Vi skal her bare ta med noen elementære sider ved regresjonsanalysen, og for å lette tilegnelsen skal vi knytte fremstillingen til et eksempel.

B. Regresjonsfunksjonen

Når en skal stille opp fôrplaner for kuer, er det nyttig å kjenne dyrenes levendevekt. Siden de færreste gårdsbruk disponerer en vekt som det kan veies kuer på, blir kuenes vekt i praksis bestemt ut fra brystomfanget. Det er som kjent en viss sammenheng mellom brystomfang og vekt. Tar vi for oss et bestemt univers av kuer, f.eks. det universet U som består av alle kuer som kan regnes å tilhøre NRF-rasen, vil vi finne at kuer med stort brystomfang som regel er tyngre enn kuer med lite brystomfang. Det er imidlertid ingen eksakt matematisk sammenheng mellom brystomfang og vekt slik at vekten kan uttrykkes som en matematisk funksjon av brystomfanget. Nei, vi vet alle at to kuer med nøyaktig samme brystomfang kan ha høyst forskjellig vekt. Likevel har vi følelsen av at det er en vekt som er typisk eller gjennomsnittlig for hvert brystomgang, og at denne typiske vekten stiger med brystomfanget. Hvorledes skal vi kunne formulere dette på en eksakt

måte? Ved å ta det statistiske begrepsapparatet i bruk kan vi sette opp en eksakt modell som gir en mulig forklaring på hvorledes sammenhengen mellom brystomfang og vekt er å oppfatte. Vi skal se litt nærmere på dette.

I universet U, hvor gjentakene er alle NRF-kuer, kan vi oppfatte brystomfanget som en random variabel, $\underline{x}_1$ og vekten som en random variabel $\underline{x}_0$. Hver enkelt ku har en bestemt verdi av $\underline{x}_1$ og en bestemt verdi av $\underline{x}_0$. La oss ta for oss et bestemt brystomfang, x_1. Alle kuer i U som har dette spesielle brystomfanget kan vi si tilhører et subunivers av U, og dette subuniverset vil vi gi betegnelsen $U\underline{x}_1=x_1$. I dette subuniverset kan det tenkes å være kuer med mange forskjellige vekter. Vi vil derfor tenke oss at det finnes en fordelingsfunksjon for vekten av kuer som tilhører dette subuniverset. En fordelingsfunksjon i et subunivers blir gjerne kalt en betinget fordelingsfunksjon, men den er av samme natur som en vanlig fordelingsfunksjon. Den tilsvarende forventningen blir kalt en betinget forventning, og vi kan si at den er betinget av at $\underline{x}_1=x_1$. Som en naturlig betegnelse på denne betingede forventningen kan vi bruke $E(\underline{x}_0 | U\underline{x}_1=x_1)$ som vi vanligvis vil avkorte til $E(\underline{x}_0 | x_1)$.

I avsnittet ovenfor var det nokså vilkårlig hvilket brystomfang vi tok for oss. Noe tilsvarende som det som er nevnt for dette brystomfanget gjelder for et hvilket som helst aktuelt brystomfang. Vi kan derfor tillate oss å oppfatte x_1 som en matematisk variabel som vi selv spesifiserer verdien av, og vi kaller den en uavhengig variabel. Siden vi bruker brystomfanget til å forklare vekten, blir brystomfanget også kalt en forklaringsvariabel. Vekten kalles avhengig variabel.

Den modellen vi vil sette opp for sammenhengen mellom

brystomfang og vekt spesifiserer at $E(\underline{x}_0 | U\underline{x}=x_1)$ er en funksjon av x_1. For hvert aktuelt brystomfang kan det tenkes å finnes mange forskjellige vekter, men bare en betinget forventning for vekten. For hver gang vi tar for oss et nytt aktuelt brystomfang får vi å gjøre med en ny betinget forventning for vekten. Vi tenker oss nå at den betingede forventningen for vekten er en funksjon av brystomfanget i vedkommende subunivers. Denne funksjonen kaller vi regresjonsfunksjonen for vekten med hensyn på brystomfanget.

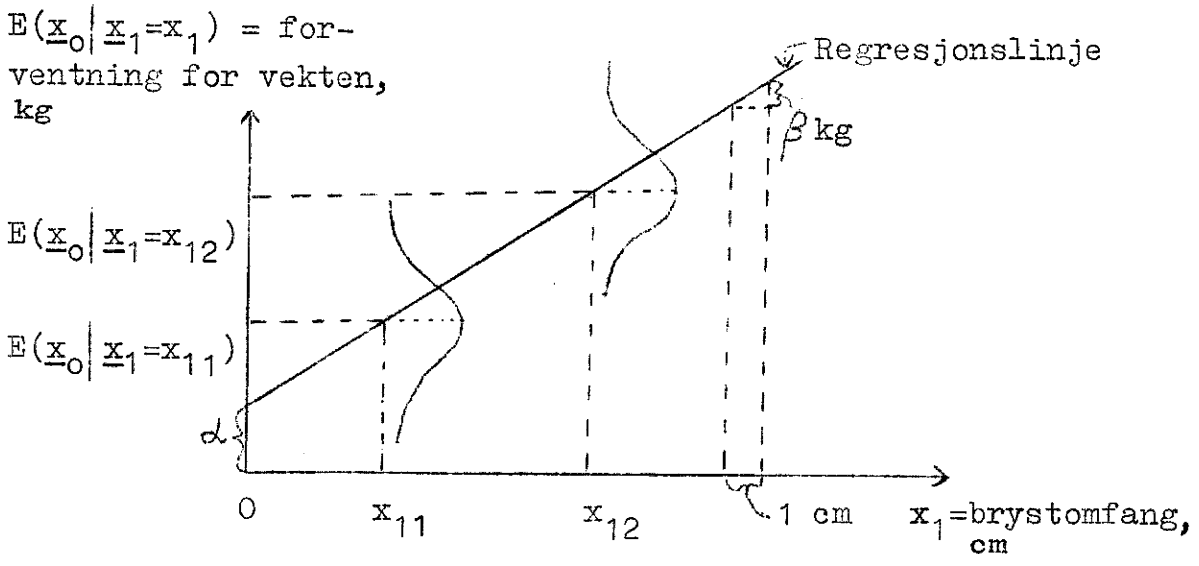


Fig. 11. Tenkt regresjonslinje for vekt med hensyn på brystomfang.

Forholdet er illustrert nærmere i fig. 11. For enkelthets skyld har vi der tenkt oss at regresjonsfunksjonen er lineær slik at den fremstiller en rett linje. En slik linje blir kalt en regresjonslinje. Som alle andre linjer har den selvsagt en matematisk likning, og denne kan skrives på følgende måte:

(98) $$E(\underline{x}_0 | x_1) = \alpha + \beta x_1.$$

Koeffisienten β blir kalt regresjonskoeffisienten.

