Per Ottestad

STATISTIKK

(Del III)

Utgave 1970

Per Ottestad

STATISTIKK

(Del III)

Utgave 1970

J. Regresjon og korrelasjon.

K. Om bruk av sampler fra gitte begrensede universer.

J. Regresjon og korrelasjon.

J.1. Lineær regresjon med en uavhengig variabel.

En foreløpig behandling av lineær regresjon med en uavhengig variabel er gitt i avsnitt B.9. I det følgende skal vi gå noe nærmere inn på saken.

La x_1 og x_o være to random variabler vi har skaffet oss observasjoner av (x_{1i} og x_{oi}) i et sampel på n gjentak. Ved regresjonsfunksjonen for x_o m.h.p. x_1 forstår vi da den funksjonelle sammenhengen mellom den betingede forventningen for x_o, oppfattet som avhengig variabel, og x_1 oppfattet som uavhengig variabel (se avsnitt D.5). Er x_o en kontinuerlig random variabel og en betegner regresjonsfunksjonen med $r(x_1)$, har en at

$$E(x_o;\ x_1) = r(x_1)$$

Her er da

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

hvor integrasjonsområdet omfatter alle de verdier som x_o kan ha når x_1 er gitt.

I det enkleste tilfelle er funksjonen lineær :

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

eller når vi innfører gjennomsnittet for x_1 (se B.9) :

$$r(x_1) = \alpha_o + \beta_{o1}\ (x_1-\bar{x}_1)$$

Konstantene eller parametrene α_o og β_{o1} er som oftest ukjente og må estimeres. Betegnes estimatorene med $\hat{\alpha}_o$ og $\hat{\beta}_{o1}$, er den estimerte regresjonsfunksjonen

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

Disse betegnelsene for estimatorene blir sjelden brukt. I stedet for $\hat{\alpha}_o$ brukes a_o (eller en annen bokstav), og i stedet for $\hat{\beta}_{o1}$ brukes som oftest b_{o1}. Den estimerte regresjonsfunksjonen skrives da slik :

$$\hat{r}(x_1) = a_0 + b_{01}\,(x_1 - \bar{x}_1)$$

Når en skal estimere de to parametrene, må en ta som utgangspunkt at observasjonene (x_{1i} og x_{0i}) av de to random variabler er gitt. Estimeringen går så ut på å tilpasse regresjonsfunksjonen til observasjonene. En bygger da på modellen

$$x_{0i} = \alpha_0 + \beta_{01}\,(x_{1i} - \bar{x}_1) + e_i$$

hvor e_i er n innbyrdes uavhengige verdier av en random variabel. Den metoden en bruker er Minste kvadraters metode, og siktepunktet er å gjøre kvadratsummen Σe_i^2 så liten som mulig. Dette vil si at en må oppfatte

$$S = \Sigma\, e_i^2 = \Sigma[x_{0i} - \alpha_0 - \beta_{01}(x_{1i} - x_1)]^2$$

som en funksjon av α_0 og β_{01}. Minimaliseringen utføres så på vanlig måte ved at en danner de to partielle deriverte av S m.h.p. α_0 og β_{01} og setter disse lik null.

En finner at

$$\frac{\partial S}{\partial \alpha_0} = \Sigma 2[x_{0i} - \alpha_0 - \beta_{01}(x_{1i} - \bar{x}_1)]\,(-1) = 0$$

og

$$\frac{\partial S}{\partial \beta_{01}} = \Sigma\, 2[x_{0i} - \alpha_0 - \beta_{01}(x_{1i} - \bar{x}_1)]\,(-1)(x_{1i} - \bar{x}_1) = 0$$

Erstattes så α_0 og β_{01} i disse to ligningene med estimatorene a_0 og b_{01}, vil en finne at

$$a_0 = \bar{x}_0$$

og

$$b_{01} = \frac{\Sigma(x_{1i} - \bar{x}_1)(x_{0i} - \bar{x}_0)}{\Sigma(x_{1i} - \bar{x}_1)^2}$$

Hvordan b_{01} skal beregnes i praksis er forklart i avsnitt B.9. Resultatet av tilpassningen er vist for et eksempel i Fig.B.2. En vil se at den estimerte regresjonslinjen skjærer tvers igjennom og gir en beskrivelse av en slags tendens i punktsvermen.

Estimeringen av β_{o1} ved b_{o1} forutsetter at den betingede forventningen for x_o , altså $E(x_o;x_1)$, faktisk er en lineær funksjon av x_1. I noen tilfelle vet en at det er slik. Et eksempel er nevnt i avsnitt D.5 (side 95). I de aller fleste tilfelle har en imidlertid ikke kjennskap til regresjonsfunksjonens form, og en må da ty til en videregående analyse som vi skal beskrive senere. Det en som oftest må falle tilbake på, er at den ukjente regresjonsfunksjonen har slike egenskaper at den lar seg erstatte av en rekkeutvikling (Taylor-rekken). Den lineære formen får en så hvis de ikke-lineære ledd i rekkeutviklingen er uten betydning.

Hensikten med en regresjonsanalyse er å finne en funksjon av x_1 som for hver verdi av x_1 gir et best mulig estimat av den betingede forventningen for x_o. Hva en skal forstå med "best mulig" kan naturligvis diskuteres. Men vi kan iallfall si at vi ønsker at estimeringen skal være forventningsrett. Vi er imidlertid også interessert i presisjonen ved estimeringen og må derfor skaffe oss et uttrykk for σ_b og en estimator av dette standardavvik, d.v.s. s_b. I de fleste tilfelle er en også interessert i å teste en null-hypotese som går ut på at $\beta_{o1} = 0$.

La oss først ta for oss modellen for x_{oi} og se på de enkelte leddene i denne. Det første leddet (α_o) er en lineær funksjon av gjennomsnittet for x_1. Vi har nemlig satt $\alpha_o = \beta_o + \beta_{o1}\,\bar{x}_1$. Siden vi betrakter α_o som en parameter må vi også oppfatte $\bar{x}_1$ som en parameter. Betydningen av dette er at det universet vi opererer med, ikke er det som de n gjentakene representerer i egenskap av et random sampel. Det er et noe endret univers vi har å gjøre med, nemlig et univers hvor forventningen for x_1 er lik $\bar{x}_1$. Vi kan si at dette betyr at vi innfører en restrik-

sjon eller begrensning på gyldigheten av de eventuelle resultater vi kommer fram til. Disse resultatene kan ikke appliseres på hele universet.

Med det samme kan vi nevne at vi også må oppfatte kvadratsummen for x_1, altså $\Sigma(x_{1i}-\bar{x}_1)^2$, som en parameter. I det universet våre eventuelle utsagn gjelder for, er ikke bare $E(x_1) = \bar{x}_1$, vi har også at $var(x_1) = \frac{1}{n-1}\Sigma(x_{1i}-\bar{x}_1)^2$. Vi har dermed to restriksjoner på gyldighetsområdet for våre utsagn.

Om leddet $\beta_{o1}(x_{1i}-\bar{x}_1)$ vil vi forutsette at det inneholder hele effekten av x_1. Dette betyr naturligvis at vi forutsetter at regresjonsfunksjonen er eksakt lineær. Denne forutsetningen er nødvendig her. Er nemlig regresjonsfunksjonen eksakt lineær, kan vi sette $E(e) = 0$ for hver verdi av x_1. Er derimot $E(e) \neq 0$, betyr det at e inneholder et element som enten hører med til α_o, til $\beta_{o1}(x_{1i}-\bar{x}_1)$ eller har en eller annen ikke-lineær sammenheng med x_1. Om e_i kan vi imidlertid ikke forutsette at den ikke har noen sammenheng med x_1. Selv om $E(e) = 0$ for hver verdi av x_1, kan var(e) være avhengig av x_1. For det eksemplet vi behandlet i avsnitt D.5 (side 95) fant vi at regresjonsfunksjonen for x_o m.h.p. x_1 var eksakt lineær. Men vi fant også at var(e) var en lineær funksjon av x_1.

Av modellen for x_{oi} finner vi at

$$\bar{x}_o = \alpha_o + \bar{e}$$

Siden $E(\bar{e}) = E(e) = 0$, vil vi ha at $E(\bar{x}_o) = \alpha_o$, d.v.s. at $\bar{x}_o$ er en forventningsrett estimator av α_o.

Videre vil vi også finne at

$$x_{oi}-\bar{x}_o = \beta_{o1}(x_{1i}-\bar{x}_1) + (e_i-\bar{e})$$

Innsettes dette i formelen for b_{o1}, finner vi at

$$b_{o1} = \beta_{o1} + \frac{\Sigma(x_{1i}-\bar{x}_1)(e_i-\bar{e})}{\Sigma(x_{1i}-\bar{x}_1)^2}$$

Forutsatt at E(e) = 0 for hver verdi av x_1 vil forventningen for det siste leddet være lik null. Dette vil så si at b_{o1} er en forventningsrett estimator av β_{o1}.

For å komme videre har en måttet forutsette at e er en random variabel med normal fordelingsfunksjon, at E(e) = 0 og at var(e) er uavhengig av x_1. En må m.a.o. forutsette at

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

Forutsettes dette, kan det bevises 1) at $\bar{x}_o$ og b_{o1} er normale random variable, 2) at de er uavhengige, 3) at $E(\bar{x}_o) = \alpha_o$ og $var(\bar{x}_o) = var(e)/n$, og 4) at $E(b_{o1}) = \beta_{o1}$ og at

$$var(b_{o1}) = \frac{var(e)}{\Sigma(x_{1i}-\bar{x}_1)^2}$$

For å kunne nyttiggjøre oss disse resultatene må vi ha en estimator av var(e). I avsnitt B.9 (side 50) innførte vi differensene $d_i = (x_{oi}-\bar{x}_o)-b_{o1}(x_{1i}-\bar{x}_1)$ og viste at

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

Forutsetter vi som foran at e har normal fordelingsfunksjon og at var(e) er uavhengig av x_1, kan vi bevise at

$$s_e^2 = \frac{1}{n-2}\,\Sigma\, d_i^2$$

er en forventningsrett estimator av var(e) og at derfor

$$s_b^2 = \frac{\Sigma d_i^2}{(n-2)\Sigma(x_{1i}-\bar{x}_1)^2}$$

er en forventningsrett estimator av $var(b_{o1})$. Det kan da også bevises at

$$t = \frac{b_{o1}-\beta_{o1}}{s_b}$$

er en Students t (se E.3) med f=n-2 frihetsgrader.

Dette kan vi så bruke til beregning av konfidensgrensene for β_{o1}. Disse grensene er $b_{o1} \mp a.s_b$ hvor verdien av a finnes i Tabell I for f=n-2 og den P-verdi en ønsker å bruke. For eksemplet i Tab.B.8 har vi n=15, b_{o1}=0,2678, $\Sigma(x_{1i}-\bar{x}_1)^2$=12,144 og $\Sigma(x_{oi}-\bar{x}_o)^2$ = 1,0031. Dette gir så

$$\Sigma d_i^2 = 1{,}0031-0{,}8707 = 0{,}1324$$

Vi finner så videre at s_e^2 = 0,0102, s_b^2 = 0,000 838 og s_b = 0,0289. Velger vi så å bruke konfidenssannsynligheten 0,95 (d.v.s.P=0,05), ser vi av Tabell I at vi (for f=13) skal sette a = 2,16. Altså er $a.s_b$ = 0,0624 og konfidensgrensene for β_{o1} blir

$$b_{o1} \mp a.s_b = 0{,}2678 \mp 0{,}0624 = \begin{cases} 0{,}21 \\ 0{,}33 \end{cases}$$

At konfidensintervallet ikke inneholder β_{o1} = 0 vil naturligvis si at null-hypotesen β_{o1} må forkastes dersom vi velger å bruke forkastningsnivået P = 0,05.

Testing av denne null-hypotesen kan naturligvis også utføres ved at en beregner

$$|t| = \frac{|b_{o1}|}{s_b}$$

og ser etter om den verdien en finner for |t| er større enn verdien av a (for f=n-2) i Tabell I. For vårt eksempel finner vi at |t| = 0,2678/0,0289 = 9,27 som er betydelig større enn a = 2,16 (for P=0,05) og også større enn a = 3,102 for P=0,01.

Av hensyn til fremstillingen i avsnittene J.3 og J.4 skal vi se hvordan vi kan gjøre bruk av korrelasjonskoeffisienten når det gjelder testen av null-hypotesen β_{o1} = 0. I avsnitt B.9 (side 50) ble denne koeffisienten definert ved formelen

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

Av dette finner vi at

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

og at

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

Her er da $r_{o1}^2\ \Sigma(x_{oi}-\bar{x}_o)^2$ den delen av kvadratsummen for x_o som stammer fra x_1 og $\Sigma\, d_i^2 = (1-r_{o1}^2)\ \Sigma(x_{oi}-\bar{x}_o)^2$ det vi kan kalle "Rest". Hvis vi så til de forutsetningene om den random variable e som vi har benyttet foran, tilføyer $\beta_{o1} = 0$, kan det vises at disse to delene av kvadratsummen for x_o er uavhengige. Dividert med antall frihetsgrader (f=1 og f=n-2) gir disse to delene to varianser V_1 og V_R slik som vist i Tab.J.1.

Tabell J.1.

	Kvadratsum	f	V
Regresjon	$r_{o1}^2\ \Sigma(x_{oi}-\bar{x}_o)^2$	1	V_1
Rest	$(1-r_{o1}^2)\Sigma(x_{oi}-\bar{x}_o)^2$	n-2	V_R

Varianskvotienten $F = V_1/V_R$ gir oss så en test av null-hypotesen $\beta_{o1} = 0$. For F finner vi at den kan skrives slik

$$F = \frac{r_{o1}^2}{1-r_{o1}^2}\,(n-2)$$

Siden teller-variansen (V_1) har f = 1 frihetsgrad, er

$$t = \frac{r_{o1}}{\sqrt{1-r_{o1}^2}}\,\sqrt{n-2}$$

en Students t med f = n-2 frihetsgrader. Det er lett å vise at denne t er den samme som den vi benyttet foran, $t = b_{o1}/s_b$.

I avsnitt D.5 (side 97) er det vist at vi kan sette

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

hvor σ_o og σ_1 er standardavvikene for x_o og x_1, og hvor ρ er korrelasjonskoeffisienten mellom de to random variabler i

universet. Siden σ_0 og σ_1 er positive og endelige, betyr $\beta_{01}=0$ også at $\rho=0$. Den test av null-hypotesen $\beta_{01}=0$ som vi refererte foran, er derfor også en test av null-hypotesen $\rho_i = 0$.

I avsnitt D.5 har vi behandlet den normale fordelingsfunksjonen for to random variabler. Det ble da vist at for dette tilfelle betyr $\rho = 0$ at de to random variablene er uavhengige. Dette gjelder imidlertid ikke alle tilfelle. En kan tenke seg at $\rho = 0$ og at modellen er $x_{0i} = \alpha_0 + e_i$. Det er da ingen korrelasjon mellom de to random variablene, men det kan likevel være avhengighet mellom dem. Avhengigheten kan gi seg utslag i at var(e) har en eller annen sammenheng med x_1. Vi nevner dette fordi det kan ha betydning at en merker seg at $\rho = 0$ ikke alltid betyr uavhengighet.

Et av formålene med en regresjonsanalyse er å finne en estimert regresjonsfunksjon $\hat{r}(x_1)$ som kan brukes til estimering av den betingede forventning for x_0. Er denne forventningen en lineær funksjon av x_1, er for en valt verdi av x_1, la oss si $x_1 = c$,

$$E(x_0;\ x_1=c) = r(c) = \alpha_0 + \beta_{01}(c-\bar{x}_1)$$

Estimatoren er

$$\hat{r}(c) = \bar{x}_0 + b_{01}(c-\bar{x}_1)$$

Vi har nå at $E(\bar{x}_0) = \alpha_0$ og $E(b_{01}) = \beta_{01}$. Følgelig er

$$E[\hat{r}(c)] = r(c)$$

d.v.s. at $\hat{r}(c)$ er forventningsrett estimator av r(c).

Når $\hat{r}(x_1)$ brukes til estimering av den betingede forventning for x_0, er en naturligvis også interessert i presisjonen. Som ved annen estimering får en uttrykk for denne ved bredden av konfidensintervallet.

Forutsettes det at regresjonsfunksjonen er eksakt lineær og at var(e) er uavhengig av x_1, kan det vises at $\bar{x}_o$ og b_{o1} er ukorrelerte. Under disse forutsetninger kan derfor formelen for var[$\hat{r}(c)$] utledes ved hjelp av den formelen for variansen for en sum som ble referert i avsnitt D.6 (side 101). Vi finner at

$$\text{var}[\hat{r}(c)] = \text{var}(\bar{x}_o) + (c-\bar{x}_1)^2 . \text{var}(b_{o1})$$

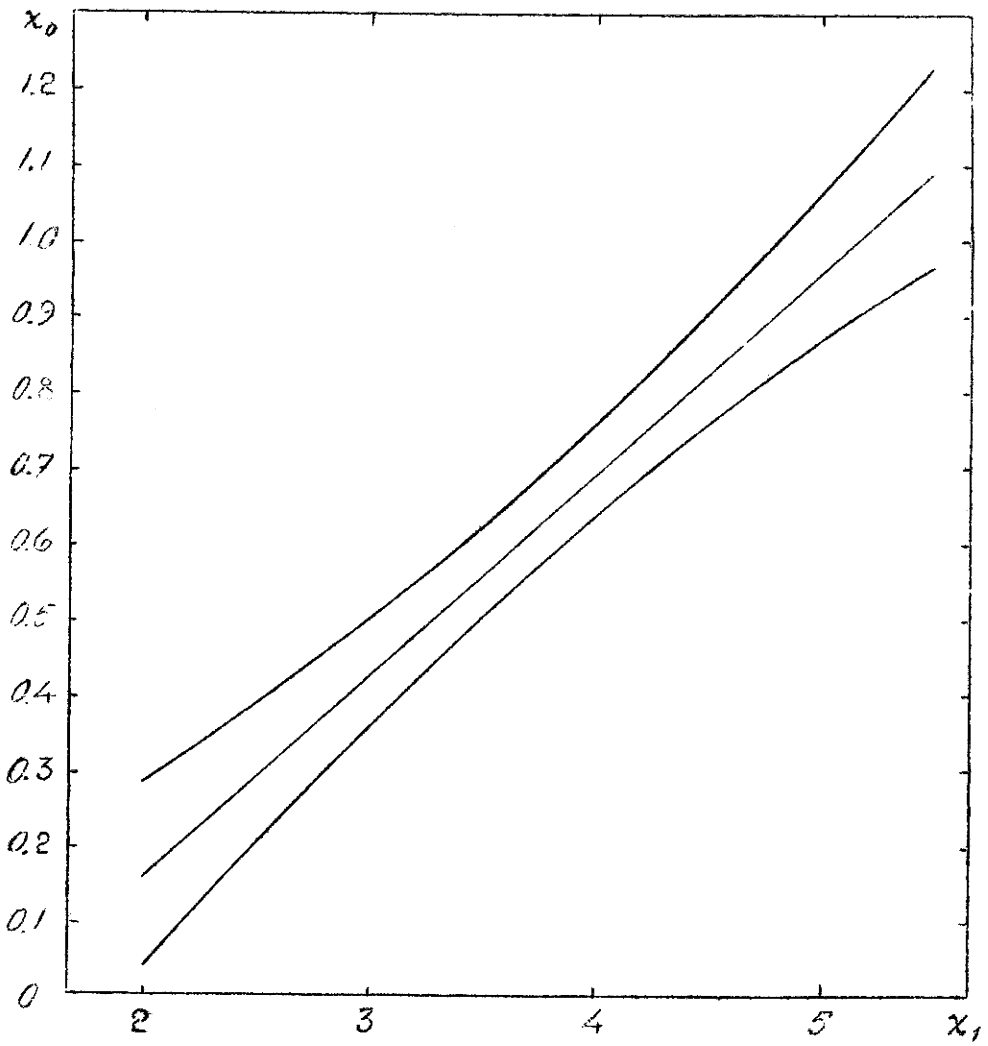
$$= \text{var}(e)\left[\frac{1}{n} + \frac{(c-\bar{x}_1)^2}{\Sigma(x_{1i}-\bar{x}_1)^2}\right]$$

Vi ser at var[$\hat{r}(c)$] er avhengig av $(c-\bar{x}_1)^2$. Den er minst for $c = \bar{x}_1$ og vokser med avstanden mellom den valte verdi c og gjennomsnittet for x_1. Estimatoren av var[$\hat{r}(c)$] som vi skal betegne med $s^2_{\hat{r}}$, får vi ved å erstatte var(e) med $s_e^2 = \frac{1}{n-2} \Sigma d_i^2$.

