NOEN PROBLEMSTILLINGER HVOR DET INNGÅR MER ENN ÉN RANDOM VARIABEL

Forelesninger ved Norges landbrukshøgskole

ved

Ivar Kristianslund

Institutt for matematiske fag

Norges landbrukshøgskole

Vollebekk 1969

NOEN PROBLEMSTILLINGER HVOR DET INNGÅR MER ENN ÉN RANDOM VARIABEL

Forelesninger ved Norges landbrukshøgskole

ved

Ivar Kristianslund

Institutt for matematiske fag
Norges landbrukshøgskole

Vollebekk 1969

I n n h o l d

I. Innledning

Dette notatet er skrevet for jordskiftestudenter ved Norges landbrukshøgskole som bruker mitt hefte MATEMATISK STATISTIKK, Vollebekk 1969, men som har behov for en mer omfattende tekst.

I hovedavsnitt II er det gitt en innføring i simultane, marginale og betingede fordelingsfunksjoner m.v. Disse emner er ikke behandlet eksplisitt i noen av de forelesningshefter som brukes for studenter ved NLH for tiden. Stoffet kan derfor være av interesse også for andre studenter. Det er nemlig en viktig del av grunnlaget for svært mange statistiske metoder.

I hovedavsnitt III har en nokså detaljert gjennomgått formlene for forventningen og variansen for en funksjon av random variable samt enkelte tilgrensende emner. Erfaringen synes å tyde på at manglende forståelse av det stoffet som er gjennomgått i hovedavsnitt II og III er en viktig grunn til at enkelte studenter har vanskelig for å bli helt fortrolig med statistikken.

Hovedavsnitt IV gir en innføring i regresjonsanalyse og en har her lagt vekt på å knytte forbindelse med hovedavsnitt II. I hovedavsnitt II har en forsøkt å vise hva begrepene korrelasjon, regresjon, osv. egentlig står for i tilknytning til et univers. I hovedavsnitt IV har en vist hvorledes en på grunnlag av et sampel kan trekke slutninger om universet.

De avsnittene som er merket med en stjerne (*) er beregnet på spesielt interesserte studenter og kan overspringes uten skade for sammenhengen.

Alle random variable er skrevet med understrekede symboler, mens verdier av de random variable er skrevet uten understrekning. Symbolbruken er ellers stort sett den som professor dr. Per Ottestad har brukt i sine forelesningshefter. Framstillingen er også på mange andre måter sterkt påvirket av professor Ottestads arbeider. Således er ideene til de fleste eksemplene i avsnitt III A1 og IV hentet fra hans oppgavesamling.

II. Simultane, marginale og betingede fordelingsfunksjoner og beslektede begreper

A. Simultane fordelingsfunksjoner

La oss betrakte to diskrete random variable, $\underline{x}_1$ og $\underline{x}_2$. Vi vil ta for oss en vilkårlig verdi x_1 blant de verdier $\underline{x}_1$ kan anta og en vilkårlig verdi x_2 blant de verdier $\underline{x}_2$ kan anta. Sannsynligheten for at $\underline{x}_1$ skal anta verdien x_1 og at $\underline{x}_2$ samtidig skal anta verdien x_2 kan skrives som $P(\underline{x}_1 = x_1 \text{ og } \underline{x}_2 = x_2)$. Prinsipielt kunne vi tenke oss å finne denne sannsynligheten ved opptelling av gjentak i et univers. La oss anta at sannsynligheten kan skrives som en funksjon, $f(x_1, x_2)$ av x_1 og x_2. Vi får da:

$$(1) \qquad P(\underline{x}_1 = x_1 \text{ og } \underline{x}_2 = x_2) = f(x_1, x_2).$$

Funksjonen $f(x_1, x_2)$ blir kalt den simultane (samtidige) fordelingsfunksjonen for $\underline{x}_1$ og $\underline{x}_2$. På tilsvarende måte kan vi operere med en simultan fordelingsfunksjon for mer enn to random variable.

Vi vil belyse det hele med et eksempel. La $\underline{x}_1$ være resultatet av et tilfeldig kast med en "riktig" terning og $\underline{x}_2$ resultatet av et tilfeldig kast med en "riktig" mynt. $\underline{x}_1$ kan da anta verdiene 1, 2, 3, 4, 5 eller 6. $\underline{x}_2$ kan anta verdien 0 som vi lar stå for mynt eller 1 som vi lar stå for krone. Hvis vi foretar et kast med terningen og mynten samtidig, kan vi tenke oss 12 forskjellige resultater med sannsynligheter som vist i tabell 1.

Tabell 1. Mulige utfall og tilhørende sannsynligheter ved kast med en mynt og en terning samtidig

Verdier av $\underline{x}_1$ (Det vil si resultatet av myntkastet)	Verdier av $\underline{x}_2$ (Dvs. resultatet av terningskastet)						
	1	2	3	4	5	6	Sum
0	1/12	1/12	1/12	1/12	1/12	1/12	1/2
1	1/12	1/12	1/12	1/12	1/12	1/12	1/2
Sum	1/6	1/6	1/6	1/6	1/6	1/6	1

Alle resultatene har samme sannsynlighet, 1/12, og vi kan derfor skrive:

$$(2) \quad P(\underline{x}_1 = x_1 \text{ og } \underline{x}_2 = x_2) = f(x_1, x_2) = \frac{1}{12}.$$

Den simultane fordelingsfunksjon for to diskrete random variable kan framstilles grafisk i et 3-dimensjonalt aksesystem som vist i fig. 1.

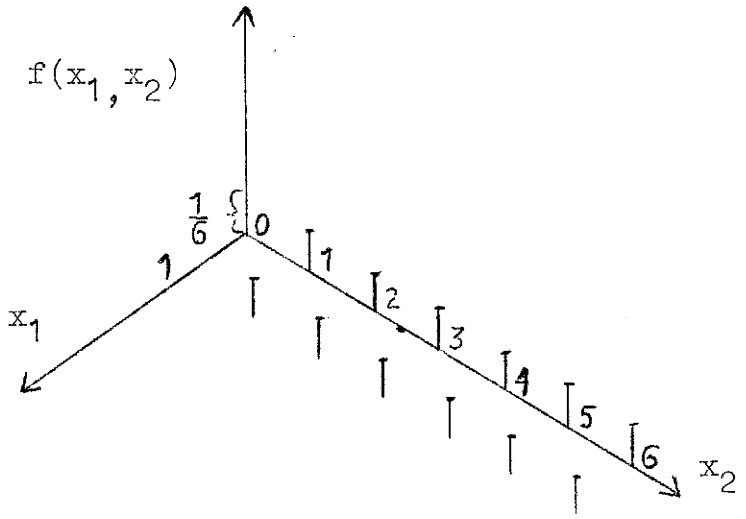


Fig. 1. Den simultane fordelingsfunksjonen i tabell 1 framstilt grafisk

Vi skal se på et annet eksempel som ikke er fullt så enkelt.

Hvis en rød korthornokse (genkonstellasjon $\frac{A}{A}$) krysses med en hvit ku (genkonstellasjon $\frac{a}{a}$), blir avkommet i første krysningsgenerasjon (F_1) skimlet (genkonstellasjon $\frac{A}{a}$). Krysses skimlete dyr ($\frac{A}{a}$) av F_1-generasjonen med hverandre, vil sannsynlighetene for rødt ($\frac{A}{A}$), skimlet ($\frac{A}{a}$) og hvitt ($\frac{a}{a}$) avkom i F_2-generasjonen være henholdsvis 1/4, 1/2 og 1/4. I et random sampel på 4 dyr av F_2 (tallet 4 er valgt nokså vilkårlig) kan en oppfatte antall røde ($\frac{A}{A}$) dyr som en random variabel, $\underline{x}_1$. Videre kan vi oppfatte antall skimlete ($\frac{A}{a}$) dyr som en random variabel $\underline{x}_2$ og antall hvite ($\frac{a}{a}$) dyr som en random variabel $\underline{x}_3$. Det kan

vises at den simultane fordelingsfunksjonen for $\underline{x}_1$, $\underline{x}_2$ og $\underline{x}_3$ er gitt ved følgende formel (som er et eksempel på den såkalte mulitinomiale fordelingsfunksjon):

$$(3) \qquad f(x_1,x_2,x_3) = \frac{4!}{x_1!x_2!x_3!}\left(\frac{1}{4}\right)^{x_1}\left(\frac{1}{2}\right)^{x_2}\left(\frac{1}{4}\right)^{x_3}.$$

I (3) kan vi for hver x sette inn et hvilket som helst helt tall fra 0 til 4, men selvsagt med den restriksjonen at summen av x-ene må være lik 4.

Siden summen av x-ene i vårt problem alltid må være lik 4, er en av x-ene alltid kjent hvis de to andre er gitt. Fordelingsfunksjonen (3) kan derfor like gjerne oppfattes som en simultan fordelingsfunksjon for bare to random variable, f.eks. for $\underline{x}_1$ og $\underline{x}_2$, og vi skriver den da på følgende måte:

$$(4) \qquad f(x_1,x_2) = \frac{4!}{x_1!x_2!(4-x_1-x_2)!}\left(\frac{1}{4}\right)^{x_1}\left(\frac{1}{2}\right)^{x_2}\left(\frac{1}{4}\right)^{4-x_1-x_2}.$$

$\underline{x}_1$ og $\underline{x}_2$ kan som nevnt anta en hvilken som helst hel verdi fra 0 til 4, men summen av x_1 og x_2 i (4) må være mindre enn eller lik 4.

I tabell 2 nedenfor har en regnet ut funksjonsverdiene $f(x_1,x_2)$ for alle mulige kombinasjoner av x_1 og x_2.

Tabell 2. Funksjonsverdiene $f(x_1,x_2)$ regnet ut etter (4)

Verdier av $\underline{x}_1$ (dvs. antall røde dyr)	Verdier av $\underline{x}_2$ (dvs. antall skimlete dyr)					$f_1(x_1)$
	$x_2=0$	$x_2=1$	$x_2=2$	$x_2=3$	$x_2=4$	
$x_1=0$	1/256	8/256	24/256	32/256	16/256	81/256
$x_1=1$	4/256	24/256	48/256	32/256	0	108/256
$x_1=2$	6/256	24/256	24/256	0	0	54/256
$x_1=3$	4/256	8/256	0	0	0	12/256
$x_1=4$	1/256	0	0	0	0	1/256
$f_2(x_2)$	16/256	64/256	96/256	64/256	16/256	1

Sannsynlighetene inne i tabell 2 er simultane sannsynligheter. Sannsynlighetene knytter seg på vanlig måte til gjentak i et univers. I vårt spesielle tilfelle er hvert gjentak et random sampel på 4 dyr fra F_2-generasjonen. Det universet U vi betrakter består altså av en tenkt uendelighet av slike sampler. (Som vi har sett tidligere, finnes det også et univers hvor hvert gjentak er et enkelt dyr fra F_2 og hvor sannsynlighetene for rød, skimlet og hvit er henholdsvis 1/4, 1/2 og 1/4, men det er ikke dette universet som er av interesse i forbindelse med tabell 2. I og med at vi opererer med to universer er eksemplet unødig komplisert, men eksemplet har andre fordeler som gjør at vi likevel vil bruke det.)

For å forenkle språkbruken, vil vi i omtalen av tabell 2 snakke om "en besetning" i stedet for den mer presise betegnelsen "et random sampel på 4 dyr", men vær på vakt mot misforståelser!

Av tabell 2 ser vi f.eks. at sannsynligheten for at $\underline{x}_1$ skal ha verdien 3 og $\underline{x}_2$ verdien 1 er 8/256. Av alle besetningene i universet er det altså en brøkdel på 8/256 som består av 3 røde og 1 skimlet dyr.

Vi har hittil bare snakket om simultane fordelingsfunksjoner for to diskrete random variable. Overgangen fra det diskrete til det kontinuerlige tilfelle når vi har å gjøre med simultane fordelingsfunksjoner har mye til felles med den tilsvarende overgang når vi har å gjøre med den vanlige (marginale) fordelingsfunksjonen for en enkelt random variabel. Vi kan ikke her gå i detaljer, men skal ganske kort forklare litt om simultane fordelingsfunksjoner for kontinuerlige random variable. Den simultane fordelingsfunksjonen $f(x_1,x_2)$ for to kontinuerlige random variable $\underline{x}_1$ (f.eks. høyden av voksne menn) og $\underline{x}_2$ (f.eks. vekten av voksne menn) kan om den er kjent framstilles grafisk som en flate i et 3-dimensjonalt rom. Hvis vi avgrenser et sammenhengende område O i x_1x_2-planet som omfat-

ter alle de kombinasjoner av verdier av $\underline{x}_1$ og $\underline{x}_2$ som kan forekomme i universet, vil volumet over dette området, men under flaten $f(x_1,x_2)$ være lik 1. Tar vi for oss et område o innen O vil sannsynligheten for at $\underline{x}_1$ og $\underline{x}_2$ skal anta verdier som er representert ved et punkt (x_1,x_2) innen o være lik volumet over o, men under flaten $f(x_1,x_2)$. (Tegn figur selv.)

Begrepet simultan fordelingsfunksjon (for diskrete eller for kontinuerlige random variable) kan også uten prinsipielle matematiske vanskeligheter utvides til å gjelde mer enn to random variable. Vi må imidlertid da stort sett gi avkall på muligheten av å framstille fenomenet grafisk.

B. Marginale fordelingsfunksjoner

Hvis vi i eksemplet i tabell 2 bare er interessert i den random variable $\underline{x}_1$ (antall røde dyr) og ikke i $\underline{x}_2$, kan vi betrakte de marginale sannsynlighetene for $\underline{x}_1$ som vi finner i margen lengst til høyre i tabell 2. Disse er framkommet ved summasjon av hver linje i tabellen og refererer seg til samme univers som de simultane sannsynlighetene. Vi ser f.eks. at sannsynligheten for at $\underline{x}_1$ skal anta verdien 1 er lik 108/256 når vi betrakter alle besetninger under ett uansett antall skimlete dyr.

De marginale sannsynlighetene for $\underline{x}_1$ kan oppfattes som verdier av den marginale fordelingsfunksjonen $f_1(x_1)$ for $\underline{x}_1$. Som allerede antydet ved vårt eksempel, framkommer den marginale fordelingsfunksjonen for $\underline{x}_1$ ved summasjon over alle verdier av x_2 av den simultane fordelingsfunksjonen for $\underline{x}_1$ og $\underline{x}_2$. Generelt kan vi derfor skrive:

$$(5) \qquad f_1(x_1) = \sum_{x_2} f(x_1,x_2).$$

Når vi har å gjøre med kontinuerlige random variable blir summasjonen erstattet av en integrasjon.

Det kan vises at $f_1(x_1)$ i vårt spesielle tilfelle blir:

(6) $f_1(x_1) = \binom{4}{x_1}(\frac{1}{4})^{x_1}(\frac{3}{4})^{4-x_1}$

(kontroller at formelen er riktig).
På tilsvarende måte kan en vise at $f_2(x_2)$ i vårt tilfelle blir:

(7) $f_2(x_2) = \binom{4}{x_2}(\frac{1}{2})^{x_2}(\frac{1}{2})^{4-x_2} = \binom{4}{x_2}.\ (\frac{1}{2})^4 = \frac{1}{16}\binom{4}{x_2}.$

De marginale fordelingsfunksjonene $f_1(x_1)$ og $f_2(x_2)$ er vanlige fordelingsfunksjoner av den typen vi har operert med tidligere.

C. Uavhengighet mellom random variable.

Vi kan nå definere begrepet uavhengighet mellom random variable.

De to(diskrete eller kontinuerlige) random variable $\underline{x}_1$ og $\underline{x}_2$ er uavhengige hvis og bare hvis den simultane fordelingsfunksjonen for $\underline{x}_1$ og $\underline{x}_2$ er lik produktet av den marginale fordelingsfunksjonen for $\underline{x}_1$ og den marginale fordelingsfunksjonen for $\underline{x}_2$, altså hvis

(8) $f(x_1,x_2) = f_1(x_1).f_2(x_2).$

Det er lett å vise at $\underline{x}_1$ og $\underline{x}_2$ i eksemplet i tabell 2 ikke er uavhengige. Av tabellen ser vi f.eks. at $f(1,2) = 48/256 \neq 108/256 \cdot 96/256 = \frac{81}{512}$. De random variable i tabell 1, derimot, er uavhengige.

Uavhengighetskriteriet, (8) for random variable svarer til uavhengighetskriteriet for kjennetegn. Det kan også utvides til å gjelde et vilkårlig antall random variable. De n (diskrete eller kontinuerlige) random variable $x_1, x_2, \ldots, x_n$ er uavhengige hvis og bare hvis den simultane fordelingsfunksjonen $f(x_1, x_2, \ldots, x_n)$ er lik produktet av de marginale fordelingsfunksjonene $f_1(x_1), f_2(x_2), \ldots, f_n(x_n)$, altså hvis:

(9) $f(x_1, x_2, \ldots, x_n) = f_1(x_1) \cdot f_2(x_2) \ldots f_n(x_n).$

Vi kommer tilbake til uavhengighetskriteriet i avsnitt II E.

D. Teoretisk kovarians og teoretisk korrelasjonskoeffisient

Hvis vi vil finne forventningene og de teoretiske variansene for $\underline{x}_1$ og $\underline{x}_2$, går vi ut fra de marginale fordelingsfunksjonene $f_1(x_1)$ og $f_2(x_2)$ og de vanlige definisjonsformlene for forventning og varians. For eksemplet i tabell 2 får vi da (se tabellen):

$$(10)\quad E(\underline{x}_1) = \sum_{x_1} f_1(x_1)x_1 = 0.84/256 + 1.108/256 + 2.54/256 + 3.12/256 + 4.1/256 = 1.$$

$$(11)\quad E(\underline{x}_2) = \sum_{x_2} f_2(x_2)x_2 = 2.$$

$$(12)\quad \text{var}(\underline{x}_1) \sum_{x_1} f_1(x_1)(x_1 - E(x_1)) = 3/4 \quad \text{dvs.} \quad \sigma_1 = \sqrt{3/4}.$$

$$(13)\quad \text{var}(\underline{x}_2) = \sum_{x_2} f_2(x_2)(x_2 - E(x_2)) = 1 \quad \text{dvs.} \quad \sigma_2 = 1.$$

Den teoretiske kovariansen, σ_{12} mellom to diskrete random variable $\underline{x}_1$ og $\underline{x}_2$ kan defineres på følgende måte:

$$(14)\quad \sigma_{12} = \sum_{x_1} \sum_{x_2} f(x_1, x_2)(x_1 - E(\underline{x}_1))(x_2 - E(\underline{x}_2)).$$

Hvis $\underline{x}_1$ og $\underline{x}_2$ er kontinuerlige random variable blir dobbeltsummasjonen i (14) erstattet av en dobbeltintegrasjon.