I fig. 11 er x_1 og $E(\underline{x}_0 | \underline{x}_1 = x_1)$ målt langs henholdsvis abscisse-aksen og ordinataksen, og regresjonslinjen er tegnet inn. Det stykket linjen skjærer av ordinataksen er lik α og linjens vinkelkoeffisient er lik β.

Tar vi for oss brystomfanget x_{11} vil noen kuer med dette brystomfanget ha store vekter, mens andre har små vekster. Dette har en forsøkt a antyde ved å skissere en fordelingsfunksjon (i en litt uvanlig stilling), idet en tenker seg at vi også måler faktiske vekter, x_2 langs ordinataksen. Forventningen for vekten i det subunivers hvor alle kuer har brystomfanget x_{11} er lik $E(\underline{x}_0 | \underline{x}_1 = x_{11})$ som også er avmerket på figuren.

Tar vi for oss et annet brystomfang, f.eks. brystomfanget x_{12} blir forholdet tilsvarende, som antydet i fig. 11. Regresjonslinjen gir oss altså alle sammenhørende par av x_1=brystomfang og $E(\underline{x}_0 | x_1)$=betinget forventning for vekten.

Når regresjonsfunksjonen antas å være lineær, sier vi at vi har a gjøre med <u>simpel lineær regresjon</u>. Det er imidlertid ingen prisipielle vanskeligheter forbundet med å operere med andre funksjoner som fremstiller krumme linjer, f.eks. følgende funksjon av 2. grad:

$$(99) \qquad E(\underline{x}_0 | x_1) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_1^2.$$

Vi sier da at vi har å gjøre med <u>krumlinjet regresjon</u>.

Vekten av en ku avhenger selvsagt ikke bare av brystomfanget, men også av en rekke andre variable som f.eks. lengden, kjøttfylden og alderen. Det er også mulig å ta hensyn til disse variable. Når vi tar med mer enn én forklaringsvariabel sier vi at vi har å gjøre med <u>multippel regresjon</u>. Vil vi f.eks. ta med lengden, x_2 i tillegg til brystomfanget, x_1, kan regresjonsfunksjonen skrives på følgende måte:

$$(100) \qquad E(\underline{x}_0 | x_1 x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Likningen (100) fremstiller et plan i et 3-dimensjonalt rom.

Ved å foreta hypotesetesting kan en avgjøre hvilke ledd som eventuelt bør fjernes i en gitt hypotetisk regresjonsfunksjon. I det følgende skal vi imidlertid bare behandle regresjonsfunksjoner av formen (98).

C. Estimering av regresjonsfunksjonen

Hvis vi har en ku med kjent brystomfang, x_{11} og skal finne ut på grunnlag av dette hva kua veier, er det rimelig å gjette på at vekten = $E(\underline{x}_o | x_{11}) = \alpha + \beta x_{11}$. Problemet i praksis er imidlertid at de konstante koeffisientene α og β praktisk talt alltid er ukjente. Den første oppgaven i regresjonsanalysen blir derfor å estimere α og β. Estimatene vil vi betegne med henholdsvis a og b, og b vil vi kalle den estimerte regresjonskoeffisienten. Vi får da en estimert regresjonsfunksjon som kan skrives på denne måten:

(101) $$\text{est. } E(\underline{x}_o | x_1) = a + b\, x_1$$

Her står est. $E(\underline{x}_o | x_1)$ for estimert betinget forventning for $\underline{x}_o$ betinget av x_1. Vi må regne med at $a \neq \alpha$ og at $b \neq \beta$. Derfor vil i alminnelighet også est. $E(\underline{x}_o | x_1) \neq E(\underline{x}_o | x_1)$. Vi ønsker selvsagt at forskjellene skal bli små. Derfor gjelder det å bruke et godt estimeringsprinsipp.

Estimeringen av α og β må bygge på sammenhørende verdier av brystomfang og vekt for et sampel av kuer. Hvis vi f.eks. er interessert i NRF-kuer, burde vårt sampel helst være et random sampel fra universet av NRF-kuer. Etter at kuene er målt og veid kan vi tenke oss at vi avsetter sammenhørende verdier av brystomfang og vekt i et såkalt spredningsdiagram eller punktdiagram som vist i fig. 12.

Det prinsippet som vanligvis blir brukt til estimering av α og β går under navnet minste kvadraters metode og kan lettest forklares i tilknytning til fig. 12. Hvis vårt sampel består av n kuer, vil spredningsdiagrammet inneholde n punkter, og disse ligger selvsagt fast. La oss tenke oss at vi har tegnet inn en linje med koeffisienter a og b gjennom punktsvermen som vist på figuren. Hvis vi lar a variere vil linjen

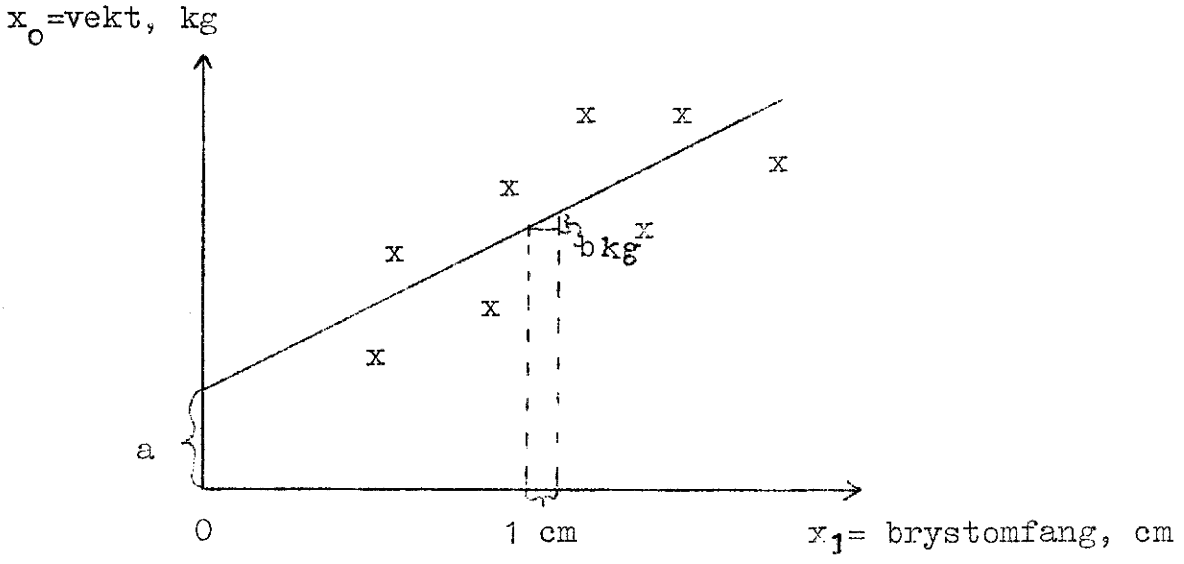


Fig. 12. Spredningsdiagram for sammenhengen mellom brystomfang og vekt med estimert regresjonslinje inntegnet.

bli parallellforskjøvet. Lar vi b variere vil linjen bli dreiet. Ved å variere a og b er det altså mulig å få fram alle tenkelige rette linjer i planet. Minste kvadraters metode går ut på å velge a og b slik at summen av kvadratene av de loddrette avstandene fra punktene til linjen blir så liten som mulig. Dette kravet er oppfylt når a og b blir bestemt etter følgende formler:

$$(102) \qquad b = \frac{\sum(x_o-\bar{x}_o)(x_1-\bar{x}_1)}{\sum(x_1-\bar{x}_1)^2} = \frac{\sum x_o x_1 - \frac{\sum x_o \sum x_1}{n}}{\sum x_1^2 - \frac{(\sum x_1)^2}{n}}$$

(103) $$a = \bar{x}_o - b\bar{x}_1$$

Minste kvadraters metode har vist seg å føre til estimatorer som har visse ønskelige egenskaper. Bl.a. vil vi ha at $E(\underline{b}) = \beta$ under nokså generelle betingelser.