Under forutsetninger som er nevnt foran, er både $\bar{x}_o$ og b_{o1} normale random variabler. Da vil også $\hat{r}(c)$ være en normal random variabel. Regner en så med at s_e^2 er uavhengig av $\bar{x}_o$ og b_{o1}, er

$$t = \frac{\hat{r}(c)-r(c)}{s_{\hat{r}}}$$

en Students t med f=n-2 frihetsgrader. Konfidensgrensene for r(c) er derfor lik $\hat{r}(c) \mp a.s_{\hat{r}}$ hvor verdien av a finnes i Tabell I for f=n-2 frihetsgrader. Det sees av formelen for $s_{\hat{r}}$ at konfidensintervallet blir større jo større avstand det er mellom den valte verdi c og $\bar{x}_1$. Dette vil bety at presisjonen er størst for verdier i nærheten av gjennomsnittet for x_1. Dette sees også av Fig.J.1. De to krumme kurvene viser konfidensgrensene for r(c) for eksemplet i Tab.B.8. Den rette linjen er grafen for $\hat{r}(x_1)$.



Det er nevnt foran at en har måttet gjøre visse forutsetninger om den random variable e. De mest vidtgående forutsetningene går ut på at e_i er innbyrdes uavhengige, at e har normal fordelingsfunksjon med E(e) = 0 og at var(e) er uavhengig av x_1. Hva avvik fra disse forutsetningene betyr for f.eks. konfidensgrensene for β_{01} og r(c) har en vel ikke full oversikt over. Men etter de erfaringer en har fra andre områder er det god grunn til å regne med at rimelige forandringer av forutsetningene ikke vil ha noen konsekvenser for den praktiske bruk av de formler en har kunnet utlede.

Formålet med en regresjonsanalyse kan være å undersøke om det er påviselig korrelasjon mellom de random variable. Er det påviselig korrelasjon, d.v.s. at null-hypotesen $\beta_{01} = 0$ er forkastet, må konklusjonen gå ut på at det eksisterer en eller annen forbindelse mellom x_o og x_1. Denne forbindelsen kan imidlertid tenkes oppstått på to forskjellige måter. Det kan være en direkte forbindelse slik at den ene variabel er årsaksvariabel og den andre effektvariabel. Men som oftest er nok situasjonen slik at denne tolkingen ikke kan forsvares. Korrelasjonen kan nemlig være formidlet av en tredje random variabel u - eller flere slike - ved at det er en forbindelse mellom u og x_1 og en forbindelse mellom u og x_o. Det er grunn til å tro at det er på denne måten korrelasjonen mellom to random variabler som oftest er kommet i stand.

Regelen synes imidlertid å være at en har bruk for en funksjon av x_1 til estimering av den betingede forventning for x_o. Problemet kan oppstå f.eks. på grunn av at x_o er en random variabel det er vanskelig og kostbart å observere direkte. Kubikkinnholdet av et grantre eller en tømmerstokk er et eksempel på en slik random variabel. En kan da i slike tilfelle bruke observasjonene av en annen random variabel (x_1) til estimeringen slik vi har beskrevet det.

Når $\hat{r}(x_1)$ brukes i et nytt tilfelle eller gjentak, må en være oppmerksom på at den er å oppfatte som en prognose. En slik prognose er her som ellers basert på de konklusjoner en er kommet til på grunnlag av et sampel. Den kan derfor bare anvendes i det universet (her med to restriksjoner) som de n gjentakene representerer i egenskap av et random sampel. Forutsetningen for at prognosen kan brukes er derfor at det nye gjentak kan oppfattes som et gjentak i dette universet.

En av forutsetningene for dette er at de x_1-verdier en bruker i nye tilfelle, ligger innenfor det variasjonsområdet en har i det samplet som danner grunnlaget for prognosen. Det finnes riktignok tilfelle hvor denne regelen kanskje ikke må tas for alvorlig, nemlig de tilfelle hvor en vet på forhånd at regresjonsfunksjonen har den form en har nyttet, f.eks. den lineære formen. Men er f.eks. linearitet noe en slutter seg til på grunnlag av observasjonene av x_1 og x_o i det aktuelle samplet,, må en være ytterst forsiktig med a bruke den estimerte lineære regresjonsfunksjonen utenfor variasjonsområdet for x_1. At regresjonsfunksjonen synes å være lineær innenfor variasjonsområdet for x_1, er ikke grunn nok til å gå ut fra at den er lineær også utenfor dette området.

Det er derfor viktig at en skaffer seg en beskrivelse av gjentakene. En må regne med at noen av de karakteristikkene som denne beskrivelsen inneholder, kan ha innflytelse på regresjonsfunksjonen. Dette betyr at de nye gjentakene en anvender den estimerte regresjonsfunksjonen på, må tilfredsstille den samme beskrivelsen. Vi kan som eksempel tenke oss at observasjonene i Tab. B.8 stammer fra gjentak på en bestemt bonitet. Dette vil si at bonitetsklassen er med i beskrivelsen. Den estimerte regresjonsfunksjonen kan derfor ikke uten videre nyttes for gjentak fra en annen bonitet.

La oss nå tenke oss at det er på det rene at en av de to random variablene (x_1) er årsaksvariabel og den andre (x_o) er effektvariabel. Det kan da være fristende å oppfatte regresjonsfunksjonen for x_o m.h.p. x_1 som uttrykk for den effekt på x_o en vil oppnå ved aktivt å endre x_1. Regresjonsfunksjonen er i

mange tilfelle blitt oppfattet slik. Men dette kan ikke uten videre forsvares. Den effekten som regresjonsfunksjonen gir uttrykk for, er nemlig ikke alltid en ene-effekt av x_1. En må regne med at en endring i x_1 som finner sted uten aktiv påvirkning ved menneskelig handling, trekker andre random variabler med seg. Disse andre variabler kan ha betydning for den effekten som er registrert gjennom regresjonsfunksjonen. Tenker vi oss at x_1 kan endres ved aktiv handling, kan en ikke regne med at handlingen om den utføres vil påvirke disse andre variablene. Følgen kan da bli at effekten på x_o blir en annen enn den regresjonsfunksjonen gir uttrykk for. Dette er en av grunnene til at en tar med flere variabler som uavhengige variabler i regresjonsanalysene.

J.2. Multippel regresjon.

Vi skal nå tenke oss at vi har et antall variabler $x_1, x_2, x_3, x_4, \ldots$ som vi kan tenke oss brukt som uavhengige variabler i en regresjonsfunksjon. Den avhengig variable vil vi også nå betegne med x_0. Det er altså forventningen for x_0, betinget av $x_1, x_2, x_3, \ldots$, vi tar sikte på å estimere. Som eksempel kan nevnes at dersom x_0 kvistmengde (gitt i prosent) hos gran, kan x_1 være årringbredden, x_2 trehøyden og x_3 diameteren i brysthøyde.

For å gjøre fremstillingen alminnelig vil vi ikke forutsette at regresjonsfunksjonen er lineær. Som oftest vet vi ikke hvilken matematisk form regresjonsfunksjonen har, og vi må derfor også i dette tilfelle forutsette at den kan erstattes av en rekkeutvikling. Denne vil i regelen inneholde alle lineære ledd, alle ledd av 2. grad som f.eks. x_1^2 og x_1x_3, alle ledd av 3.grad som f.eks. x_1^3, $x_1^2x_2$, $x_2x_3^2$ o.s.v.

I det følgende skal vi innskrenke antallet av ledd til tre, x_1, x_2 og x_3. Vi vil imidlertid ikke forutsette at alle tre ledd er observerte random variabler. Teknikken er den samme om x_1 og x_2 er observerte variabler og x_3 er en matematisk funksjon av x_1 og x_2 som f.eks. x_1^2 eller x_1x_2. Fremstillingen vil også dekke det tilfelle at x_1 er en observert random variabel, x_2 og x_3 er funksjoner av x_1 som f.eks. $x_2=x_1^2$ og $x_3=x_1^3$. Det er også mange andre muligheter. Vi kan f.eks. ha at $x_1 = \log u$ hvor u er en observert random variabel som kanskje selv ikke blir benyttet i regresjonsfunksjonen.

Den regresjonsfunksjonen vi skal beskjeftige oss med, kan derfor skrives slik :

$$r(x_1,x_2,x_3) = \beta_0 + \beta_{01\cdot 23}\, x_1 + \beta_{02\cdot 13}\, x_2 + \beta_{03\cdot 12}\, x_3$$

$$= \alpha_0 + \beta_{01\cdot 23}(x_1-\bar{x}_1) + \beta_{02\cdot 13}(x_2-\bar{x}_2) + \beta_{03\cdot 12}(x_3-\bar{x}_3)$$

hvor

$$\alpha_0 = \beta_{01\cdot 23}\,\bar{x}_1 + \beta_{02\cdot 13}\,\bar{x}_2 + \beta_{03\cdot 12}\,\bar{x}_3$$

Her er $\beta_{01\cdot 23}$ regresjonskoeffisienten for x_0 m.h.p. x_1 i en regresjonsfunksjon hvor også x_2 og x_3 er med, $\beta_{02\cdot 13}$ er regresjonskoeffisienten for x_0 m.h.p. x_2 i en regresjonsfunksjon hvor også x_1 og x_3 er med og $\beta_{03\cdot 12}$ er regresjonskoeffisienten for x_0 m.h.p. x_3 i en regresjonsfunksjon hvor også x_1 og x_2 er med. Estimatorene for α_0 og β-koeffisientene vil vi betegne med a_0 (for α_0), $b_{01\cdot 23}$, $b_{02\cdot 13}$ og $b_{03\cdot 12}$. Den estimerte regresjonsfunksjonen er derfor

$$\hat{r}(x_1,x_2,x_3) = a_0 + b_{01\cdot 23}(x_1-\bar{x}_1) + b_{02\cdot 13}(x_2-\bar{x}_2) + b_{03\cdot 12}(x_3-\bar{x}_3)$$

Modellen for x_{0i} $(i=1,2,3,\ldots n)$ er nå

$$x_{0i} = r(x_{1i},x_{2i},x_{3i}) + e_i$$

hvor e_i forutsettes å være n innbyrdes uavhengige verdier av en random variabel e. Med samme begrunnelse som er gitt i J.1 kan vi også her sette $E(e) = 0$.

Til estimering av α_0 og β-koeffisientene bruker vi også i dette tilfelle Minste kvadraters metode. Vi skal derfor minimalisere

$$S = \Sigma[x_{0i} - \alpha_0 - \beta_{01\cdot 23}(x_{1i}-\bar{x}_1) - \beta_{02\cdot 13}(x_{2i}-\bar{x}_2) - \beta_{03\cdot 12}(x_{3i}-\bar{x}_3)]^2$$

Vi deriverer først partielt m.h.p. α_0 og setter resultatet lik null. Vi finner da at

$$\frac{\partial S}{\partial \alpha_0} = -\Sigma 2[x_{0i} - \alpha_0 - \beta_{01\cdot 23}(x_{1i}-\bar{x}_1) - \beta_{02\cdot 13}(x_{2i}-\bar{x}_2) - \beta_{03\cdot 12}(x_{3i}-\bar{x}_3)]$$

Settes så $\frac{\partial S}{\partial \alpha_0} = 0$, finner vi at estimatoren av α_0 er $a_0 = \bar{x}_0$.

Deriverer vi så partielt m.h.p. $\beta_{01\cdot 23}$ og setter resultatet lik null, finner vi at

$$\Sigma[x_{0i}-\alpha_0-\beta_{01\cdot 23}(x_{1i}-\bar{x}_1)-\beta_{02\cdot 13}(x_{2i}-\bar{x}_2)-\beta_{03\cdot 12}(x_{3i}-\bar{x}_3)](x_{1i}-\bar{x}_1) = 0$$

De to partielle deriverte av S m.h.p. $\beta_{02\cdot 13}$ og $\beta_{03\cdot 12}$ fører til to tilsvarende ligninger.

Resultatet er følgende tre ligninger som estimatorene skal tilfredsstille*:

$$\begin{aligned} S_{11}b_{01\cdot 23} + S_{12}b_{02\cdot 13} + S_{13}b_{03\cdot 12} &= S_{01} \\ S_{12}b_{01\cdot 23} + S_{22}b_{02\cdot 13} + S_{23}b_{03\cdot 12} &= S_{02} \\ S_{13}b_{01\cdot 23} + S_{23}b_{02\cdot 13} + S_{33}b_{03\cdot 12} &= S_{03} \end{aligned}$$

Her er da S_{11}, S_{12} o.s.v. kvadratsummer og produktsummer. Vi har f.eks. at $S_{22} = \Sigma(x_{2i}-\bar{x}_2)^2$ og $S_{13} = \Sigma(x_{1i}-\bar{x}_1)(x_{3i}-\bar{x}_3)$.

La oss ta for oss et eksempel. For et sampel på n=208 flater i en granskog ble følgende fire random variabler observert: x_1 = bonitet (middelhøyde ved 50 års alder), x_2 = middeldiameter, x_3 = diametertilvekst og x_0 = volum. Gjennomsnittene er : $\bar{x}_0 = 19{,}9364$, $\bar{x}_1 = 14{,}4327$, $\bar{x}_2 = 17{,}6913$ og $\bar{x}_3 = 2{,}6749$.

De videre beregningene førte til følgende ligninger :

$$\begin{aligned} 3267\, b_{01\cdot 23} + 323\, b_{02\cdot 13} + 564\, b_{03\cdot 12} &= 2450 \\ 323\, b_{01\cdot 23} + 6738\, b_{02\cdot 13} + 319\, b_{03\cdot 12} &= 5608 \\ 564\, b_{01\cdot 23} + 319\, b_{02\cdot 13} + 233\, b_{03\cdot 12} &= -165 \end{aligned}$$

Løsningene er

$$b_{01\cdot 23} = 1{,}7463,\ b_{02\cdot 13} = 1{,}0503 \text{ og } b_{03\cdot 12} = -6{,}3733$$

*Vi kan ikke her komme inn på hvordan slike ligninger løses i praksis og forutsetningene for at det finnes en løsning. Vi må nøye oss med å vise til litteratur som behandler lineære ligninger.

Ved innsetting av gjennomsnittene og b-verdiene i regresjonsfunksjonen finner en at

$$\hat{r}(x_1,x_2,x_3) = 1{,}7463\,x_1 + 1{,}0503\,x_2 - 6{,}3733\,x_3 - 6{,}8007$$

Når en regresjonsfunksjon er estimert, viser det seg praktisk talt alltid at det er en del av variasjonen i x_o som ikke blir beskrevet av funksjonen. For differensene

$$d_i = x_{oi} - \hat{r}(x_{1i},x_{2i},x_{3i})$$

vil en praktisk talt alltid finne at $\Sigma d_i^2 > 0$.

Det kan vises at kvadratsummen for x_o kan deles opp i to ledd som vi vil betegne med "Reduksjon" og "Rest", altså at

$$S_{oo} = \Sigma(x_{oi}-\bar{x}_o)^2 = \text{Reduksjon} + \text{Rest}$$

hvor

$$\text{Rest} = \Sigma d_i^2$$

og $$\text{Reduksjon} = b_{o1\cdot 23}\,S_{o1} + b_{o2\cdot 13}\,S_{o2} + b_{o3\cdot 12}\,S_{o3}$$

hvor S_{o1}, S_{o2} og S_{o3} er høyresidene i de tre ligningene.

Vi kan så innføre en korrelasjonskoeffisient $R_{o\cdot 123}$ som en parallell til den r_{o1} vi benyttet i avsnitt J.1. Denne koeffisienten som kalles den multiple korrelasjonskoeffisienten, er definert ved

$$R^2_{o\cdot 123} = \frac{\text{Reduksjon}}{S_{oo}} = 1 - \frac{\text{Rest}}{S_{oo}}$$

Siden det leddet vi har kalt Rest er en positiv størrelse, mindre eller i høyden lik S_{oo}, må $R^2_{o\cdot 123}$ være en positiv størrelse mindre eller i høyden lik enheten, d.v.s. at

$$0 \leq R^2_{o\cdot 123} \leq 1$$

Vi ser at $R^2 = 0$ når Rest = S_{oo}, og $R^2 = 1$ når Rest = 0, d.v.s. når Reduksjon = S_{oo}.

For vårt eksempel har vi at S_{oo} = 14120, og vi finner at

Reduksjon = 1,7463.2450 + 1.0503.5608 + 6,3733.165 =

11220,1119.

Altså er $R^2_{0\cdot 123} = 11220,1119/14120 = 0,7946$ og $R = 0,89$.

Vi ser at de to leddene i oppdelingen av kvadratsummen for x_0 kan skrives slik

$$\text{Reduksjon} = R^2_{0\cdot 123} \cdot S_{00}$$

og

$$\text{Rest} = (1-R^2_{0\cdot 123}) \cdot S_{00}$$

Forutsettes det så 1) at et i modellen for x_{0i} (side 216) er n innbyrdes uavhengige verdier av en random variabel e, 2) at e har normal fordelingsfunksjon, 3) at var(e) er uavhengig av x_1, x_2 og x_3, og 4) at $\beta_{01\cdot 23} = \beta_{02\cdot 13} = \beta_{03\cdot 12} = 0$, kan det bevises at

$$\frac{\text{Reduksjon}}{3} \quad \text{og} \quad \frac{\text{Rest}}{n-4}$$

er uavhengige varianser. Derfor har

$$F = \frac{\frac{1}{3}\,\text{Reduksjon}}{\frac{\text{Rest}}{n-4}} = \frac{R^2_{0\cdot 123}}{1-R^2_{0\cdot 123}} \cdot \frac{n-4}{3}$$

den fordelingsfunksjonen som F-testen bygger på, jfr. avsnitt G.3, side 152. Vi kan derfor bruke denne varianskvotienten til test av en null-hypotese som går ut på at alle tre regresjonskoeffisientene (β-ene) er lik null. Legg imidlertid merke til at det er bare en null-hypotese som testes, ikke tre som testes samtidig.

For vårt eksempel finner vi at

$$F = \frac{0,7946}{0,2054} \cdot \frac{204}{3} = 268,22$$

med f = 3 frihetsgrader for teller-variansen og f = 204 for nevner-variansen. Vi kan derfor trygt forkaste null-hypotesen. Konklusjonen blir at x_0 er påviselig avhengig av de tre andre variablene. Legg merke til at konklusjonen ikke går ut på at x_0 er påviselig avhengig av hver enkelt av de tre andre variablene. Situasjonen kan være slik at det f.eks. er to av de

uavhengig variablene som yter det aller meste til Reduksjonen, altså at bidraget fra den tredje er ubetydelig. Det er nødvendig at dette blir brakt på det rene. Dette skal vi imidlertid komme tilbake til i neste avsnitt.

Vi har foran behandlet det multiple regresjonsproblemet ved å forutsette at det er tre mulige uavhengige variabler i regresjonsfunksjonen. Alle de resultatene vi er kommet til, kan imidlertid generaliseres til et hvilket som helst antall. Setter vi antallet av mulige uavhengige variabler til m, vil Minste kvadraters metode føre til m ligninger som de m regresjonskoeffisientene skal tilfredsstille. Betegner vi så korrelasjonskoeffisienten med $R_{0.12\ldots m}$, vil vi også nå få en Reduksjon som er lik $R^2.S_{oo}$ og en rest som er lik $(1-R^2).S_{oo}$. En null-hypotese som går ut på at alle regresjonskoeffisientene (d.v.s. β-ene) er lik null, kan så testes ved varianskvotienten

$$F = \frac{R^2}{1-R^2} \quad \frac{n-1-m}{m}$$

med f=m frihetsgrader for tellervariansen og f=n-1-m frihetsgrader for nevnervariansen.