Vi ser at $\underline{x}_1$ og $\underline{x}_2$ inngår på en helt symmetrisk måte i (14). Følgelig er kovariansen mellom $\underline{x}_1$ og $\underline{x}_2$ lik kovariansen mellom $\underline{x}_2$ og $\underline{x}_1$, dvs. $\sigma_{12} = \sigma_{21}$.

Utregningen av σ_{12} for eksemplet i tabell 2 er vist i tabell 3. Den første faktoren i hver celle i tabellen er sannsynligheten som knytter seg til vedkommende kombinasjon av $\underline{x}_1$-verdi og $\underline{x}_2$-verdi. Den neste faktoren er $x_1 - E(\underline{x}_1)$ og den siste faktoren er $x_2 - E(\underline{x}_2)$.

Tabell 3. Utregning av kovariansen mellom $\underline{x}_1$ og $\underline{x}_2$ for eksemplet i tabell 2.

x_1	x_2				
	0	1	2	3	4
0	$\frac{1}{256}.(-1).(-2)$	$\frac{8}{256}.(-1).(-1)$	$\frac{24}{256}.(-1).0$	$\frac{32}{256}.(-1).1$	$\frac{16}{256}.(-1).2$
1	$\frac{4}{256}.0.(-2)$	$\frac{24}{256}.0.(-1)$	$\frac{48}{256}.0.0$	$\frac{32}{256}.0.1$	
2	$\frac{6}{256}.1.(-2)$	$\frac{24}{256}.1.(-1)$	$\frac{24}{256}.1.0$		
3	$\frac{4}{256}.2.(-2)$	$\frac{8}{256}.2.(-1)$			
4	$\frac{1}{256}.3.(-2)$				

$\sigma_{12}=-\frac{1}{2}$

Produktet av disse tre faktorene summert over alle cellene i tabellen er lik kovariansen som i vårt tilfelle blir lik $-\frac{1}{2}$.

Kovariansen er et mål for samvariasjonen mellom $\underline{x}_1$ og $\underline{x}_2$. I celler hvor både x_1 er stor i forhold til $E(\underline{x}_1)$ og samtidig $\underline{x}_2$ er stor i forhold til $E(\underline{x}_2)$ vil produktet $(x_1-E(x_1))(x_2-E(x_2))$ være positivt. Dette produktet vil også bli positivt i celler hvor x_1 er liten i forhold til $E(\underline{x}_1)$ og samtidig x_2 er liten i forhold til $E(\underline{x}_2)$ siden begge faktorene vil være negative i slike celler. Hvis celler av de to typer vi nå har nevnt har store sannsynligheter vil dette bidra til å gjøre kovariansen positiv. At slike celler har stor sannsynlighet betyr imidlertid at store $\underline{x}_1$-verdier har en tendens til å opptre sammen med store $\underline{x}_2$-verdier og at små $\underline{x}_1$-verdier fortrinnsvis opptrer sammen med små $\underline{x}_2$-verdier. Vi ser altså at en positiv kovarians gir uttrykk for en positiv samvariasjon mellom $\underline{x}_1$ og $\underline{x}_2$. På liknende måte kan vi resonnere oss til at en negativ samvariasjon mellom $\underline{x}_1$ og $\underline{x}_2$ fører til at kovariansen blir negativ. (En negativ samvariasjon vil si at store x_1 helst forekommer sammen med små x_2 og at små x_1 oftest opptrer sammen med store x_2.)

Som et mål for graden av samvariasjonen mellom to random variable $\underline{x}_1$ og $\underline{x}_2$ kan en i stedet for den teoretiske kovariansen σ_{12} bruke den teoretiske korrelasjonskoeffisienten ρ_{12} som er lik ρ_{21}. Denne har den fordelen at den er en ubenevnt størrelse. For ρ_{12} gjelder alltid følgende ulikhet:

$$(15) \quad -1 \leq \rho_{12} \leq 1.$$

Betegner vi standardavviket for $\underline{x}_1$ (i den marginale fordelingsfunksjonen for $\underline{x}_1$) med σ_1 og standardavviket for $\underline{x}_2$ (i den marginale fordelingsfunksjonen for $\underline{x}_2$) med σ_2, kan ρ_{12} defineres på følgende måte:

$$(16) \quad \rho_{12} = \frac{\sigma_{12}}{\sigma_1 \cdot \sigma_2}.$$

Definisjonen gjelder både for det diskrete og det kontinuerlige tilfelle. For eksemplet i tabell 2 får vi (se (12), (13) og tabell 3):

$$(17) \quad \rho_{12} = \frac{-0.5}{\sqrt{\frac{3}{4} \cdot 1}} = -\frac{1}{\sqrt{3}} = -0.58.$$

At den teoretiske korrelasjonskoeffisienten mellom $\underline{x}_1$ og $\underline{x}_2$ i vårt eksempel er negativ vil altså si at når det er mange røde dyr i en "besetning" så er det som regel få skimlete, og omvendt. Dette er slik som vi skulle vente det.

Hvis to random variable $\underline{x}_1$ og $\underline{x}_2$ er uavhengige, kan det vises at kovariansen $\sigma_{12} = 0$, og dermed også at korrelasjonskoeffisienten $\rho_{12} = 0$. Vi sier da at $\underline{x}_1$ og $\underline{x}_2$ er ukorrelerte. Hvis $\rho_{12} \neq 0$ sier vi at $\underline{x}_1$ og $\underline{x}_2$ er korrelerte. (For ordens skyld gjør vi oppmerksom på at to random variable kan være ukorrelerte uten å være uavhengige.)

E. Betingede fordelingsfunksjoner

Vi har nå gjennomgått litt om simultane og marginale fordelingsfunksjoner. En tredje type fordelingsfunksjoner som det er naturlig å nevne i denne sammenheng er betingede fordelingsfunksjoner. La oss igjen ty til vårt eksempel i tabell 2. Sett at vi bare betrakter "besetninger" (dvs. random sampler på 4 dyr av F_2) hvor antall skimlete dyr ($\underline{x}_2$) er lik 1. Vi kunne spørre hvorledes disse besetningene fordeler seg med hensyn til antall røde dyr. Det er klart at vi da må betrakte den kolonnen i tabell 2 hvor $x_2 = 1$. Sannsynlighetene i denne kolonnen gir oss forsåvidt den søkte fordelingen, men summen av disse sannsynlighetene er 64/256 og ikke 1 som vi ville foretrekke i denne sammenheng. Hvis vi imidlertid dividerer hvert tall i denne kolonnen med summen av tallene i kolonnen som er lik 64/256, får vi et nytt sett av sannsynligheter som vi kaller betingede sannsynligheter, nemlig 8/64, 24/64, 24/64, 8/64 og 0. Summen av disse sannsynlighetene er, som vi ser lik 1. Den betingede fordelingsfunksjonen for $\underline{x}_1$ betinget av at $\underline{x}_2$ antar verdien 1, kan skrives på følgende måte:

$$(18)\quad g_1(x_1 \mid \underline{x}_2 = 1) = \frac{f(x_1, 1)}{f_2(1)} = \frac{\frac{4!}{x_1!1!(4-x_1-1)!}\left(\frac{1}{4}\right)^{x_1}\left(\frac{1}{2}\right)\left(\frac{1}{4}\right)^{4-x_1-1}}{\frac{1}{16}\binom{4}{1}}$$

Funksjonene f og f_2 er her de samme som (4) og (7).
Hvis vi i (18) setter inn etter tur verdiene 0, 1, 2 og 3 for x_1, finner vi sannsynlighetene 8/64, 24/64, 24/64 og 8/64.

Helt generelt definerer vi den betingede fordelingsfunksjonen for $\underline{x}_1$ betinget av at $\underline{x}_2$ antar verdien x_2 på følgende måte:

$$(19)\quad g_1(x_1 \mid \underline{x}_2 = x_2) = \frac{f(x_1, x_2)}{f_2(x_2)} \qquad (f_2(x_2) > 0)$$

Definisjonen gjelder både for det diskrete og for det kontinuerlige tilfelle. Sannsynlighetene $g_1(x_1 \mid \underline{x}_2 = x_2)$ refererer seg til

et subunivers hvor verdien av $\underline{x}_2$ er lik et bestemt (oppgitt) tall x_2 for alle gjentak. Det finnes en betinget fordelingsfunksjon for $\underline{x}_1$ for hver verdi $\underline{x}_2$ kan anta. For vårt eksempel i tabell 2 har vi altså 5 betingede fordelingsfunksjoner for x_1. På tilsvarende måte har vi for vårt eksempel 5 betingede fordelingsfunksjoner for $\underline{x}_2$.

Likningen (19) kan omskrives på følgende måte:

$$(20) \quad f(x_1,x_2) = f_2(x_2).g_1(x_1 \mid \underline{x}_2 = x_2).$$

(20) svarer til både... og setningen i sannsynlighetsregningen som er gitt ved (21) nedenfor:

$$(21) \quad P(E_1E_2 \mid U) = P(E_2 \mid U).\ P(E_1 \mid UE_2).$$

Uavhengighetskriteriene for kjennetegn og for random variable svarer også til hverandre. Hvis E_1 og E_2 er uavhengige kjennetegn, har vi at $P(E_1E_2 \mid U) = P(E_2 \mid U).\ P(E_1 \mid U)$. Hvis de random variable $\underline{x}_1$ og $\underline{x}_2$ er uavhengige er (8) oppfylt. I dette tilfelle er altså $g_1(x_1 \mid \underline{x}_2 = x_2) = f_1(x_1)$. Alle de betingede fordelingsfunksjonene for $\underline{x}_1$ er da like og lik den marginale fordelingsfunksjonen for $\underline{x}_1$. En slik situasjon har vi for eksemplet i tabell 1.

Hvis vi betrakter en betinget fordelingsfunksjon for $\underline{x}_1$, f.eks. fordelingsfunksjonen (18), har vi på vanlig måte en forventning for $\underline{x}_1$ som vi nå kaller en betinget forventning, og en varians for $\underline{x}_1$ som vi kaller en betinget varians.

F. Regresjonsfunksjonene i universet

1. Regresjonsfunksjonen for $\underline{x}_1$ med hensyn på $\underline{x}_2$

I tabell 4 nedenfor har en på grunnlag av tabell 2 tabulert 5 forskjellige betingede fordelingsfunksjonene for $\underline{x}_1$ i 5 kolonner. Under hver kolonne har en ført opp de betingede forventningene for $\underline{x}_1$. Disse vil vi betegne med $E(\underline{x}_1 \mid \underline{x}_2=x_2$ el-

ler $E(\underline{x}_1 | Ux_2)$. For hver verdi x_2 av $\underline{x}_2$ har vi nemlig et subunivers av "besetninger". Den tilsvarende betingede fordelingsfunksjonen for $\underline{x}_1$ angir fordelingen etter x_1 innen dette subuniverset.

Tabell 4. De betingede fordelingsfunksjoner og de betingede forventninger for $\underline{x}_1$ (eksemplet i tabell 2)

Verdier av $\underline{x}_1$ (Dvs. antall røde dyr)	Verdier av $\underline{x}_2$ (Dvs. antall skimlete dyr)				
	$x_2 = 0$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$
$x_1 = 0$	1/16	8/64	24/96	32/64	1
$x_1 = 1$	4/16	24/64	48/96	32/64	
$x_1 = 2$	6/16	24/64	24/96	0	
$x_1 = 3$	4/16	8/64	0	0	
$x_1 = 4$	1/16	0	0	0	
Sum	1	1	1	1	1
$E(x_1 \mid Ux_2)$	2	1,5	1	0.5	0

De betingede forventningene for $\underline{x}_1$ kan uttrykkes som en funksjon av x_2. Denne funksjonen kalles regresjonsfunksjonen for $\underline{x}_1$ med hensyn på $\underline{x}_2$. I vårt tilfelle ser vi uten videre av tabell 4 at regresjonsfunksjonen er lineær. $E(\underline{x}_1 | Ux_2)$ synker nemlig med 1/2 enhet for hver enhet x_2 stiger.

Når regresjonsfunksjonen for en random variabel $\underline{x}_1$ med hensyn på en random variabel $\underline{x}_2$ er lineær, kan den i alminnelighet skrives på følgende måte:

$$(22) \quad E(\underline{x}_1 | Ux_2) = E(\underline{x}_1) + \beta_{12} (x_2 - E(\underline{x}_2)).$$

Det som her er sagt gjelder også for kontinuerlige random variable. Den betingede forventningen $E(\underline{x}_1 | Ux_2)$ er altså lik (den marginale) forventningen $E(\underline{x}_1)$ pluss et ledd som avhenger av to ting, nemlig av hvor mye x_2 avviker fra $E(\underline{x}_2)$ og av en konstant koeffisient β_{12} som blir kalt regresjonskoeffisienten for $\underline{x}_1$ med hensyn på $\underline{x}_2$. Regresjonskoeffisienten β_{12} kan generelt

beregnes etter følgende formel:

$$(23)\quad \beta_{12} = \frac{\sigma_{12}}{\sigma_2^2}.$$

For vårt eksempel får vi da (se (13) og tabell 3):

$$(24)\quad \beta_{12} = \frac{\sigma_{12}}{\sigma_2^2} = \frac{0,5}{1} = -0,5.$$

Regresjonsfunksjonen (22) kan derfor skrives på følgende måte (se (10), (24) og (11)):

$$(25)\quad E(\underline{x}_1 | Ux_2) = 1 - 0,5(x_2 - 2)$$

eller

$$(26)\quad E(\underline{x}_1 | Ux_2) = 2 - 0,5x_2.$$

Ved å sette inn etter tur $x_2 = 0$, $x_2 = 1$, $x_2 = 2$ osv. i (26), får en de betingede forventningene som er gjengitt nederst i tabell 4.

2. Regresjonsfunksjonen for $\underline{x}_2$ med hensyn på $\underline{x}_1$

Tabell 5 nedenfor svarer til tabell 4, men $\underline{x}_1$ og $\underline{x}_2$ har her byttet roller. I tabell 5 har en tabulert de 5 betingede fordelingsfunksjonene for $\underline{x}_2$ på 5 linjer. Til høyre på hver linje har en ført opp de betingede forventningene $E(x_2 | Ux_1)$ for x_1.

Tabell 5. De betingede fordelingsfunksjoner og de betingede forventninger for $\underline{x}_2$ (eksemplet i tabell 2).

Verdier av $\underline{x}_1$ (Dvs. antall røde dyr)	Verdier av $\underline{x}_2$ (dvs. antall skimlete dyr)					Sum	$E(x_2 \mid Ux_1)$
	$x_2 = 0$	$x_2 = 1$	$x_2 = 2$	$x_2 = 3$	$x_2 = 4$		
$x_1 = 0$	1/81	8/81	24/81	32/81	16/81	1	2,67
$x_1 = 1$	4/108	24/108	48/108	32/108	0	1	2,00
$x_1 = 2$	6/54	24/54	24/54	0	0	1	1,33
$x_1 = 3$	4/12	8/12	0	0	0	1	0,66
$x_1 = 4$	1	0	0	0	0	1	0

Vi får ikke egentlig fram noe nytt ved å snu om problemet på denne måten, men vi skal likevel utføre de tilsvarende beregningene som ovenfor.

Av tabell 5 ser vi at også regresjonsfunksjonen for $\underline{x}_2$ med hensyn på $\underline{x}_1$ er lineær. Regresjonskoeffisienten β_{21} for $\underline{x}_2$ med hensyn på $\underline{x}_1$ finner vi av (23) idet vi bytter om indeksene 1 og 2. Vi får da (se (12) og tabell 3):

$$(27)\quad \beta_{21} = \frac{\sigma_{12}}{\sigma_1^2} = \frac{-0,5}{3/4} = -\frac{2}{3} = -0,67.$$

Regresjonsfunksjonen for $\underline{x}_2$ med hensyn på $\underline{x}_1$ kan da skrives slik:

$$(28)\quad E(\underline{x}_2|Ux_1) = E(\underline{x}_2) + \beta_{21}(x_1 - E(\underline{x}_1))$$

eller

$$(29)\quad E(\underline{x}_2|Ux_1) = 2 - 0,67(x_1 - 1)$$

eller

$$(30)\quad E(\underline{x}_2|Ux_1) = 2,67 - 0,67x_1.$$

3. Størrelsen av korrelasjonskoeffisienten og regresjonskoeffisientene. Sammenhengen mellom regresjonslinjene

Vi ser av (23) og (27) at både β_{12} og β_{21} blir lik 0 når kovariansen σ_{12} er lik 0, dvs. når korrelasjonskoeffisienten ρ_{12} er lik 0.

I fig. 2 nedenfor har en framstilt de to regresjonsfunksjonene (26) og (30) grafisk i samme aksesystem.

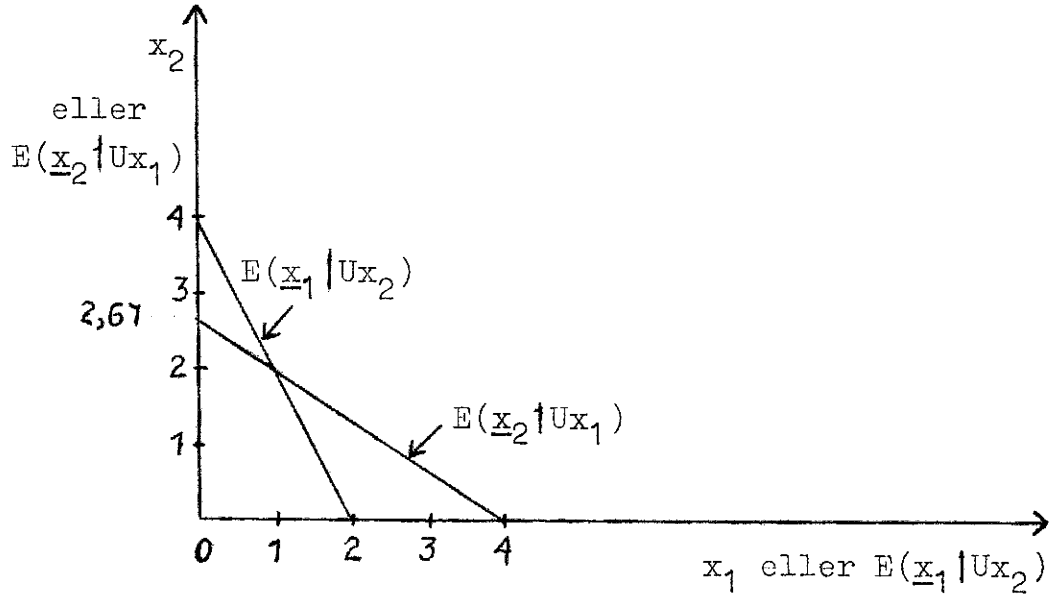


Fig. 2.

Av figuren ser vi bl.a. at de to regresjonslinjene skjærer hverandre i punktet (1,2), altså i punktet $(E(\underline{x}_1), E(\underline{x}_2))$. (Vis ved hjelp av (22) og (28) at skjæringspunktet er nettopp dette punktet.)