D. Korrelasjonskoeffisienten

Når en har estimert en regresjonslinje (101), er det vanlig å regne ut en størrelse r som gjerne blir kalt korrelasjonskoeffisienten. Det ville være mer korrekt å bruke betegnelsen den estimerte korrelasjonskoeffisienten eller den empiriske korrelasjonskoeffisienten. r blir nemlig regnet ut på grunnlag av observasjonene i samplet, og den kan i mange tilfelle betraktes som et estimat av en tilsvarende størrelse i universet som vi betegner med symbolet ρ (gresk r, leses ro). Det er ρ som bør kalles korrelasjonskoeffisienten. For å skille den fra r kan en eventuelt bruke betegnelsen den teoretiske korrelasjonskoeffisienten. Vi skal i det følgende ikke oppholde oss ved ρ, men se litt nærmere på r.

Den estimerte korrelasjonskoeffisienten r har alltid samme fortegn som b, og den tilfredsstiller alltid følgende uttrykk:

(104) $$-1 \leqq r \leqq 1$$

Formelen for r kan skrives på følgende måte:

(105) $$r = \frac{\Sigma(x_o-\bar{x}_o)(x_1-\bar{x}_1)}{\sqrt{\Sigma(x_o-\bar{x}_o)^2\Sigma(x_1-\bar{x}_1)^2}} = \frac{\Sigma x_o x_1 - \frac{\Sigma x_o \Sigma x_1}{n}}{\sqrt{\left[\Sigma x_o^2 - \frac{(\Sigma x_o)^2}{n}\right]\left[\Sigma x_1^2 - \frac{(\Sigma x_1)^2}{n}\right]}}$$

r kan oppfattes som et mål for hvor tett punktene i fig. 12 ligger inn til den estimerte regresjonslinjen. Hvis punktene ligger tett inn til linjen er r nær 1 hvis b er positiv og nær -1 hvis b er negativ. Hvis punktene ligger spredt helt vilkårlig i planet vil både b og r være lik 0. To estimerte regresjonslinjer som har samme a og samme b og som er estimert på grunnlag av det samme antall observasjoner vil ha forskjellig r hvis punktenes spredning omkring linjen ikke er like stor i de to tilfelle. I fig. 13 har en antydet noen situasjoner med forskjellig r. (r er ikke utregnet eksakt.)

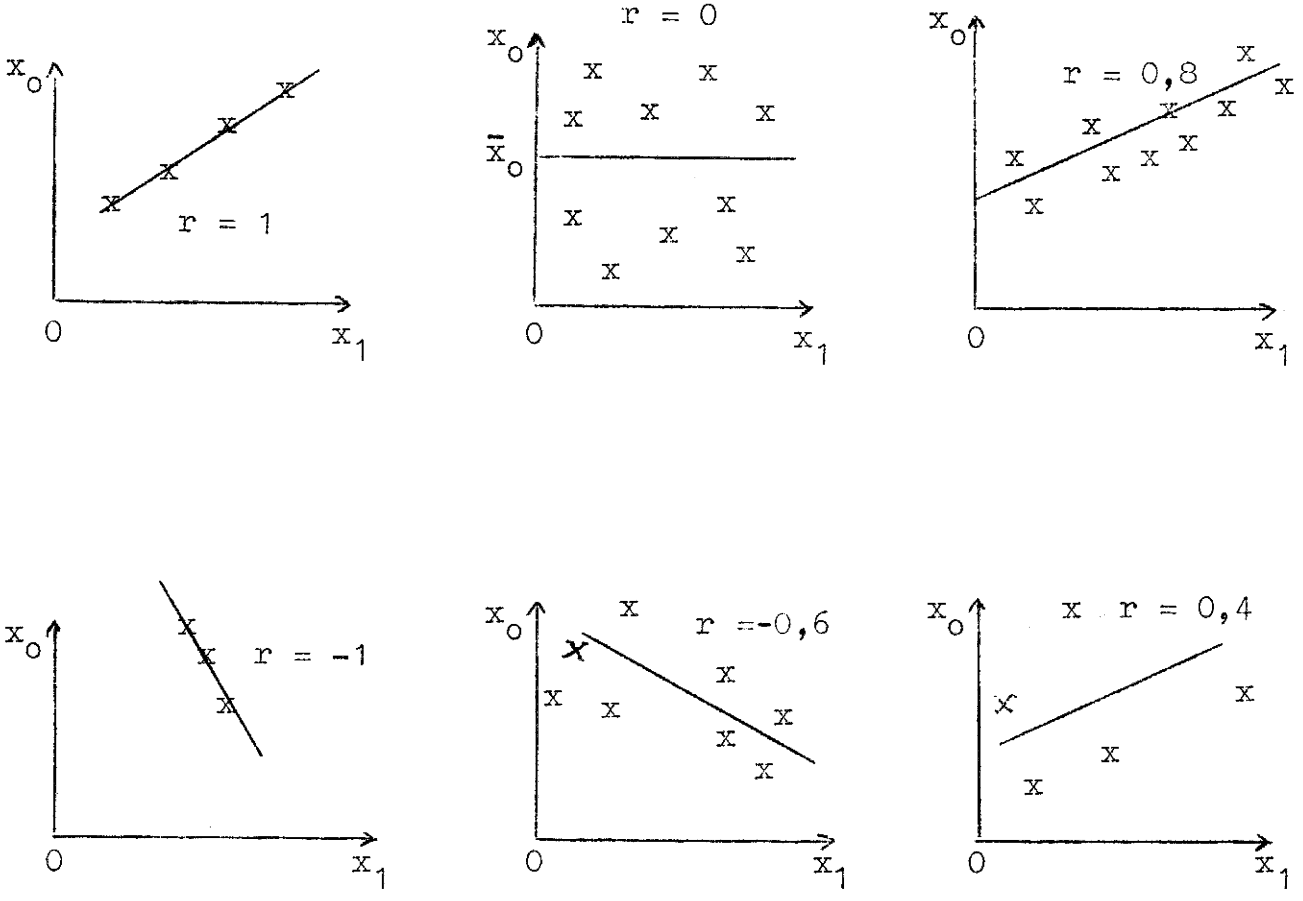


Fig. 13. Eksempler på spredningsdiagram med estimerte regresjonslinjer inntegnet og korrelasjonskoeffisienter oppgitt.