Denne testen har liten interesse i praksis. Når den nevnes her, er det fordi den går igjen i det aller meste av den litteratur som handler om regresjonsanalyse. Det er ikke gitt at alle de variabler som er med fra begynnelsen, bør være med i den regre regresjonsfunksjonen en aksepterer til slutt. En er derfor interessert i hvilke uavhengige variabler en bør ha med. De metoder en må bruke for å kunne ta standpunkt til dette, skal vi ta for oss i neste avsnitt.

J.3. Bruk av ortogonale funksjoner.

Vi vil også nå tenke oss at vi har tre uavhengige variabler (x_1,x_2,x_3) i regresjonsfunksjonen. Vi skal så føre inn lineære funksjoner (y_1,y_2,y_3) av disse og utføre regresjonsanalysen med disse nye variablene som uavhengige variabler. Til disse nye variablene vil vi stille som krav at

$$\Sigma(y_{1i}-\bar{y}_1)(y_{2i}-\bar{y}_2) = \Sigma(y_{1i}-\bar{y}_1)(y_{3i}-\bar{y}_3) = \Sigma(y_{2i}-\bar{y}_2)(y_{3i}-\bar{y}_3) = 0$$

Det kan dannes flere sett av slike nye variabler. Ett av settene er

$$y_1 = x_1-\bar{x}_1$$
$$y_2 = x_2-\bar{x}_2-b_{21}(x_1-\bar{x}_1)$$
$$y_3 = x_3-\bar{x}_3-b_{32\cdot1}(x_2-\bar{x}_2)-b_{31\cdot2}(x_1-\bar{x}_1)$$

Her er b_{21} regresjonskoeffisienten for x_2 m.h.p. x_1, $b_{31\cdot2}$ og $b_{32\cdot1}$ regresjonskoeffisientene for x_3 m.h.p. x_1 og x_2. Disse koeffisientene forutsettes å være estimatorer som vi kommer til ved å bruke Minste kvadraters metode. Gjennomsnittene er naturligvis $\bar{y}_1 = \bar{y}_2 = \bar{y}_3 = 0$.

At produktsummen av to og to y-er er lik null er lett å se. La oss ta for oss summen $\Sigma(y_{1i}-\bar{y}_1)(y_{3i}-\bar{y}_3) = \Sigma y_{1i}\cdot y_{3i}$. For denne finner vi

$$\Sigma y_{1i}\cdot y_{3i} = \Sigma[x_{3i}-\bar{x}_3-b_{32\cdot1}(x_{2i}-\bar{x}_2)-b_{31\cdot2}(x_{1i}-\bar{x}_1)](x_{1i}-\bar{x}_1)$$
$$= S_{13} - b_{32\cdot1}\cdot S_{12} - b_{31\cdot2}\cdot S_{11}$$

Setter vi

$$S = \Sigma[x_{3i}-\alpha_3-\beta_{32\cdot1}(x_{2i}-\bar{x}_2)-\beta_{31\cdot2}(x_{1i}-\bar{x}_1)]^2$$

og

$$\frac{\partial S}{\partial \beta_{31\cdot2}} = 0$$

finner vi at

$$S_{11}\cdot b_{31\cdot2} + S_{12}\cdot b_{32\cdot1} = S_{13}$$

Vi ser derfor at $\Sigma y_{1i}\cdot y_{3i} = 0$.

På samme måte kan det vises at $\Sigma y_{1i} \cdot y_{2i} = \Sigma y_{2i} \cdot y_{3i} = 0$.

Den regresjonsfunksjonen vi nå tar sikte på å estimere er

$$E(x_0; y_1, y_2, y_3) = r(y_1, y_2, y_3) = \alpha_0 + \beta_{01} y_1 + \beta_{02} y_2 + \beta_{03} y_3$$

Det er ikke nødvendig nå å bruke slike betegnelser for regresjonskoeffisientene som f.eks. $\beta_{02 \cdot 13}$ fordi koeffisienten til f.eks. y_1 nå er den samme uansett hvilke og hvor mange andre uavhengige variabler som er tatt med.

Modellen for x_{0i} er

$$x_{0i} = \alpha_0 + \beta_{01} y_{1i} + \beta_{02} y_{2i} + \beta_{03} y_{3i} + e_i$$

Parametrene α_0 og β-ene estimeres så ved Minste kvadraters metode, d.v.s. ved minimalisering av $S = \Sigma e_i^2$. På grunn av at $\Sigma y_{1i} = \Sigma y_{2i} = \Sigma y_{3i} = 0$ og at y-ene er ortogonale, vil vi nå finne at

$$\frac{\partial S}{\partial \alpha_0} = -2 \Sigma (x_{0i} - \alpha_0)$$

$$\frac{\partial S}{\partial \beta_{01}} = -2 \Sigma (x_{0i} - \alpha_0 - \beta_{01} y_{1i}) \cdot y_{1i}$$

og tilsvarende uttrykk for de to andre partielle deriverte. Setter vi så de fire partielle deriverte lik null, finner vi at estimatorene er

$$a_0 = \bar{x}_0$$

$$b_{01} = \frac{\Sigma y_{1i}(x_{0i} - \bar{x}_0)}{\Sigma y_{1i}^2} = \frac{\Sigma y_{1i} \cdot x_{0i}}{\Sigma y_{1i}^2}$$

$$b_{02} = \frac{\Sigma y_{2i}(x_{0i} - \bar{x}_0)}{\Sigma y_{2i}^2} = \frac{\Sigma y_{2i} \cdot x_{0i}}{\Sigma y_{2i}^2}$$

$$b_{03} = \frac{\Sigma y_{3i}(x_{0i} - \bar{x}_0)}{\Sigma y_{3i}^2} = \frac{\Sigma y_{3i} \cdot x_{0i}}{\Sigma y_{3i}^2}$$

Den estimerte regresjonsfunksjonen er nå

$$\hat{r}(y_1, y_2, y_3) = \bar{x}_\bullet + b_{01} y_1 + b_{02} y_2 + b_{03} y_3$$

Setter vi så som i foregående avsnitt

$$d_i = x_{oi} - \hat{r}(y_{1i}, y_{2i}, y_{3i})$$

finner vi at

$$\Sigma d_i^2 = S_{oo} - [b_{o1}\Sigma y_{1i}x_{oi} + b_{o2}\Sigma y_{2i}x_{oi} + b_{o3}\Sigma y_{3i}x_{oi}]$$

Som i foregående avsnitt setter vi så

$$S_{oo} = \text{Reduksjon} + \text{Rest}$$

hvor $\text{Rest} = \Sigma d_i^2$.

Det kan vises at den oppdeling av kvadratsummen for x_o vi får på denne måten, er den samme som vi får ved hjelp av den teknikken som ble beskrevet i foregående avsnitt. Det vi oppnår ved hjelp av y-ene, er at de tre leddene i Reduksjon er ukorrelerte og gir den reduksjon som skyldes hver enkelt y. Vi har f.eks. at $b_{o2}\Sigma y_{2i}x_{oi}$ er det bidrag som skyldes y_2.

Beregningen av disse reduksjonsbidragene kan i praksis utføres på flere måter. Den teknikken som vel er den enkleste og som de fleste programmer for elektronisk databehandling er basert på, er følgende.

La oss forutsette at settet av y-variable er valt, f.eks. det settet som er nevnt i begynnelsen av dette avsnittet (side 221). Det reduksjonsbidraget som skyldes $y_1 = x_1 - \bar{x}_1$ er da (jfr. avsnitt J.1) lik

$$R_{o\bullet 1}^2 \Sigma(x_{oi} - \bar{x}_o)^2 = R_{o\bullet 1}^2 S_{oo}$$

hvor $R_{o\bullet 1} = r_{o1}$. Bruker vi så på neste trinn både x_1 og x_2, vil den reduksjon som skyldes begge disse være

$$R_{o\bullet 12}^2 \cdot S_{oo}$$

Det bidrag som skyldes y_2, må da være lik differensen

$$[R_{o\bullet 12}^2 - R_{o\bullet 1}^2] S_{oo}$$

På det tredje trinnet bruker vi så alle tre x-ene. Den samlede

reduksjon blir da $R^2_{0\cdot 123} \cdot S_{00}$, og det reduksjonsbidraget som skyldes y_3 blir derfor lik

$$[R^2_{0\cdot 123} - R^2_{0\cdot 12}]\, S_{00}$$

Vi ser at summen av de tre reduksjonsbidragene er $R^2_{0\cdot 123} \cdot S_{00}$.

Bruker vi det vanlige skjema for variansanalyse, vil vi få den oppstillingen som er vist i Tab.J.2.

Tabell J.2.

Uavhengig variabel	Kvadratsum	f	V
y_1	$R^2_{0\cdot 1} \cdot S_{00}$	1	V_1
y_2	$[R^2_{0\cdot 12} - R^2_{0\cdot 1}] \cdot S_{00}$	1	V_2
y_3	$[R^2_{0\cdot 123} - R^2_{0\cdot 12}] \cdot S_{00}$	1	V_3
Rest	$[1 - R^2_{0\cdot 123}] \cdot S_{00}$	n-4	V_R

Det sees også av denne oppstillingen at den totale reduksjon med f = 3 frihetsgrader er oppdelt i tre reduksjonsbidrag, hvert med f = 1 frihetsgrad.

Forutsettes det så 1) at e_i i modellen for x_{0i} er n innbyrdes uavhengige verdier av en random variabel e, 2) at e har normal fordelingsfunksjon, 3) at var(e) er uavhengig av x_1, x_2 og x_3, og 4) at $\beta_{01} = \beta_{02} = \beta_{03} = 0$, vil en finne at

$$F_1 = V_1/V_R, \quad F_2 = V_2/V_R \quad \text{og} \quad F_3 = V_3/V_R$$

er varianskvotienter med den fordelingsfunksjon som F-testen bygger på, jfr.avsnitt G.3 (side 152). De tre kvotientene kan derfor brukes til test av hver av de tre null-hypotesene : 1) $\beta_{01} = 0$, 2) $\beta_{02} = 0$, og 3) $\beta_{03} = 0$.

På grunn av den felles nevnervariansen (V_R) er imidlertid disse tre varianskvotientene ikke uavhengige av hverandre. De er positivt korrelerte. Konsekvensen av dette er at vi kan ikke bruke som kritisk verdi det tallet vi finner i f.eks. Tabell III

for f = 1 og f = n-4 frihetsgrader. Vi må stille strengere krav.

Spørsmålet om hvilket krav en må stille er ennå under diskusjon. Vi kan imidlertid anbefale et krav som går ut på at alle F-verdiene må være større enn den verdi en finner i f.eks. Tabell III når en bruker f = 1 frihetsgrad for hver av tellervariansene og $f = \frac{n-4}{3}$ for den felles nevnervariansen. En vil da operere på et forkastningsnivå som er meget nær likt det nivå tabellen er beregnet på. Bruker en derfor Tabell III, vil en operere på nivået P = 0,05 for hver av varianskvotientene.

I Tabell III, IV og V er antall frihetsgrader (f) hele tall. Bruker en $f = \frac{n-4}{3}$, vil en få et bruddent tall i mange tilfelle og må derfor interpolere i tabellen. Er antall frihetsgrader, altså $f = \frac{n-4}{3}$, større enn 5, vil lineær interpolasjon gi nøyaktig nok verdi. For et mindre antall er det utarbeidet tabeller hvor antall frihetsgrader for nevnervariansen er gitt for hver tiendedel. En slik tabell er gjengitt i Tabell VI. Her er da antall frihetsgrader for hver av tellervariansene lik f = 1. Verdiene er gitt for P = 0,05, P = 0,025 og P = 0,01.

Når det her er sagt at en må bruke $f = \frac{n-4}{3}$, er dette naturligvis under forutsetning av at en har tre uavhengige variabler. Er antall uavhengige variabler i reg.funksjonen lik m, må en bruke $f = \frac{n-m-1}{m}$.

La oss bruke denne teknikken på eksemplet fra foregående avsnitt, hvor x_0 er volum, x_1 bonitet, x_2 middeldiameter, og x_3 er diametertilvekst. Nytter vi det settet av y-variabler som er gitt i begynnelsen av dette avsnittet, vil en finne at

$$R^2_{0\cdot 1} = 0,1301, \quad R^2_{0\cdot 12} = 0,4342 \quad \text{og} \quad R^2_{0\cdot 123} = 0,7946$$

Resultatet av de videre beregningene er gitt i Tab.J.3.

Tabell J.3.*

Uavhengig variabel	Kvadratsum	f	V	F
y_1	0,1301.14120 = 1837,0120	1	1837,01	129,21
y_2	0,3041.13120 = 4293,8920	1	4293,89	302,03
y_3	0,3604.14120 = 5088,8480	1	5088,85	357,94
Rest	0,2054.14120 = 2900,2480	204	14,22	

Av Tabell III (P=0,05) ser vi at til f=1 frihetsgrad for teller-variansen og $f = \frac{204}{3} = 68$ frihetsgrader for nevnervariansen svarer en kritisk verdi for F lik ca.4. Bruker vi Tabell IV (P=0,025) er den kritiske verdi ca.5,28. De varianskvotienter vi har fått for dette eksemplet er meget større. Vi forkaster derfor alle tre null-hypotesene.

Spørsmålet blir imidlertid hvordan vi skal gå fram dersom en av de testede null-hypotesene ikke kan forkastes. La oss for å forklare dette, bruke vårt eksempel på nytt med følgende forandring : la $R^2_{0\bullet 1}$ og $R^2_{0\bullet 123}$ være uforandret, og la $R^2_{0\bullet 12} = 0,1331$ (i stedet for 0,4342). Vi vil da få de kvadratsummer, varianser og F-verdier som er gitt i Tab. J.4.

Tabell J.4.

Uavhengig variabel	Kvadratsum	f	V	F	ny F
y_1	1837,0120	1	1837,01	129,21	127,98
y_2	42,3600	1	42,36	2,98	
y_3	9340,3800	1	9340,38	656,99	650,71
Rest	2900,2480	204	14,22		
Ny Rest	2942,6080	205	14,35		

* For testingen er det naturligvis ikke nødvendig at det multipliseres med $S_{00} = 14120$.

Vi ser nå at F-verdien (2,98) for y_2 er så liten at vi kan ikke forkaste null-hypotesen $\beta_{02} = 0$. Hvis vi vil ta avgjørelsen på forkastningsnivået P = 0,05, må vi jo også nå kreve en F-verdi lik eller større enn 4. Det at $\beta_{02} = 0$ ikke kan forkastes betyr naturligvis ikke at $\beta_{02} = 0$. Det betyr bare at y_2 ikke gir et påviselig bidrag til estimeringen av x_0. Vi er imidlertid da stilt overfor en ny situasjon, nemlig den å estimere regresjonsfunksjonen for x_0 m.h.p. y_1 og y_3. På grunn av ortogonaliteten y-ene imellom er bidragene til reduksjonen som skyldes y_1 og y_3 uforandret. Det at vi tar ut y_2 betyr imidlertid at bidraget fra den må regnes med i det som vi kan kalle en ny Rest. Kvadratsummen for denne blir derfor 42,3600 + 2900,2480 = 2942,6080. Samtidig vil antall frihetsgrader for Rest (altså ny Rest) bli økt med en, til f = 205. Vi får derfor en ny restvarians lik 2942,6080/205 = 14,3542 og nye F-verdier for y_1 og y_3. Disse nye F-verdiene må så sammenlignes med den verdi vi finner i F-tabellen (f.eks.Tabell III) for f = 1 frihetsgrad for hver av de to tellervariansene og f = $\frac{205}{2}$ = 102,5 frihetsgrader for nevnervariansen. Av Tabell III ser vi at denne verdien ligger mellom 4,00 og 3,92. De to F-verdiene vi har funnet, er betydelig større. Konklusjonen blir derfor at null-hypotesene $\beta_{01}=0$ og $\beta_{02} = 0$ forkastes.

I regresjonsfunksjonen vil nå y_2 mangle, og den estimerte funksjon blir

$$\hat{r}(y_1,y_3) = \bar{x}_0 + b_{01}\, y_1 + b_{03}\, y_3$$

hvor

$$y_1 = x_1 - \bar{x}_1 \quad \text{og} \quad y_3 = x_3 - \bar{x}_3 - b_{32\cdot 1}(x_2 - \bar{x}_2) - b_{31\cdot 2}(x_1 - \bar{x}_1)$$

Her er da $b_{01} = S_{01}/S_{11}$. For å kunne beregne b_{03} må vi først

beregne de to regresjonskoeffisientene $\beta_{32\cdot 1}$ og $\beta_{31\cdot 2}$. Disse finnes av ligningene

$$S_{11}\, b_{31\cdot 2} + S_{12}\, b_{32\cdot 1} = S_{31}$$
$$S_{12}\, b_{31\cdot 2} + S_{22}\, b_{32\cdot 1} = S_{32}$$

Videre har en at

$$\Sigma y_{3i}x_{0i} = S_{03} - S_{02}\cdot b_{32\cdot 1} - S_{01}\cdot b_{31\cdot 2}$$
$$\Sigma y_{3i}^2 = S_{33}\cdot(1-R_{3\cdot 12}^2)$$

og til slutt

$$b_{03} = \frac{\Sigma y_{3i}x_{0i}}{\Sigma y_{3i}^2}$$

Den testingsteknikk som er antydet med dette, er en teknikk som tar sikte på en trinnvis eliminasjon av y-ene, en om gangen. La oss nå tenke oss at vi til å begynne med har m mulige uavhengige variabler ($x_1, x_2, \ldots\ldots x_m$) og dermed også m y-variabler. Vi skal senere komme inn på hvilket av de mange sett av y-variabler vi bør velge. Her vil vi gå ut fra at valget er gjort.

La oss så tenke oss at det er utført testing av null-hypoteser i p trinn og at resultatet er at p y-variabler er tatt ut. Vi står da overfor den situasjon at vi har (m-p) y-variabler igjen og dermed (m-p) varianser, hver med f=1 frihetsgrad. Kvadratsummen for den Rest vi må bruke på dette trinnet er da lik den opprinnelige $S_{00}(1-R_{0\cdot 123\ldots m}^2)$ pluss summen av kvadratsummene for de uttatte y-variablene. Antall frihetsgrader for denne Rest er derfor $f = n-1-m+p$. Vi beregner så varianskvotienten for hver av de resterende (m-p) y-ene og sammenligner med den verdien vi finner i f.eks. Tabell III for f=1 frihetsgrad for hver av tellervariansene og $f = \frac{n-1-m+p}{m-p}$ frihetsgrader for den felles nevnervarians. Resultatet på dette trinnet er da

ett av følgende to :

1) Alle F-verdiene er så store at vi aksepterer dem som signifikante. I så fall beholder vi de (m-p) y-variablene i regresjonsfunksjonen. Analysen er dermed avsluttet.

2) I det minste en av de (m-p) F-verdiene er så liten at vi ikke kan akseptere den som signifikant. I så fall tar vi ut den y-variable som har den minste F-verdien og går til neste trinn med (m-p-1) y-variabler.

Det er klart at denne fremgangsmåten kan bli meget arbeidskrevende. Er antallet av potensielle uavhengige variabler (m) større enn 4-5, blir arbeidet i praksis så omfattende at en bør benytte elektronisk regnemaskin.

Spørsmålet om hvordan en skal velge settet av y-variabler er vanskelig og noe endelig svar har vi vel ennå ikke fått. Svaret vil også avhenge av bl.a. hva hensikten med regresjonsanalysen er. Hvis hensikten utelukkende er å estimere forventningen for den avhengige variable, kan vi legge visse praktiske og økonomiske kriterier til grunn for valget. Dette vil bl.a. føre til krav om at antallet av de x-variabler som tas med i regresjonsfunksjonen skal være minst mulig og at de skal være enkle å skaffe seg observasjoner av. Selv om det ikke er noe fullt samsvar mellom antall x-variabler og antall y-variabler, er det i regelen slik at jo færre y-variabler, jo mindre er antall x-variabler.