Hvis $\underline{x}_1$ og $\underline{x}_2$ er ukorrelerte er som nevnt både β_{21} og β_{12} lik 0. De to regresjonslinjene står da loddrett på hverandre. Regresjonslinjen for $\underline{x}_1$ med hensyn på $\underline{x}_2$ er parallell med x_2-aksen og går gjennom punktet $(E(\underline{x}_1),0)$ mens regresjonslinjen for $\underline{x}_2$ med hensyn på $\underline{x}_1$ er parallell med x_1-aksen og går gjennom punktet $(0, E(\underline{x}_2))$.

Vi skal ta for oss et eksempel hvor de to regresjonslinjene faller sammen. Sett at sannsynligheten for "oksekalv" ved en storfefødsel er 0,52 og at sannsynligheten for "kvigekalv" er 0,48. Vi vil betrakte et univers hvor hvert gjentak er et random sampler på 4 storfefødsler. La den random variable $\underline{x}_1$ være antall oksekalver i et slikt sampel. $\underline{x}_1$ kan altså ha verdiene 0, 1, 2, 3 eller 4. Fordelingsfunksjonen for $\underline{x}_1$ er da en binomial fordelingsfunksjon:

$$(31)\quad f_1(x_1) = \binom{4}{x_1} 0{,}52^{x_1} 0{,}48^{4-x_1}$$

med forventning $E(\underline{x}_1) = 4.0{,}52 = 2{,}08$ og varians $\text{var}(\underline{x}_1) = 4.0{,}52.0{,}48 = 0{,}9984$.
Fordelingsfunksjonen (31) kan også skrives som en simultan fordelingsfunksjon for to random variable $\underline{x}_1$ og $\underline{x}_2$ hvor $\underline{x}_2 = 4-\underline{x}_1$ = antall kvigekalver. Den random variable $\underline{x}_2$ er altså her en eksakt lineær funksjon av $\underline{x}_1$. Den simultane fordelingsfunksjonen for $\underline{x}_1$ og $\underline{x}_2$ blir:

$$(32)\quad f(x_1,x_2) = \binom{4}{x_1} 0{,}52^{x_1} 0{,}48^{x_2} \qquad (x_1 + x_2 = 4).$$

Funksjonen (31) er den marginale fordelingsfunksjonen for $\underline{x}_1$. Den marginale fordelingsfunksjonen for $\underline{x}_2$ er gitt ved (33).

$$(33)\quad f_2(x_2) = \binom{4}{x_2} 0{,}48^{x_2} 0{,}52^{4-x_2}.$$

Forventningen for $\underline{x}_2$ er $\text{var}(\underline{x}_2) = 4.0{,}48.0{,}52 = 0{,}9984$. Den simultane fordelingsfunksjonen (32) er tabulert i tabell 6 nedenfor.

Tabell 6. Funksjonsverdiene $f(x_1,x_2)$ regnet ut etter (27)

Verdier av $\underline{x}_1$ (Dvs. antall oksekalver	Verdier av $\underline{x}_2$ (Dvs. antall kvigekalver					
	0	1	2	3	4	Sum
0	0	0	0	0	0,0531	0,0531
1	0	0	0	0,2300	0	0,2300
2	0	0	0,3738	0	0	0,3738
3	0	0,2700	0	0	0	0,2700
4	0,0731	0	0	0	0	0,0731
Sum	0,0731	0,2700	0,3738	0,2300	0,0531	1,0000

Ved hjelp av formelen (10) s. 8, finner vi at kovariansen mellom $\underline{x}_1$ og $\underline{x}_2$ i dette eksemplet blir:

(34) $\sigma_{12} = 0,0731(4-2,08)(0-1,92) + 0,2700(3-2,08)(1-1,92)$
$+ 0,3738(2-2,08)(2-1,92) + 0,2300(1-2,08)(3-1,92)$
$+ 0,0531(0-2,08)(4-1,92) = -0,2695 - 0,2285 - 0,0024$
$-0,2683 - 0,2297 = -0,9984.$

Korrelasjonskoeffisienten ρ_{12} finner vi da ved hjelp av (16) s. 10:

(35) $\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \cdot \sigma_2} = \frac{-0,9984}{0,9984} = -1.$

Regresjonskoeffisienten β_{12} blir i følge formelen (23) s. 14:

(36) $\beta_{12} = \frac{\sigma_{12}}{\sigma_2^2} = \frac{-0,9984}{0,9984} = -1.$

Regresjonsfunksjonen for $\underline{x}_1$ med hensyn på $\underline{x}_2$ blir da (se (22) s. 13):

(37) $E(\underline{x}_1 | Ux_2) = 2,08-1(x_2-1,92)$

eller

(38) $E(\underline{x}_1 | Ux_2) = 4-x_2$

som er et nokså selvfølgelig resultat.
På tilsvarende måte kan vi finne regresjonsfunksjonen for $\underline{x}_2$ med hensyn på $\underline{x}_1$. Vi får i følge (27) s. 15:

(39) $\beta_{21} = \frac{\sigma_{12}}{\sigma_1^2} = \frac{-0,9984}{0,9984} = -1.$

Ved å bruke (28) s. 15 , får vi da:

(40) $E(\underline{x}_2 | Ux_1) = 1,92-1(x_1-2,08)$

eller

(41) $E(\underline{x}_2 | Ux_1) = 4-x_1.$

Som vi ser, faller de to regresjonslinjene (38) og (41) sammen. (Tegn diagram.) I alle tilfelle hvor $\rho_{12} = 1$ faller de to regresjonslinjene sammen. For vårt eksempel ser vi av tabell 6 at alle sammenhørende observasjoner av x_1 og x_2 kan avmerkes som et punkt på regresjonslinjen. Observasjonspar som ikke representerer et punkt på regresjonslinjen (38) (eller (41)) har sannsynlighet 0 og vil altså ikke forekomme.

I et tilfelle som det vi behandlet i vårt siste eksempel er vi vanligvis ikke interessert i å operere med en simultan fordelingsfunksjon. Den ene random variable er jo en eksakt funksjon av den andre (funksjonens form er den samme som den tilsvarende regresjonsfunksjonen) og det er da mer naturlig å operere med en enkelt random variabel. Vi kan imidlertid i slike tilfelle av og til være interessert i å finne formler for forventningen eller variansen for den ene random variable når forventningen og variansen for den andre er gitt. Dette skal vi komme tilbake til senere.

I det foregående har vi for det meste brukt diskrete random variable som eksempler. Teorien er i prinsippet den samme når vi har å gjøre med kontinuerlige random variable. I praktiske korrelasjons- og regresjonsanalyser er det kontinuerlige random variable vi får å gjøre med i de fleste sammenhenger. Vi har brukt diskrete random variable i eksemplene ovenfor fordi det er enklere å illustrere teorien på denne måten.

I praksis er det sjelden at vi kjenner verdiene av μ_1, μ_2, σ_1, σ_2 og σ_{12}. Vi har derfor sjelden anledning til å beregne størrelsene ρ_{12}, β_{12} og β_{21} slik som vi har gjort i dette hovedavsnittet. Derimot er vi ofte interessert i å estimere disse størrelsene og å teste hypoteser om dem. Dette skal vi behandle i detalj i hovedavsnitt IV.

III. Forventning og varians for en funksjon av random variable

A. En lineær funksjon av en enkelt random variabel

1. Definisjon og eksempler

La oss ta for oss to (kontinuerlige eller diskrete) random variable $\underline{x}$ og $\underline{y}$. Vi vil tenke oss at det til hver verdi x av den random variable $\underline{x}$ svarer en enkelt verdi y av den random variable $\underline{y}$ og at hver y er en lineær funksjon av den tilsvarende x. I slike tilfelle kan vi skrive:

(42) $y = a + bx$

hvor a og b er konstante koeffisienter.
Vi vil også bruke skrivemåten

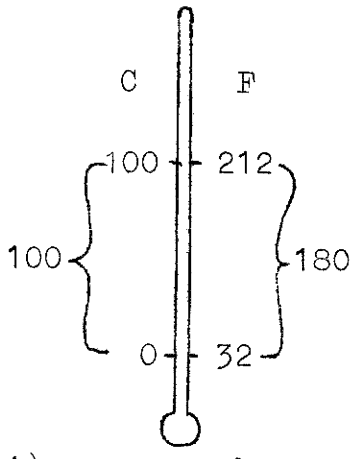
(43) $\underline{y} = a + b\underline{x}$.

Vi sier da at den random variable $\underline{y}$ er en lineær funksjon av den random variable $\underline{x}$, og med skrivemåten (43) mener vi at (42) gjelder for alle verdier x som $\underline{x}$ kan anta. Likningen (43) blir også ofte kalt en transformasjonslikning, og vi sier da at vi har foretatt en transformasjon (i vårt tilfelle en lineær transformasjon) fra den random variable $\underline{x}$ til den random variable $\underline{y}$.

La oss forsøke å forklare dette ved noen eksempler.

Eks. 1. La $\underline{x}$ være temperaturen målt i ^{o}C et sted i Ås den 1. januar kl. 12.00. Hvert år kan vi skaffe oss en ny observasjon, x av denne random variable. Anta at vi kan tenke oss et univers av år hvor $\underline{x}$ har en bestemt fordelingsfunksjon, f(x) med $E(\underline{x}) = \mu_{\underline{x}}$ og $\text{var}(\underline{x}) = \sigma^2_{\underline{x}}$.

Sett at det termometeret vi brukte hadde to skalaer, én med Celsiusgrader og én med Farenheitgrader. Hvis vi et år målte temperaturen x på Celsiusskalaen ville vi da måle temperaturen



$y = 32 + \frac{180}{100} x = 32 + \frac{9}{5} x$ på Fahrenheitskalaen. Dette ville gjelde for alle x. Vi kan derfor si at $\underline{y}$ er en ny random variabel som er en lineær funksjon av $\underline{x}$, og vi kan skrive:

(44) $\underline{y} = a + b\,\underline{x}$ eller $\underline{y} = 32 + \frac{9}{5}\,\underline{x}$

Det er lett å innse at også $\underline{y}$ kan tenkes å ha en eller annen fordelingsfunksjon $f_y(y)$ med en eller annen forventning $E(\underline{y}) = \mu_{\underline{y}}$ og en eller annen varians $\text{var}(y) = \sigma^2_{\underline{y}}$

Eks. 2. Sett at vi tar for oss et univers som består av alle skogsarbeidere i USA. Disse arbeideres årlige inntekt (målt i dollar) kan betraktes som en random variabel $\underline{x}$. Det er selvsagt ikke noe i veien for å måle inntekten i kroner i stedet for i dollar. Dette ville f.eks. være naturlig om vi ville foreta en sammenlikning med skogsarbeidere i Norge. Amerikanske skogsarbeideres årsinntekt målt i kroner er en random variabel som vi vil betegne med $\underline{y}$. Vi har da

(45) $\underline{y} = a + b\underline{x}$ eller hvis \$1,00 = kr. 5,00: $\underline{y} = 0 + 5\underline{x} = 5\underline{x}$

Også i dette tilfelle har $\underline{y}$ en eller annen fordelingsfunksjon som er karakterisert ved en bestemt forventning og en bestemt varians.

Eks. 3. Sett at vi tar for oss et univers hvor gjentakene er 5-barnsfamilier og betrakter en random variabel $\underline{x}$ som er antall gutter. I hver familie kan vi si at antall gutter utgjør en viss prosent av antall barn. Hvis f.eks. antall gutter i en besemt familie er x, vil prosent gutter, y være

(46) $y = \frac{x.100}{5} = 20x$

Dette gjelder for alle x. Hvis vi i dette universet be-

trakter det prosentiske antall gutter, kan dette prosentiske antall oppfattes som en random variabel $\underline{y}$ som er en lineær funksjon av den random variable $\underline{x}$.

(47) $\underline{y} = a + b\underline{x}$ eller $\underline{y} = 20\underline{x}$

Eks. 4. La oss igjen betrakte universet av 5-barnsfamilier i eks. 3 hvor $\underline{x}$ var antall gutter. I stedet for å la $\underline{y}$ være det prosentiske antall gutter kunne vi også la $\underline{y}$ være antall jenter. Igjen ville $\underline{y}$ bli en lineær funksjon av $\underline{x}$, nemlig

(48) $\underline{y} = a + b\underline{x}$ eller $\underline{y} = 5-\underline{x}$

2. Formler

I alle eksemplene ovenfor er det lett å forestille seg at $\underline{y}$ har en eller annen fordelingsfunksjon $f_y(y)$. Et problem som en ofte står overfor i statistikken i liknende tilfelle er å utlede fordelingsfunksjonen for $\underline{y}$. Vi skal ikke gå inn på dette problemet. Vi skal her bare presentere noen formler som kan brukes til å løse den mer begrensede oppgaven å finne forventningen og variansen for $\underline{y}$ direkte uten å gå veien om fordelingsfunksjonen for $\underline{y}$ og om definisjonsformlene for forventning og varians.

Nedenfor har en gjengitt noen viktige setninger som gjelder generelt for situasjoner av den typen som er beskrevet i eksemplene.

Setning 1. La $\underline{x}$ være en random variabel med forventning $E(\underline{x})=\mu_{\underline{x}}$ og teoretisk varians $\text{var}(\underline{x}) = \sigma^2_{\underline{x}}$ og la $\underline{y}$ være en random variabel som er en lineær funksjon av $\underline{x}$:

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

La oss betegne forventningen og den teoretiske variansen for $\underline{y}$ med $E(\underline{y}) = \mu_{\underline{y}}$ og $\text{var}(\underline{y}) = \sigma^2_{\underline{y}}$. Det kan da bevises at følgende likheter gjelder generelt:

(50) $\mu_{\underline{y}} = a + b\mu_{\underline{x}}$

(51) $\sigma^2_{\underline{y}} = b^2 \sigma^2_{\underline{x}}$ dvs. $\sigma_y = b\,\sigma_{\underline{x}}$

(50) kan vi uttrykke i ord på følgende måte: Når $\underline{y}$ er en lineær funksjon av $\underline{x}$, er μ_y den samme lineære funksjon av μ_x.

Vi skal demonstrere bruken av (50) og (51) ved hjelp av eksempel 1 ovenfor. Ved å sammenlikne (49) med (44), ser vi at a i dette eksemplet er 32 og at b er 9/5. Sett at vi vet at $E(\underline{x}) = 15^oC$ og at $var(\underline{x}) = 25$. Hvis vi vil finne $E(\underline{y})$ og $var(\underline{y})$, kan dette gjøres på følgende enkle måte ved hjelp av (50) og (51):

(52) $E(\underline{y}) = a + b\,E(\underline{x}) = 32 + \frac{9}{5} \cdot (-15) = 5$

(53) $Var(\underline{y}) = b^2\, var(\underline{x}) = (\frac{9}{5})^2 \cdot 25 = 81$

På tilsvarende måte kan vi finne $E(\underline{y})$ og $var(\underline{y})$ for de andre eksemplene hvis vi kjenner $E(\underline{x})$ og $var(\underline{x})$. Det eneste vi trenger i tillegg er de numeriske verdiene av a og b, og disse finner vi ved å se på den transformasjonslikningen som gjelder i hvert enkelt tilfelle.

Formlene (50) og (51) er meget nyttige, og vi kommer til å referere til dem i det følgende. Vi har også tilsvarende formler for gjennomsnittet og den empiriske variansen.

3. Tilsvarende formler for sampelstørrelser

I eksemplene ovenfor ser vi at $\underline{y}$ og $\underline{x}$ refererer seg til samme univers, og videre at y og x knytter seg til samme gjentak. Hvis vi tar et random sampel på n gjentak fra det felles univers og observerer x og y, får vi altså n observasjoner av $\underline{x}$ og n observasjoner av $\underline{y}$.

Disse observasjonene er da selvsagt knyttet sammen ved følgende likning:

(54) $y_i = a + bx_i$ $\qquad (i = 1,2,\ldots,n)$

Hvis vi kjenner gjennomsnittet ($\bar{x}$) og middelavviket (s_x) for x_i i dette samplet, kan vi finne gjennomsnittet ($\bar{y}$) og middelavviket (s_y) for y_i ved hjelp av likninger som tilsvarer (50) og (51) uten å gå veien om de enkelte observasjonene y_i av $\underline{y}$. Vi skal bevise dette. Vi har

$$(55)\quad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{\sum_{i=1}^{n} (a + bx_i)}{n} = \frac{n\,a + b\sum_{i=1}^{n} x_i}{n} = a + b\,\bar{x}$$

og

$$(56)\quad s_y^2 = \frac{\sum_{i=1}^{n} (y_i-\bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} (a+bx_i-a-b\bar{x})^2}{n-1} = \frac{b^2\sum_{i=1}^{n} (x_i-\bar{x})^2}{n-1} = b^2 s_x^2$$

Det første resultatet (55) kan vi uttrykke i ord på følgende måte: <u>Når y_i er en lineær funksjon (54) av x_i, er $\bar{y}$ den samme lineære funksjon av $\bar{x}$.</u>

Formlene (50)-(51) og (55)-(56) svarer til hverandre. De første har å gjøre med universet, mens de siste gjelder et sampel. Forutsetningen for å bruke (50)-(51) er at (49) gjelder, mens forutsetningen for å bruke (55)-(56) er at (54) gjelder.

4. Mer om transformasjoner

For å kaste mer lys over det som skjer ved en transformasjon av typen (49) skal vi se litt nærmere på eksempel 1 ovenfor. La oss anta som før at $\mu_x = 15^{\circ}C$ og at $\sigma_x = 5^{\circ}C$. Da finner vi som vist tidligere at $\mu_y = 5^{\circ}F$ og at $\sigma_y = 9^{\circ}F$.

Til en videre illustrasjon vil vi tenke oss at $\bar{x}$ har normal fordelingsfunksjon. Siden $\underline{y}$ er en lineær funksjon av $\underline{x}$, kan det da bevises at også $\underline{y}$ har normal fordelingsfunksjon (beviset tas ikke med her). Fordelingsfunksjonene for $\underline{x}$ og $\underline{y}$ finnes da uten videre ved innsetting i formelen som uttrykker den normale fordelingsfunksjon. Fordelingsfunksjonene er følgende:

$$(57)\quad f(x) = \frac{1}{5\sqrt{2\pi}}\, e^{-\frac{(x+15)^2}{2.25}} \qquad f(y) = \frac{1}{9\sqrt{2\pi}}\, e^{-\frac{(y-5)^2}{2.81}}$$

De to fordelingsfunksjonene er framstilt grafisk i fig. 3 nedenfor.