Av (102) og (105) ser vi at b og r har samme teller. Hvis r=0 er også b=0, og da er $a=\bar{x}_c$ i følge (103).

Det hender ofte at en regner ut r uten å regne ut a og b. Hvis det er en sterk positiv eller negativ samvariasjon mellom x_0 og x_1 vil r bli nær 1 i tallverdi. Hvis det er liten samvariasjon i de to rekker av tall vil r være et tall i nærheten av 0.

E. Konfidensintervall for β. Hypotesetesting

På grunnlag av visse forutsetninger er det mulig å beregne et konfidensintervall for β. Disse forutsetningene innebærer bl.a. at de betingede fordelingsfunksjonene som er omtalt i tilknytning til fig. 11 alle er normale og at parameteren σ er den samme i dem alle.

Konfidensintervallet kan beregnes ved hjelp av følgende formler:

(106) Konfidensgrenser for β: $b \mp a\, s_b$

(107) $f = n - 2$

(108) $$s_b^2 = \frac{1-r^2}{n-2} \frac{\sum(x_0-\bar{x}_0)^2}{\sum(x_1-\bar{x}_1)^2} = \frac{1-r^2}{n-2} \frac{\sum x_0^2 - \frac{(\sum x_0)^2}{n}}{\sum x_1^2 - \frac{(\sum x_1)^2}{n}}$$

Tallet a finnes i tabell I når vi går inn med den ønskede konfidenssannsynligheten Q og f = n-2.

Konfidensintervallet kan brukes på vanlig måte til å teste hypoteser om β. Hvis vi har en hypotese, H_0 som går ut på at $\beta = \beta_0$, kan vi teste hypotesen på følgende måte: Hvis β_0 viser seg å ligge mellom konfidensgrensene kan hypotesen ikke forkastes. Ligger β_0 utenfor konfidensintervallet,

må H_0 forkastes. Signifikansnivået, P er lik 1 minus konfidenssannsynligheten, Q. (H_A forutsettes her å være to-sidig.)

En hypotese av spesiell interesse er nullhypotesen $H_0: \beta = 0$. Når vi har beregnet et konfidensintervall bør vi derfor alltid legge merke til om konfidensgrensene har forskjellige fortegn. Hvis dette er tilfelle, er det ingen signifikant sammenheng mellom x_1 og forventningen (betinget forventning) for $\underline{x}_0$. Hvis x_1, som i vårt eksempel, er å oppfatte som verdien av en random variabel $\underline{x}_1$, sier vi da også at det ikke er noen påviselig korrelasjon mellom $\underline{x}_1$ og $\underline{x}_0$.

Hvis $\rho = 0$ sier vi at det ikke er noen korrelasjon mellom $\underline{x}_1$ og $\underline{x}_0$. I et slikt tilfelle vil også β være lik 0.

Siden r er et estimat av ρ, vil en r som ligger i nærheten av 0 kunne tyde på at $\rho = 0$ og $\beta = 0$. En r som har stor tallverdi vil derimot tyde på at nullhypotesen må forkastes. r gir oss en god pekepinn, men før vi trekker en konklusjon bør vi foreta hypotesetesting som vist ovenfor.

F. Mer om bruken av regresjonsanalysen

I praksis viser det seg at en ofte får forbausende gode resultater med simpel lineær regresjon. En lineær regresjonsfunksjon må imidlertid oftest betraktes som en tilnærmelse til den egentlige regresjonsfunksjonen som kan være krumlinjet. Tilnærmelsen kan bli meget god når en har å gjøre med et relativt begrenset intervall av x_1-verdier. Tar vi for oss regresjonslinjen i fig. 11, er det klart at den ikke har noen praktisk mening for brystomfang på 0 cm. Kuer med dette brystomfang vil i følge figuren veie α kg.

Når en anvender en estimert regresjonsfunksjon til å beregne est. $E(\underline{x}_0 | x_1)$ bør en avstå fra å bruke x_1-verdier som ligger utenfor det intervallet av x_1-verdier som er representert i det samplet som estimeringen av regresjonsfunksjonen bygde på. En skal med andre ord helst ikke ekstrapolere en estimert regresjonsfunksjon. Det kan nemlig tenkes at den egentlige regresjonsfunksjonen avviker sterkt fra å være lineær utenfor det nevnte intervallet.

Vi har hittil tenkt oss at x_1 står for verdier av en random variabel $\underline{x}_1$. Regresjonsanalysen kan også brukes i mange tilfelle da x_1 står for tall som vi selv har valt og som ikke er å oppfatte som verdier av en random variabel. Dette er typisk i mange eksperimentelle situasjoner. La oss tenke oss at vi har en rekke forsøksdyr av et bestemt slag og at vi ønsker å studere sammenhengen mellom $\underline{x}_0$ = tilvekst i en bestemt periode (kg) og x_1 = tilførsel av et bestemt vitamin (internasjonale enheter pr. dag). I en slik situasjon er det opp til oss selv å avgjøre hvilke vitamindoser vi vil anvende. Det er derfor ikke naturlig eller hensiktsmessig å oppfatte x_1 som verdier av en random variabel $\underline{x}_1$. I vårt foregående eksempel var forholdet annerledes. Der hadde vi ikke selv noen innflytelse på hvilke brystomfang vi kom til å observere da vi skaffet oss vårt sampel.

Regresjonsanalysen blir den samme enten x_1 er verdier av en random variabel eller ikke. Vi skal imidlertid merke oss at tolkningen av r blir litt annerledes enn før i det sistnevnte tilfellet. r blir fremdeles et mål for punktenes spredning omkring den estimerte regresjonslinjen, men kan ikke lenger oppfattes som et estimat av ρ. Størrelsen

av r vil bli påvirket av hvilke x_1-verdier vi velger. Derfor bør vi være litt forsiktige med å trekke slutninger om β på grunnlag av r i en slik situasjon.

For å være sikker på å oppdage eventuell tendens til krumming, bør en velge x_1-verdiene slik at de dekker hele det aktuelle intervallet for x_1. Hvis en kunne være sikker på at regresjonen var lineær, burde en velge x_1-verdier som lå i ytterpunktene av intervallet. Da ville en få den beste estimeringen av β.

Regresjonsanalysen brukes ofte til å estimere en størrelse på grunnlag av en annen, men den brukes også til å studere årsakssammenhenger. En bør være litt reservert når det gjelder å oppdage nye årsakssammenhenger ved hjelp av regresjonsanalyse. Tilfeldigheter kan føre til at en "oppdager" ting som ikke er der. Regelen bør vel være at regresjonsanalysen brukes til å bekrefte og tallfeste årsakssammenhenger som en på faglig grunnlag hadde mistanke om var der før data ble samlet inn.