En må også ta et visst hensyn til at en ny y-variabel som tas med, vil føre med seg restriksjoner når det gjelder gyldighetsområdet for bruken av regresjonsfunksjonen. Dette viser også at det vil være fornuftig å finne fram til det settet som inne-

holder det minste antall y-variabler. Det er iallfall god utsikt til at en vil oppnå dette på følgende måte.

Vi vil nå gå ut fra at vi ikke har informasjoner som kan gi holdepunkt for valget av sett av y-variabler. Det er da praktisk å gå fram trinn for trinn og velge på hvert trinn den nye y-variable en vil ta med. Et prinsipp en kan gjøre bruk av, er å velge som ny y-variabel, blant dem en har til rådighet, den som gir det største reduksjonsbidraget. Dette vil da si at en på første trinn må velge som y_1 den x-variable som er sterkest korrelert med x_0.

For det eksemplet vi har benyttet foran, har vi at

$$R^2_{0\cdot 1} = 0{,}1301, \quad R^2_{0\cdot 2} = 0{,}3306 \quad \text{og} \quad R^2_{0\cdot 3} = 0{,}0083$$

Vi setter derfor $y_1 = x_2 - \bar{x}_2$. Reduksjonsbidraget fra denne blir så $R^2_{0\cdot 2} S_{00} = 0{,}3306 \cdot 14120 = 4668{,}0720$.

På 2.trinn har vi så valget mellom

1) $y_2 = x_1 - \bar{x}_1 - b_{12}(x_2 - \bar{x}_2)$ hvor $b_{12} = S_{12}/S_{22}$

2) $y_2 = x_3 - \bar{x}_3 - b_{32}(x_2 - \bar{x}_2)$ hvor $b_{32} = S_{32}/S_{22}$

Reduksjonsbidragene fra disse er

1) $(R^2_{0\cdot 12} - R^2_{0\cdot 2})\, S_{00}$ og 2) $(R^2_{0\cdot 23} - R^2_{0\cdot 2})\, S_{00}$

Det som blir avgjørende for valget, er derfor hvilken av $R^2_{0\cdot 12}$ og $R^2_{0\cdot 23}$ er størst. For vårt eksempel har vi at $R^2_{0\cdot 12} = 0{,}4342$ og $R^2_{0\cdot 23} = 0{,}3908$. Vi velger derfor det første alternativet som gir reduksjonsbidraget

$$(0{,}4342 - 0{,}3306) \cdot 14120 = 1462{,}8320$$

Da vi i dette eksemplet har bare tre uavhengige variabler, blir naturligvis

$$y_3 = x_3 - \bar{x}_3 - b_{31\cdot 2}(x_1 - \bar{x}_1) - b_{32\cdot 1}(x_2 - \bar{x}_2)$$

med det reduksjonsbidraget som er gitt i Tab.J.3.

For Rest har vi naturligvis samme resultat som tidligere nemlig $(1-R^2_{0\cdot 123})\ S_{00} = 2900{,}2480$. Utføres så testingen slik som vist i Tab.J.3 (for et annet sett y-variabler), vil vi finne at null-hypotesene $\beta_{01} = 0$, $\beta_{02} = 0$ og $\beta_{03} = 0$ må forkastes. Resultatet er at alle x-variablene skal tas med i regresjonsfunksjonen. Resultatet er m.a.o. det sammen som det som ble funnet i avsnitt J.2.

Program for denne analyseteknikken er utarbeidet eller kan bli utarbeidet for alle elektroniske regnemaskiner. Må arbeidet utføres på bordmaskin, har denne teknikken en vesentlig fordel fremfor andre. Det er nemlig mulig å undersøke på hvert trinn om det har en hensikt å fortsette analysen.

La oss, for å forklare dette, ta for oss vårt eksempel på nytt. Valget av $y_1 = x_2 - \bar{x}_2$ fører til reduksjonen $R^2_{0\cdot 2}S_{00}$. Den reduksjon vi vil oppnå ved å sette inn y_2 og y_3 - det blir det samme hvilket sett vi bruker - må da bli

$$(R^2_{0\cdot 123} - R^2_{0\cdot 2})\ S_{00} = (0{,}7946 - 0{,}3306) \cdot 14120 = 6551{,}6800.$$

Stilt sammen i det vanlige variansanalyseskjema får vi den oppstillingen som er vist i Tab.J.5.

Tabell J.5.

Uavhengig variabel	Kvadratsum	f	V	F
y_1	4668,0720	1	4668,07	328,27
y_2 og y_3	6551,6800	2 (1)	6551,68	460,74
Rest	2900,2480	204	14,22	

Til kvadratsummen for y_2 og y_3 svarer f = 2 frihetsgrader. En må imidlertid regne med den muligheten at det aller meste av dette reduksjonsbidraget skyldes en av de to vari-

ablene, enten y_2 eller y_3. Derfor bør en bruke f = 1 frihetsgrad for dette sammensatte reduksjonsbidraget, slik som vist i tabellen. De to F-verdiene skal så sammenlignes med den verdi en finner i f.eks. Tabell III for f=1 frihetsgrad for hver av tellervariansene og $f = \frac{204}{2} = 102$ for den felles nevnervarians. Velges P = 0,05, ser en av Tabell III at den kritiske verdien for F er lik ca. 4. Resultatet viser da at en bør utvide analysen, men det viser heller ikke noe mer enn det. Hadde vi for y_2 og y_3 fått $F \approx 2$, ville dette være nok til å fortelle oss at en videregående analyse ikke har noen hensikt.

Det er nevnt i innledningen til avsnitt J.2 at det ikke alltid er slik at de random variabler vi benytter som uavhengige variabler i regresjonsfunksjonen (altså x_1, x_2, x_3 ....) er direkte observerte random variabler. Situasjonen kan f.eks. være slik at en som uavhengige variabler vil bruke de observerte random variablene x_1 og x_2 og at en så som en tredje variabel vil bruke $x_3 = x_1 \cdot x_2$. I det eksemplet vi har benyttet hvor x_0 er volum og x_2 er middeldiameter, kunne en overveie å bruke x_2^2 i stedet for x_2 eller i tillegg til x_2. Siden x_0 er volum, ville et slikt valg kanskje vise seg å være vellykt. I litteraturen kan en finne mange eksempler som viser at slike uavhengige variabler har vært benyttet.

Dette gir da også svar på det spørsmålet som ble stilt i avsnitt J.1 om hvordan vi skal kunne ta standpunkt til om vi kan nøye oss med en lineær regresjonsfunksjon. Vi forutsatte da at regresjonsfunksjonens form ikke var kjent på forhånd.

Vi må da tenke oss at vi har en observert random variabel (x_1) som vi vil bruke som uavhengig variabel. For å undersøke om vi kan nøye oss med en lineær funksjon av x_1, kan vi så ta med $x_2 = x_1^2$, $x_3 = x_1^3$ o.s.v. og prøve om disse gir signifikante bidrag til estimeringen av forventningen for x_0 i tillegg til det bidraget som x_1 gir. I praksis vil dette som oftest si at vi i tillegg til x_1 også tar med $x_2 = x_1^2$. Det vi da må undersøke er om $y_1 = x_1 - \bar{x}_1$ og $y_2 = x_2 - \bar{x}_2 - b_{21}(x_1 - \bar{x}_1)$ samtidig gir signifikante reduksjonsbidrag. Til dette kan vi bruke den teknikken vi har beskrevet i dette avsnittet.

La oss ta for oss eksemplet i Tab.B.8 (side 44). For dette eksemplet finner vi at $R^2_{0\cdot 1} = 0{,}8680$ og $R^2_{0\cdot 12} = 0{,}8872$. Kvadratsummen for x_0 er $S_{00} = 1{,}0031$. Vi har derfor følgende resultat :

Uavhengig variabel	Kvadratsum	f	V	F
y_1	0,8707	1	0,8707	92,38
y_2	0,0193	1	0,0193	2,05
Rest	0,1131	12	0,0094	

Vi konstatere så at F = 2,05 ikke er signifikant. Dette vil så si at vi ikke oppnår signifikant bedre estimering av forventningen for x_0 ved å ta med x_1^2 enn ved å bruke x_1 alene.

J.4. Om $\hat{r}(x_1,x_2, \ldots)$ som estimator av $r(x_1,x_2, \ldots)$.

La oss tenke oss at vi har m potensielle uavhengige variabler, x_j $(j = 1,2,3,\ldots m)$, i en regresjonsfunksjon for x_o m.h.p. x_j. Innfører vi et sett ortogonale y-variabler, vil regresjonsfunksjonen være

$$r(y_1,y_2, \ldots y_m) = \alpha_o + \beta_{o1}\, y_1 + \beta_{o2} y_2 + \ldots + \beta_{om}\, y_m$$

Forutsettes det at alle y-variabler gir påviselig bidrag til estimeringen av forventningen for x_o, er den estimerte regresjonsfunksjonen

$$\hat{r}(y_1,y_2, \ldots y_m) = \bar{x}_o + b_{o1}y_1 + b_{o2}y_2 + \ldots + b_{om}y_m$$

Spørsmål som så melder seg, er om $\hat{r}$ er en forventningsrett estimator av r og hvilken presisjon $\hat{r}$ i tilfelle har.

La oss om e_i i modellen

$$x_{oi} = r(y_{1i},y_{2i},\ldots y_{mi}) + e_i$$

forutsette følgende :

1) e_i er n innbyrdes uavhengige verdier av en random variabel e med forventningen $(E(e_i) = 0$.
2) var(e) er uavhengig av y-variablene.
3) e er en normal random variabel.

Er disse forutsetningene tilfredsstilt, kan det vises

1) at $\bar{x}_o$ og b_{oj} $(j=1,2,..m)$ er innbyrdes uavhengige,
2) at $E(\bar{x}_o) = \alpha_o$ og $E(b_{oj}) = \beta_{oj}$ for alle j,
3) at for alle j er b_{oj} en normal random variabel,
4) at
$$\mathrm{var}(b_{oj}) = \frac{\mathrm{var}(e)}{\Sigma\, y_{ji}^2}$$
5) at
$$V_R = \frac{1}{n-m-1}\, \Sigma(x_{oi}-\bar{x}_o-b_{o1}y_{1i}-b_{o2}y_{2i}-\ldots-b_{om}y_{mi})^2$$
er en forventningsrett estimator av var(e).

Vi vil så tenke oss at vi velger et sett av verdier av $x_1, x_2, x_3, \ldots x_m$ og beregner de tilsvarende verdier av y-ene. La os betegne resultatene med $y_1 = c_1$, $y_2 = c_2$, ... $y_m = c_m$. Disse verdier må oppfattes som gitte tall.

Estimatoren av $r(c_1, c_2, \ldots c_m)$ er da

$$\hat{r}(c_1, c_2, \ldots c_m) = \bar{x}_o + b_{o1}c_1 + b_{o2}c_2 + \ldots + b_{om}c_m$$

Under de forutsetningene vi har nevnt foran, er denne estimatoren en sum av ikke-korrelerte ledd, og derfor er

$$E[\hat{r}(c_1, c_2, \ldots c_m)] = E(\bar{x}_o) + c_1E(b_{o1}) + c_2E(b_{o2}) \ldots c_mE(b_{om})$$

$$= \alpha_o + \beta_{o1}c_1 + \beta_{o2}c_2 + \ldots + \beta_{om}c_m$$

$$= r(c_1, c_2, \ldots c_m)$$

og

$$\text{var}[\hat{r}(c_1, c_2, \ldots c_m)] = \text{var}(\bar{x}_o) + c_1^2\text{var}(b_{o1}) + c_2^2\text{var}(b_{o2}) + \ldots + c_m^2\text{var}(b_{om})$$

$$= \text{var}(e)\left\{ \frac{1}{n} + \frac{c_1^2}{\Sigma y_{1i}^2} + \frac{c_2^2}{\Sigma y_{2i}^2} + \ldots + \frac{c_m^2}{\Sigma y_{mi}^2} \right\}$$

I praksis erstattes så var(e) med V_R, og den siste formelen blir da formelen for $s_{\hat{r}}^2$.

Under de samme forutsetningene er da

$$t = \frac{\hat{r}(c_1, c_2, \ldots c_m) - r(c_1, c_2, \ldots c_m)}{s_{\hat{r}}}$$

en Students t med f = n-m-1 frihetsgrader. Dette vil da si at vi kan estimere forventningsrett den betingede forventningen for x_o, altså $E(x_o; c_1, c_2, \ldots c_m)$ ved $\hat{r}(c_1, c_2, \ldots c_m)$, og vi kan beregne konfidensgrensene for denne forventningen ved

$$\hat{r}(c_1, c_2, \ldots c_m) \pm a.s_{\hat{r}}$$

hvor verdien av a finnes i Tabell I for f = n-m-1 frihetsgrader og den P-verdi en ønsker å bruke.

Hvis noen av y-variablene ikke gir påviselig bidrag til estimeringen av forventningen for x_o, bør vi naturligvis ta ut disse variablene. Dette vil si at i formelen for $\hat{r}(c_1,c_2,..c_m)$ og i formelen for $s_{\hat{r}}$ settes $c = 0$ for disse y-ene. Dette fører imidlertid til at kvadratsummen for V_R må settes lik den opprinnelige $(1-R^2_{o\cdot 12..m})\, S_{oo}$ pluss kvadratsummene for de y-ene som er tatt ut. Samtidig må så antall frihetsgrader for Rest (eller den nye V_R) økes med antallet av de uttatte y-variablene.

La oss anvende dette for det eksemplet vi har gjort bruk av foran. Vi skal da benytte det siste settet av y-variabler, altså

$$y_1 = x_2-\bar{x}_2$$

$$y_2 = x_1-\bar{x}_1- b_{12}(x_2-\bar{x}_2)$$

$$y_3 = x_3-\bar{x}_3- b_{31\cdot 2}(x_1-\bar{x}_1) - b_{32\cdot 1}(x_2-\bar{x}_2)$$

Vi må da først beregne regresjonskoeffisientene i de to siste funksjonene. De tallene vi har bruk for til dette, vil man finne i de tre ligningene vi benyttet i avsnitt J.2 (side 217). Vi finner da først at

$$b_{12} = \frac{S_{12}}{S_{22}} = \frac{323}{6738} = 0{,}04794$$

Til beregning av regresjonskoeffisientene $b_{32\cdot 1}$ og $b_{31\cdot 2}$ har vi ligningene

$$S_{11}\, b_{31\cdot 2} + S_{12}\, b_{32\cdot 1} = S_{13}$$
$$S_{12}\, b_{31\cdot 2} + S_{22}\, b_{32\cdot 1} = S_{23}$$

som gir $b_{31\cdot 2} = 0{,}16876$ og $b_{32\cdot 1} = 0{,}03925$.
En finner så videre at $R^2_{1\cdot 2} = 0{,}00474$ og $R^2_{3\cdot 12} = 0{,}46224$ (jfr.J.2,side 218). Vi har derfor at

$$\Sigma y_{1i}^2 = S_{22} = 6738$$

$$\Sigma y_{2i}^2 = S_{11}(1-R_{1\cdot 2}^2) = 3251,51442$$

$$\Sigma y_{3i}^2 = S_{33}(1-R_{3\cdot 12}^2) = 125,29808$$

Videre finnes at

$$\Sigma y_{1i}\cdot x_{oi} = S_{o2} = 5608$$

$$\Sigma y_{2i}\cdot x_{oi} = S_{o1} - b_{12}\cdot S_{o2} = 2181,15248$$

$$\Sigma y_{3i}\cdot x_{oi} = S_{o3} - b_{31\cdot 2}\cdot S_{o1} - b_{32\cdot 1}\cdot S_{o2} = -798,57600$$

Vi har derfor at

$$b_{o1} = \frac{5608}{6738} = 0,83229$$

$$b_{o2} = \frac{2181,15248}{3251,51442} = 0,67081$$

$$b_{o3} = \frac{-798,57600}{125,29808} = -6,37341$$

Fra før (avsnitt J.2,side 217) har vi at $\bar{x}_o = 19,9364$ og derfor er

$$\hat{r}(y_1,y_2,y_3) = 19,9364 + 0,83229\ y_1 + 0,67081\ y_2 - 6,37341\ y_3$$

Fører vi så x_1,x_2 og x_3 inn i denne funksjonen, vil vi få samme resultat som i avsnitt J.2 (side 218). At resultatet må bli det samme kommer naturligvis av at alle y-variabler gir påviselige bidrag til estimeringen.

Innsettes i formlene for y_1,y_2 og y_3 $\bar{x}_1 = 14,4327$, $\bar{x}_2 = 17,6913$ og $\bar{x}_3 = 2,6749$, finner vi at

$$y_1 = x_2 - 17,6913$$

$$y_2 = x_1 - 0,04794\ x_2 - 13,5846$$

$$y_3 = x_3 - 0,16876\ x_1 - 0,03925\ x_2 + 0,4551$$

La oss så som eksempel velge $x_1=12$, $x_2=20$ og $x_3=3$. Ved innsetting i formlene for y-ene vil vi da finne at

$$c_1 = 2,3087, \quad c_2 = -2,5434 \quad \text{og} \quad c_3 = 0,6450$$

som gir

$$\hat{r}(c_1,c_2,c_3) = 16,0410$$

og

$$\frac{1}{n} + \frac{c_1^2}{\Sigma y_{1i}^2} + \frac{c_2^2}{\Sigma y_{2i}^2} + \frac{c_3^2}{\Sigma y_{3i}^2} = 0,01091$$

Fra før har vi at $V_R = 14,22$, og derfor er

$$s_{\hat{r}} = \sqrt{14,22 \cdot 0,01091} = 0,394$$

Velges så konfidensgrenser som svarer til konfidenssannsynligheten 0,95, skal vi etter Tabell I for f = 204 frihetsgrader sette a = 1,97 (ca.). Altså er $a \cdot s_{\hat{r}} = 0,7762$. Konfidensgrensene blir derfor

$$\hat{r}(c_1,c_2,c_3) \mp a \cdot s_{\hat{r}} = 16,0410 \mp 0,7762 = \begin{cases} 15,26 \\ 16,82 \end{cases}$$

Når en skal bruke denne metoden, må en passe på at det ikke brukes andre verdisammensetninger av $x_1, x_2, x_3, \ldots$ enn slike som kan forekomme i det universet som er representert av samplet. I praksis vil dette bety at observasjonene av $x_1, x_2, x_3, \ldots$ i samplet må vise at den verdisammensetningen en ønsker å bruke, er realistisk . Vi vil imidlertid i denne sammenhengen nøye oss med å vise til det som er sagt om saken i avsnitt J.1, side 212-213.

J.5. Om bruk av en estimert regresjonsfunksjon.

Regresjonsanalyse er tatt i bruk som metode for flere formål. Et av disse er å estimere forventningen for en random variabel x_o ved hjelp av observasjoner av andre random variabler. Grunnen kan f.eks. være at x_o er vanskelig å observere direkte. Volumet av de grantrær som finnes på en flate i skogen, er et eksempel på en slik random variabel. Som vi har vist kan en da observere et antall andre variabler som er korrelerte med x_o og så bruke regresjonsfunksjonen for x_o m.h.p. disse andre random variabler som estimator av forventningen for x_o.

For å kunne estimere regresjonsfunksjonen må vi naturligvis ha et sampel av gjentak og for hvert gjentak observasjoner av x_o og et antall potensielle uavhengige variabler. Hensikten med regresjonsanalysen er da å estimere en funksjon av de uavhengig variablene som gir en så presis estimering av forventningen for x_o at vi kan bruke den i nye tilfelle. Dette vil si at vi bruker den estimerte regresjonsfunksjonen som grunnlag for en prognose. Formålet med selve analysen er da å finne ut om alle de random variabler vi under planleggingen av undersøkelsen har regnet med som potensielle uavhengig variabler, skal være med og hvilke koeffisienter (eller vekter) de skal ha. Dessuten tar analysen sikte på å finne ut hvilken form for funksjonen en bør bruke. Det er f.eks. ikke alltid at en lineær funksjon er tilfredsstillende. Den funksjonen vi stanser ved til slutt, kan f.eks. inneholde annen grads ledd eller/og ledd

av høyere grad.