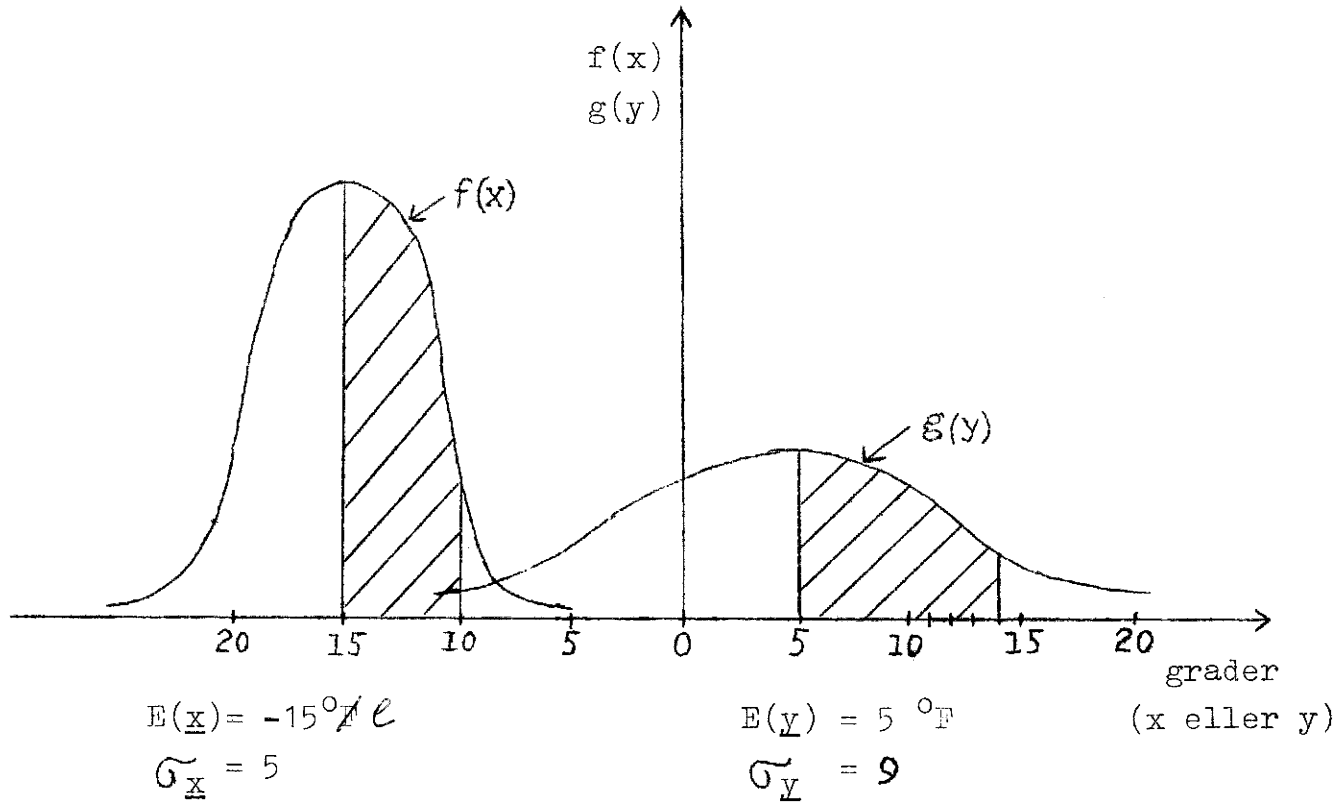


Fig. 3.

Øvelse 1. Forklar hvorfor $P(-15 < \underline{x} < -10) = P(5 < \underline{y} < 14)$, dvs. forklar hvorfor de to observerte arealene i fig. 3 må være like store.

Øvelse 2. En random variabel $\underline{x}$ har forventning μ og standardavvik σ. Finn forventningen og standardavviket for $\underline{y}$ når $\underline{y}$ er definert ved følgende formel: $\underline{y} = \frac{\underline{x}-\mu}{\sigma}$.

Øvelse 3. Vis at $\frac{Z}{N_{\mu}}$ (se MATEMATISK STATISTIKK, Vollebekk 1969, s. 102) har forventningen p og standardavviket $\sqrt{\frac{pq}{n}}$.

B. En vilkårlig funksjon av en random variabel *

La $\underline{x}$ være en random variabel med fordelingsfunksjonen f(x) og la $\underline{y} = h(\underline{x})$ være en funksjon av $\underline{x}$. Det kan da bevises at forventningen for $\underline{y}$ (dvs. forventningen for h(x)) er

(58) $E(h(\underline{x})) = \sum_x h(x)f(x)$ hvis $\underline{x}$ er diskret

og

(59) $E(h(\underline{x}) = \int_S h(x)f(x)dx$ hvis $\underline{x}$ er kontinuerlig.

Vi skal ikke gjøre noe forsøk på å bevise (58) og (59). Formlene i seg selv er meget enkle. Hvis vi erstatter h(x) med x i de to formlene får vi som vi ser de vanlige definisjonsformlene for $E(\underline{x})$.

Øvelse 4.*

a) Bruk (58) og (59) til å bevise (50). b) La $\underline{y}$ være følgende funksjon av $\underline{x}$: $\underline{y} = h(\underline{x})=c$ hvor c er en konstant. Vis at $E(h(\underline{x}))=c$ og kommenter resultatet. c) La c være en konstant og la $\underline{y}=h(\underline{x})$ være en funksjon av $\underline{x}$. Vis at $E(ch(x))=cE(h(x))$.

Hvis vi har k forskjellige funksjoner $h_1, h_2, \ldots, h_k$ kan det vises at følgende setning gjelder:

(60) $E(h_1(\underline{x})+h_2(\underline{x})+ \ldots +h_k(\underline{x}))=E(h_1(\underline{x}))+E(h_2(\underline{x}))+ \ldots +E(h_k(\underline{x}))$

Vi skal se på en anvendelse av (59) som kaster lys over definisjonsformelen for den teoretiske variansen og som vi skal dra nytte av senere. (I det diskrete tilfelle bruker vi (58) i stedet for (59), men utledningen er analog.)

Den teoretiske variansen $var(\underline{x})$ er definert som $\int_S f(x)(x-\mu)^2dx$. Hvis vi oppfatter $(\underline{x}-\mu)^2$ som en funksjon, $\psi(\underline{x})$ av $\underline{x}$, ser vi at den teoretiske variansen kan oppfattes som forventningen for $(\underline{x}-\mu)^2$. Vi har altså

(61) $var(\underline{x})=\sigma^2 = \int_S f(x)(x-\mu)^2dx = E(\underline{x}-\mu)^2$

C. En lineær funksjon av flere random variable

1. En funksjon av to random variable

Sett at vi har to random variable, $\underline{x}_1$ og $\underline{x}_2$ med forventninger $E(\underline{x}_1)=\mu_1$ og $E(\underline{x}_2)=\mu_2$ og varianser $var(\underline{x}_1)=\sigma_1^2$ og $var(\underline{x}_2)=\sigma_2^2$.

Vi vil ikke utelukke den mulighet at observasjonene av $\underline{x}_1$ og $\underline{x}_2$ knytter seg parvis til samme gjentak i et felles univers. I så fall kan det tenkes at $\underline{x}_1$ og $\underline{x}_2$ er korrelerte. La oss betegne den teoretiske korrelasjonskoeffisienten mellom $\underline{x}_1$ og $\underline{x}_2$ med ρ_{12}. Vi vil foreløpig ikke gjøre noen spesielle forutsetninger om ρ_{12}. Denne kan altså ha en hvilken som helst verdi mellom -1 og 1, deriblant også 0.

Hvis observasjonene av henholdsvis $\underline{x}_1$ og $\underline{x}_2$ knytter seg til gjentak fra to forskjellige univers, har det ingen mening å snakke om korrelasjonskoeffisienten mellom $\underline{x}_1$ og $\underline{x}_2$. I slike tilfelle setter vi automatisk $\rho_{12} = 0$ i formlene nedenfor.

Bortsett fra det som allerede er sagt, spesifiserer vi ingen ting om fordelingsfunksjonene for $\underline{x}_1$ og $\underline{x}_2$. De to random variable behøver altså ikke å ha normal fordelingsfunksjon eller noen annen kjent fordelingsfunksjon.

La oss nå definere en ny random variabel, $\underline{y}$ som er en lineær funksjon av de to random variable $\underline{x}_1$ og $\underline{x}_2$:

$$(62) \quad \underline{y} = a+b_1\underline{x}_1+b_2x_2$$

Her er a, b_1 og b_2 gitte konstante tall.

Vi er nå interessert i forventningen og variansen for $\underline{y}$. Det kan vises at disse er gitt ved følgende formler:

$$(63) \quad E(\underline{y}) = \mu_{\underline{y}} = a+b_1\mu_1+b_2\mu_2$$

$$(64) \quad \text{var}(\underline{y}) = \sigma_{\underline{y}}^2 = b_1^2\sigma_1^2 + b_2^2\sigma_2^2 + 2b_1b_2\rho_{12}\sigma_1\sigma_2$$

(63) er helt analog med (50). Vi ser at μ_y er den samme lineære funksjon av μ_1 og μ_2 som $\underline{y}$ er av $\underline{x}_1$ og $\underline{x}_2$. Når vi ser bort fra det siste leddet er også (64) analog med (51). Når $\rho = 0$ faller selvfølgelig det siste leddet bort.

2. En funksjon av et vilkårlig antall random variable

(63) og (64) kan generaliseres til et vilkårlig antall random variable. La den random variable $\underline{y}$ være definert som føl-

gende lineære funksjon av de random variable $\underline{x}_1$, $\underline{x}_2$,...,$\underline{x}_n$:

(65) $\underline{y} = a + b_1\underline{x}_1 + b_2\underline{x}_2 + \ldots + b_n\underline{x}_n$

Det kan da vises at forventningen $\mu_{\underline{y}}$ for $\underline{y}$ blir

(66) $E(\underline{y}) = \mu_{\underline{y}} = a + b_1\mu_1 + b_2\mu_2 + \ldots + b_n\mu_n$

hvor μ_i er forventningen for x_i (i=1,2,..,^{n}). Formelen for variansen for $\underline{y}$ skal vi ta med bare for det tilfelle at de n $\underline{x}$-ene er innbyrdes ukorrelerte. Det kan da bevises at variansen for $\underline{y}$ er uttrykt ved følgende formel:

(67) $\text{var}(\underline{y}) = \sigma_y^2 = b_1^2\sigma_1^2 + b_2^2\sigma_2^2 + \ldots + b_n^2\sigma_n^2$

Her er σ_i^2 variansen for $\underline{x}_i$ (i = 1,2,...,n).

Formelen (65) gjelder også i de tilfelle hvor a = 0 og $b_1=b_2= \ldots = b_n = 1$. På grunnlag av (60) og (65) kan vi derfor si helt generelt at <u>forventningen for en sum er lik summen av forventningene for addendene</u>.

3. <u>Forventning og teoretisk varians for et gjennomsnitt</u>

Vi skal se på et meget viktig eksempel på bruk av formlene (66) og (67). Sett at vi har en random variabel, $\underline{x}$ med forventning μ og standardavvik σ og med en vilkårlig fordelingsfunksjon, f(x). La oss ta for oss et random sampel på n observasjoner av $\underline{x}$. Før vi har skaffet oss samplet vet vi naturligvis ikke hvilke verdier $\underline{x}$ vil anta for hvert av de n gjentakene.

Det første gjentaket kan gi en verdi av $\underline{x}$ som ligger hvor som helst innen variasjonsområdet for $\underline{x}$, og sannsynligheten for de forskjellige verdier er gitt ved fordelingsfunksjonen, f(x) for $\underline{x}$. Det samme gjelder for hvert av de øvrige gjentakene. Følgelig kan vi tenke oss at vi har n random variable, $\underline{x}_1$, $\underline{x}_2$, ..., $\underline{x}_n$ som alle har samme fordelingsfunksjon, nemlig fordelingsfunksjonen for $\underline{x}$. Siden samplet er et <u>random</u> sampel følger

det også at de random variable $\underline{x}_1$, $\underline{x}_2$, ..., $\underline{x}_n$ er innbyrdes ukorrelerte. Det som ligger i at f.eks. $\underline{x}_1$ og $\underline{x}_2$ er ukorrelerte er følgende: Om den random variable $\underline{x}_1$ skulle anta en stor verdi (med stor menes her stor i forhold til forventningen, μ for $\underline{x}_1$) så gir ikke dette grunnlag for å gjette på om $\underline{x}_2$ vil anta en stor eller liten verdi (med stor eller liten menes her stor eller liten i forhold til forventningen μ for $\underline{x}_2$).

Tar vi for oss gjennomsnittet $\bar{x}$ av de n observasjonene, så har dette selvsagt en bestemt verdi for et bestemt sampel som vi har skaffet oss. Men før vi har skaffet oss samplet kan gjennomsnittet av verdiene av den random variable for n gjentak oppfattes som en random variabel som vi vil betegne med $\underline{\bar{x}}$.

Den random variable $\underline{\bar{x}}$ er følgende funksjon av de random variable $\underline{x}_1$, $\underline{x}_2$, ..., $\underline{x}_n$:

$$(68)\quad \underline{\bar{x}} = \frac{\underline{x}_1 + \underline{x}_2 + \ldots + \underline{x}_n}{n} = \frac{1}{n}\underline{x}_1 + \frac{1}{n}\underline{x}_2 + \ldots + \frac{1}{n}\underline{x}_n$$

Sammenlikner vi (68) med (65) ser vi at $a=0$ og $b_1=b_2=\ldots=b_n=\frac{1}{n}$. Ved hjelp av (66) og (67) finner vi så forventningen og variansen for $\underline{\bar{x}}$.

$$(69)\quad E(\underline{\bar{x}})=\mu_{\underline{\bar{x}}} = \underbrace{\frac{1}{n}\mu + \frac{1}{n}\mu + \ldots + \frac{1}{n}\mu}_{n \text{ ledd}} = \frac{n\mu}{n} = \mu$$

$$(70)\quad \mathrm{var}(\underline{\bar{x}}) = \sigma^2_{\underline{\bar{x}}} = \underbrace{\frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \ldots + \frac{1}{n^2}\sigma^2}_{n \text{ ledd}} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

(69) og (70) er to av de viktigste formlene i statistikken. Som vi har sett, gjelder de uansett hvilken fordelingsfunksjon $\underline{x}$ har.

4. Forventningen for den empiriske variansen*

På samme måte som vi i foregående avsnitt oppfattet gjennomsnittet som en random variabel, kan vi også oppfatte den

empiriske variansen som en random variabel. Som før tar vi utgangspunkt i en random variabel $\underline{x}$ med forventning μ og teoretisk varians σ^2. Vi tenker oss så at et random sampel på n observasjoner av $\underline{x}$ gir oss en verdi av hver av n random variable $\underline{x}_i$. Disse random variable har alle samme fordelingsfunksjon f med forventning μ og varians σ^2. Den random variable $\underline{s}^2$ kan derved oppfattes som en funksjon av n random variable:

$$(71) \quad s^2 = \frac{\sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})}{n-1}$$

Vi vil bevise at forventningen for den empiriske variansen er lik den teoretiske variansen, altså at

$$(72) \quad E(s^2) = \sigma^2$$

Vi vil begynne med å skrive om telleren i formelen (71). Dette kan gjøres på følgende måte:

$$(73) \quad \sum_{i=1}^{n} (\underline{x}_i - \bar{\underline{x}})^2 = \sum_{i=1}^{n} ((\underline{x}_i-\mu)-(\bar{\underline{x}}-\mu))^2$$

$$= \sum_{i=1}^{n} ((\underline{x}_i-\mu)^2 - 2(\underline{x}_i-\mu)(\bar{\underline{x}}-\mu)+(\bar{\underline{x}}-\mu)^2)$$

$$= \sum_{i=1}^{n} (\underline{x}_i-\mu)^2 - 2(\bar{\underline{x}}-\mu) \sum_{i=1}^{n} (\underline{x}_i-\mu) + \sum_{i=1}^{n} (\bar{\underline{x}}-\mu)^2$$

$$= \sum_{i=1}^{n} (\underline{x}_i-\mu)^2 - 2(\bar{\underline{x}}-\mu)(\sum_{i=1}^{n} \underline{x}_i - n\mu) + n(\bar{\underline{x}}-\mu)^2$$

$$= \sum_{i=1}^{n} (\underline{x}_i-\mu)^2 - 2n(\bar{\underline{x}}-\mu)^2 + n(\bar{\underline{x}}-\mu)^2$$

$$= \sum_{i=1}^{n} (\underline{x}_i-\mu)^2 - n(\bar{\underline{x}}-\mu)^2$$

Siden $\underline{x}_i$ har identisk samme fordelingsfunksjon som $\underline{x}$, kan vi i følge (61) skrive

$$(74)\quad E(\underline{x}_i-\mu)^2 = \sigma^2$$

og dermed ved å bruke (60):

$$(75)\quad E\left(\sum_{i=1}^{n} (\underline{X}_i - \mu)^2\right) = n\sigma^2.$$

I følge (61) og (70) vet vi at

$$(76)\quad \mathrm{var}(\bar{\underline{x}}) = E(\bar{\underline{x}}-\mu)^2 = \frac{\sigma^2}{n}$$

Av (73) får vi da, ved å bruke (62)-(63), (75), resultatet av øvelse 4c og (76):

$$(77)\quad E\left(\sum_{i=1}^{n} (\underline{x}_i-\bar{\underline{x}})^2\right) = E\left(\sum_{i=1}^{n} (\underline{x}_i-\mu)^2\right) - nE\,(\bar{\underline{x}}-\mu)^2$$

$$= n\sigma^2 - n \cdot \frac{\sigma^2}{n} = (n-1)\sigma^2$$

Dermed får vi ved å bruke resultatet av øvelse 4c:

$$(78)\quad E(\underline{s}^2) = E\,\frac{\sum_{i=1}^{n} (\underline{x}_i-\bar{\underline{x}})^2}{n-1} = \frac{1}{n-1} E\left(\sum_{i=1}^{n} (\underline{x}_i-\bar{\underline{x}})^2\right) = \frac{1}{n-1}(n-1)\sigma^2 = \sigma^2$$

Vi ser altså at s^2 er en forventningsrett estimator av σ^2. Hvis vi ikke hadde brukt n-1 i nevneren for s^2 ville vi ikke ha fått dette resultatet. Her har vi altså forklaringen på det "mystiske" forholdet at vi bruker n-1 i stedet for n i nevneren for s^2.

IV. Innføring i regresjonsanalyse

A. Innledning

Regresjonsteori og regresjonsanalyse er en viktig del av statistikken og omfatter en mangfoldighet av problemstillinger. Regresjonsteknikken er dessuten meget smidig og er av stor praktisk nytte innen en rekke fagområder. Også det vi forbinder med ordet korrelasjon kan naturlig behandles som en del av eller i tilknytning til regresjonsanalysen.

Hovedlinjene i regresjonsanalysen er nokså greie og lettfattelige. Det er detaljene som av og til kan volde vanskeligheter. De detaljene det her siktes til er imidlertid nokså viktige hvis en skal ha den fulle nytte av analysen.