XIII. Kji-kvadrat test

Den fordelingsfunksjonen som kalles kji-kvadrat-fordelingen har mange anvendelser i forbindelse med hypotesetesting. Vi skal ved hjelp av eksempler beskrive bruken av et par testmetoder som går ut på å teste hypoteser om sannsynligheter.

Navnet kji-kvadrat kommer av at den testvariabelen vi bruker gjerne får betegnelsen χ^2, altså kvadratet av χ. Den greske bokstaven χ svarer til vår ch og uttales kji.

A. Hypotetiske sannsynligheter i et enkelt univers

I et krysningsforsøk ble maisplanter med røde og melne korn krysset med planter med hvite og glasne korn. Hos avkommet fant en da planter med E_1= røde og melne korn, E_2= røde og glasne korn, E_3= hvite og melne korn og E_4= hvite og glasne korn. Antall planter med disse kjennetegn var:

E_1	E_2	E_3	E_4
447	140	132	49

Vi ønsker å teste en hypotese som går ut på at $P(E_1|U) = 9/16$, $P(E_2|U) = P(E_3|U) = 3/16$ og $P(E_4|U) = 1/16$.

Testingsteknikken går ut på at vi regner ut en størrelse χ^2 som rent praktisk sett kan oppfattes som et mål for hvor dårlig våre data synes å stemme overens med hypotesen. Er χ^2 et stort tall, forkaster vi hypotesen. Utregningen av χ^2 kan settes opp i tabellform på følgende måte:

Tabell 5. Utregning av χ^2 for et krysningsforsøk med mais

(1) Kjennetegn, E_i	(2) Z_i	(3) p_i	(4) np_i	(5) Z_i-np_i	(6) $(Z_i-np_i)^2$	(7) $\frac{(Z_i-np_i)^2}{np_i}$
E_1=røde og melne	447	9/16	432	15	225	0,521
E_2=røde og glasne	140	3/16	144	-4	16	0,111
E_3=hvite og melne	132	3/16	144	-12	144	1,000
E_4=hvite og glasne	49	1/16	48	1	1	0,021
	n=768	1	n=768	0		χ^2= 1,653

I kolonne (1) i tabell 5 har en ført opp de forskjellige kategorier av avkom som kan forekomme. Disse kategoriene må alltid være uttømmende og ikke overlappende. Ethvert avkom må altså kunne henføres til en og bare en kategori. Antall kategorier (antall kjennetegn) vil vi betegne med m som er lik 4 i vårt eksempel. Kolonne (2) inneholder den observerte absolutte frekvensen Z_i for kjennetegnet E_i (i=1,2,...,m). Summen $\sum_{i=1}^{m} Z_i$ vil vi betegne med n som er lik 768 i eksemplet. n er altså sampelstørrelsen. De 768 avkom må nemlig oppfattes som et sampel fra et univers U av avkom etter denne spesielle type krysning. Kolonne (3) inneholder den hypotetiske sannsynligheten p_i for E_i i universet U (i=1,2,...,m). Summen av tallene i kolonne (3) må alltid være lik 1. Vi kan si at kolonne (2) beskriver vårt sampel, mens kolonne (3) beskriver universet slik det er i følge hypotesen.

I kolonne (4) har vi ført opp forventningen np_i for antall avkom med kjennetegnet E_i (i=1,2,...,m). I vårt eksempel er np_i et helt tall for alle i, men slik er det ikke alltid. At np_i er en forventning er kanskje ikke umiddelbart innlysende for alle, men intuitivt er det ikke vanskelig å være med på dette. Hvis vi

tenker oss en situasjon med bare to kjennetegn, f.eks. E_1 og iE_1 vil det vi har gjennomgått om binomialfordelingen gi oss det nødvendige bakgrunnsstoff for forståelsen.

Legg merke til at summen av kolonne (4) alltid blir lik n. Kolonne (2) og (4) er altså helt sammenliknbare. Kolonne (2) gir oss den observerte fordelingen av de 768 avkom mens (4) gir oss den forventede fordelingen, dvs. den fordelingen vi ville få i "gjennomsnitt" for en uendelighet av sampler, alle på n=768 avkom. En testing av hypotesen vil derfor kunne foretas ved å sammenlikne kolonne (2) og kolonne (4). Det ville f.eks. ikke være unaturlig å se på differensene mellom disse kolonnene. Dette er gjort i kolonne (5).

Hvis hypotesen synes å stemme dårlig overens med våre data, vil avvikene mellom observert og forventet resultat i kolonne (5) bli store i tallverdi. En nærliggende første tanke kunne være å summere disse avvikene og bruke summen som et mål for hvor dårlig hypotesen synes å stemme med våre data. Denne summen vil imidlertid alltid bli 0. Hvis det blir for mange avkom i en kategori i forhold til det forventede antall, vil det nemlig alltid bli et tilsvarende antall for få i de øvrige kategorier.

For å eliminere dette fortegnsproblemet kunne vi kvadrere avvikene som vist i kolonne (6). Men heller ikke summen av tallene i kolonne (6) er noe godt mål for hypotesens tilsynelatende "dårlighet". I første linje i tabell 5 ser vi f.eks. at $Z_i - np_i = 15$ mens $np_i = 432$. Avviket på 15 er kanskje ikke særlig stort, men hvis np_i hadde vært f.eks. 20, ville de fleste av oss ha sagt at et avvik på 15 er stort. Det er altså rimelig å se avvikene og dermed også de kvadrerte avvikene i forhold til det forventede resultat, np_i. Dette er gjort i kolonne (7) hvor en har dividert tallene i kolonne (6) med np_i.

Summen av tallene i kolonne (6) er det søkte kji-kvadrat. Dette vil selvsagt alltid være positivt, men forøvrig kan det ha en hvilken som helst verdi. Hypotesen kan tenkes å være riktig uansett hvor stort tallet χ^2er. Men det skulle være klart at et stort χ^2 gir oss større grunn til skepsis overfor hypotesen enn et lite χ^2. (Hvis χ^2 er et tall meget nær 0, kan det være grunn til mistanke om regnefeil, falske data e.l., men vi ser bort fra slike ting her.)

Spørsmålet er hvor stort χ^2 må være for at vi skal bestemme oss for å forkaste hypotesen. Kji-kvadratfordelingen gir oss et hjelpemiddel til å avgjøre dette. Under forutsetning av at hypotesen er riktig kan det nemlig bevises at tallet $\chi^2 = 1{,}653$ som vi har regnet ut er en verdi av en random variabel $\underline{\chi}^2$ som er tilnærmet fordelt etter kji-kvadratfordelingen. Grafen til kji-kvadratfordelingen har et liknende forløp som grafen til F-fordelingen, med en lang "hale" til høyre. Kji-kvadratfordelingen har en parameter f som kalles antall frihetsgrader. I den type problem som vi har å gjøre med i dette avsnittet blir f bestemt etter følgende formel:

(109) $$f = m-1$$

Vi skal ikke gjenta her prinsippene for hypotesetesting, men viser til tidligere avsnitt. I vårt eksempel blir f= 4-1 = 3. Hvis vi har bestemt oss for å bruke signifikansnivået P=0,05, finner vi i tabell V at den kritiske verdien av χ^2 er a=7,815. Siden tallet $\chi^2 = 1{,}653$ som vi har funnet er mindre enn a kan vi ikke forkaste hypotesen. Siden dette ikke er en null-hypotese, men en naturvitenskapelig hypotese som er stilt opp på grunnlag av Mendels lover, kan det da være grunn til å akseptere hypotesen.