Vi har nevnt foran at det oftest er nødvendig å passe på at de verdier av de uavhengige variabler som vi bruker i funksjonen, er slike som samplet viser er mulige kombinasjoner av verdier av de variable. Dette kan resultere i at funksjonen må karakteriseres som lite tilfredsstillende fordi det kan være andre kombinasjoner en ønsker å bruke den for. Det kan da bli spørsmål om det er forsvarlig å bruke funksjonen for andre kombinasjoner.

Svar på dette spørsmål kan en neppe gi uten at en har informasjoner i tillegg til dem samplet kan gi. Men det som er viktigst, er at en under planleggingen av undersøkelsen tar sikte på a få et så representativt sampel som mulig. Dette betyr at den som er ansvarlig for undersøkelsen, har visse opplysninger om det universet som samplet skal representere i egenskap av et random sampel.

La oss på ny ta for oss eksemplet fra foregående avsnitt og tenke oss at vi har et kart som visergrensene for granskogområdene i et større geografisk område. På dette kartet kan vi så tenke oss granskogområdene inndelt i N flater av den størrelsen vi ønsker å bruke. Disse flatene kan nummereres med nummerne 1 - N, og vi kan så ved hjelp av et eller annet teknisk hjelpemiddel ta ut et random sampel på n flater. På disse flatene som vi forutsetter at vi kan lokalisere i skogen, tar vi så observasjoner av x_0, x_1, x_2, x_3, ... og benytter disse til estimering av regresjonsfunksjonen for x_0 m.h.p. de andre variablene. Samplet er da en random representant for et konkret univers som består av de N flatene. I mange tilfelle vil imidlertid også

disse N flatene bli oppfattet som et sampel som representerer et abstrakt univers. Men siden samplet på n flater er tatt ut ved loddtrekning blant de N flatene, vil det også representere det abstrakte universet i egenskap av et random sampel.

I teorien er dette meget enkelt. Men enhver med litt kjennskap til skog, vil vel uten videre innse at å praktisere teorien kan by på mange vanskер. For bare å nevne ett forhold : granskogen innen området skal deles i N flater, og dette kan naturligvis gjøres på mange forskjellige måter.

At samplet er tatt ut slik vi har beskrevet, gir ingen sikker garanti for at det vil omfatte alle de kombinasjoner av verdier av de random variablene som finnes på de N flatene. Når vi skal benytte den estimerte regresjonsfunksjonen til estimering av forventningen for x_o på en av de N-n flatene som ikke er med i samplet, kan vi derfor finne for enkelte flater at verdien av f.eks. x_2 faller utenfor variasjonsområdet for x_2 i samplet. Som vi har nevnt foran, kan vi da strengt tatt ikke bruke regresjonsfunksjonen for disse flatene.

En utvei en kan ty til i en slik situasjon, er å ta et random sampel blant de (N-n) gjentak og addere dette til det første samplet. En må da gjøre regnearbeidet om igjen og får naturligvis en ny regresjonsfunksjon. Dette gir imidlertid fremdeles ingen sikker garanti for at det innen det geografiske området ikke finnes flater som funksjonen ikke bør brukes for.

Dette viser at en ikke må utgi en estimert regresjonsfunksjon som et teknisk middel til estimering av forventningen for den avhengig variable, uten at det følger med en anvisning på hvilke verdisammensetninger av de uavhengig variabler den kan brukes for. Å gi en slik anvisning er imidlertid ikke enkelt. Som oftest er det nemlig slik at når verdien av x_1 er gitt eller valt, er verdiområdet for x_2 kortet inn i sammenligning med hele verdiområdet. I praksis blir det vel sjelden tatt noe hensyn til dette, men det burde bli det. Denne saken er imidlertid så pass komplisert at vi skal nøye oss med disse merknader.

Alt dette viser at det er viktig å komme fram til en estimeringsfunksjon med så lite antall uavhengige variabler som mulig og som likevel gir estimater med tilstrekkelig presisjon. Og dette betyr da igjen at det er viktig at en er meget omhyggelig med valget av uavhengige variabler.

Et annet formål med en regresjonsanalyse er å undersøke hvilke variabler ($x_1, x_2, x_3, \ldots$) det er som har betydning for verdien av x_0. For at en slik undersøkelse skal ha mening, må da $x_1, x_2, x_3, \ldots$ i relasjon til x_0 kunne oppfattes som årsaksvariabler. Det finnes naturligvis situasjoner hvor de kan oppfattes på denne måten. La oss f.eks. tenke oss at x_0 er mengde avling pr. arealenhet for en bestemt sort poteter. Random variabler som vi da kan oppfatte som årsaksvariabler er gjennomsnitstemperaturen i vekstperioden (x_1), nedbørsmengden i vekstperioden (x_2) og en karakteristikk av dyrkingsjorda (x_3). I et slikt

tilfelle kan en ved hjelp av regresjonsanalyse skaffe seg holdepunkter for hvilken betydning x_1, x_2, og x_3 har for avlingsmengden.

Også da må en benytte seg av ortogonale funksjoner, altså et sett av y-variabler slik disse ble definert i avsnitt J.3. Men det settet en skal bruke, må velges etter andre kriterier enn når det gjelder a finne en god esterimeringsfunksjon. Den regresjonsfunksjonen våre beregninger som oftest
fører til, er nemlig/avhengig av hvilket sett av y-variabler vi benytter oss av. Benytter vi settet

$$y_1 = x_1 - \bar{x}_1$$

$$y_2 = x_2 - \bar{x}_2 - b_{21}(x_1 - \bar{x}_1)$$

vil vi kanskje finne at y_2 ikke gir påviselig bidrag til reduksjonen av kvadratsummen (S_{oo}) for x_o. Og konklusjonen vil bli at x_2 ikke er en påviselig årsaksvariabel eller at effekten av den er tatt med i x_1.

Benytter vi imidlertid settet

$$y_1 = x_2 - \bar{x}_2$$

$$y_2 = x_1 - \bar{x}_1 - b_{12}(x_2 - \bar{x}_2)$$

kan vi komme til samme konklusjon med den forandring at x_1 og x_2 bytter rolle.

Det er derfor nødvendig at en på forhånd har bestemt seg for i hvilken rekkefølge årsaksvariablene skal tas inn når en danner de ortogonale funksjonene. En må m.a.o. ha en hypotese om hvordan de potensielle årsaksvariablene virker inn på verdien av x_o. Hvordan en skal gå fram når en har dette problemet å hanskes med, kan det neppe gis regler for. Problemet oppstår vel også som oftest i en serie

av undersøkelser slik at planen for undersøkelsen avhenger av hvilke resultater en har kommet til i de foregående undersøkelser.

Ofte er det imidlertid ikke mulig å avgjøre hvilke variabler kan oppfattes som årsaksvariabler og hvilke en kan oppfatte som effektvariabler. Som vi allerede har nevnt, er det ofte slik at samvariasjonen mellom random variabler ikke beror på direkte årsakssammenheng. I regelen er det kanskje slik at samvariasjonen skyldes bakenfor liggende årsaksvariabler (eller faktorer) som en ikke er i stand til å identifisere og som en derfor heller ikke kan skaffe seg observasjoner for. Spørsmålet kan da bli om det er mulig å finne ut hvor mange slike bakenfor liggende årsaksvariabler en bør regne med og om observasjonene indikerer hvilke variabler dette er. Metoder for slike undersøkelser ligger utenfor rammen for denne fremstilling.

I avsnitt J.1 (side 213-214) er det nevnt at det kan være fristende, når det er på det rene at x_1 er årsaksvariabel og x_0 effektvariabel, å oppfatte regresjonsfunksjonen for x_0 m.h.p. x_1 som uttrykk for den effekt en vil oppnå på x_0 ved aktivt a forandre x_1. Vi advarte der mot å bruke regresjonsfunksjonen som uttrykk for dette. Har vi imidlertid mange uavhengige variabler i regresjonsfunksjonen som alle er årsaksvariabler i relasjon til x_0, og x_1 ikke har betydning for verdiene av $x_2, x_3, \ldots$, kan regresjonsfunksjonen brukt med forsiktighet kunne gi oss opplysning om hva som vil skje med x_0 som en følge av at vi forandrer verdien av x_1.

I dette tilfelle må nemlig både situasjonen x_1=a og situasjonen x_1=b sammen med samme sett verdier av $x_2, x_3, \ldots$ finnes realisert i det observasjonsmateriale vi bygger på, forutsatt naturligvis at a og b er verdier innenfor variasjonsområdet for x_1. Regresjonsfunksjonen skulle derfor kunne gi beskjed om hva som vil hende med x_0 også når en aktivt forandrer verdien av x_1. Risikoen for konklusjonsfeil skyldes da at en neppe noen gang kan være helt sikker på at det ikke finnes random variabler som har betydning for x_0, som påvirkes av forandringer i x_1, og som en ikke har observasjoner for.

I aktuelle tilfelle vil en som oftest finne at det er korrelasjon mellom årsaksvariablene, slik at en forandring av x_1 følges av forandring i en eller flere av de andre. Det spørsmål som da melder seg i praksis, er om det er mulig å finne ut hvordan de årsaksvariablene som er avhengige av x_1, skal endres aktivt samtidig med x_1 slik at den totale forandringen fører til en realistisk situasjon. Å takle dette problemet er naturligvis meget vanskelig. Her må vi nøye oss med å nevne det.

J.6. Regresjonsfunksjonen som beskrivelse av effekten av en kvantitativ forsøksfaktor.

I avsnitt H.1 (side 158) er det nevnt at hvis forsøksfaktoren er kvantitativ, kan en være interessert i å skaffe seg en matematisk beskrivelse av hvordan effekten avhenger av forsøksfaktoren. Som eksempel kan vi tenke oss et forsøk med smågriser og hvor forsøksleddene er et standardfor tilsatt valte mengder A-vitamin. Hensikten kan da være å gi en beskrivelse av hvordan forventningen for vekstøkningen avhenger av vitaminmengden.

Vi skal nå betegne effekten (f.eks. vektøkningen) med x_0 og forsøksfaktoren (vitaminmengden) med x_1. Den betingede forventningen er i avsnittene H.2 og H.3 betegnet med μ_j eller med $\mu + a_j$. Oppgaven går derfor ut på å beskrive den eventuelle avhengighet mellom μ_j (som avhengig variabel) og x_{1j} som uavhengig variabel, d.v.s. å estimere en funksjon

$$\mu_j = r(x_{1j})$$

hvor $j=1,2,3,....k$ og k er antall forsøksledd.

I noen tilfelle har en kanskje visse informasjoner om den matematiske form for $r(x_1)$. Men i regelen er funksjonsformen ukjent. Som oftest må en anta at den er komplisert og den kan være forskjellig fra det ene tilfelle til det andre. Vi kan imidlertid forutsette at funksjonen har slike egenskaper at den kan erstattes av en rekkeutvikling, slik at vi kan sette

$$r(x_1) = \beta_0 + \beta_{01\cdot 23\cdots}x_1 + \beta_{02\cdot 13\cdots}x_1^2 + \cdots\cdots$$

Det innsees vel uten videre at hvis vi tar med alle

ledd fra x_1 til x_1^{k-1}, kan koeffisientene i $r(x_1)$ tilpasses slik at funksjonen gir en nøyaktig beskrivelse av gjennomsnittene $\bar{x}_{0j}$. Antall ledd i rekkeutviklingen, konstantleddet regnet med, må derfor være mindre eller i høyden lik antall forsøksledd (k).

Den variable x_1 er naturligvis i dette tilfelle ikke en random variabel. Verdiene av x_1 (altså x_{1j}) er kvantiteter som er vedtatt brukt som forsøksledd under planleggingen av forsøket. Men siden ingen er i stand til å gjenta en handling eksakt, må en regne med at x_1 også inneholder et random element. En må imidlertid kunne gå ut fra at i tilfredsstillende utførte forsøk har dette random element liten betydning. Vi kan derfor si at den uavhengige variable i disse tilfelle er en tilnærmet kontrollert variabel.

For at det skal bli overensstemmelse med de symboler vi har brukt foran, skal vi betegne x_1^2 med x_2, x_1^3 med x_3 o.s.v. Vi skal også føre inn gjennomsnittene $\bar{x}_1$, $\bar{x}_2$, $\bar{x}_3$ .... og sette

$$r(x_1) = \alpha_0 + \beta_{01\cdot 23\cdot\cdot}(x_1-\bar{x}_1) + \beta_{02\cdot 13\cdot\cdot}(x_2-\bar{x}_2)+\ldots$$

Vi vil også her benytte ortogonale funksjoner (se avs. J.3) og først tenke oss at vi bruker settet

$$y_1 = x_1-\bar{x}_1$$
$$y_2 = x_2-\bar{x}_2-b_{21}(x_1-\bar{x}_1)$$
$$y_3 = x_3-\bar{x}_3-b_{32\cdot 1}(x_2-\bar{x}_2)-b_{31\cdot 2}(x_1-\bar{x}_1)$$
$$\text{o.s.v.}$$

og sette

$$r(x_1) = \alpha_0 + \beta_{01}\, y_1 + \beta_{02}\, y_2 + \beta_{03}\, y_3 + \ldots.$$

Anvendelse av Minste kvadraters metode fører så til $\bar{x}_0$ som estimator av α_0 og

$$b_{01} = \frac{\Sigma y_{1j}(\bar{x}_{0j} - \bar{x}_0)}{\Sigma y_{1j}^2} = \frac{\Sigma y_{1j}\bar{x}_{0j}}{\Sigma y_{1j}^2}$$

$$b_{02} = \frac{\Sigma y_{2j}(\bar{x}_{0j} - \bar{x}_0)}{\Sigma y_{2j}^2} = \frac{\Sigma y_{2j}\bar{x}_{0j}}{\Sigma y_{2j}^2}$$

o.s.v.

som estimatorer av β_{01} , β_{02} o.s.v.

Den kvadratsummen en tar sikte på å få delt opp ved hjelp av y-ene, er $n\Sigma(\bar{x}_{0j}-\bar{x}_0)^2 = nS_{00}$ med f = k-1 frihetsgrader. Dette gjelder både et lokalt forsøk etter prinsippet fri randomisering når antall gjentak (n) er det samme for alle forsøksledd,* for et lokalt blokkforsøk med n blokker eller gjentak, og for det forsøket vi i avsnitt H.8 har kalt et utvidet forsøk med n gjentak.

Nyttes korrelasjonskoeffisientene $R_{0\cdot 1}$, $R_{0\cdot 12}$ o.s.v. på samme måte som i avsnitt J.3, blir kvadratsummen nS_{00} delt opp slik som vist i Tab. J.6 hvor det er forutsatt at k = 5. Antall y-er er derfor under denne forutsetningen lik k-1 = 4 og $R_{0\cdot 1234} = 1$. Det leddet som er betegnet med Rest (2), er restkvadratsummen for det tilfelle at forsøket er utført etter prinsippet fri randomisering. Er forsøket

* Er antall gjentak for T_j lik n_j, finner en at estimatoren av α_0 er $\bar{x}_0 = \frac{1}{N}\Sigma n_j\bar{x}_{0j}$ hvor $N = \Sigma n_j$. Estimatoren av β_{01} er

$$b_{01} = \frac{\Sigma n_j y_{1j}(\bar{x}_{0j}-\bar{x}_0)}{\Sigma n_j y_{1j}^2}$$

og tilsvarende for de andre koeffisientene.

et blokkforsøk med n gjentak eller et utvidet forsøk med n gjentak, er Rest(2) = $\Sigma\Sigma[x_{oji}-\bar{x}_{oi}-\bar{x}_{oj}+\bar{x}_o]^2$ med et antall frihetsgrader lik f = (n-1)(k-1).

Tabell J.6.

Uavhengig variabel	Kvadratsum	f	V
y_1	$R^2_{o\bullet 1} \cdot nS_{oo}$	1	V_1
y_2	$(R^2_{o\bullet 12}-R^2_{o\bullet 1}) \cdot nS_{oo}$	1	V_2
y_3	$(R^2_{o\bullet 123}-R^2_{o\bullet 12}) \cdot nS_{oo}$	1	V_3
y_4	$(1-R^2_{o\bullet 123}) \cdot nS_{oo}$	1	V_4
Rest(2)	$\Sigma\Sigma[x_{oji}-\bar{x}_{oj}]^2$	k(n-1)	V_R

De null-hypotesene som skal testes, er $\beta_{o1} = 0$, $\beta_{o2} = 0$, $\beta_{o3} = 0$ og $\beta_{o4} = 0$. Disse testes samtidig ved hjelp av varianskvotientene V_1/V_R, V_2/V_R, V_3/V_R og V_4/V_R Testingen utføres trinnvis slik det ble forklart i avsnitt J.3 (side 231). Forskjellen er at vi nå har en estimator V_R for var(e) som det ikke skal gjøres noe med. Kvadratsummen for en ikke-signifikant varians skal i dette tilfelle ikke adderes til Restkvadratsummen foran neste trinn i testingen.

Hvilken fremgangsmåte en bør bruke i praksis, vil avhenge av hva slags hjelpemiddel en har til rådighet. Har en anledning til å bruke en elektronisk regnemaskin, vil det som oftest lønne seg å ta med alle variabler x_1, x_1^2, x_1^3.... x_1^{k-1} med det samme. Settet av ortogonale funksjoner kan så velges slik det er beskrevet i avsnitt J.3.

Har en imidlertid ikke anledning til å bruke dette hjelpemidlet, blir denne fremgangsmåten for arbeidskrevende, iallfall dersom k > 5. Dersom informasjoner en har

på forhånd, ikke peker ut et annet sett, bør en velge det settet av ortogonale funksjoner vi har benyttet foran. En kan så gå fram skrittvis. Først prøves førstegradsleddet alene. En har da følgende analyse :

Uavhengig variabel	Kvadratsum	f	V
y_1	$R^2_{0\cdot 1}.nS_{00}$	1	V_1
Rest(1)	$(1-R^2_{0\cdot 1}).nS_{00}$	k-2	V_2

Kvadratsummen for Rest(1) er naturligvis her summen av kvadratsummene for $y_2, y_3, \ldots . y_{k-1}$. Denne kvadratsummen kan da skyldes hovedsakelig en av disse k-2 variablene. Den bør derfor brukes med f = 1 frihetsgrad, slik at $F_2 = (1-R^2_{0\cdot 1}).nS_{00}/V_R$. En bruker så f.eks. Tabell III med f = 1 frihetsgrad for hver av de to teller-variansene og $f = \frac{1}{2}f_R$ for den felles nevner-varians. En ikke-signifikant F_2 sier oss at det ikke er noen grunn til å gå videre med analysen. En må da nøye seg med en lineær funksjon, d.v.s. forutsatt naturligvis at $F_1 = V_1/V_R$ er signifikant på det valte signifikansnivået. Er derimot F_2 signifikant, bør en gå videre med analysen ved å ta med også $x_2 = x_1^2$ og da i formen y_2. Analysen blir da

Uavhengig variabel	Kvadratsum	f	V
y_1	$R^2_{0\cdot 1}.nS_{00}$	1	V_1
y_2	$(R^2_{0\cdot 12}-R^2_{0\cdot 1}).nS_{00}$	1	V_2
Rest(1)	$(1-R^2_{0\cdot 12}).nS_{00}$	k-3	V_3

Her må vi så bruke Rest(1) med f = 1 frihetsgrad. Vi har da tre varianskvotienter: $F_1 = V_1/V_R$, $F_2 = V_2/V_R$ og $F_3 = (1-R^2_{0\cdot 12}).nS_{00}/V_R$. Kravet til signifikans må nå skjerpes

ved at en bruker f.eks. Tabell III med f = 1 frihetsgrad for hver av de tre teller-variansene og $\frac{1}{3} f_R$ for den felles nevner-variansen.