Det som kanskje først og fremst faller en i tankene i forbindelse med ordet regresjon er et problem som i første omgang kan synes meget enkelt, nemlig det å trekke en rett linje mest mulig "midt igjennom" en sverm av punkter i planet. Vi skal se på et eksempel.

Det er ofte nyttig å kunne anslå slaktevekten av en gris på grunnlag av brystomfanget. Dette er mulig fordi det er en viss sammenheng mellom brystomfang og vekt. Jo større brystomfang en gris har, desto større slaktevekt kan vi regne med at den har. Det er imidlertid ikke noen eksakt matematisk sammenheng mellom brystomfang og slaktevekt. Griser med samme brystomfang kan ha nokså forskjellige slaktevekter, men i en viss gjennomsnittlig forstand stiger slaktevekten med brystomfanget.

La oss tenke oss at vi vil konstruere en kurve som kan brukes til å lese av slaktevekten x_2 når brystomfanget x_1 er kjent. Anta at vi er interessert i et bestemt univers av griser, (f.eks. norske griser i alderen 4 - 10 måneder). Brystomfanget og slaktevekten i dette universet kan oppfattes som to random variable $\underline{x}_1$ og $\underline{x}_2$. Vi vil tenke oss at vi har et random sampel på n = 10 griser fra dette universet. For hver gris har vi observert brystomfanget og slaktevekten.
Observasjoner (tenkte tall) er gjengitt i tabell 7.

Tabell 7. Brystomfang (x_1) og slaktevekt (x_2) i et random sampel på 10 griser.

x_1 (cm)	80	94	88	83	96	92	96	103	109	86
x_2 (kg)	42	75	50	50	72	61	64	75	100	50

På grunnlag av de 10 observasjonene (parobservasjonene) i tabell 7 kan vi tegne et såkalt spredningsdiagram som vist i fig. 4. Hver gris er her representert ved et punkt (kryss) som angir brystomfanget og vekten for vedkommende gris.

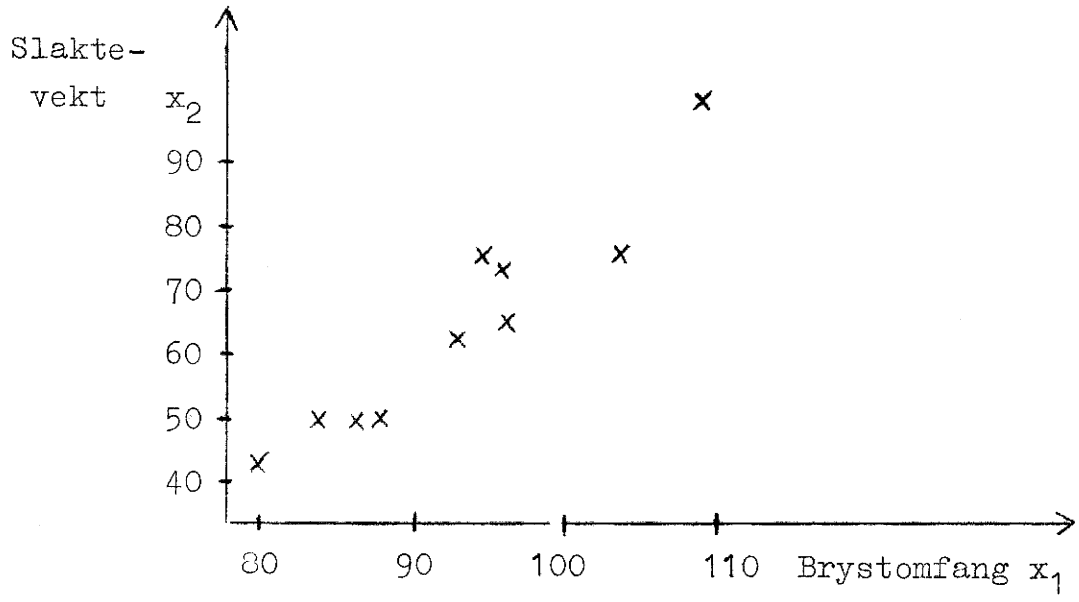


Fig. 4

Hvis vi kan anta at den underliggende tendensen er lineær blir oppgaven å trekke en rett linje på beste måte igjennom punktsvermen i fig. 4. Vi sier at vi "trekker en rett linje", men en slik linje har selvsagt en matematisk likning. Det vi gjør i første omgang er derfor å finne likningen for linjen. Dette eksemplet (som vi skal komme tilbake til senere) kan tjene som et utgangspunkt for en innføring i hva regresjonsanalysen dreier seg om. Problemet i vårt eksempel kan for det første presiseres nærmere ved hjelp av statistiske begreper. Om nødvendig kan det også generaliseres på forskjellige måter. Vi kan f.eks. bruke en krum linje i stedet for en rett linje.

I såfall snakker vi om krumlinjet regresjon. Slaktevekten kan også avhenge av andre variable enn brystomfanget, f.eks. av rygglengden, fethetsgraden, o.l. Tar vi dette i betraktning, kan vi generalisere problemet slik at det f.eks. går ut på å legge et plan gjennom en punktsverm i rommet. I slike tilfelle har vi å gjøre med multipel regresjon. Vårt problem tatt i vid betydning kan ha flere løsninger, og disse må vurderes på forskjellig vis. Dette bringer oss inn på hypotesetesting og intervallestimering i tilknytning til regresjonsanalysen. Endelig skal løsningene brukes til å besvare praktiske spørsmål, bl.a. ved hjelp av prediksjoner. Alt dette og mere til blir behandlet i regresjonsteorien.

B. Regresjonsanalysens tilknytning til hovedavsnitt II.

I hovedavsnitt II foran forklarte vi i detalj hva som menes med en regresjonsfunksjon idet vi tok utgangspunkt i den simultane fordelingsfunksjonen for to random variable, $\underset{\sim}{x}_1$ og $\underset{\sim}{x}_2$. I avsnittet ovenfor derimot var vårt utgangspunkt et random sampel på n gjentak. I det følgende skal vi vise hvilken forbindelse det er mellom de to problemstillingene.

Vi sa ovenfor at vi ønsket å konstruere en kurve som kan brukes til å lese av slaktevekten når brystomfanget x_1 er kjent. Hva slags kurve skal dette bli når det ikke er noen eksakt matematisk sammenheng mellom x_1 og x_2? Vi skal her presisere dette nærmere. Den ideelle kurve ville være en kurve som viser forventningen for $\underset{\sim}{x}_2$ som en funksjon av x_1, altså regresjonsfunksjonen for $\underset{\sim}{x}_2$ med hensyn på $\underset{\sim}{x}_1$.

Situasjonen kan skisseres nærmere på følgende måte. Til et bestemt brystomfang, f.eks. 95 cm svarer det en hel serie av slaktevekter. Vi kan faktisk si at vi har å gjøre med en betinget fordelingsfunksjon for slaktevekten betinget av at brystomfanget er 95 cm. Denne betingede fordelingsfunksjonen beskriver fordelingen av slaktevekten i det subuniverset av griser som alle har brystomfanget 95 cm. Til denne betingede fordelingsfunksjonen svarer det også en betinget forventning for slakte-

vekten, la oss si at denne er 65 kg. Tar vi for oss andre brystomfang, er situasjonen tilsvarende. Til hvert brystomfang svarer det en bestemt betinget forventning for slaktevekten. Denne betingede forventningen kan antas å være en funksjon av brystomfanget, og denne funksjonen er altså den samme regresjonsfunksjonen.

Når vi har for oss en bestemt gris og skal anslå slaktevekten, kan vi neppe foreta oss noe bedre enn å bruke den betingede forventningen for slaktevekten som svarer til denne grisens brystomfang. Vi må altså bruke regresjonsfunksjonen. Denne er imidlertid ukjent. Oppgaven blir derfor i første omgang å estimere koeffisientene i regresjonsfunksjonen på grunnlag av observasjonene av $\underset{\sim}{x}_1$ og $\underset{\sim}{x}_2$ i et random sampel av griser. Når dette er gjort, kan vi sette opp kurver eller tabeller av den typen som finnes i Hejes Lommealmanakk og som kan benyttes av de praktiserende bønder.

Det er svært mange situasjoner hvor en på liknende måte finner det formålstjenlig å estimere en regresjonsfunksjon. La oss generelt tenke oss at vi er interessert i to random variable $\underset{\sim}{x}_1$ og $\underset{\sim}{x}_2$ som knytter seg til et og samme univers, U. Vi regner med at det kan være en sammenheng mellom verdiene av de to random variable og vi er interessert i å få klarlagt denne sammenhengen, men vi vet lite eller ingenting om den simultane, de marginale og de betingede fordelingsfunksjonene for de to random variable. I slike tilfelle tyr vi ofte til regresjonsanalyse. Det vi da gjør i første omgang er å estimere (koeffisientene i) den ene eller begge regresjonsfunksjonene. Dessuten estimerer vi gjerne korrelasjonskoeffisienten. Selv om vi altså ikke går så langt som til å estimere (parameterne i) selve fordelingsfunksjonene, kan vi på denne måten skaffe oss en vesentlig informasjon om samvariasjonen mellom de to random variable.

Det første vi gjør er å skaffe oss et random sampel på n gjentak fra universet U. Et vilkårlig gjentak vil vi på vanlig måte betegne som gjentak nr. i, idet vi tenker oss gjentakene nummerert på en vilkårlig måte (f.eks. i den rekkefølgen de er observert) fra 1 til n ($i = 1,2,\ldots,n$). Hos hvert gjentak noterer

vi verdien x_{1i} av den random variable $\underline{x}_1$ og verdien x_{2i} av den random variable $\underline{x}_2$.

Som regel er det vanskelig å vite om en regresjonsfunksjon er lineær eller om den har en annen funksjonell form. I dette hovedavsnittet vil vi hele tiden forutsette at regresjonsfunksjonene er lineære. (Holdbarheten av denne forutsetningen kan undersøkes nærmere ved hypotesetestingsteknikk, men dette er noe som hører inn under krumlinjet regresjon som ikke vil bli behandlet her.)

Hvis regresjonsfunksjonene kan forutsettes å være lineære, estimerer vi gjerne parameterne og koeffisientene i universet ved hjelp av størrelser som er avledet av observasjonene i samplet, og som finnes ved hjelp av formler som svarer til de formlene vi brukte for universet i hovedavsnitt II.

Nedenfor har en satt opp en oversikt over noen størrelser som refererer seg til et univers og de tilsvarende sampelstørrelser.

Parametre eller koeffisienter i universet:	μ_1	μ_2	σ_1	σ_2	σ_{12}	β_{12}	β_{21}	ρ_{12}
Estimater av størrelsene ovenfor:	$\bar{x}_1$	$\bar{x}_2$	s_1	s_1	s_{12}	b_{12}	b_{21}	r_{12}

Størrelsen $\bar{x}_1$, $\bar{x}_2$, s_1 og s_2 (gjennomsnitt og middelavvik) er velkjente fra før, og vi trenger ikke å presentere formlene for disse om igjen.

Størrelsen s_{12} blir kalt den empiriske kovariansen. Formlen for denne svarer til formlen (14) s. 8 for den teoretiske kovariansen og skrives på følgende måte:

$$(79) \quad s_{12} = \frac{\sum(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)}{n-1} = \frac{\sum x_{1i}x_{2i} - \frac{\sum x_{1i} \sum x_{2i}}{n}}{n-1}$$

(For å forenkle skrivearbeidet vil vi i hele dette hovedavsnittet la være å skrive på nedre og øvre summasjonsgrense under og over tegnet $\sum$. Dette kan ikke føre til forvekslinger, da alle summeringer går fra i=1 til i=n. Vi vil imidlertid beholde fotindeksene i på x-ene.)

Ved hjelp av de 5 størrelsene $\bar{x}_1$, $\bar{x}_2$, s_1, s_2 og s_{12} kan r_{12}, b_{12} og b_{21} regnes ut etter formler som svarer helt ut til formlene (16) s. 10, (23) s.14 og (27) s. 15 for ρ_{12}, β_{12} og β_{21}.

Formlene er følgende:

$$(80)\quad r_{12}=\frac{s_{12}}{s_1 s_2}=\frac{\sum(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)}{\sqrt{\sum(x_{1i}-\bar{x}_1)^2\sum(x_{2i}-\bar{x})^2}}=\frac{\sum x_{1i}x_{2i}-\frac{\sum x_{1i}\sum x_{2i}}{n}}{\sqrt{\left[\sum x_{1i}^2-\frac{(\sum x_{1i})^2}{n}\right]\left[\sum x_{2i}^2-\frac{(\sum x_{2i})^2}{n}\right]}}$$

$$(81)\quad b_{12}=\frac{s_{12}}{s_2^2}=\frac{\sum(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)}{\sum(x_{2i}-\bar{x}_2)^2}=\frac{\sum x_{1i}x_{2i}-\frac{\sum x_{1i}\sum x_{2i}}{n}}{\sum x_{2i}^2-\frac{(\sum x_{2i})^2}{n}}$$

$$(82)\quad b_{21}=\frac{s_{12}}{s_1^2}=\frac{\sum(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)}{\sum(x_{1i}-\bar{x}_1)^2}=\frac{\sum x_{1i}x_{2i}-\frac{\sum x_{1i}\sum x_{2i}}{n}}{\sum x_{1i}^2-\frac{(\sum x_{1i})^2}{n}}$$

Formlene (80) - (82) får vi svært ofte bruk for. Legg merke til strukturen i formlene og sammenlikn formlene med hverandre. Formlene er meget lette å huske til tross for at de ser noe kompliserte ut ved første øyekast.

Ved praktiske beregninger bruker vi nesten alltid formlene lengst til høyre. Disse formlene (samt formel (79)) gjelder for øvrig like fullt om vi overalt erstatter x_{1i} med $(x_{1i}-c_1)$ og x_{2i} med $(x_{2i}-c_2)$ hvor c_1 og c_2 er to konstante tall som vi kan velge fritt for hvert sampel på en slik måte at regningen forenkles. Denne fremgangsmåten kan av og til være nyttig, f.eks. hvis x-ene er store tall.

De estimerte regresjonslinjene for $\underline{x}_1$ med hensyn på $\underline{x}_2$ og for $\underline{x}_2$ med hensyn på x_1 kan skrives på følgende måte som svarer helt til (22) s. 13 og (28) s. 15:

$$(83)\quad \text{est. } E(\underline{x}_1 \mid U x_2) = \bar{x}_1 + b_{12}(x_2-\bar{x}_2)$$

$$(84)\quad \text{est. } E(\underline{x}_2 \mid U x_1) = \bar{x}_2 + b_{21}(x_1-\bar{x}_1)$$

På venstre side av likhetstegnene har vi skrevet "est." som viser at det her er snakk om estimerte betingede forventninger.

Vi skal illustrere det hele ved hjelp av det hovedeksemplet vi arbeidet med på s. 3-16. Dette er ikke noe godt eksempel på hvilken nytte vi har av regresjonsanalyse i praksis, men det er en fordel å kunne belyse hele problemkomplekset ved hjelp av et enkelt eksempel. Vi skal senere ta for oss et eksempel som det er lett å se det praktiske siktemålet med.

Vi betrakter som før et univers U hvor hvert gjentak er en "besetning" (dvs. egentlig et random sampel på 4 dyr av F_2-generasjonen). Til hver "besetning" knytter det seg en verdi av den random variable $\underline{x}_1$ = antall røde dyr og en verdi av den random variable $\underline{x}_2$ = antall skimlete dyr.

La oss nå tenke oss at vi ikke hadde noe som helst kjennskap til Mendels lover. Vi vil med andre ord "glemme" alle de tallmessige opplysningene vi har om eksemplet i hovedavnsitt II. Utfra en slik situasjon vil vi tenke oss at vi skal estimere de to regresjonslinjene og korrelasjonskoeffisienten.

Det første vi må gjøre er å skaffe oss et random sampel av gjentak fra U. La oss f.eks. anta at vi vil arbeide med en sampelstørrelse på n = 10 "besetninger". I de to første kolonnene i tabell 8 har en gjengitt observasjonene i et slikt sampel.

Tabell 8. Observasjoner av $\underline{x}_1$ og $\underline{x}_2$ i et random sampel på n = 10 "besetninger", samt noen beregninger

x_{1i}	x_{2i}	x_{1i}^2	x_{2i}^2	$x_{1i}x_{2i}$
2	1	4	1	2
1	3	1	9	3
2	2	4	4	4
1	0	1	0	0
0	3	0	9	0
2	1	4	1	2
1	2	1	4	2
3	0	9	0	0
0	2	0	4	0
0	1	0	1	0

$\sum x_{1i}=12$ $\sum x_{2i}=15$ $\sum x_{1i}^2=24$ $\sum x_{2i}^2=33$ $\sum x_{1i}x_{2i}=13$

Nedenfor har en vist hvorledes beregningene kan utføres på grunnlag av tallene i tabell 8 og formlene (80) - (82).

$\bar{x}_1 = 1,2$ $\bar{x}_2 = 1,5$

$(\sum x_{1i})^2 = 144$ $(\sum x_{2i})^2 = 225$ $\frac{\sum x_{1i}\sum x_{2i}}{n} = \frac{12.15}{10} = \frac{180}{10} = 18$

$\frac{(\sum x_{1i})^2}{n} = 14,4$ $\frac{(\sum x_{2i})^2}{n} = 22,5$

$$r_{12} = \frac{13 - 18}{\sqrt{(24-14,4)(33-22,5)}} = \frac{-5}{\sqrt{9,6.10,5}} = \frac{-5}{\sqrt{100,8}} = \frac{-5}{10,04} = -0,498$$

$$b_{12} = \frac{-5}{10,5} = -0,47619$$

$$b_{21} = \frac{-5}{9,6} = -0,52083$$

Ved hjelp av vårt sampel har vi altså funnet $r_{12} = -0,498$. Dette tallet er et estimat av den sanne eller teoretiske korrelasjonskoeffisienten ρ_{12} som er lik -0,58 (s.10). På tilsvarende måte er $b_{12} = -0,47619$ et estimat av $\beta_{12} = -1,5$ (s.14) mens $b_{21} = -0,52083$ er et estimat av $\beta_{21} = -0,67$ (s.15).

Ved innsetting i formelen (83) får vi:

(85) est. $E(\underline{x}_1 | Ux_2) = 1,2 - 0,47619(x_2 - 1,5)$

eller

(86) est. $E(\underline{x}_1 | Ux_2) = 1,914285 - 0,47619x_2$.

Og ved innsetting i (84) får vi:

(87) est. $E(\underline{x}_2 | Ux_1) = 1,5 - 0,52083(x_1 - 1,2)$

eller

(88) est. $E(\underline{x}_2 | Ux_1) = 2,124996 - 0,52083x_1$.