La oss til slutt ta med et lite forbehold. Den random variable vi har å gjøre med er bare tilnærmet kji-kvadrat fordelt.

Tilnærmelsen blir gjerne ansett for å være god nok så lenge alle tallene np_i er i det minste lik 5. Hvis kravet ikke er oppfylt kan vi bøte på dette ved å slå sammen visse kategorier og beregne et kji-kvadrat på grunnlag av et mindre antall kategorier. Hvis f.eks. np_i i tabell 5 hadde vært 4 i stedet for 48, kunne vi ha lagt sammen Z_i og p_i for E_3 og E_4 og beregnet et nytt χ^2 hvor m var lik 3.

B. Ukjente sannsynligheter som i følge hypotesen er de samme i flere universer

Vi skal se på en type til av χ^2-test som det er svært nyttig å ha kjennskap til.

I følgende oppstilling angir tallene Z_{ij} antall vaksinerte personer angrepet av kopper, ordnet etter antall år som har gått siden de ble vaksinert og etter graden av angrepet.

Antall år siden vaksinasjon	Hardt angrepne E_1	Lett angrepne E_2	Sum
0-25	Z_{11}=163	Z_{21}=324	n_1=487
25-45	Z_{12}=483	Z_{22}=449	n_2=932
Sum	646	773	1419

Vi ønsker å undersøke om det på grunnlag av disse data kan påvises at tiden som har gått fra vaksinasjonen til sykdomsangrepet, har noen betydning for graden av angrepet.

Universet U av alle personer som er angrepet av kopper og som er vaksinert for 0-45 år siden kan deles i to subuniverser etter hvor mange år det er gått siden vaksinasjonen. I universet U_1 er det gått 0-25 år og i U_2 25-50 år. Fra U_1 har vi et (for-

håpentlig random) sampel på n_1=487 personer og fra U_2 et sampel på n_2=932 personer. For å få besvart vårt spørsmål, stiller vi opp en null-hypotese som går ut på at $P(E_1|U_1) = P(E_1|U_2)$. (Hypotesen innebærer selvfølgelig også at $P(E_2|U_1) = P(E_2|U_2$.). E_1 og E_2 er nemlig motsatte kjennetegn. Det følger videre at $P(E_1|U_2) = P(E_1|U)$ og at $P(E_2|U_2) = P(E_2|U)$. Alt i alt kan vi si noe upresist at hypotesen innebærer at tallene i de tre linjene i oppstillingen ovenfor kan ventes å ha en tendens til å være proporsjonale.)

Utregningen av χ^2 for dette eksemplet kan foretas på liknende måte som i tabell 5, men vi må nå ha en dobbel tabell. Første del av denne tabellen bygger på tallene i første linje i oppstillingen (0-25 år), og andre del bygger på tallene i annen linje (25-45 år).

Vi mangler sannsynligheter å sette inn i tabellene, men disse estimeres på grunnlag av sumtallene i siste linje i oppstillingen. Vi får: Est. $P(E_1|U) = \hat{p}_1 = 646/1419 = 0{,}45525$ og est. $P(E_2|U) = \hat{p}_2 = 773/1419 = 0{,}54475 = 1-\hat{p}_1$.
Utregningen for å finne χ^2 er vist i tabell 6.

Tabell 6. Utregning av χ^2 for data over vaksinerte personer angrepet av kopper.

0-25 år:

Kjennetegn E_i	Z_{i1}	$\hat{p}_i$	$n_1\hat{p}_i$	$Z_{i1}-n_1\hat{p}_i$	$(Z_{i1}-n_1\hat{p}_i)^2$	$\frac{(Z_{i1}-n_1\hat{p}_i)^2}{n_1\hat{p}_i}$
E_1	163	0,45525	221,70675	-58,70675	3446,48	15,545
E_2	324	0,54475	265,29325	58,70675	3446,48	12,991
n_1=	487	1,00000	487,00000	0		

25-45 år:

Kjennetegn,E_i	Z_{i2}	$\hat{p}_i$	$n_2\hat{p}_i$	$Z_{i2}-n_2\hat{p}_i$	$(Z_{i2}-n_2\hat{p}_i)^2$	$\frac{(Z_{i2}-n_2\hat{p}_i)^2}{n_2\hat{p}_i}$
E_1	483	0,45525	424,29300	58,70700	3446,51	8,123
E_2	449	0,54475	507,70700	-58,70700	3446,51	6,782
n_2=	932	1,00000	932,00000	0		χ^2=43,447

Utregningen av χ^2 skulle ikke trenge nærmere kommentar. Også i dette tilfelle er det slik at et stort χ^2 gjør oss mistenkelige overfor hypotesen. Et stort χ^2 er nemlig et tegn på at linjene i den opprinnelige oppstillingen av våre data avviker sterkt fra å være proporsjonale. Dette kan leseren overbevise seg om ved et logisk resonnement eller ved å konstruere kunstige sett av data. Et sett av data hvor linjene er eksakt proporsjonale vil f.eks. gi et χ^2 som er eksakt lik 0.

Det skulle være klart at jo mer de nevnte linjene avviker fra å være proporsjonale, desto dårligere kan vi si at våre data synes å stemme overens med hypotesen.

Hvor stort χ^2 må være for at vi skal forkaste hypotesen blir avgjort på grunnlag av tabell V. Det kan nemlig vises at under forutsetning av at hypotesen er riktig er tallet χ^2 en verdi av en random variabel $\underline{\chi}^2$ som tilnærmet følger kji-kvadrat-fordelingen.

Den metodikken vi nå har gjennomgått kan brukes også i tilfelle hvor vi har flere enn to kjennetegn E_1 og E_2 og/eller flere enn to subuniverser U_1 og U_2. Har vi f.eks. m kjennetegn E_i og k universer, U_j, vil en tabell som svarer til tabell 6 måtte inneholde k deltabeller og hver av disse deltabellene vil inneholde m linjer.

I oppgaver av denne type kan antall frihetsgrader bestemmes etter følgende formel:

(110) $$f = (k-1)(m-1)$$

I vårt eksempel får vi $f = (2-1)(2-1) = 1$. Bruker vi signifikansnivået $P = 0,05$, finner vi i tabell V at $a = 3,841$. Vår χ^2 som er lik 43,447 er derfor signifikant og nullhypotesen må følgelig forkastes. Av den opprinnelige dataoppstillingen trekker

vi da den konklusjonen at det er relativt flest hardt angrepne i den gruppen som er vaksinert for 25-45 år siden.