Det vi oppnår ved å bruke denne fremgangsmåten, er at vi på hvert trinn i analysen kan avgjøre om det har noen hensikt å gå videre. Har vi f.eks. ført inn x_1 (som y_1) og $x_2 = x_1^2$ (som y_2), vil en ikke-signifikant F_3 bety at vi ikke vil oppnå noe ved å utvide analysen. På den annen side vil ikke en signifikant F_3 bety at vi kan regne en forbedring av regresjonsfunksjonen ved å utvide analysen, men den er da en oppfordring til å gå videre.

Som oftest fører divisjonen av f_R med antallet av teller-varianser til et bruddent tall. Vi må da interpolere i F-tabellen eller for $f_R < 5$ bruke Tabell VI som er omtalt på side 228.

La oss ta for oss et eksempel. I Tab.J.7 er gitt observasjonene av x_o (mengde avling) fra et forsøk med fire kvanta fullgjødsel til eng. Forsøksleddene er 0 kg (T_1), 25 kg (T_2), 50 kg (T_3) og 75 kg (T_4) gjødsel pr.dekar. De n = 2 gjentak er to lokaliteter i Norge.

Tabell J.7.

Gjentak	T_1	T_2	T_3	T_4	S_i
1	54	76	108	112	350
2	76	106	120	128	430
S_j	130	182	228	240	780
$\bar{x}_{oj}$	65	91	114	120	

Variansanalysen gir følgende resultat :

	Kvadratsum	f	V
Gjentak	800	1	
T	3754	3	1251,33
Rest	92	3	30,67

Varianskvotienten er altså lik $F = 1251,33/30,67 = 48,04$ som er signifikant.

Vi finner så at $\Sigma(x_{1j}-\bar{x}_1)^2 = S_{11} = 3125$, $\Sigma(\bar{x}_{0j}-\bar{x}_0)^2 = S_{00} = 1877$ og $\Sigma(x_{1j}-\bar{x}_1)(\bar{x}_{0j}-\bar{x}_0) = S_{01} = 2350$. Følgelig er

$$R_{0\cdot 1}^2 = \frac{S_{01}^2}{S_{11}\cdot S_{00}} = 0,9415$$

Ahalysen på første trinn blir derfor :

Uavhengig variabel	Kvadratsum	f
y_1	$0,9415 \cdot 3754 = 3534,39$	1
Rest(1)	$0,0585 \cdot 3754 = 219,61$	2
Rest(2)	92,00	3

Vi finner derfor at $F_1 = 3534,39/30,67 = 115,24$ og $F_2 = 219,61/30,67 = 7,16$. Til f = 1 for hver av de to teller-variansene og f = 3/2 for nevner-variansen svarer etter Tabell VI og P = 0,05 a = 36,2. Den siste av de to varianskvotientene er derfor ikke signifikant. Dette betyr at vi ikke oppnår signifikant forbedring av estimeringen av den betingede forventningen for x_0 ved å ta med nye ledd i rekkeutviklingen. Vi må nøye oss med en lineær funksjon.

Gjennomsnittene er $\bar{x}_0 = 97,5$ og $\bar{x}_1 = 37,5$, og regresjonskoeffisienten er $b_{01} = S_{01}/S_{11} = 0,752$. Vi har derfor at

$$\hat{r}(x_1) = 97,5 + 0,752\ (x_1-37,5) = 69,3 + 0,752\ x_1$$

Det er nevnt i begynnelsen av dette avsnittet at hensikten med en slik analyse som dette, er å få tak i en funksjon av x_1 som gir en beskrivelse av den avhengighet som eventuelt finnes mellom forsøksfaktoren (x_1) og forventningen (μ_j) for x_0. Den regresjonsfunksjonen $\hat{r}(x_1)$ som er resultatet av analysen, er da å oppfatte som et tilnærmet riktig uttrykk for dette avhengighetsforhold. Som oftest er en neppe interessert i å bruke resultatet til beregning av f.eks. konfidensgrenser for μ_j. Men er en det, vil en finne metoden beskrevet i avsnittene J.1 og J.4.

Vi har her gått ut fra at den matematiske form for $r(x_1)$ er ukjent og at en derfor må ty til rekkeutviklingen. I flere tilfelle har det imidlertid vist seg formålstjenlig å innføre en ny variabel, altså en ny variabel som er en kjent funksjon av x_1, f.eks. $u = \sqrt{x_1}$ eller $u = \log x_1$. For vårt eksempel fører begge disse variablene til resultater som iallfall tilsynelatende er bedre enn det vi får ved å bruke x_1 selv. Korrelasjonskoeffisienten mellom x_1 og $\bar{x}_0$ er lik $\sqrt{0,9415} = 0,970$, mellom $\sqrt{x_1}$ og $\bar{x}_0$ er den 0,987 og mellom $\log x_1$ og $\bar{x}_0$ er den lik 0,994. I dette tilfelle gir derfor en lineær funksjon av $\log x_1$ en tilsynelatende noe bedre beskrivelse av forventningen for x_0 enn en lineær funksjon av x_1.

K. OM BRUK AV SAMPLER FRA GITTE BEGRENSEDE UNIVERSER.

K. Innledning.

Vi har hittil betraktet universet som en abstrakt mengde gjentak som vi kan skaffe oss kunnskaper om bare gjennom sampler. Regelen er at vi må oppfatte universet som den mengde gjentak som samplet representerer i egenskap av et random sampel.

Vi har imidlertid også oppgaver som går ut på å beskrive en gitt konkret mengde gjentak ved å bruke sampler. Vi kan til eksempel tenke oss et univers som består av alle skogeiendommene i Hedmark i 1964, eventuelt begrenset til de eiendommene som har et visst minste areal produktiv skogsmark. Vi tenker oss at vi er interessert i det gjennomsnittlige og det totale barskogareal i dette universet. Eller, det kan være den totale lengde skogsbilveier vi ønsker å få greie på. Har vi penger nok og et tilstrekkelig antall kvalifiserte personer å sette til arbeidet, ville det ikke være uoverkommelig å få nøyaktige opplysninger om både barskogarealet og lengden av bilveier.

Slike undersøkelser av hele universet eller totaltellinger, som slike undersøkelser i regelen blir kalt, vil naturligvis bli kostbare. Det er derfor også i slike tilfelle behov for sampler som kan gi tilstrekkelig nøyaktige opplysninger. Bruk av sampler kalles ofte representative tellinger.

Et univers og et sampel består av gjentak eller telleenheter. I de totale tellinger innen jordbruks- og skogbrukssektoren som er gjennomført av Statistisk Sentralbyrå, er det det enkelte gardsbruk eller den enkelte skogeiendom som er gjentak. For hvert gjentak skaffer en seg data (observasjoner) av areal brukt til f.eks. potet-dyrking, antall melkekyr, antall dekar produktiv skogsmark, antall

traktorer, lengden av skogsbilveier osv. En får på denne måten en totalbeskrivelse av hele universet, og de feil som hefter ved denne beskrivelsen, skyldes utelukkende unøyaktige data eller observasjoner. Av praktiske grunner er mange av observasjonene skjønnsmessige.

Taksering av en skog har samme målsetting. Det gjelder å få en mest mulig nøyaktig bestemmelse av kubikkmasse, tilvekst osv. Ser en bort fra observasjonsfeil, vil en total taksering gi nøyaktige opplysninger om slike ting for hele universet. På grunn av kostnadene har en imidlertid også i disse tilfelle blitt tvunget til å skaffe seg opplysninger ved hjelp av sampler. Som gjentak eller telleenhet velger en da kanskje sirkulære flater av valt størrelse. Det er nøyaktig det samme som at en når det gjelder jordbrukstellingene, har gått over til å bruke sampler av gardsbruk.

Den forskjell det er mellom disse undersøkelser og de vi tidligere har beskjeftiget oss med er følgende. Når universet som samplet tas fra, er en abstrakt mangfoldighet, fører det ikke til noen som helst forandring av universet at vi tar ut et sampel. Universet er det samme før og etter at samplet er tatt ut. Når derimot universet er en konkret begrenset mengde gjentak, kan det at vi tar ut et sampel bety en forandring av universet. Dette gjør at vi ikke uten enkelte korreksjoner kan bruke de setningene om sampler som vi har gjennomgått foran.

Den forskjell det er mellom de to situasjonene, kan vi illustrere ved å ta for oss på ny det spørsmål som ble behandlet i avsnitt C.5 (side 72). La oss tenke oss et univers på N gjentak hvor H gjentak har kjennetegnet E. De andre (N - H) gjentakene har da kjennetegnet ikke-E. Sannsynligheten for E i dette universet er $P(E|U) = p = H/N$.

La oss så tenke oss at vi tar ut et random sampel på n gjentak. Da gir (se side72) sannsynligheten for at samplet skal bestå av z gjentak med E og (n-z) gjentak med iE lik

$$P_z = \frac{\binom{H}{z}\binom{N-H}{n-z}}{\binom{N}{n}}$$

hvor til eksempel $\binom{H}{z} = \frac{H!}{z!(H-z)!}$. Dette kalles den hypergeometriske funksjon eller lov. Det kan også vises at når N og H er meget store i forhold til n (teoretisk at N og H $\to \infty$, samtidig som H/N = p), går funksjonen over til den binomiale:

$$P_z = \binom{n}{z} p^z \, (1-p)^{n-z}$$

Det kan også vises at forventningen for z er

$$E(z) = \Sigma P_z \cdot z = \frac{nH}{N} = np$$

og at

$$\text{var}\,(z) = \sigma^2 = \Sigma P_z (z-Ez)^2 = n\,\frac{H}{N}\left(1 - \frac{H}{N}\right)\frac{N-n}{N-1}$$

$$= np(1-p)\,\frac{N-n}{N-1}$$

Vi ser at når N $\to \infty$, er Gr. $\frac{N-n}{N-1} = 1$ og

$$\text{var}\,(z) = np(1-p)$$

som er formelen for var (z) når P_z er gitt ved binomialfunksjonen.

Forventningen for z er altså den samme i de to tilfelle. Men når samplet (n) er så stort i forhold til universet (N) at uttakingen av samplet betyr et merkbart innhugg i universet, må vi for σ^2 bruke faktoren $\frac{N-n}{N-1} < 1$.

La oss som eksempel tenke oss at universet består av de N = 52

kortene i en kortstokk og at E = spar. Da er H = 13, N-H = 39 og $p = \frac{H}{N} = \frac{1}{4}$. Sannsynligheten for z sparkort i et random sampel på n = 13 kort er da

$$P_z = \frac{\binom{13}{z}\binom{39}{13-z}}{\binom{52}{13}}$$

Vi har at

$$E(z) = np = 13/4$$

og

$$\text{var.}(z) = np(1-p)\frac{N-n}{N-1} = 13\,\frac{1}{4}\cdot\frac{3}{4}\cdot\frac{39}{51} = 13\cdot\frac{1}{4}\cdot\frac{3}{4}\cdot 0{,}765$$

$$= 2{,}44 \cdot 0{,}765 = 1{,}87$$

Lager vi oss en kortstokk av 100 enkelte kortstokker, har vi N = 5200 og H = 1300. Er samplet som før på n = 13 kort, finner vi at

$$\frac{N-n}{N-1} = \frac{5187}{5199} = 0{,}997..$$

og derfor at var.(z) er praktisk talt lik np(1-p) = 2,44.

Denne faktoren $\frac{N-n}{N-1}$, eller oftest $\frac{N-n}{N}$, forekommer derfor i mange av de formlene vi har bruk for når vi skal bruke sampler fra begrensede universer. Naturligvis kan N også i slike tilfelle være så stor i forhold til n at faktoren ikke har noen betydning.

Siden antall gjentak (N) i universet er et endelig tall, kan vi sette nummer på gjentakene. Numrene er i = 1,2,3...N. Det kan vises at antall sampler på n gjentak er $\binom{N}{n}$. Er f.eks. N = 4 og n = 2, er antall sampler lik $\binom{4}{2} = 6$. Dette er de samplene der gjentakene har numrene: 1 og 2, 1 og 3, 1 og 4, 2 og 3, 2 og 4 og 3 og 4. Et rent tilfeldig sampel eller et sampel uten restriksjoner er et sampel som er tatt ut på en slik måte at sannsynligheten for uttak av et bestemt sampel er den samme for alle sampler, lik $1/\binom{N}{n}$. Er N = 4 og n = 2,

er denne sannsynligheten lik 1/6.

Vi skal ikke komme inn på den teknikk som brukes i praksis (f.eks. i jordbrukstellingene) for uttaking av et slikt sampel. Vi må nøye oss med å si at en må bruke et apparat for loddtrekning som er konstruert slik at de $\binom{N}{n}$ samplene har samme sannsynlighet for å bli trukket ut. Vi har også tabeller over såkalte tilfeldige tall for dette formålet og det er vel disse som blir brukt oftest.

De data vi skaffer oss gjennom samplet er også her observasjoner av random variable (x), diskrete og kontinuerlige, og relative frekvenser (z/n) for konstante kjennetegn. En relativ frekvens er da også her en estimator av en sannsynlighet $p = H/N$ i det universet samplet er tatt fra. Når det gjelder denne estimeringen skal vi nøye oss med å konstatere at hvis samplet er et random sampel uten restriksjoner, kan vi ta utgangspunkt i den hypergeometriske funksjon. Vi ser da at z/n er en forventningsrett estimator av $p = H/N$ og at

$$\operatorname{var}\left(\frac{z}{n}\right) = \frac{p(1-p)}{n}\,\frac{N-n}{N-1}$$

Vi skal i det følgende ta for oss noen av de problemene vi står overfor når det gjelder estimering av størrelser eller parametrer i universet på basis av observasjoner av random variable. I en representativ jordbrukstelling gjelder til eksempel estimering av det totale og det gjennomsnittlige areal som brukes til dyrking av poteter. I en skogtakst gjelder det f.eks. estimering av den totale og den gjennomsnittlige diametertilvekst hos gran i de siste 10 år. I det første eksempel er det gardsbruket som er gjentak eller telleenhet, i det siste flater av en valt størrelse og form.

Vi skal først ta for oss de estimeringsproblemer vi står overfor når samplet er et random sampel uten restriksjoner.

K.2. Bruk av random sampel uten restriksjoner.

La N være antall gjentak i universet og x_i (i = 1,2,3...N) verdiene av en random variabel x. Gjennomsnittet for hele universet er da

$$\bar{X} = \frac{1}{N} \sum_{1}^{N} x_i$$

og summen er

$$X = \sum_{1}^{N} x_i = N \bar{X}$$

Her er $\bar{X}$ det samme som vi foran (kap. D) har betegnet med E(x) eller μ. Variansen for x vil vi definere ved formelen

$$\text{var}(x) = \sigma^2 = \frac{1}{N-1} \sum_{1}^{N} (x_i - \bar{X})^2$$

I samplet har vi observasjonene x_i (i = 1,2,3...n) med gjennomsnittet

$$\bar{x} = \frac{1}{n} \sum_{1}^{n} x_i$$

og middelavviket s gitt ved

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

Det kan vises at når samplet er et random sampel uten restriksjoner, er $\bar{x}$ en forventningsrett estimator av $\bar{X}$ og s^2 en forventningsrett estimator av σ^2. Vi bruker derfor også i disse tilfelle $\bar{x}$ som estimator av $\bar{X}$. Vil vi også beregne konfidensgrensene for $\bar{X}$, kan vi bruke den metoden som er beskrevet side 122 for beregning av konfidensgrensene for θ. Men på grunn av at uttaket av et sampel på n gjentak betyr et merkbart innhugg i universet, må vi ta hensyn til faktoren $\frac{N-n}{N-1}$. Det kan nemlig vises at variansen i fordelingsfunksjonen for $\bar{x}$ ikke er σ^2/n som for ubegrensede universer. Vi vil nå finne at

$$\text{Var.}(\bar{x}) = \frac{\sigma^2}{n}\,\frac{N-n}{N}$$

med estimatoren $\frac{s^2}{n}\,\frac{N-n}{N}$. Konfidensgrensene for $\bar{X}$ blir derfor

$$\bar{x} \mp a\,\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$$

Verdien av a finnes i Tabell I for n-1 frihetsgrader og den P-verdi en ønsker å bruke, f.eks. P = 0,05.

Konfidensgrensene for $X = N\bar{X}$ får vi så ved å multiplisere med N, d.v.s. at de er

$$N\left\{\bar{x} \mp a\,\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}\right\}$$

Vi ser at beregningen av konfidensgrensene forutsetter at N er kjent. Vi må derfor utføre en selvstendig telling av antallet gjentak i universet. Men det hender naturligvis at en har kjennskap til størrelsen av universet eller, en nøyaktig nok verdi for N, fra andre kilder. Gjelder det til eksempel representative jordbrukstellinger og universet omfatter alle gardsbruk i et herred eller et fylke, vil vi ha en noenlunde riktig verdi av N fra den siste totaltellingen.

I alle slike undersøkelser må en prøve å gjøre seg opp en mening om hvor stort sampel en har bruk for. Under planleggingen av en representativ jordbrukstelling må en bestemme hvor stort sampel en vil ha fra hvert herred i landet. Tellingen foregår her herredsvis. Å ta standpunkt til hvor stort sampel en vil komme til å få bruk for, er en meget vanskelig sak. Det er alltid flere hensyn som må veies mot hverandre.

For det første må en naturligvis ta hensyn til kostnadene. Et

lite sampel er billigere enn et stort. Men det er også to andre viktige momenter å ta hensyn til. Det ene er at en ønsker $\bar{X}$ estimert med en rimelig presisjon, det andre at teknikken for beregningen av konfidensgrensene har visse forutsetninger som må respekteres.

Presisjonen er omvendt proporsjonal med bredden av konfidensintervallet, $2a \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N}}$. (Dette er iallfall den enkleste måten å definere presisjonen på.) I regelen vil vedkommende som har ansvaret for undersøkelsen, allerede på planleggingsstadiet kunne oppgi en øvre grense for den bredden av konfidensintervallet (svarende til f.eks. Q = 0,95) som kan tolereres. La oss si at dette er 2 d. Av samplets størrelse (n) må vi da kreve at

$$2a \frac{s}{\sqrt{n}} \sqrt{\frac{N-n}{N}} \leq 2d$$

Bruker vi her likhetstegnet, vil vi ved løsningen av ligningen m.h.p. n få et tall for minstestørrelsen av samplet. Vi ser imidlertid at vi for å kunne løse denne ligningen må ha kjennsakp til a, s og N.

Når det gjelder a regner en vel som oftest med at samplet (n) blir så stort at en kan bruke den verdi som i Tabell I svarer til uendelig mange frihetsgrader. Brukes Q = 0,95, skal vi sette a = 1,96 som vel oftest rundes av til a = 2.

Verdien av s vil vi naturligvis ikke ha kjennskap til før etter at samplet er tatt og vi har observasjonene x_i (i = 1,2,3...n). Vi må derfor prøve å skaffe oss en noenlunde pålitelig forhåndsvurdering av s. Gjelder det representative jordbrukstellinger, kan en bruke den verdi av s som er funnet i tidligere brukte sampler eller den verdi av σ en har funne i siste totaltelling. Gjelder det taksasjon av en skog, må en også ta til takke med en verdi av s som er funnet ved tidligere og lignende takseringer.

For N vil vi, når det gjelder representative jordbrukstellinger, ha en nokså nær riktig verdi fra siste totaltelling. Nå i 1964 og de nærmeste år framover må vi bruke antall gardsbruk i herred eller fylke fra totaltellingen i 1959. Skal vi taksere en skog, vil vi vel som oftest ha et noenlunde riktig tall for totalarealet. Har vi så bestemt oss for størrelsen av takstflaten, vil vi også kunne angi en noenlunde pålitelig anslagsverdi for N.