Likningen (85) er den estimerte regresjonslikningen som svarer til regresjonslikningen (25) s. 14 i universet. Skriveformen (86) for den estimerte regresjonslikningen svarer til skrivemåten (26) for regresjonslikningen i universet. Tallet 1,914285 er således et estimat av konstantleddet 2 i (26).

På tilsvarende måte svarer (87) til (29) s. 15, mens (88) svarer til (30).

Øvelse 5.

Estimer regresjonsfunksjonene og korrelasjonskoeffisienten for eksemplet i tabell 7. Vi vil forutsette at regresjonslinjene er lineære. Tegn inn de to regresjonslinjene og punktsvermen i samme figur.

Øvelse 6.

Bevis at telleren i det siste uttrykket i (79), (80), (81) eller (82) er lik telleren i det nestsiste uttrykket. (Det tilsvarende beviset for nevnerne i (80)-(82) finnes på s. 58 i heftet for veterinærstudenter.)

Øvelse 7.

Regn ut kovariansen for eksemplet i tabell 7 etter formelen til venstre i (79), men foreta først en ordning av de 10 observasjonene i rekkefølge etter stigende x_1. (I praksis bruker vi verken denne formelen eller foretar en slik ordning.) Sett opp beregningene i en tabell av liknende type som tabell 8. Studer tabellen, spesielt kolonnen $(x_1-\bar{x}_1)(x_2-\bar{x}_2)$ og forsøk å forklare hvorfor den empiriske kovariansen og den empiriske korrelasjonskoeffisienten gir uttrykk for typen og graden av samvariasjon mellom x_1 og x_2 i samplet.

C. Estimering av koeffisientene i en regresjonsfunksjon ved hjelp av minste kvadraters metode

Den eneste begrunnelsen vi har gitt hittil for å bruke formlene (81) - (84) når vi skal estimere regresjonslinjene er at disse formlene har en viss likhet med formlene (23), (27), (22) og (28) som gjelder for universet. Dette er selvsagt en svak begrunnelse. Vi skal nå vise at formlene kan begrunnes ut fra en metode som kalles minste kvadraters metode. (Det er også

andre måter å begrunne formlene på, men vi kan ikke komme inn på dette her.)

Alle formler og likninger i forbindelse med regresjonsfunksjonen for $\underset{\sim}{x}_1$ med hensyn på $\underset{\sim}{x}_2$ blir nøyaktig av samme form som de vi har å gjøre med når vi betrakter regresjonsfunksjonen for $\underset{\sim}{x}_2$ med hensyn på $\underset{\sim}{x}_1$, bortsett fra at fotindeksene 1 og 2 må byttes om. I resten av dette hovedavsnittet skal vi derfor innskrenke oss til å behandle regresjonen for x_2 med hensyn på x_1.

Det er imidlertid stadig underforstått at det ikke er noe i veien for at $\underset{\sim}{x}_1$ og $\underset{\sim}{x}_2$ kan bytte roller. I praksis er situasjonen likevel den at vi vanligvis bare er interessert i den ene av de to regresjonsfunksjonene.

La oss igjen vende tilbake til eksemplet i tabell 7 og figur 4. Problemet er å trekke en rett linje på beste måte gjennom punktsvermen i figur 4. Det store spørsmålet er da hvilket prinsipp vi skal følge for å komme fram til en linje av den typen vi søker. (Kom med forslag.) Det kunne sies meget både om de krav som bør stilles til linjen og om tenkelige prinsipper. Vi skal her, i hvert fall foreløpig, nøye oss med å slå fast at det prinsippet som blir brukt i de fleste tilfelle er et prinsipp som går under navnet av minste kvadraters metode. Vi skal nå gjennomgå denne metoden i tilknytning til vårt eksempel. Det er da hensiktsmessig å tenke seg at linjen allerede er funnet, og at den er tegnet inn i spredningsdiagrammet slik som vist på fig. 5. (For å gjøre figuren mer oversiktlig har vi i fig. 5 ikke tegnet inn alle 10 punktene, men bare noen få.)

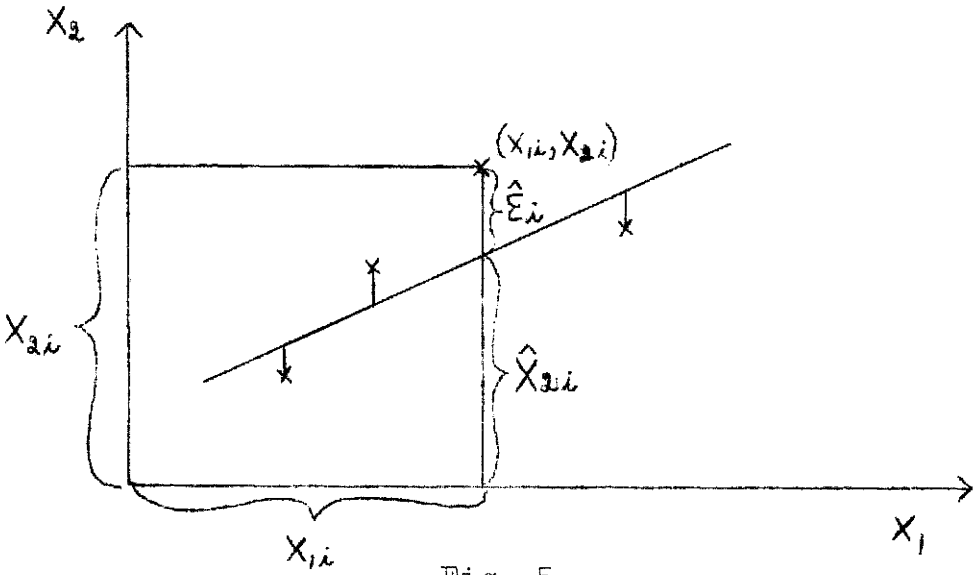


Fig. 5.

På figuren har vi trukket loddrette linjestykker fra hvert enkelt punkt opp til eller ned til linjen. Hvert linjestykke representerer avstanden fra vedkommende punkt til linjen. Minste kvadraters metode går ganske enkelt ut på at vi tenker oss hver av disse avstandene kvadrert og kvadratene summert for alle n punkter. Linjen skal velges på en slik måte at denne kvadratsummen blir minst mulig.

Vi skal nå vise hvorledes vi ved å bruke differansialregning kan komme fram til likningen for linjen. La likningen for linjen være

$$(89) \quad \hat{x}_2 = a_{21} + b_{21}(x_1 - \bar{x}_1)$$

En gris med brystomfang x_1 vil altså ha en anslått slaktevekt $\hat{x}_2$ som er gitt ved (89). For å komme fram til den konkrete linjen vi er ute etter, må vi finne formler for a_{21} og b_{21}. I disse formlene må bare observerte størrelser eller størrelser som kan avledes fra observasjonene inngå.

Før vi utleder formlene vil vi innføre noen nye symboler. La $\hat{x}_{2i}$ være den vekten som i følge linjen svarer til brystomfanget x_{1i} for gris nr. i. (Se figur 5.)Da har vi at

$$(90) \quad \hat{x}_{2i} = a_{21} + b_{21}(x_{1i} - \bar{x}_1)$$

Differensen $x_{2i} - \hat{x}_{2i}$ vil vi betegne med $\hat{\varepsilon}_i$. Denne differensen er det samme som den loddrette avstanden (positiv eller negativ) fra punktet (x_{1i}, x_{2i}) til linjen. Følgelig kan vi si at minste kvadraters metode går ut på å velge en linje med en verdi av a_{21} og b_{21} som gjør S minst mulig når S er gitt ved følgende formel:

$$(91) \quad S = \sum \hat{\varepsilon}_i^2 = \sum (x_{2i} - \hat{x}_{2i})^2 = \sum (x_{2i} - a_{21} - b_{21}(x_{1i} - \bar{x}_1))^2$$

Det siste likhetstegnet følger av (90).

Ved å variere a_{21} og b_{21} kan vi få fram alle tenkelige rette linjer. S vil være forskjellig for forskjellige linjer. S er altså en funksjon av a_{21} og b_{21}. I og med at vi betrakter et gitt sampel kan x_{2i} og x_{1i} $(i = 1,2,\ldots,n)$ betraktes som konstanter. Ved å derivere uttrykket S partielt med hensyn på a_{21}

og b_{21} og sette de deriverte lik 0 får vi likninger som kan brukes til å bestemme de verdier av a_{21} og b_{21} som minimaliserer S. Disse veridene er altså koeffisientene i den linjen vi søker.

Vi vil først derivere S partielt med hensyn på a_{21} og får da ved å bruke derivasjonsreglene for en sum, en funksjonsfunksjon, etc.

$$(92)\quad \frac{\partial S}{\partial a_{21}} = \sum 2(x_{2i} - a_{21} - b_{21}\,(x_{1i} - \bar{x}_1))(-1)$$

Ved å sette (92) lik 0 får vi videre:

$$(93)\quad \sum(x_{2i} - a_{21} - b_{21}(x_{1i} - \bar{x}_1)) = 0$$

$$(94)\quad x_{2i} - na_{21} - b_{21}\sum(x_{1i} - \bar{x}_1) = 0$$

Siden det siste leddet på venstre side i (94) er lik 0, får vi da ved å løse (94) med hensyn på a_{21}:

$$(95)\quad a_{21} = \frac{\sum x_{2i}}{n} = \bar{x}_2$$

Før vi deriverer (91) partielt med hensyn på b_{21} vil vi sette resultatet (95) i (91). Vi får da:

$$(96)\quad S = \sum((x_{2i} - \bar{x}_2) - b_{21}(x_{1i} - \bar{x}_1))^2$$

Dette gir:

$$(97)\quad \frac{\partial S}{\partial b} = \sum 2((x_{2i} - \bar{x}_2) - b_{21}(x_{1i} - \bar{x}_1))(-1)(x_{1i} - \bar{x}_1)$$

Settes dette lik 0, får vi:

$$(98)\quad \sum((x_{2i} - \bar{x}_2)(x_{1i} - \bar{x}_1) - b_{21}(x_{1i} - \bar{x}_1)^2) = 0$$

$$(99)\quad \sum(x_{2i} - \bar{x}_2)(x_{1i} - \bar{x}_1) = b_{21}\sum(x_{1i} - \bar{x}_1)^2$$

$$(100)\quad b_{21} = \frac{\sum(x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}2)}{\sum(x_{1i} - \bar{x}_1)^2}$$

Som vi ser, er uttrykket (100) for b_{21} identisk med (82). Også estimatet $a_{21} = \bar{x}_2$ er det samme som vi har brukt tidligere (se (84) og (89)).

Hvis vi nå ville la x_1 og x_2 bytte roller, måtte vi bytte om x_1 og x_2-aksene i fig. 4 og 5, eller vi kunne beholde aksene som de er og minimalisere summer av kvadratene av de vannrette avstandene fra punktene til linjen i stedet for som før de loddrette avstandene. Vi ville da finne formler for koeffisientene a_{12} og b_{12} i den estimerte regresjonsfunksjonen for $\underline{x}_1$ med hensyn på $\underline{x}_2$. Også disse formlene er identiske med de vi har brakt tidligere.

Vi ser derfor at formlene (81) - (84) kan begrunnes ut fra minste kvadraters metode.

I vårt problem slik det opprinnelig var formulert kaller vi slaktevekten x_2 den avhengige variable (den variable som skal forklares) og brystomfanget x_1 den uavhengige variable. I stedet for uavhengig variabel kan vi også bruke uttrykket forklaringsvariabel.

D. Forskjellige måter å skrive en estimert regresjonsfunksjon på

Hvis vi setter (95) inn i (89), kan den estimerte regresjonslinjen skrives på følgende måte som skiller seg fra (84) bare ved at vi har brukt et annet symbol for venstresiden:

$$(101) \quad \hat{x}_2 = \bar{x}_2 + b_{21}(x_1 - \bar{x}_1).$$

Hvis vi for x_1 setter inn $\bar{x}_1$, blir venstresiden lik $\bar{x}_2$. Dette viser at den estimerte regresjonslinjen går gjennom punktet $(\bar{x}_1, \bar{x}_2)$. Det er lett å vise at regresjonslinjen for $\underline{x}_2$ med hensyn på $\underline{x}_1$ og regresjonslinjen for $\underline{x}_1$ med hensyn på $\underline{x}_2$ skjærer hverandre nettopp i dette punktet.

(101) kan selvsagt også skrives på følgende måte:

$$(102) \quad \hat{x}_2 = \bar{x}_2 - b_{21}\bar{x}_1 + b_{21}x_1$$

Konstantleddet i den estimerte regresjonslikningen når x_1 måles fra sitt opprinnelige 0-punkt er altså

(103) $c_{21} = \bar{x}_2 - b_{21}\bar{x}_1$.

Hvis vi spesielt har i tankene en vilkårlig gris som finnes i vårt sampel (gris nr. i) kan det være naturlig å skrive (101) på følgende måte:

(104) $\hat{x}_{2i} = \bar{x}_2 + b_{21}(x_{1i} - \bar{x}_1)$

Her er $(x_{1i} - \bar{x}_1)$ avviket mellom brystomfanget for gris nr. i og det gjennomsnittlige brystomfanget i samplet. (104) er illustrert nærmere i fig. 6. Vi kan nesten si det slik at når vi skal finne den beregnede vekten $\hat{x}_{2i}$ som svarer til brystomfanget x_{1i} så bruker vi gjennomsnittsvekten $\bar{x}_2$ for hele samplet som en første tilnærmelse, og deretter korrigerer vi ved å legge til et ledd $b_{21}(x_{1i} - \bar{x}_1)$ som er positivt hvis gris nr. i har brystomfang over det gjennomsnittlige brystomfanget i samplet og negativt hvis grisens brystomfang ligger under gjennomsnittet.

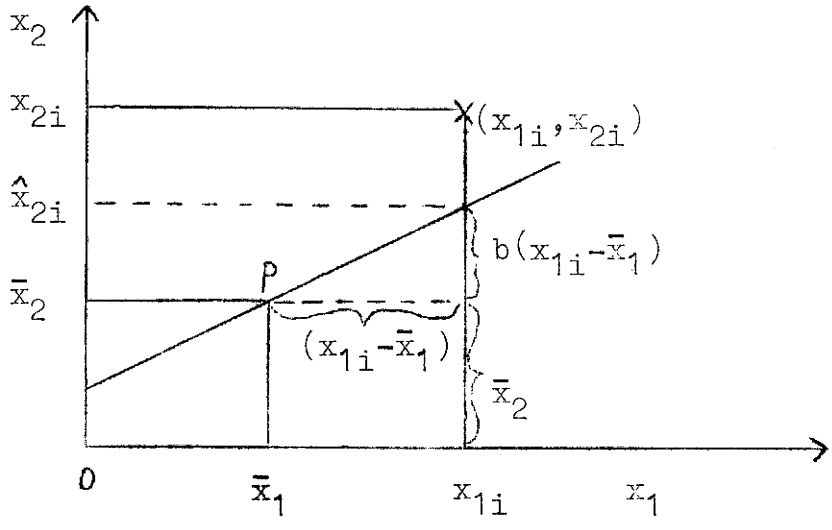


Fig. 6.

Likningen (104) kan også skrives på andre måter som av og til kan være nyttige. Siden $\hat{x}_{2i} = x_{2i} - \hat{\varepsilon}_i$ får vi ved innsetting av dette i (104):

(105) $x_{2i} - \hat{\varepsilon}_i = \bar{x}_2 + b_{21}(x_{1i} - \bar{x}_1)$

eller

(106) $x_{2i} = \bar{x}_2 + b_{21}(x_{1i} - \bar{x}_1) + \hat{\varepsilon}_i$

eller

(107) $(x_{2i} - \bar{x}_2) = b_{21}(x_{1i} - \bar{x}_1) + \hat{\varepsilon}_i$

La oss nå innføre følgende betegnelser:

(108) $u_{2i} = x_{2i} - \bar{x}_2$

$u_{1i} = x_{1i} - \bar{x}_1$

eller generelt

(109) $u_{hi} = x_{hi} - \bar{x}_h$

hvor h kan være nummeret til en hvilken som helst variabel vi måtte finne på å bruke. (107) kan da skrives på følgende måte:

(110) $u_{2i} = bu_{1i} + \hat{\varepsilon}_i$

I (110) kan vi om vi vil tenke oss at våre målinger er u-ene og at 0-punktet for u-ene er punktet P i fig. 6. Med disse variable og med dette aksesystemet (u-aksene parallelle med x-aksene) ser vi at konstantleddet i likningen faller bort.

La oss nå til sammenfatning sammenlikne følgende tre måter å skrive observasjonene x_{2i} på:

(111) $x_{2i} = (\bar{x}_2 - b_{21}\bar{x}_1) + b_{21}x_{1i} + \hat{\varepsilon}_i$

$x_{2i} = \bar{x}_2 + b_{21}(x_{1i} - \bar{x}_1) + \hat{\varepsilon}_i$

$u_{2i} = b_{21}u_{1i} + \hat{\varepsilon}_i$

Merk at vi har med leddet $\hat{\varepsilon}_i$ i alle tre formuleringer. Vi ser nå følgende: Når x-ene måles fra sine opprinnelge nullpunkter er konstantleddet $\bar{x}_2 - b_{21}\bar{x}_1$. Når x_1-verdiene uttrykkes som avvik fra gjennomsnittet blir konstantleddet x_2. Når både x_1-verdiene og x_2-verdiene uttrykkes i avviksform faller konstantleddet bort.

I alle tre tilfelle har vi imidlertid å gjøre med samme linje.

E. Mer om regresjonsmodeller

Vi har hittil tenkt oss at vi har å gjøre med to random variable $\underline{x}_1$ og $\underline{x}_2$, og at vi har skaffet oss et random sampel av gjentak fra det universet U som de to random variable knytter seg til. Når vi har en situasjon som denne kan den empiriske korrelasjonskoeffisienten r oppfattes som et estimat av den teoretiske korrelasjonskoeffisienten ρ. ($\underline{r}$ kan oppfattes som en estimator for ρ.) Vi kan da si at vi har å gjøre med en korrelasjonsmodell.

I mange viktige situasjoner er de n tallene x_{1i} å oppfatte som valgte verdier mens de n tilhørende tallene x_{2i} kan betraktes som verdier av en random variabel $\underline{x}_2$. Også i slike situasjoner kan vi på liknende måte som før snakke om de betingede forventningene for $\underline{x}_2$ betinget av x_1-verdiene og dermed også om en regresjonsfunksjon for $\underline{x}_2$ med hensyn på x_1-verdiene. Vi kan f.eks. være interessert i å estimere regresjonen for $\underline{x}_2$ = byggavling pr. dekar med hensyn på x_1 = mengden av nitrogengjødsel pr. dekar. Vi vil da vanligvis velge bestemte nitrogenmengder og anvende disse på forsøksruter slik at vi får n samhørende observasjoner av nitrogenmengde og byggavling. Disse observasjonene kan avmerkes i et spredningsdiagram på vanlig måte. I en slik situasjon, hvor x_{1i} ($i = 1,2,\ldots,n$) er å oppfatte som n konstanter, kan vi til atskillelse fra det foregående tilfelle si at vi har å gjøre med en regresjonsmodell.