Også ved denne typen χ^2-test må en stille det krav at $n_j \cdot p_i$ alltid skal være i det minste lik 5.

Den beregningsmåten for χ^2 som er vist i tabell 6 er nok ikke den enklest mulige, men den har den fordelen at den bygger på det samme grunnskjemaet som vi brukte i tabell 5 og som vi senere skal bruke i tabell 7. Som vi har sett gir beregningsmåten også et innblikk i hva χ^2 gir uttrykk for uten at vi behøver å innføre nye statistiske begreper.

C. En hypotese om at en random variabel følger en spesiell fordelingsfunksjon

Den neste testmetoden vi skal se på minner sterkt om metoden som ble gjennomgått under punkt A. Forskjellen er at de hypotetiske sannsynlighetene må regnes ut på grunnlag av en hypotetisk fordelingsfunksjon og at frihetsgradene derfor bestemmes anderledes.

Vi vil bruke oppgaven på s. 61 og s. 68 som eksempel. På s. 61 stilte vi opp en frekvensfordeling for antall celler av et bestemt slag pr. prøve av en vevsvæske. På s. 68 undersøkte vi om antall celler pr. prøve kunne antas å være en random variabel som fulgte Poissons fordelingsfunksjon. Vi sammenliknet derfor de relative frekvensene i samplet med de tilsvarende sannsynlighetene som ble regnet ut ved hjelp av Poissons fordelingsfunksjon. For å kunne regne ut sannsynlighetene måtte vi ha et tall for parameteren m. Denne ble estimert ved å sette $m=\bar{X}$. Problemet var at vi ikke visste hvor god overensstemmelsen mellom de relative frekvensene og de

tilsvarende sannsynlighetene måtte være for å kunne karakteriseres som tilfredsstillende. Dette problemet skal vi løse nå.

Vi setter opp en hypotese, H_0 som går ut på at antall celler følger Poissons fordeling. Det kan bevises at denne hypotesen kan testes ved å regne ut et kji-kvadrat på en måte som svarer helt til det som ble gjort i tabell 5. Utregningen er vist i tabell 7.

Tabell 7. Utregning av χ^2 for data over antall celler av et bestemt slag i prøver av en vevsvæske.

X	Z_i	p_i	np_i	Z_i-np_i	$(Z_i-np_i)^2$	$\frac{(Z_i-np_i)^2}{np_i}$
0	6	0,0665	6,65	-0,65	0,4225	0,0635
1	16	0,1803	18,03	-2,03	4,1209	0,2286
2	24	0,2443	24,43	-0,43	0,1849	0,0076
3	22	0,2207	22,07	-0,07	0,0049	0,0002
4	23	0,1495	14,95	8,05	64,8025	4,3346
5	6	0,0810	8,10	-2,10	4,4100	0,5444
6 og større	3	0,0577	5,77	-2,77	7,6729	1,3298
n =100		1,0000	100,00	0		$\chi^2=6,5087$

Antall frihetsgrader bestemmes nå etter en formel som likner på formel (109), s. 143, nemlig

(111) $$f = m-1-c$$

hvor m som før er antall frekvenser vi har brukt (antall linjer i tabellen) og c er antall parametre i den hypotetiske fordelingsfunksjonen som vi har måttet estimere ved hjelp av observasjonene i samplet. I vårt eksempel har vi måttet estimere 1 ukjent parameter, nemlig m. Følgelig er c=1 og f=m-1-c=7-1-1=5. Hvis vi velger signifikansnivået P=0,05 finner vi i tabell V at a=11,070. Det χ^2 vi har

funnet er altså ikke signifikant. Hypotesen H_0 om at antall celler følger Poissons fordelingsfunksjon (i hele universet av prøver) kan altså ikke forkastes.

Også ved denne type av χ^2-test er det et krav av np_i ikke skal være mindre enn 5. I tabell 7 er den minste verdien av np_i lik 5,77. Hvis vi ikke hadde slått sammen alle X-verdier fra 6 og oppover, ville imidlertid np_i ha blitt mindre enn 5 for disse verdiene.

Denne testmetoden kan også brukes når vi har å gjøre med en kontinuerlig random variabel. I slike tilfelle vil tallene i 1. kolonne i tabell 7 ikke bestå av X-verdier, men av intervaller (klasser) for X-verdiene. Sannsynlighetene p_i vil være arealer som vi kan beregne ved integrering eller ved numeriske metoder (i noen tilfelle kan vi også bruke publiserte tabeller).

FACIT TIL ØVELSENE OG OPPGAVENE

Øvelse 2, s.13: Ja. Øvelse 3, s. 13: To kjennetegn E_1 og E_2 som kan forekomme i et univers, U er uavhengige hvis og bare hvis $P(E_1E_2|U) = P(E_1|U)\ P(E_2|U)$. (Uavhengighetskriteriet kan også formuleres på andre måter.) Øvelse 4, s. 13: Både-og setningen, $P(E_1E_2E_3|U) = P(E_1|U)\ P(E_2|UE_1)\ P(E_3|UE_1E_2)$. Uavhengighets-kriteriet, $P(E_1E_2E_3|U) = P(E_1|U)\ P(E_2|U)\ P(E_3|U)$. Øvelse 5, s. 14: $P(E_1|U)=1/5$. $P(E_2|U)=11/20$. $P(E_1|UE_2)=3/11$. $P(E_2|UE_1)=3/4$. $P(E_1E_2|U)=3/20$. $P(E_2E_1|U)=3/20$. Vi får 3/20=3/20 i begge tilfelle. E_1 og E_2 er ikke uavhengige kjennetegn. P(enten E_1 eller E_2 eller $E_1E_2|U) = P(E_1|U)+P(E_2|U)-P(E_1E_2|U)$. Øvelse 6, s. 14: Vi har uavhengighet mellom alle de nevnte par av kjenne-tegn. Hvis vi har uavhengighet mellom et av de nevnte par av kjennetegn i en slik situasjon vil alle linjene i tabellen bli proporsjonale og alle kolonnene i tabellen vil bli proporsjo-nale. Dermed vil vi automatisk få alle de uavhengighetene som er nevnt. Øvelse 7, s. 15: a) 1/2. b) 1/6. c) 3/4. Hvis ett og kun ett av m mulige alternativer må inntreffe og hvis alle m alternativene har samme sannsynlighet ut fra a priori betraktninger vil sannsynligheten for et kjennetegn være antall alternativer som er "gunstige" for dette kjennetegnet dividert med m (den klassiske definisjonen av en sannsynlighet). Øvelse 8, s. 16: Ta ut et sampel. Samplet bør være random. Øvelse 9, s. 17: Nei. Sannsynligheten for å gjette riktig ville bli 5/9 som er mindre enn 2/3 som er sannsynligheten for å gjette riktig hvis jeg alltid sier at han er fra landet. Øvelse 10, s. 23: Ved 1. trekning er $P(\text{okse}|U)=13/25=0,52$. Ved 2. trekning er $P(\text{okse}|U \text{ okse 1. gang})=12/24=0,50$ mens $P(\text{okse}|U \text{ kvige 1. gang})=13/24=0,55$. Øvelse 11, s. 24: a) $P_3=0,2700$. b) $P_0=0,0531$. $P_1=0,2300$. $P_2=0,3738$. $P_4=0,0731$. c) Summen er 1 fordi det ikke finnes andre muligheter. Øvelse 12, s. 25: 1/8. Øvelse 13, s. 25: 0,0729. Øvelse 14, s. 25: a) 1/324. b) 1/36. c) 38/324. Oppgave, s. 37: a) 1/48. b) 7/24. c) 3/4. Øvelse 15, s. 43: 2,08.