Løsningen av ligningen er

$$n = \frac{a^2 s^2}{d^2 + a^2 s^2 / N}$$

Et eksempel. Den totale jordbrukstelling i 1949 viste for Møre og Romsdal fylke N = 21495 gardsbruk og for x = jordbruksarealet (i dekar) σ = 35,44 (dekar). Vi har derfor

$$a^2 s^2 = 4 \cdot 35{,}44^2 = 5021{,}32$$

og

$$\frac{a^2 s^2}{N} = 0{,}2336$$

Altså er

$$n = \frac{5021{,}32}{d^2 + 0{,}2336}$$

La oss så tenke oss at vi for dette fylke vil foreta en ny telling ved hjelp av et random sampel uten restriksjoner. Og, la oss tenke oss at vi ønsker å estimere det gjennomsnittlige jordbruksareal ($\bar{X}$) med en presisjon som svarer til d = 2 dekar. Vi vil da finne at vi har bruk for et sampel

$$n \geq \frac{5021{,}32}{4{,}2336} = 1186$$

(Det er her forutsatt brukt en konfidenssannsynlighet på Q = 0,95).

Senker vi kravet til presisjonen til d = 3 dekar, finner vi at vi har bruk for et sampel på minst 544 gjentak. Senker vi kravet ytterligere til d = 5 dekar, finner vi n ≧ 199.

I teorien er det forutsatt for bruk av t-til beregning av konfidensgrensene, at fordelingsfunksjonen for den observerte random variable (x) ikke er altfor avvikende fra den normale (se avs.F.2). Erfaringene fra de totale jordbruks- og skogbrukstellingene er at fordelingsfunksjonen for mange av de random variable det er tale om, er sterkt avvikende fra denne formen. Mange av fordelingsfunksjonene er sterkt høyreskjeve, d.v.s. at tyngdepunktet ligger langt til venstre for midtpunktet. I følgende eksempel er x areal (i dekar) produktiv skogsmark i Østfold etter tellingen i 1949. Legg merke til at de klasser for x-verdiene som er brukt i tabellen, ikke har samme bredde overalt.

x	P(x|u)
0,1 - 100	0,4231
100,1 - 250	0,2568
250,1 - 500	0,1658
500,1 - 1000	0,0923
1000,1 - 2000	0,0391
2000,1 - 5000	0,0157
5000,1 - 10000	0,0042
10000,1 - 20000	0,0025
20000,1 - 50000	0,0006
	1,0000

En har funnet at for å kunne bruke teknikken for beregning av konfidensgrensene, må det i slike tilfelle settes visse minstekrav

til størrelsen av samplet. En av de regler en har kommet fram til er et krav som går ut på

$$n > 25\,\gamma^2$$

hvor

$$\gamma = \frac{\frac{1}{N}\sum(x_i - \bar{X})^3}{\sigma^3}$$

For vårt eksempel har vi $\sigma = 35{,}44$ og $\frac{1}{N}\Sigma(x_i - \bar{X})^3 = 143599$. Altså er

$$\gamma = \frac{143599}{35{,}44^3} = 3{,}23$$

og $25\,\gamma^2 = 261$. Vi må derfor i dette tilfelle kreve $n \geq 261$.

Et sampel på $n = 261$ vil gi en presisjon ved estimeringen av $\bar{X}$ som svarer til

$$d = \frac{as}{\sqrt{n}}\sqrt{\frac{N-n}{N}} = 4{,}4 \text{ (dekar)}$$

Er dette tilstrekkelig for formålet med estimeringen av $\bar{X}$, kan vi altså nøye oss med et sampel på $n = 261$ eller noe mer. Ønsker vi større presisjon, må samplet velges større.

En slik forhåndsberegning av n kan naturligvis vise seg å gi for lite sampel fordi en må bruke noe usikre tall for N, s og $\Sigma(x_i - \bar{X})^3$. Når vi har skaffet oss samplet og observasjonene x_i, kan det vise seg at presisjonen ved estimeringen av $\bar{X}$ ikke er så stor som vi hadde regnet med. I noen tilfelle kan vi da øke samplet ved å ta med noen flere gjentak. Men dette er ikke alltid mulig.

Spørsmålet om samplets størrelse kompliseres av at det praktisk talt alltid er flere størrelser ($\bar{X}$) som skal estimeres. Og da kan det hende at de forskjellige størrelsene stiller ulike krav til samplets størrelse. Gjelder det til eksempel taksering av en skog,

kan det hende at estimeringen av den gjennomsnittlige kubikkmasse pr. gjentak (flate) setter større krav til n enn estimeringen av den gjennomsnittlige tilvekst i de siste 10 år. I slike tilfelle må en kanskje foreta en rangering av $\bar{X}$'ene etter den betydning de har. Det kan tenkes at en vil stille svakere krav til presisjon for de mindre viktige $\bar{X}$'ene enn for de mer betydningsfulle. Men til slutt blir en vel stående ved en sampelstørrelse som er et kompromiss.

K.3.Stratifisering.

Et univers som det som består av alle gardsbrukene er naturligvis meget heterogent i mange av de variable (x) som observeres for å gi grunnlag for en beskrivelse. Vi kan tenke på slike variable som f.eks. x = areal anvendt til potetdyrking eller x = antall melkekyr. Dette skyldes for det første den geografiske variasjon i klimaet og i forhold ellers som har avgjørende innflytelse på driften. Skulle vi derfor utføreen representativ telling med ett random sampel fra hele universet av gardsbruk i landet, ville vi måtte skaffe oss et meget stort sampel for å oppnå en rimelig presisjon ved estimeringen av f.eks. det gjennomsnittlige antall melkekyr.

Av denne grunn deles universet i subuniverser eller strata og en tar så et random sampel uten restriksjoner fra hvert stratum. Når det gjelder de representative jordbrukstellingene i Norge, bruker en herredene som strata. Dessuten stratifiserer en også etter bruksstørrelsen.

Når det gjelder taksering av en skog søker en å minske heterogeniteten ved å stratifisere etter f.eks. bonitet.

La oss nå tenke oss at universet deles opp i k strata ($j = 1,2,3...k$). La N_j være antall gjentak i det j'te stratum og $N = \Sigma N_j$ antall gjentak i hele universet. Fra det j'te stratum tas så et random sampel uten restriksjoner på n_j gjentak. Hele samplet blir da på $n = \Sigma n_j$ gjentak.

I den representative jordbrukstelling som ble utført i 1960 ble det stratifisert etter herreder og innen herredene i to etter bruksstørrelsen (5,1 - 100 dekar og > 100 dekar). Etter den sammenslåing av herreder som har skjedd i de siste årene, vil dette si at antallet av strata er av størrelsesorden k = 1000. Når det gjelder en skogtakst er naturligvis antallet av strata et meget lite tall.

La nå $\bar{X}$ være det gjennomsnitt i hele universet som det tas sikte på å estimere. Det er f.eks. det gjennomsnittlige antall melkekyr eller den gjennomsnittlige diametertilvekst i de siste 10 år.

For det j'te stratum er da

$$\bar{X}_j = \frac{1}{N_j} \sum_{1}^{N_j} x_{ij}$$

og i hele universet

$$\bar{X} = \frac{1}{N} \sum_{1}^{k} N_j \bar{X}_j$$

Estimatoren av $\bar{X}_j$ er

$$\bar{x}_j = \frac{1}{n_j} \sum_{1}^{n_j} x_{ij}$$

og estimatoren av $\bar{X}$ er

$$\bar{x}_{st} = \frac{1}{N} \sum_{1}^{k} N_j \bar{x}_j$$

hvor fotskriften st betyr "stratifisert". Beregningen av $\bar{x}_{st}$ forut-

setter naturligvis at N_j er kjent.

Siden samplet på n_j gjentak er et random sampel uten restriksjoner, er $\bar{x}_j$ en forventningsrett estimator av $\bar{X}_j$. Derfor er $\bar{x}_{st}$ en forventningsrett estimator av $\bar{X}$. Vi ser nemlig at

$$E(\bar{x}_{st}) = \frac{1}{N} \sum_{1}^{k} N_j E\bar{x}_j = \frac{1}{N} \sum_{1}^{k} N_j \bar{X}_j = \bar{X}.$$

Har vi vært heldige med stratifiseringen vil en større del av heterogeniteten være fjernet. Dette vil si at gjentakene skiller seg mindre fra hverandre innen det enkelte stratum enn innen universet sett under ett. Konsekvensen er at var (x) er mindre innen det enkelte stratum enn innen hele universet. Effekten av stratifiseringen kan sammenlignes med den effekt det har på var (ε) at vi deler et forsøksfelt i blokker.

Samtidig kan stratifiseringen virke til at fordelingsfunksjonen for den observerte random variable (x) ligner mer på den normale innen de enkelte strata enn innen hele universet sett under ett.

Som vi skal se senere har en heldig stratifisering den effekt at estimeringen av $\bar{X}$ blir mer presis enn om vi bruker ett sampel fra hele universet, når størrelsen (n) av samplet er den samme i de to tilfelle.

Fra foregående avsnitt har vi at

$$\text{var}(\bar{x}_j) = \frac{\sigma_j^2}{n_j} \frac{N_j - n_j}{N_j}$$

hvor

$$\sigma_j^2 = \frac{1}{N_j - 1} \sum_{1}^{N_j} (x_{ij} - \bar{X}_j)^2$$

For $\bar{x}_{st}$ finner vi så at

$$\text{var}(\bar{x}_{st}) = \frac{1}{N^2} \sum_1^k N_j^2 \text{ var}(\bar{x}_j) = \frac{1}{N^2} \Sigma N_j^2 \frac{\sigma_j^2}{n_j} \frac{N_j - n_j}{N_j}$$

$$= \frac{1}{N^2} \Sigma N_j (N_j - n_j) \sigma_j^2 / n_j$$

Estimatoren av $\text{var}(\bar{x}_{st})$ får vi ved å erstatte σ_j^2 med

$$s_j^2 = \frac{1}{n_j - 1} \sum_1^{n_j} (x_{ij} - \bar{x}_j)^2$$

Vi ser at $\text{var}(\bar{x}_{st})$ er avhengig både av σ_j og n_j. Har vi valt en bestemt måte å stratifisere på, er de k σ_j^2 gitt. Da vil $\text{var}(\bar{x}_{st})$ være avhengig av n_j alene.

Under planleggingen av en telling må en ta standpunkt til hvor stort sampel en vil ta fra hvert stratum. Å bestemme størrelsen av samplene (n_j) kalles å allokere, eller - rettere - å allokere betyr å fordele hele samplet (n) på de k strata.

Når vi skal bestemme oss for hvordan vi vil allokere, står vi meget fritt. Men vi har også visse prinsipper som vi kan gjøre bruk av.

a) Vi kan velge n_j konstant. Med et sampel på n gjentak blir da $n_j = n/k$. Forutsetningen er at $\frac{n}{k} \leq N_j$ for alle strata.

b) Proporsjonal allokering betyr at n_j velges proporsjonal med N_j, d.v.s. at

$$n_j = \frac{n}{N} N_j$$

Da er

$$\text{var}(\bar{x}_{st}) = \frac{N-n}{N^2 n} \Sigma N_j \sigma_j^2$$

c) Optimal allokering betyr at n_j velges slik at $\text{var}(\bar{x}_{st})$ blir minst mulig for gitt n. Det kan lett vises at da er

$$n_j = \frac{N_j\sigma_j}{\Sigma N_j\sigma_j} n$$

og

$$\text{var}(\bar{x}_{st}) = \frac{1}{N^2 n}(\Sigma N_j\sigma_j)^2 - \frac{1}{N^2}\Sigma N_j\sigma_j^2$$

Det finnes også et fjerde allokeringsprinsipp som tar hensyn til at kostnadene for en observasjon kan variere mellom strataene. Dette skal vi ikke komme inn på.

Spørsmålet om hvilket allokeringsprinsipp en skal bruke i en bestemt situasjon er meget vanskelig. Svaret avhenger også i stor grad av problemstillingen.

Gjelder det estimering av en $\bar{X}$, f.eks. det gjennomsnittlige antall traktorer (pr. bruk) i universet av alle gardsbruk i Norge, er det rimelig å bruke optimal allokering fordi det er dette som fører til størst presisjon for estimeringen av $\bar{X}$ ved $\bar{x}_{st}$. I regelen er det imidlertid mer enn en $\bar{X}$ som skal estimeres. Den allokering som er optimal for en av disse, behøver ikke være optimal for noen av de andre $\bar{X}$. I en representativ jordbrukstelling behøver ikke optimal allokering for estimering av det gjennomsnittlige antall melkekyr være optimal for estimering av det gjennomsnittlige areal brukt til potetdyrking. Optimal allokering for estimering av grunnflatesummen behøver ikke være optimal for estimering av diametertilveksten de siste 10 år.

Meget ofte er det vel også slik at en er like meget interessert i estimeringen av hver av de k $\bar{X}_j$ som i $\bar{X}$. Det kan til eksempel være like viktig å estimere det gjennomsnittlige antall melkekyr innen hvert

fylke som å estimere gjennomsnittet for hele landet. Hvis det er dette som er situasjonen, er det rimelig å bruke proporsjonal allokering.

Vi bør merke oss at optimal allokering forutsetter at σ_j og N_j, proporsjonal allokering at N_j, er kjent. Under planleggingen har en imidlertid vanligvis bare foreløpige og approksimative verdier av disse størrelsene. En sampelundersøkelse som er planlagt som optimal, kan derfor vise seg ikke å være det, iallfall ikke eksakt optimal.

I noen tilfelle kan N_j for flere strata være så små tall at en like godt kan foreta en fullstendig telling for disse strata.

La oss så tenke oss at vi har skaffet oss observasjoner x_{ij} ($i = 1,2,3...n_j$, $j = 1,2,3...k$). Vi har da at $\bar{x}_j = \frac{1}{n_j} \sum_1^{n_j} x_{ij}$ er en forventningsrett estimator av $\bar{X}_j$ og $\bar{x}_{st} = \frac{1}{N} \Sigma N_j \bar{x}_j$ en forventningsrett estimator av $\bar{X}$. Forutsatt at n_j er store nok kan vi beregne konfidensgrenser for $\bar{X}_j$ slik som vist i foregående avsnitt. Vi kan dessuten beregne konfidensgrenser for $\bar{X}$. Disse er

$$\bar{x}_{st} \mp a \cdot s(\bar{x}_{st})$$

der

$$s(\bar{x}_{st}) = \sqrt{\frac{1}{N^2} \Sigma N_j (N_j - n_j) \, s_j^2 / n_j}$$

med

$$s_j^2 = \frac{1}{n_j - 1} \sum_1^{n_j} (x_{ij} - \bar{x}_j)^2$$

Er det brukt proporsjonal allokering, d.v.s. $n_j = \frac{N_j}{N} n$, vil formelen for $s(\bar{x}_{st})$ forenkles til

$$s(\bar{x}_{st}) = \sqrt{\frac{N-n}{N^2 n} \Sigma N_j s_j^2}$$

Vi vil nå tenke oss at vi under planleggingen har hatt riktige verdier av N_j å gå ut fra og at vi når det gjelder estimeringen av $\bar{X}$ setter et minstekrav til presisjonen uttrykt ved en valt d (se foregående avsnitt) slik at

$$2a\ s(\bar{x}_{st}) \leq 2d$$

Brukes proporsjonal allokering, vil vi da finne at

$$n \geq \frac{a^2 \Sigma w_j s_j^2}{d^2 + a^2 \Sigma w_j s_j^2/N}$$

hvor $w_j = N_j/N$. Siden vi må ta standpunkt til størrelsen av n under planleggingen, må vi allerede da ha noenlunde pålitelige informasjoner om både s_j og N_j.

Til slutt må vi også ta for oss spørsmålet: hva kan vi i et bestemt tilfelle vente å oppnå med stratifisering i sammenligning med fri randomisring? Svar på dette spørsmålet vil vi få ved å sammenligne den bredde av konfidensintervallet for $\bar{X}$ vi kan vente når vi bruker et random sampel uten restriksjoner på n gjentak og når vi stratifiserer og bruker et sampel på $n = \Sigma n_j$ gjentak. Vi må da naturligvis sammenligne under den forutsetningen at sampelstørrelsen (n) er den samme i de to tilfelle.

Som vi har sett er den bredde av konfidensintervallet vi kan vente når vi bruker et random sampel uten restriksjoner, proporsjonal med

$$\sqrt{\text{var}\,(\bar{x})} = \sqrt{\frac{N-n}{Nn}\sigma^2}$$

Bruker vi stratifisering og proporsjonal allokering har vi tilsvarende

$$\sqrt{\text{var}(\bar{x}_{st})} = \sqrt{\frac{N-n}{N^2 n}\,\Sigma N_j \sigma_j^2}$$

La oss ta for oss et eksempel. Hvis vi stratifiserer etter bruksstørrelsen (skogareal + jordbruksareal + areal til andre formål) i 4 strata som vist i følgende tabell, vil vi etter den totale jordbruksstelling i 1949 for Buskerud ha følgende data for x = jordbruksarealet. Bruk på 5 dekar eller mindre bruksstørrelse er ikke tatt med.

Stratum	N_j	$\bar{X}_j$	σ_j^2
5,1 - 35	4771	18	63
35,1 - 75	2835	53	100
75,1 - 200	1978	122	966
> 200	247	371	8010
	N = 9831		

Vi har dessuten at $\sigma^2 = 4524$.

Tenker vi oss så at vi tar ut et random sampel uten restriksjoner på n = 1000 bruk, finner vi at

$$\sqrt{\text{var}(\bar{x})} = \sqrt{\frac{8831}{9831 \cdot 1000} \cdot 4524} = 2{,}01$$

Bruker vi proporsjonal allokering, vil vi finne at

$$\sqrt{\text{var}(\bar{x}_{st})} = \sqrt{\frac{8831}{9831^2 \cdot 1000} \cdot 4473291} = 0{,}64$$

Dette vil si at presisjonen ved estimeringen av $\bar{X}$ er omtrent 3 ganger så stor når vi bruker stratifisering og proporsjonal allokering som når vi bruker fri randomisering. Eller rettere: det er dette vi har grunn til å vente.

Den ventede effekt av stratifiseringen kan også måles ved beregning av den minstestørrelsen av samplet som må kreves for å

tilfredsstille et bestemt krav til presisjonen. Setter vi for vårt eksempel dette krav til d = 2 (dekar), vil vi finne (se foregående avsnitt) at et random sampel uten restriksjoner må være på $n \geq 3098$ bruk. Bruker vi stratifiseringen slik som vist i tabellen, og proporsjonal allokering, finner vi at samplet må være på $n \geq 435$ bruk. Ved å gjøre bruk av stratifiseringen vil vi m.a.o. kunne regne med å klare oss med et sampel som er omtrent syvende-parten av det sampel vi måtte ha dersom det tas ut som et random sampel uten restriksjoner. Vi kan imidlertid ikke regne med at effekten av stratifisering i alminnelighet er så stor som i dette eksemplet.

K.4. Regresjonsestimering.

Hensikten med stratifisering er å få mer presis estimering enn ved et random sampel uten restriksjoner, uten å øke samplets størrelse. Vi har også andre metoder som tar sikte på dette. En av disse er regresjonsestimering. La $\bar{X}_o$ og X_o være de to størrelsene som skal estimeres. I et random sampel uten restriksjoner har vi observasjonene x_{oi} (i = 1,2,3..n) med gjennomsnittet $\bar{x}_o$. La oss så tenke oss at x_1 er en random variabel som er korrelert med x_o og slik at regresjonen for x_o m.h.p. x_1 er lineær. Observasjonen av denne variable i samplet er x_{1i}.

Om x_1 vil vi så forutsette at vi kjenner $\bar{X}_1$ i universet eller at vi lettvint kan skaffe oss observasjoner av den for hvert av de N gjentak og dermed gjennomsnittet $\bar{X}_1$. Som et eksempel kan vi tenke oss at x_1 er areal jord brukt til potetdyrking og at x_{1i} (i = 1,2,3...N) er observasjoner vi har fått for hvert gardsbruk i Buskerud under den totale jordbrukstelling i 1959. I 1964 tar vi ut et random sampel av

dette universet og skaffer oss observasjonene av areal jord brukt til potetdyrking i 1964. Dette er da observasjonene x_{oi} (i = 1,2,3...n). Fra tellingen i 1959 har vi da de tilsvarende observasjoner x_{1i} (i = 1,2,3...n). Vi skal så utnytte det kjennskap til x_1 vi har fra totaltellingen i 1959, d.v.s. $\bar{X}_1$, og korrelasjonen mellom x_1 og x_o slik vi finner den for samplet, til å estimere $\bar{X}_o$.