La oss, noe kunstig, tenke oss at vi i en korrelasjonsmodell får nøyaktig de samme n tallene x_{1i} i alle tenkelige sampler på n gjentak, slik at det bare er de n tallene x_{2i} som variererer fra sampel til sampel. Hvis dette er tilfelle, blir de n tallene x_{1i} å oppfatte som n konstanter, og korrelasjonsmodellen kan da oppfattes som en regresjonsmodell. I det følgende skal vi bevise en del resultater hvorav enkelte kun har gyldighet for regresjonsmodeller. Hvis vi imidlertid tenker oss en korrela-

sjonsmodell av den typen vi nettopp har nevnt, blir resultatene gyldige også for en korrelasjonsmodell. Heretter skal vi, når ikke noe annet er sagt, betrakte x_{1i} som n tall som ikke forandrer seg fra sampel til sampel.

I (28) s. 15 skrev vi regresjonsfunksjonen for $\underline{x}_2$ med hensyn på $\underline{x}_1$ på følgende måte:

$$(112)\quad E(\underline{x}_2 | Ux_1) = E(\underline{x}_2) + \beta_{21}(x_1 - E(\underline{x}_1))$$

Når vi nå skal betrakte tilfelle hvor tallene x_1 er valgte konstanter, har det liten mening å operere med en forventning $E(\underline{x}_1)$. Vi kunne selvsagt velge en eller annen definisjon for $E(\underline{x}_1)$ i slike tilfelle, f.eks. $E(\underline{x}_1) = \bar{x}_1$, men hvis vi setter dette inn i (112), finner vi at regresjonsfunksjonen (112) i såfall må gå gjennom punktet $(\bar{x}_1, E(\underline{x}_2))$, men det er det jo ingen grunn til å anta at den gjør hvis $E(\underline{x}_2)$ som før oppfattes som den marginale forventningen for $\underline{x}_2$. Vi ser derfor at det er hensiktsmessig å velge en annen skrivemåte for regresjonsfunksjonen $E(\underline{x}_2 | Ux_1)$. Vi vil heretter bruke en skrivemåte som er like anvendelig enten vi har å gjøre med en korrelasjonsmodell eller en regresjonsmodell.

Siden vi hele tiden forutsetter at regresjonsfunksjonen er lineær, er det uten videre klart at regresjonsfunksjonen kan skrives på følgende måte:

$$(113)\quad E(\underline{x}_2 | Ux_1) = \gamma_{21} + \beta_{21} x_1$$

(Sammenlikn (30) s. 15 hvor $\gamma_{21} = 2{,}67$ og $\beta_{21} = -0{,}67$.)
Siden $\bar{x}_1$ nå betraktes som en konstant, er det ikke noe i veien for å bruke en skrivemåte hvor $\bar{x}_1$ forekommer i regresjonsfunksjonen i universet Ofte er det hensiktsmessig å bruke følgende skrivemåte:

$$(114)\quad E(\underline{x}_2 | Ux_1) = \alpha_{21} + \beta_{21}(x_1 - \bar{x}_1)$$

Sammenhengen mellom γ_{21} og α_{21} er som en lett ser følgende:

$$(115)\quad \gamma_{21} = \alpha_{21} - \beta_{21}\bar{x}_1$$

(Sammenlikn med (103) s. 48.) Både α_{21} og γ_{21} er konstanter.

Til sammenfatning kan vi nå si at vi ønsker å estimere koeffisientene i regresjonsfunksjonen (114). Ved å bruke minste kvadraters metode har vi tidligere kommet fram til den estimerte regresjonslikningen (101), s. 44. Estimeringsproblemet er altså forsåvidt løst. Vi ønsker imidlertid å vite om de estimatorene vi har brukt er gode estimatorer. Spesielt ønsker vi å vite om de er forventningsrette. Videre er vi interessert i å kunne teste hypoteser om β_{21} og i å lage konfidensgrenser for β_{21}. For å kunne løse disse problemene er det nødvendig å presisere modellen nærmere. Vi skal gjøre dette ved hjelp av det hovedeksemplet vi arbeidet med på s. 3-16 og s. 38. Dette eksemplet har bl.a. den fordelen at regresjonsfunksjonene i universet er kjent, noe som vanligvis ikke er tilfelle i praksis, men som gjør det lettere å illustrere modellen.

Den estimerte regresjonsfunksjonen (101) kan også skrives på formen (104) eller (106). Vi vil nå anvende (106) på hvert enkelt av de (10) gjentakene vi arbeidet med i tabell 8, s. 38. Først vil vi skrive opp de generelle symbolene, og deretter vil vi sette inn de spesielle koeffisientene vi beregnet tidligere på grunnlag av tabell 8. Vi får da:

$$
\begin{array}{lll}
x_{21} = \bar{x}_2 + b_{21}(x_{11} - \bar{x}_1) + \hat{\varepsilon}_1 & & 1 = 1{,}5 - 0{,}52083(2-1{,}2) + \hat{\varepsilon}_1 \\
x_{22} = \bar{x}_2 + b_{21}(x_{12} - \bar{x}_1) + \hat{\varepsilon}_2 & & 3 = 1{,}5 - 0{,}52083(1-1{,}2) + \hat{\varepsilon}_2 \\
\vdots & & \vdots \\
x_{2i} = \bar{x}_2 + b_{21}(x_{1i} - \bar{x}_1) + \hat{\varepsilon}_i & \text{eller} & x_{2i} = 1{,}5 - 0{,}52083(x_{1i} - \bar{x}_1) + \hat{\varepsilon}_i \\
\vdots & & \vdots \\
x_{2n} = \bar{x}_2 + b_{21}(x_{1n} - \bar{x}_1) + \hat{\varepsilon}_n & & 1 = 1{,}5 - 0{,}52083(0-1{,}2) + \hat{\varepsilon}_{10}
\end{array}
\qquad (116)
$$

I (116) har en for å spare skrivearbeidet, bare tatt med gjentak nr. 1, 2 og 10, og dessuten et vilkårlig gjentak (nr. i) som kan stå for et hvilket som helst gjentak fra nr. 1 til nr. 10. Tallene $\hat{\varepsilon}_i$ (i=1,2,...,n) kan vi kalle restledd. De er ukjente, men kan lett beregnes av (116). Oppstillingen (116) kan vi kalle en estimert regresjonsmodell. Den viser hvorledes vi i følge våre estimater, mener at de n tallene x_{2i} er framkommet.

Vi skal nå stille opp en tilsvarende regresjonsmodell på grunnlag av regresjonsfunksjonene i universet. I vårt spesielle tilfelle er disse kjente, så vi er i stand til å sette inn numeriske verdier for å illustrere modellen nærmere. I (118) nedenfor har vi til venstre skrevet en generell lineær regresjonsmodell, mens vi til høyre har satt inn de kjente verdiene for vårt eksempel. Legg merke til at selve regresjonslinjene i (118) gjelder universet mens de spesielle x-verdiene som er satt inn er tatt fra et bestemt sampel, nemlig det samme samplet som vi opererte med i (116). Siden slike sampelverdier generelt sett ikke representerer punkter på regresjonslinjen, har vi måttet føye til forstyrrelsesledd ε_i.

Regresjonsfunksjonen for vårt eksempel er gitt i (30) på s. 15. Vi vil skrive denne på formen (114). Siden $\bar{x}_1$ i vårt eksempel er lik 1,2, får vi da:

$$
\begin{aligned}
(117)\quad E(\underline{x}_2 | Ux_1) &= 2{,}67-0{,}67x_1+0{,}67\bar{x}_1-0{,}67\bar{x}_1 \\
&= 2{,}67+0{,}67.1{,}2-0{,}67(x_1-\bar{x}_1) \\
&= 3{,}474-0{,}67(x_1-\bar{x}_1)
\end{aligned}
$$

Ved å sammenlikne dette med (114), ser vi at α_{21} i vårt eksempel er lik 3,474, mens β_{21} som tidligere nevnt er lik -0,67.

I (118) nedenfor har vi skrevet opp den regresjonsmodellen som vi har det estimerte motstykke til i (116).

$$
(118)\quad
\begin{array}{ll}
x_{21} = \alpha_{21}+\beta_{21}(x_{11}-\bar{x}_1)+\varepsilon_1 & 1 = 3{,}474-0{,}67(2-1{,}2)+\varepsilon_1 \\
x_{22} = \alpha_{21}+\beta_{21}(x_{12}-\bar{x}_1)+\varepsilon_2 & 3 = 3{,}474-0{,}67(1-1{,}2)+\varepsilon_2 \\
\vdots & \vdots \\
x_{2i} = \alpha_{21}+\beta_{21}(x_{1i}-\bar{x}_1)+\varepsilon_i \quad \text{eller} & x_{2i} = 3{,}474-0{,}67(x_{1i}-1{,}2)+\varepsilon_i \\
\vdots & \vdots \\
x_{2n} = \alpha_{21}+\beta_{21}(x_{1n}-\bar{x}_1)+\varepsilon_n & 1 = 3{,}474-0{,}67(0-1{,}2)+\varepsilon_{10}
\end{array}
$$

I vårt tilfelle kjenner vi α og β. Det er derfor mulig å beregne størrelsen av hver ε. I alminnelighet er dette ikke mulig. Tallene blir derfor ofte kalt latente eller uobserverbare variable.

Vi kunne sløyfe tallene ε i (118), men da måtte vi på venstre side av likhetstegnene sette $E(\underline{x}_2|Ux_1)$ i stedet for de observerte verdiene av $\underline{x}_2$ i vårt sampel.

La oss understreke at det vi har på venstre side i (116) er nøyaktig det samme som det vi har på venstre side i (118). Det er på høyre side vi finner en forskjell. Parentesene $(x_{1i}-\bar{x}_1)$ er nøyaktig de samme i begge tilfelle. Forskjellen, derimot er følgende: Mens vi i (118) har brukt α_{21} og β_{21}, har vi i (116) brukt estimater av disse, nemlig estimatet $\bar{x}_2$ i stedet for α_{21} og b_{21} i stedet for β_{21}. Vi har med andre ord brukt regresjonslinjen $E(\underline{x}_2|Ux_1) = \alpha_{21}+\beta_{21}(x_1-\bar{x}_1)$ i (118) og den estimerte regresjonslinjen $\hat{x}_2 = \bar{x}_2+b_{21}(x_1-\bar{x}_1)$ i (116). I et diagram med x_1 som abscisse og x_2 som ordinat er ε_i (i=1,2,...,n) de loddrette avstandene (positive eller negative) til den første linjen og $\hat{\varepsilon}_i$ (i = 1,2,...,n) de loddrette avstandene til den andre linjen. Siden de to linjene i alminnelighet er forskjellige, er også i alminnelighet $\hat{\varepsilon}_i \neq \varepsilon_i$ (i = 1,2,...,n).

Hvis vi i praksis mener å kunne forutsette at regresjonsfunksjonen er lineær, setter vi opp en modell slik som vi har gjort i (114) eller (hvis vi spesielt har et sampel i tankene) til venstre i (118). Tallene α og β oppfatter vi da som konstante koeffisienter eller parametre som beskriver virkeligheten. Med utgangspunkt i denne modellen og et sampel foretar vi så en estimering. Hvis vi bruker minste kvadraters metode kommer vi da fram til den estimerte regresjonsmodellen (101) eller, hvis vi spesielt har samplet i tankene, (116). Tallene $\bar{x}_2$ og b_{21} oppfattes da som estimerte koeffisienter eller parametre.

Modellen (118) er ikke en fullstendig regresjonsmodell uten at vi også gjør visse forutsetninger om forstyrrelsesleddene ε_i. Modellen (118) oppfattes nemlig ofte som en hypotetisk modell, og vi kan være interessert i å betrakte andre alternative modeller, f.eks. modeller hvor regresjonsfunksjonen er ikke-lineær eller modeller hvor vi har flere forklaringsvariable enn x_1 (multipel regresjon). Hvis vi ikke gjør noen som helst forutsetninger om ε_i vil et hvilket som helst sett av data på n observasjoner være i skjønneste overensstemmelse med modellen (118). Modellen er da med andre ord helt triviell..

Før vi går over til forutsetningene om ε_i, skal vi utdype modellen litt nærmere. Vi vil da ta for oss oppstillingen til venstre i (118). La oss tenke oss en hel serie av random sampler, alle på n gjentak. Vi vil forutsette at de n tallene x_{1i} $(i = 1,2,\ldots,n)$ blir de samme i alle samplene. Konkret er dette lett å tenke seg hvis f.eks. x_{1i} er valgte mengder av et gjødselslag. Vi kan f.eks. tenke oss at vi gjentatte ganger anvender de samme n gjødselmengdene på et random sampel av n forsøksruter. De n forventede avlingsmengdene $\alpha_{21}+\beta_{21}(x_{1i}-\bar{x}_1)$ vil da bli de samme i alle sampler, men det er klart nok for enhver praktiker at de n faktisk observerte avlingsmengdene x_{2i} vil variere fra sampel til sampel. Dermed skulle det også være klart at vi i en slik situasjon har å gjøre med n random variable $\underline{x}_{2i}$ $(i = 1,2,\ldots,n)$. Men da følger det av (118) at vi på høyre side i (118) i en slik situasjon har å gjøre med n random variable ε_i $(i = 1,2,\ldots,n)$. Det må jo være slik all den stund venstresidene i (118) betraktes som random variable, mens leddene $\alpha_{21}+\beta_{21}(x_{1i}-\bar{x}_1)$ er konstanter.

Til sammenfatning kan vi altså si at det går ann å skrive om (118) på en slik måte at vi understreker de n tallene x_{2i} og de n tallene ε_i. Vi har da i tankene en uendelighet av sampler hvor vi alltid har de samme n x_1-verdiene. Når vi anlegger en slik betraktningsmåte kan modellen (118) skrives i sammentrengt form på følgende måte:

$$(119)\qquad \underline{x}_{2i} = \alpha_{21} + \beta_{21}(x_{1i}-\bar{x}_1) + \underline{\varepsilon}_i \qquad (i = 1,2,\ldots,n).$$

Vi er nå kommet til et punkt hvor vi er i stand til å spesifisere de forutsetningene vi vil gjøre om de n random variable $\underline{\varepsilon}_i$.

For det første vil vi forutsette at forventningen for $\underline{\varepsilon}_i$ er lik 0 for alle i.

$$(120)\qquad E(\underline{\varepsilon}_i) = 0 \qquad (i = 1,2,\ldots,n)$$

Denne forutsetningen innebærer (gansle) enkelt at $E(\underline{x}_{2i}) = \alpha_{21}+\beta_{21}(x_{1i}-\bar{x}_1)$. Dette ser vi ved å bruke reglen (50) s. 23 på (119). Av (119) ser vi nemlig at $\underline{x}_{2i}$ er en lineær funksjon av $\underline{\varepsilon}_i$ for alle i.

Den neste forutsetningen vi vil gjøre er at var($\underline{\varepsilon}_i$) er den samme, nemlig en konstant σ_ε^2, for alle i.

(121) $\text{var}(\underline{\varepsilon}_i) = \sigma_\varepsilon^2 \qquad (i = 1,2,\ldots,n).$

Videre vil vi forutsette at de enkelte random variable $\underline{\varepsilon}_i$ alle er innbyrdes ukorrelerte (dvs. den teoretiske kovariansen mellom to av dem er lik 0 uansett hvilke to av de n random variable vi tar for oss).

Når vi i det følgende skal lage konfidensintervaller og teste hypoteser vil vi bygge på den forutsetningen at hver enkelt av de random variable $\underline{\varepsilon}_i$ er normalt fordelt.

(122) $\varepsilon_i : N(0_1, \sigma_\varepsilon) \qquad (i = 1,2,\ldots,n).$

Øvelse 8.

Bruk det hovedeksemplet vi arbeidet med på s. 3-16 og det tilhørende samplet i tabell 8 s.38 og utfør følgende: (1) Tegn inn regresjonslinjen (30) s. 15 og den tilsvarende estimerte regresjonslinjen (88) s.39 i et og samme diagram. (2) Tegn også inn punktsvermen som svarer til tabell 8 i samme diagram. (3) Lokaliser de n avstandene ε_i og de n avstandene $\hat{\varepsilon}_i$ i diagrammet. (Tegn ikke for liten figur.)

F. Litt om estimatorenes egenskaper

1. Estimatoren for konstantleddet

Som estimator for konstantleddet α_{21} har vi brukt $\bar{\underline{x}}_2$. Denne estimatoren kan omskrives på følgende måte:

$$(123)\quad \bar{\underline{x}}_2 = \frac{1}{n}\sum \underline{x}_{2i} = \frac{1}{n}\sum\left[\alpha_{21} + \beta_{21}(x_{1i} - \bar{x}_1) + \underline{\varepsilon}_i\right]$$

$$= \frac{1}{n}\left[n\alpha_{21} + \beta_{21}\sum(x_{1i} - \bar{x}_1) + \sum \underline{\varepsilon}_i\right] = \alpha_{21} + \frac{1}{n}\sum \underline{\varepsilon}_i$$

Hvis vi nå bruker reglen (66) s. 28 finner vi i følge (120) at forventningen for $\bar{\underline{x}}_2$ blir følgende:

$$(124)\quad E(\bar{\underline{x}}_2) = \alpha_{21} + \frac{1}{n}\sum 0 = \alpha_{21}$$

Altså er $\bar{\underline{x}}_2$ en forventningsrett estimator for α_{21}.