Øvelse 18, s. 46: $\mu=1,6$. $\sigma^2=0,640$. b) 0,027. Øvelse 19, s. 46: $\sigma^2=0,9984$. Øvelse 19b, s. 46: Forventningen er 1 i alle tre tilfelle. Standardavviket er henholdsvis $\sqrt{2/3}$, $\sqrt{1/2}$ og 0. Standardavviket er forskjellig på grunn av at sannsynlighetene er forskjellige i de tre universene. Øvelse 19c, s. 48: $P(\underline{X}\leq 160 \text{ eller } \underline{X}\geq 192)\leq 0,25$. Øvelse 20b, s. 48: $P(\mu-a \leq \underline{X} \leq \mu+a) \geq 1-1/a^2$. Her må $a>1$. Oppgave s. 61: $\bar{X}=2,71$. $s=1,47$. Se også oppgave s. 68. Oppgave s. 68:

X	0	1	2	3	4	5	6	7
z/n	0,06	0,16	0,24	0,22	0,23	0,06	0,02	0,01
f(X)	0,0665	0,1803	0,2443	0,2207	0,1495	0,0810	0,0366	0,0142

Øvelse 24, s. 92: 31,8 og 44,8. Øvelse 26, s. 97: Konfidensgrensene for forventningen for differensen mellom vekten av venstre og høyre nyre er 0,34 og 2,32. Øvelse 27, s. 98: Det kan være ønskelig at vektene ikke varierer for mye (jevn kvalitet). Øvelse 28, s. 102: -0,18 og 7,18. Øvelse 30, s. 104: 0,4862 og 0,5502. Øvelse 31II, s. 104: 0,0168 og 0,0530. Øvelse 31III, s. 104: -0,57 og 0,82. Øvelse 31 IV, s. 104: 0,1 og 7,1. Øvelse 32, s. 121: H_0: Forventningen for lengden er den samme i alle tre gruppene. $F=5,924$. $a=F_{0,05}(2,30)=3,32$. H_0 må forkastes på 5%-nivået. Øvelse 33, s. 128: $F=69,3$. $a=F_{0,05}(3,12)=3,49$. F er signifikant på 5%-nivået. H_0 forkastes. Det er påviselig forskjell på forventningen for blomstringstiden for de 4 artene.

Tabell I

Students t

f	P	
	0,05	0,01
1	12.706	63.657
2	4.303	9.925
3	3.182	5.841
4	2.776	4.604
5	2.571	4.032
6	2.447	3.707
7	2.365	3.499
8	2.306	3.355
9	2.262	3.250
10	2.228	3.169
11	2.201	3.106
12	2.179	3.055
13	2.160	3.012
14	2.145	2.977
15	2.131	2.947
16	2.120	2.921
17	2.110	2.898
18	2.101	2.878
19	2.093	2.861
20	2.086	2.845
21	2.080	2.831
22	2.074	2.819
23	2.069	2.807
24	2.064	2.797
25	2.060	2.787
26	2.056	2.779
27	2.052	2.771
28	2.048	2.763
29	2.045	2.756
30	2.042	2.750
40	2.021	2.704
60	2.000	2.660
120	1.980	2.617
	1.960	2.576

Tabell II. Varianskvotienten F. P = 0,05

v_2 \ v_1	f for teller.									
	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,88	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,80	3,41	3,18	3,02	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,38	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,22	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,44	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,54	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,52	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83

f for nevner

Tabell II. Varianskvotienten F. P = 0,05

v_2 \ v_1	f for teller.								
	12	15	20	24	30	40	60	120	∞
1	243,9	245,9	248,0	249,0	250,1	251,1	252,2	253,3	254,3
2	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2.54
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2,50	2,47	2,43	2,38	2,34	2,30
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,20	2,13	2,05	2,00	1,96	1,91	1,86	1,81	1,76
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1.00

f for nevner

Tabell IV.. Varianskvotienten F. P = 0,01

v_2 \ v_1	f for teller.								
	1	2	3	4	5	6	7	8	9
1	4052	4999,5	5403	5625	5764	5859	5928	5982	6022
2	98,49	99,01	99,17	99,25	99,30	99,33	99,36	99,37	99,39
3	34,12	30,81	29,46	28,71	28,24	27,91	27,67	27,49	27,35
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16
6	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94
11	9,65	7,20	6,22	5,67	5,32	5,07	4,89	4,74	4,63
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39
13	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19
14	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68
18	8,28	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40
22	7,94	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26
25	7,77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,22
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56
∞	6,64	4,60	3,78	3,32	3,02	2,80	2,64	2,51	2,41

f for nevner

Tabell IV, Varianskvotienten F. P = 0,01

v_2 \ v_1	f for teller.									
	10	12	15	20	24	30	40	60	120	∞
1	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

f for nevner.

Tabell V. Kji-kvadrat

f	P			
	0,99	0,95	0,05	0,01
1	0,000	0,004	3,841	6,635
2	0,020	0,103	5,991	9,210
3	0,115	0,352	7,815	11,341
4	0,297	0,711	9,488	13,277
5	0,554	1,145	11,070	15,086
6	0,872	1,635	12,592	16,812
7	1,239	2,167	14,067	18,475
8	1,646	2,733	15,507	20,090
9	2,088	3,325	16,919	21,666
10	2,558	3,940	18,307	23,209
11	3,053	4,575	19,675	24,725
12	3,571	5,226	21,026	26,217
13	4,107	5,892	22,362	27,688
14	4,660	6,571	23,685	29,141
15	5,229	7,261	24,996	30,578
16	5,812	7,962	26,296	32,000
17	6,408	8,672	27,587	33,409
18	7,015	9,390	28,869	34,805
19	7,633	10,117	30,144	36,191
20	8,260	10,851	31,410	37,566
21	8,897	11,591	32,671	38,932
22	9,542	12,338	33,924	40,289
23	10,196	13,091	35,172	41,638
24	10,856	13,848	36,415	42,980
25	11,524	14,611	37,652	44,314
26	12,198	15,379	38,885	45,642
27	12,879	16,151	40,113	46,963
28	13.565	16,928	41,337	48,278
29	14,256	17,708	42,557	49,588
30	14,953	18,493	43,773	50,892