Variansen for $\bar{\underline{x}}_2$ blir i følge (121), (123) og (67) s. 28:

$$(125)\quad \operatorname{var}(\bar{\underline{x}}_2) = \frac{1}{n^2}\sum \sigma_\varepsilon^2 = \frac{\sigma_\varepsilon^2}{n}$$

2. Estimatoren for regresjonskoeffisienten

Som estimator for regresjonskoeffisienten β_{21} har vi brukt $\underline{b}_{21}$:

$$(126)\quad \underline{b}_{21} = \frac{\sum (x_{1i}-\bar{x}_1)(\underline{x}_{2i}-\bar{\underline{x}}_2)}{(x_{1i}-\bar{x}_1)^2}$$

Telleren kan omskrives på følgende måte:

$$(127)\quad \sum (x_{1i}-\bar{x}_1)(\underline{x}_{2i}-\bar{\underline{x}}_2) = \sum \underline{x}_{2i}(x_{1i}-\bar{x}_1) - \underbrace{\bar{\underline{x}}_2 \sum (x_{1i}-\bar{x}_1)}_{=0}$$

$$= \sum (\alpha_{21}+\beta_{21}(x_{1i}-x_1)+\underline{\varepsilon}_i)(x_{1i}-\bar{x}_1)$$

$$= \underbrace{\alpha_{21}\sum (x_{1i}-\bar{x}_1)}_{=0} + \beta_{21}\sum (x_{1i}-\bar{x}_1)^2 + \sum \underline{\varepsilon}_i (x_{1i}-\bar{x}_1)$$

Settes dette inn i (126), får vi:

$$(128)\quad \underline{b}_{21} = \frac{\beta_{21}\sum (x_{1i}-\bar{x}_1) + \sum \underline{\varepsilon}_i (x_{1i}-\bar{x}_1)}{\sum (x_{1i}-\bar{x}_1)^2} = \beta_{21} + \frac{\sum \underline{\varepsilon}_i (x_{1i}-\bar{x}_1)}{\sum (x_{1i}-\bar{x}_1)^2}$$

Vi ser at den random variable $\underline{b}_{21}$ er en lineær funksjon av de n random variable $\underline{\varepsilon}_i$. Siden disse er normalt fordelt er også $\underline{b}_{21}$ normalt fordelt. Forventningen og variansen for $\underline{b}_{21}$ finner vi av (66) og (67) s. 28. Vi får:

$$(129)\quad E(\underline{b}_{21}) = \beta_{21} + \frac{\sum 0(x_{1i}-\bar{x}_1)}{\sum (x_{1i}-\bar{x}_1)^2} = \beta_{21}$$

$$(130)\quad \operatorname{var}(\underline{b}_{21}) = \frac{\sum \sigma_\varepsilon^2 (x_{1i}-\bar{x}_1)^2}{(\sum (x_{1i}-\bar{x}_1)^2)^2} = \frac{\sigma_\varepsilon^2}{\sum (x_{1i}-\bar{x}_1)^2}$$

Av (129) ser vi at $\underline{b}_{21}$ er en forventningsrett estimator for β_{21}. Sammenfattende kan vi si at $\underline{b}_{21}$ under våre forutsetninger er normalt fordelt med forventning β_{21} og varians

$$\frac{\sigma_\varepsilon^2}{\sum(x_{1i}-\bar{x}_1)^2} .$$

Det er selvsagt en fordel at en forventningsrett estimator har liten varians. Av (130) ser vi at variansen for $\underline{b}_{21}$ er liten når kvadratsummen i nevneren for var($\underline{b}_{21}$) er stor. Når x_1-verdiene velges er det derfor en fordel å velge verdier som gjør denne kvadratsummen stor. Da en ofte også er interessert i å undersøke om regresjonsfunksjonen virkelig er lineær er det ofte best å fordele x_1-verdiene noenlunde jevnt over det intervallet for x_1 som en er interessert i. En sammenklumpning av x_1-verdiene over et lite intervall er naturlig nok lite hensiktsmessig.

3. En forventningsrett estimator for σ_ε^2

Uten å framføre noe bevis, skal vi her presentere en forventningsrett estimator $\underline{s}^2$ for σ_ε^2. Estimatoren $\underline{s}^2$ kan skrives på følgende måte:

$$(131) \quad \underline{s}^2 = \frac{1}{n-2}\sum \underline{\varepsilon}_i^2 = \frac{1}{n-2}\sum(\underline{x}_{2i}-\bar{\underline{x}}_2-\underline{b}_{21}(x_{1i}-\bar{x}_1))^2$$

Telleren i estimatet s^2 er den samme kvadratsummen som vi minimaliserer når vi bruker minste kvadraters metode.

G. En identitet mellom kvadratsummer.

Vi vil se litt på variasjonen i tallene x_{2i} (i=1,2,...,n) i et sampel. Vi er vant til å bruke den empiriske variansen som et mål for denne variasjonen. Telleren i denne variansen er

(132) $$\sum(x_{2i}-\bar{x}_2)^2.$$

Vi vil heretter kalle denne kvadratsummen den totale kvadratsummen. Det ville ikke være så urimelig å bruke denne kvadratsummen som et mål for variasjonen i tallene x_{2i}. I et spredningsdiagram med x_1 som abscisse og x_2 som ordinat er den totale kvadratsummen lik summen av kvadratene av de loddrette avstandene fra de enkelte punktene til en vannrett linje gjennom punktet $(0, \bar{x}_2)$.

Hvis vi legger inn en estimert regresjonslinje i det samme diagrammet vil summen av kvadratene av avstandene fra punktene i spredningsdiagrammet til denne linjen bli

(133) $$\sum\hat{\varepsilon}_i^2.$$

Kvadratsummen (133) vil vi heretter kalle restkvadratsummen. Det er lett å se at restkvadratsummen er mindre enn eller i høyden lik den totale kvadratsummen. Den er lik den totale kvadratsummen når den estimerte regresjonslinjen er vannrett, dvs. når $b_{21}=0$.

Vi skal nå vise hvilken sammenheng det er mellom restkvadratsummen og den totale kvadratsummen. Vi tar da utgangspunkt i restkvadratsummen og omskriver denne:

(134)
$$\sum\hat{\varepsilon}_i^2 = \sum(x_{2i}-\hat{x}_{2i})^2 = \sum\left\{x_{2i}-\left[\bar{x}_2+b_{21}(x_{1i}-\bar{x}_1)\right]\right\}^2$$
$$= \sum\left[(x_{2i}-\bar{x}_2)-b_{21}(x_{1i}-\bar{x}_1)\right]^2$$
$$= \sum\left[(x_{2i}-\bar{x}_2)^2 - 2b_{21}(x_{2i}-\bar{x}_2)(x_{1i}-\bar{x}_1)+b_{21}^2(x_{1i}-\bar{x}_1)^2\right]$$
$$= \sum(x_{2i}-\bar{x}_2)^2 - 2b\sum(x_{2i}-\bar{x}_2)(x_{1i}-\bar{x}_1) + b_{21}^2\sum(x_{1i}-\bar{x}_1)^2$$

Det nest siste leddet i (134) kan skrives om slik at de to siste leddene kan trekkes sammen. Vi får da:

$$(135) \quad \sum \hat{\xi}_i^2 = \sum (x_{2i} - \bar{x}_2)^2 - b_{21}^2 \sum (x_{1i} - \bar{x}_1)^2$$

Vi ser av (135) at restkvadratsummen er lik den totale kvadratsummen minus det siste leddet i (135). Dette siste leddet gir derfor uttrykk for hvor mye mindre kvadratsummen blir når vi tar avstanden fra punktene i spredningsdiagrammet til den estimerte regresjonslinjen, enn når vi tar avstandene til den vannrette linjen gjennom punktet $(0, \bar{x}_2)$. En liten rest-kvadratsum er et uttrykk for at punktene i spredningsdiagrammet ligger tett inntil den estimerte regresjonslinjen. Når restkvadratsummen er mye mindre enn den totale kvadratsummen kan vi si at den estimerte regresjonslinjen "forklarer" en stor del av variasjonen i tallene x_{2i}. Det er derfor rimelig å si at den siste kvadratsummen i (135) gir uttrykk for hvor stor del av den totale kvadratsummen vi har "forklart" ved å legge inn en estimert regresjonslinje. Den siste kvadratsummen i (135) vil vi i det følgende kalle regresjonskvadratsummen. Vi har altså følgende identitet:

(136) Den totale kvadratsummen = Regresjonskvadratsummen + Restkvadratsummen.

Øvelse 9.

Vis at regresjonskvadratsummen også kan skrives på følgende måter:

a) Regresjonskvadratsum = $b_{21} \sum (x_{21} - \bar{x}_2)(x_{1i} - \bar{x}_1)$

b) Regresjonskvadratsum = $\dfrac{\left[\sum (x_{2i} - \bar{x}_2)(x_{1i} - \bar{x}_1)\right]^2}{\sum (x_{1i} - \bar{x}_1)^2}$

H. Mer om den empiriske korrelasjonskoeffisienten

Formelen for den empiriske korrelasjonskoeffisienten r er gitt ved (80) s. 37. . Vi skal nå vise at kvadratet av r kan skrives på en måte som gir ytterligere innblikk i hva r står for. Vi får:

$$(137)\quad r^2 = \frac{\left[\sum(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)\right]^2}{\sum(x_{1i}-\bar{x}_1)^2\sum(x_{2i}-\bar{x}_2)^2} = \left[\frac{\sum(x_{1i}-\bar{x}_1)(x_{2i}-\bar{x}_2)}{\sum(x_{1i}-\bar{x}_1)^2}\right]^2 \cdot \frac{\sum(x_{1i}-\bar{x}_1)^2}{\sum(x_{2i}-\bar{x}_2)^2}$$

eller

$$(138)\quad r^2 = \frac{b^2\sum(x_{1i}-\bar{x}_1)^2}{\sum(x_{2i}-\bar{x}_2)^2} = \frac{\text{Regresjonskvadratsum}}{\text{Total kvadratsum}}$$

I følge (135) er regresjonskvadratsummen lik den totale kvadratsummen minus restkvadratsummen. Hvis vi setter dette inn i (138), får vi:

$$(139)\quad r^2 = \frac{\sum(x_{2i}-\bar{x}_2)^2 - \sum\varepsilon_i^2}{(x_{2i}-\bar{x}_2)^2} = 1 - \frac{\text{Restkvadratsum}}{\text{Total kvadratsum}}$$

r^2 kan selvsagt aldri bli negativ, og spørsmålet er nå hvilke grenser r^2 må ligge mellom. Av (138) ser vi at r^2 er lik 0 hvis regresjonskvadratsummen er lik 0. I følge (135) er restkvadratsummen da lik den totale kvadratsummen. Dette inntreffer hvis den estimerte regresjonslinjen faller sammen med den horisontale linjen gjennom punktet $(0,\bar{x}_2)$, dvs. hvis $b_{21}=0$ (se (101) s. 44). I en slik situasjon er det ingen tendens til samvariasjon mellom x_1 og x_2 i samplet. Punktene i spredningsdiagrammet ligger da spredt tilfeldig i planet.

Av (139) ser vi at r^2 ikke kan bli større enn 1. Vi har dermed vist at r^2 alltid befinner seg innenfor følgende grenser:

(140) $\quad 0 \leq r^2 \leq 1.$

I følge (139) er r^2 lik 1 når restkvadratsummen er lik 0. Dette inntreffer hvis alle punktene i spredningsdiagrammet ligger på den estimerte regresjonslinjen.

Vi ser av denne diskusjonen at r^2 er stor hvis vi har å gjøre med et sampel hvor punktene i spredningsdiagrammet slutter seg tett inntil den estimerte regresjonslinjen. Hvis samplet er slik at punktene i spredningsdiagrammet ligger svært spredt omkring linjen, derimot, blir r^2 liten.

Den empiriske korrelasjonskoeffisienten r gir uttrykk for graden av lineær samvariasjon i samplet på samme måte som r^2, men selvsagt i en annen skala. Dessuten kan vi av r også se om samvariasjonen i samplet er positiv eller negativ, dvs. om regresjonslinjen stiger eller faller. Av (80) - (82) ser vi nemlig at r har samme fortegn som b_{21} (og b_{12}). Hvis vi noen gang skal beregne r ut fra r^2, må vi bestemme fortegnet for r ved å se på telleren i (80). Nevneren i (80) skal pr. definisjon være positiv.

På grunnlag av (140) og det som her er sagt om fortegnet for r, ser vi at r alltid ligger mellom følgende grenser:

(141) $\quad -1 \leq r \leq 1$

r^2 eller r kan alltid brukes som et mål for en punktsverms spredning omkring en rett linje. Hvis samvariasjonen i samplet (og universet) ikke er lineær, er imidlertid et slikt mål av liten interesse.

Den empiriske korrelasjonskoeffisienten r blir ofte oppfattet som et estimat av den teoretiske korrelasjonskoeffisienten ρ. Forat ρ skal ha noen mening må vi imidlertid ha å gjøre med to random variable $\underline{x}_1$ og $\underline{x}_2$. Tallene x_1 kan med andre ord ikke være valgte konstanter. Skal $\underline{r}$ være en god estimator for ρ, må vi dessuten skaffe oss et random sampel av gjentak fra det felles universet som $\underline{x}_1$ og $\underline{x}_2$ knytter seg til. Endelig vil vi nevne at

den teoretiske korrelasjonskoeffisienten ρ er av liten interesse som et mål for samvariasjonen mellom $\underline{x}_1$ og $\underline{x}_2$ i universet dersom regresjonsfunksjonene i universet ikke er lineære.

La oss se litt mer på r^2. Med utgangspunkt i (138) kan vi si at r^2 viser hvor stor brøkdel av den totale kvadratsummen den innlagte regresjonslinjen har forklart. Dette er en tolkning av r^2 som en støter på av og til. Hvis f.eks. $r^2 = 0{,}86$, sier en da at den estimerte regresjonslinjen har forklart 86 % av variasjonen i den avhengige variable i samplet. Vi skal imidlertid ikke legge altfor stor vekt på dette, da prosentsatsen er avhengig av det målet vi her bruker for variasjon.

Vi har ofte bruk for å beregne størrelsen av restkvadratsummen. Med utgangspunkt i (139) kan vi lett finne en formel som egner seg til dette bruk. Vi ser bort fra det mellomste leddet i (139) og løser den likningen vi da får med hensyn på restkvadratsummen. Vi får da:

(142) Restkvadratsum = $(1-r^2)$ • Total kvadratsum.

Øvelse 10.

Skisser en punktsverm og en estimert regresjonslinje for en situasjon hvor a) $r = 1$, b) $r = -1$ og c) $r = 0$. d) Skisser også en punktsverm hvor alle punktene ligger på en krum linje. Forklar med utgangspunkt i (139) hvorfor r^2 ikke blir lik 1 i dette tilfelle.

Øvelse 11.

Vis at $r^2 = b_{21} \cdot b_{12}$.

I. Konfidensgrenser for regresjonskoeffisienten og hypotesetesting.

Vi har tidligere kommet fram til en formel (130) for var($\underline{b}_{21}$).

Videre har vi i (131) presentert en estimator $\underline{s}^2$ for σ_ε^2. Hvis vi bruker (142) hvor vi nå oppfatter den empiriske korrelasjonskoeffisienten som en random variabel $\underline{r}^2$, kan estimatoren (131) omskrives på følgende måte:

$$(143)\quad \underline{s}^2 = \frac{(1-\underline{r}^2)\sum(x_{2i}-\bar{\underline{x}}_2)^2}{n-2}$$

Setter vi inn estimatoren for σ_ε^2, altså $\underline{s}^2$, i stedet for σ_ε^2 i formelen (130) for var ($\underline{b}_{21}$), får vi en estimator $\underline{s}^2_{b_{21}}$ for var ($\underline{b}_{21}$):

$$(144)\quad \underline{s}^2_{b_{21}} = \frac{(1-\underline{r}^2)\sum(\underline{x}_{21}-\bar{\underline{x}}_2)^2}{(n-2)\sum(x_{1i}-\bar{\underline{x}}_1)^2} = \frac{(1-\underline{r}^2)}{n-2}\,\frac{\sum \underline{x}_2^2 - \frac{(\sum \underline{x}_{2i})^2}{n}}{\sum x_{1i}^2 - \frac{(\sum x_{1i})^2}{n}}$$

La oss nå forme en random variabel $\underline{t}$ på følgende måte:

$$(145)\quad \underline{t} = \frac{\underline{b}_{21} - \beta_{21}}{\underline{s}_{b_{21}}}$$

Hvis de forutsetninger vi tidligere har gjort om $\underline{\varepsilon}_i$ er oppfylt, kan det bevises at $\underline{t}$ er fordelt etter Students t-fordeling. Antall frihetsgrader er:

$$(146)\quad f = n-2.$$

Den random variable $\underline{t}$ i (145) kan på vanlig måte brukes til å teste hypoteser om β_{21}. Vi skal her vise hvorledes den kan brukes til å konstruere et konfidensintervall for β_{21}.

I fig. 8 på s. 88 i heftet for veterinærstudenter er en t-fordeling framstilt. Vi kjenner fra før den definisjonsmessige sammenheng mellom a, P og Q. Siden $\underline{t}$ i (145) er t-fordelt, ser vi umiddelbart av den nevnte figur at følgende sannsynlighetsutsagn er oppfylt:

$$(147)\quad P\left(-a \leq \frac{\underline{b}_{21} - \beta_{21}}{\underline{s}_{b_{21}}} \leq a\right) = Q$$

Dette kan lett omformes slik at vi får følgende:

$$(148)\quad P\left(\underline{b}_{21} - a\underline{s}_{b_{21}} \leq \beta_{21} \leq \underline{b}_{21} + a\underline{s}_{b_{21}}\right) = Q$$

De to grensene $\underline{b}_{21} \mp a\underline{s}_{b_{21}}$ er som vi ser random variable som vil variere fra sampel til sampel. Sannsynligheten for at de to grensene vil falle slik at det konstante tallet β_{21} blir liggende mellom dem er lik Q.

Konfidensgrensene for β_{21} kan derfor beregnes ved hjelp av følgende formler:

$$(149)\quad \begin{aligned} &\text{Nedre grense:} && b_{21} - as_{b_{21}} \\ &\text{Øvre grense:} && b_{21} + as_{b_{21}} \end{aligned}$$

a finnes av tabellen over t-fordelingen for den konfidenssannsynligheten Q en har valgt å bruke og for f = n-2. Videre finnes $s_{b_{21}}$ av (144) idet en sløyfer understrekningene i formelen. Den estimerte regresjonskoeffisienten b_{21} regnes ut på vanlig måte etter (82).

Konfidensintervallet kan også brukes til å teste hypoteser. En hypotetisk verdi for β_{21} som viser seg å falle utenfor de konfidensgrensene som vi finner på grunnlag av et sampel ved bruk av konfidenssannsynligheten Q må forkastes på sannsynlighetsnivået P = 1-Q. Hvis derimot den hypotetiske verdien viser seg å ligge

innenfor konfidensintervallet, kan hypotesen ikke forkastes på dette nivået. Denne måten å teste hypoteser på gir nøyaktig samme konklusjon som bruk av (145) direkte.

Hvis vi ønsker å teste en hypotsese om at $\beta_{21} = 0$, kan vi bare se etter om tallet 0 ligger mellom konfidensgrensene. I såfall kan hypotesen ikke forkastes.

I en situasjon hvor det gir mening å snakke om en korrelasjonskoeffisient ρ i universet og hvor vi kan forutsette at regresjonsfunksjonene er lineære, vil et test av hypotesen $\beta_{21}=0$ samtidig også være et test av hypotesen $\rho = 0$ og av hypotesen $\beta_{12} = 0$. Hvis en av de 3 størrelsene β_{21}, ρ og β_{12} er lik 0, må nemlig også kovariansen være lik 0, og dermed må alle tre størrelsene være lik 0. (Se (16), (23) og (27).)

Øvelse 12.

Bruk tallene i tabell 7 s. 33 og beregn konfidensgrensene for β_{21} for dette eksemplet.