Regresjonsestimatoren av $\bar{X}_o$ er da

$$\bar{x}_{or} = \bar{x}_o + b(\bar{X}_1 - \bar{x}_1)$$

hvor b er den vanlige regresjonskoeffisienten

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

Fotskriften r betyr at det gjelder regresjonsestimatoren.

Denne estimeringsmetoden er også blitt brukt i slike tilfelle der en lettvint og billig kan skaffe seg observasjoner av x_1 for hele universet ved skjønn. Metoden er til eksempel brukt til estimering av det gjennomsnittlige tømmervolum ($\bar{X}_o$) pr. arealenhet (flate) i en skog. En delte skogen i N like store flater og tok ut et random sampel uten restriksjoner på n flater. En observerte så tømmervolumet skjønnsmessig på hver av de N flatene, og da naturligvis også på hver av de n flatene i samplet. De skjønnsmessige observasjonene i samplet er x_{1i} (i = 1,2,3..n). Tømmervolumet ble så bestemt med mer eksakte metoder på hver flate i samplet, observasjonene x_{oi} (i = 1,2,3...n). På grunnlag av observasjonene i samplet beregnet en så $\bar{x}_o$, $\bar{x}_1$ og b. Verdien av $\bar{X}_1$ fikk en så som gjennomsnittet av de N skjønnsmessige observasjonene.

Et lignende eksempel er nevnt av W.G. Cochran (Sampling Techniques. 2. utg. 1963). I en frukthage var det N = 200 ferskentre. Vekten av frukten ble bedømt skjønnsmessig (x_1) på hver tre. En fant da X_1 = 11600 lb. (pounds) og $\bar{X}_1$ = 11600/200 = 58. I et random sampel uten restriksjoner på n = 10 gjentak (trær) ble så frukten veid. Dette ga n observasjoner x_{oi} og en har de tilsvarende n skjønns-observasjoner er x_{1i}.

En fant da $\bar{x}_o$ = 543/10 = 54,3 og $\bar{x}_1$ = 569/10 = 56,9. Regresjonskoeffisienten er b = 1,0545. Regresjonsestimatet av $\bar{X}_o$ er derfor

$$\bar{x}_{or} = 54,3 + 1,0545\ (58 - 56,9) = 54,3 + 1,16 = 55,46$$

Korrelasjonen mellom x_1 og x_o er høy i dette tilfelle. Korrelasjonskoeffisienten er r = 0,97.

Det synes kanskje litt eiendommelig at en kan komme til et brukbart resultat på denne måten. Skjønnsmessige observasjoner er erfaringsmessig svært ofte beheftet med systematiske feil. I vårt eksempel har observatøren påviselig overvurdert vekten. Denne systematiske observasjonsfeilen vil imidlertid finnes både i $\bar{X}_1$ og i $\bar{x}_1$ og derfor bli borte i differensen $(\bar{X}_1 - \bar{x}_1)$.

Det kan vises at når regresjonen for x_o m.h.p. x_1 er lineær, er $\bar{x}_{or}$ en forventningsrett estimator av $\bar{X}_o$. Lineær regresjon er imidlertid en forutsetning for dette. Før en bruker denne estimatoren i praksis må en derfor undersøke om denne forutsetningen er tilfredsstilt

En eksakt riktig formel for var.$(\bar{x}_{or})$ har vi ikke. Men det kan vises at

$$\text{var.}(\bar{x}_{or}) = \frac{(N-n)(n+1)}{Nn^2}(1-\rho^2)\,\text{var.}(x_o)$$

er riktig nok for praktiske formål. Her er ρ korrelasjonskoeffisienten mellom x_o og x_1 i universet,

Som estimator av var$(\bar{x}_{or})$ brukes

$$s^2(\bar{x}_{or}) = \frac{(N-n)(n+1)}{Nn^2} s_\varepsilon^2$$

hvor

$$s_\varepsilon^2 = \frac{1}{n-2} \Sigma[x_{oi} - \bar{x}_o - b(x_{1i} - \bar{x}_1)]^2$$

$$= \frac{1-r^2}{n-2} \Sigma(x_{oi} - \bar{x}_o)^2 = \frac{n-1}{n-2} s_o^2(1-r^2)$$

Konfidensgrensene for $\bar{X}_o$ er

$$\bar{x}_{or} \mp a\, s(\bar{x}_{or}) = \bar{x}_o + b(\bar{X}_1 - \bar{x}_1) \mp a\, s(\bar{x}_{or})$$

hvor a finnes i Tabell I for n - 2 frihetsgrader.

For vårt eksempel har vi at $(n-1)s_o^2 = 998{,}1$ og $1-r^2 = 0{,}052$. Følgelig er

$$s^2(\bar{x}_{or}) = \frac{(200-10)11}{200 \cdot 10^2} \quad \frac{998{,}1 \cdot 0{,}052}{8} = 0{,}678$$

og $s(\bar{x}_{or}) = 0{,}823$. Av Tabell I ser vi at til P = 0,05 og n-2 = 8 frihetsgrader er a = 2,306. Altså er $a \cdot s(\bar{x}_{or}) = 1{,}90$ og de konfidensgrensene for $\bar{X}_o$ som svarer til konfidenssannsynligheten Q = 0,95, er

$$\bar{x}_{or} \mp a\, s(\bar{x}_{or}) = 55{,}46 \mp 1{,}90 = \begin{cases} 53{,}56 \\ 57{,}36 \end{cases}$$

Konfidensgrensene for X_o, d.v.s. for vekten av hele fruktavlingen, er da

$$N[\bar{x}_{or} \mp a\, s(\bar{x}_{or})] = 200(55{,}46 \mp 1{,}90) = \begin{cases} 10712 \\ 11472 \end{cases}$$

For sampler som ikke er altfor små, vil vi finne at

$$\frac{s(\bar{x}_{or})}{s(\bar{x}_{o})} \approx \sqrt{1-r^2}$$

Dette viser at det vi vinner i økt presisjon ved å gjøre bruk av observasjonene av x_1, avhenger hovedsakelig av styrken av korrelasjonen mellom x_1 og x_o.

Regresjonsestimering brukes også sammen med statifisering, men teknikken for dette har vi ikke tid til å ta med noe om her.

K.5. Om konkrete universer som gjentak i et abstrakt univers.

De metodene for estimering av parametrer i konkrete universer som er beskrevet i de foregående avsnitt, er sentrale. Men de representerer bare et utvalg av de metoder en nå har til rådighet. Vi har ikke nevnt de metodene som er arbeidet ut for estimering av sannsynligheten (eller den relative frekvens) for et konstant kjennetegn, f.eks. sannsynligheten for kjennetegnet "i det minste en traktor" på norske gardsbruk. Det faller utenfor rammen for denne fremstillingen å komme inn på de metodene en nå har for slike formål.

Å estimere parametrer i konkrete universer er en viktig oppgave i mange tilfelle. Det kan ha stor interesse i mange situasjoner at en har et godt estimat av f.eks. areal dyrket jord innen et geografisk område. Det kan også være av betydning at en har et godt estimat av volumet av grantrærne på en skogeiendom eller i et distrikt.

I mange situasjoner må en imidlertid oppfatte et slikt konkret univers som et gjentak i et abstrakt univers. Dette kan komme av at det kan være altfor arbeidskrevende å skaffe seg observasjon av verdien av en random variabel eller frekvensen for et kjennetegn i gjentakene i et abstrakt univers. En kan da ta ut et eller flere random sampler som representanter for et enkelt gjentak. Gjentaket som jo er gitt og begrenset, kan så oppfattes som et univers. Et par eksempler vil vise at det må være slik.

I det eksemplet vi benyttet i avsnitt K.4 (side 275), har vi betraktet de 200 ferskentrærne i en frukthage som et univers. Men i en større sammenheng vil en kanskje oppfatte disse trærne (eller frukthagen selv) som et gjentak i et abstrakt univers. Gjennomsnittsvekten av frukten pr. tre i et random sampel er da en observasjon for dette gjentaket. Det samme er regresjonsestimatet.

La oss, for å nevne et annet eksempel, tenke oss at det er utført et feltforsøk med et antall potetsorter. La oss videre tenke os at det en er ute etter, er mulige ulikheter sortene imellom når det gjelder angrepssannsynligheten for en sykdom. De observasjoner en da har bruk for,er det relative antall syke planter for hver rute. For å skaffe seg slike observasjoner, måtte en imidlertid undersøke hver enkelt plante innen hver rute, noe som naturligvis er nokså arbeidskrevende. I stedet for å undersøke alle plantene, kan en ta ut et random sampel av planter blant de planter som finnes på ruten og undersøke disse. Det relative antall syke planter i dette samplet er da en forventningsrett estimator av det relative antall syke planter på hele ruten. Ruten eller forsøksenheten spiller derfor her en dobbelt rolle. Den er et konkret univers under observasjonsprosessen. Men i en videre sammenheng er den et gjentak i et abstrakt univers.

I slike tilfelle er differensen mellom parameterverdien og estimatet en observasjonsfeil. Gjelder det et forsøk, utgjør denne differensen en del av restleddet (e) i modellen for forsøket. Har derfor estimatoren lav presi-

sjon, kan observasjonsfeilen gi et ikke uvesentlig bidrag i ugunstig retning til presisjonen av en kontrastestimator. Men som oftest er det nok heterogeniteten i forsøksmaterialet og samspillet mellom forsøksfaktoren og heterogenitetsfaktorene som er avgjørende for presisjonen.

At det kan bli altfor arbeidskrevende å skaffe seg observasjoner uten å bruke et eller flere random sampler innen hvert gjentak, er ikke ualminnelig. For i slike tilfelle å skaffe seg mest mulig presise estimater, kan en ta i bruk en av de metodene vi har beskrevet foran eller andre metoder vi ikke har kunnet ta med i dette kurset. I aktuelle tilfelle kan det imidlertid bli vanskelig å ta standpunkt til hvor meget arbeid en bør legge i å skaffe seg et godt estimat. Dette spørsmålet kan vi imidlertid ikke komme nærmere inn på.

Tabell I

Students t

f	P		
	0,05	0,02	0,01
1	12.706	31.821	63.657
2	4.303	6.965	9.925
3	3.182	4.541	5.841
4	2.776	3.747	4.604
5	2.571	3.365	4.032
6	2.447	3.143	3.707
7	2.365	2.998	3.499
8	2.306	2.896	3.355
9	2.262	2.821	3.250
10	2.228	2.764	3.169
11	2.201	2.718	3.106
12	2.179	2.681	3.055
13	2.160	2.650	3.012
14	2.145	2.624	2.977
15	2.131	2.602	2.947
16	2.120	2.583	2.921
17	2.110	2.567	2.898
18	2.101	2.552	2.878
19	2.093	2.539	2.861
20	2.086	2.528	2.845
21	2.080	2.518	2.831
22	2.074	2.508	2.819
23	2.069	2.500	2.807
24	2.064	2.492	2.797
25	2.060	2.485	2.787
26	2.056	2.479	2.779
27	2.052	2.473	2.771
28	2.048	2.467	2.763
29	2.045	2.462	2.756
30	2.042	2.457	2.750
40	2.021	2.423	2.704
60	2.000	2.390	2.660
120	1.980	2.358	2.617
∞	1.960	2.326	2.576

Tabell II. Kji-kvadrat

f	P			
	0,99	0,95	0,05	0,01
1	0,000	0,004	3,841	6,635
2	0,020	0,103	5,991	9.210
3	0,115	0,352	7,815	11,341
4	0,297	0,711	9,488	13,277
5	0,554	1,145	11,070	15,086
6	0,872	1,635	12,592	16,812
7	1,239	2,167	14,067	18,475
8	1,646	2,733	15,507	20,090
9	2,088	3,325	16,919	21,666
10	2,558	3,940	18,307	23,209
11	3,053	4,575	19,675	24,725
12	3,571	5,226	21,026	26,217
13	4,107	5,892	22,362	27,688
14	4,660	6,571	23,685	29,141
15	5,229	7,261	24,996	30,578
16	5,812	7,962	26,296	32,000
17	6,408	8,672	27,587	33,409
18	7,015	9,390	28,869	34,805
19	7,633	10,117	30,144	36,191
20	8,260	10,851	31,410	37,566
21	8,897	11,591	32,671	38,932
22	9,542	12,338	33,924	40,289
23	10,196	13,091	35,172	41,638
24	10,856	13,848	36,415	42,980
25	11,524	14,611	37,652	44,314
26	12,198	15,379	38,885	45,642
27	12,879	16,151	40,113	46,963
28	13.565	16,928	41,337	48,278
29	14,256	17,708	42,557	49,588
30	14,953	18,493	43,773	50,892

Tabell III. Varianskvotienten F. P = 0,05

v_2 \ v_1	f for teller.									
	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,88	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,80	3,41	3,18	3,02	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
25	4,24	3,38	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24
26	4,22	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20
28	4,20	3,34	2,95	2,71	2,56	2,44	2,36	2,29	2,24	2,19
29	4,18	3,33	2,93	2,70	2,54	2,43	2,35	2,28	2,22	2,18
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
60	4,00	3,15	2,76	2,52	2,37	2,25	2,17	2,10	2,04	1,99
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91
∞	3,84	2,99	2,60	2,37	2,21	2,09	2,01	1,94	1,88	1,83

f_2 for nevner

Tabell III. Varianskvotienten F. P = 0,05

v_2 \ v_1	f for teller.								
	12	15	20	24	30	40	60	120	∞
1	243,9	245,9	248,0	249,0	250,1	251,1	252,2	253,3	254,3
2	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	2,69	2,62	2,54	2,50	2,47	2,43	2,38	2,34	2,30
13	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	2,20	2,13	2,05	2,00	1,96	1,91	1,86	1,81	1,76
24	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

v_2 for nevner

Tabell IV. Varianskvotienten F. P = 0,025

v_2 \ v_1	f for teller.									
	1	2	3	4	5	6	7	8	9	10
1	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,7	963,3	968,6
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77
21	5,83	4,42	3,82	3,48	3,25	3,09	2,97	2,87	2,80	2,73
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70
23	5,75	4,35	3,75	3,41	3,18	3,02	2,90	2,81	2,73	2,67
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64
25	5,69	4,29	3,69	3,35	3,13	2,97	2,85	2,75	2,68	2,61
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59
27	5,63	4,24	3,65	3,31	3,08	2,92	2,80	2,71	2,63	2,57
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55
29	5,59	4,20	3,61	3,27	3,04	2,88	2,76	2,67	2,59	2,53
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27
120	5,15	3,80	3,23	2,89	2,67	2,52	2,39	2,30	2,22	2,16
∞	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05

f for nevner.

Tabell IV. Varianskvotienten f. P = 0,025

v_2 \ v_1	f for teller.								
	12	15	20	24	30	40	60	120	∞
1	976,7	984,9	993,1	997,2	1001.	1006	1010	1014	1018
2	39,41	39,43	39,45	39,46	39,46	39,47	39,48	39,49	39,50
3	14,34	14,25	14,17	14,12	14,08	14,04	13,99	13,95	13,90
4	8,75	8,66	8,56	8,51	8,46	8,41	8,36	8,31	8,26
5	6,52	6,43	6,33	6,28	6,23	6,18	6,12	6,07	6,02
6	5,37	5,27	5,17	5,12	5,07	5,01	4,96	4,90	4,85
7	4,67	4,57	4,47	4,42	4,36	4,31	4,25	4,20	4,14
8	4,20	4,10	4,00	3,95	3,89	3,84	3,78	3,73	3,67
9	3,87	3,77	3,67	3,61	3,56	3,51	3,45	3,39	3,33
10	3,62	3,52	3,42	3,37	3,31	3,26	3,20	3,14	3,08
11	3,43	3,33	3,23	3,17	3,12	3,06	3,00	2,94	2,88
12	3,28	3,18	3,07	3,02	2,96	2,91	2,85	2,79	2,72
13	3,15	3,05	2,95	2,89	2,84	2,78	2,72	2,66	2,60
14	3,05	2,95	2,84	2,79	2,73	2,67	2,61	2,55	2,49
15	2,96	2,86	2,76	2,70	2,64	2,59	2,52	2,46	2,40
16	2,89	2,79	2,68	2,63	2,57	2,51	2,45	2,38	2,32
17	2,82	2,72	2,62	2,56	2,50	2,44	2,38	2,32	2,25
18	2,77	2,67	2,56	2,50	2,44	2,38	2,32	2,26	2,19
19	2,72	2,62	2,51	2,45	2,39	2,33	2,27	2,20	2,13
20	2,68	2,57	2,46	2,41	2,35	2,29	2,22	2,16	2,09
21	2,64	2,53	2,42	2,37	2,31	2,25	2,18	2,11	2,04
22	2,60	2,50	2,39	2,33	2,27	2,21	2,14	2,08	2,00
23	2,57	2,47	2,36	2,30	2,24	2,18	2,11	2,04	1,97
24	2,54	2,44	2,33	2,27	2,21	2,15	2,08	2,01	1,94
25	2,51	2,41	2,30	2,24	2,18	2,12	2,05	1,98	1,91
26	2,49	2,39	2,28	2,22	2,16	2,09	2,03	1,95	1,88
27	2,47	2,36	2,25	2,19	2,13	2,07	2,00	1,93	1,85
28	2,45	2,34	2,23	2,17	2,11	2,05	1,98	1,91	1,83
29	2,43	2,32	2,21	2,15	2,09	2,03	1,96	1,89	1,81
30	2,41	2,31	2,20	2,14	2,07	2,01	1,94	1,87	1,79
40	2,29	2,18	2,07	2,01	1,94	1,88	1,80	1,72	1,64
60	2,17	2,06	1,94	1,88	1,82	1,74	1,67	1,58	1,48
120	2,05	1,94	1,82	1,76	1,69	1,61	1,53	1,43	1,31
∞	1,94	1,83	1,71	1,64	1,57	1,48	1,39	1,27	1,00

f for nevner.

Tabell V. Varianskvotienten F. P = 0,01

v_2 \ v_1 f for teller	1	2	3	4	5	6	7	8	9
1	4052	4999,5	5403	5625	5764	5859	5928	5982	6022
2	98,49	99,01	99,17	99,25	99,30	99,33	99,36	99,37	99,39
3	34,12	30,81	29,46	28,71	28,24	27,91	27,67	27 49	27,35
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10.16
6	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94
11	9,65	7,20	6,22	5,67	5,32	5,07	4,89	4,74	4,63
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39
13	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19
14	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89
16	8,53	6,23	5,29	4,77	4.44	4,20	4,03	3,89	3,78
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68
18	8,28	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40
22	7,94	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26
25	7,77	5,57	4,68	4,18	3,86	3,63	3,46	3,32	3,22
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56
∞	6,64	4,60	3,78	3,32	3,02	2,80	2,64	2,51	2,41

Tabell V, Varianskvotienten F. P = 0,01

V_2 \ V_1	f for teller.									
	10	12	15	20	24	30	40	60	120	∞
1	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

f for nevner.

Tabell VI. Varianskvotienten F. $f_1=1$.

f_2	P = 0,05	P = 0,025	P = 0,01
1.0	161.4	648	4052
1.1	105.9	375	1992
1.2	74.7	240	1110
1.3	56.1	166	682
1.4	44.4	121	452
1.5	36.2	92.9	308
1.6	30.4	74.0	235
1.7	26.2	60.8	181
1.8	23.0	51.2	144
1.9	20.5	44.0	118
2.0	18.5	38.5	98.5
2.1	16.9	34.2	84.0
2.2	15.6	30.7	72.8
2.3	14.5	27.9	64.1
2.4	13.6	25.6	57.0
2.5	12.8	23.7	51.3
2.6	12.1	22.0	46.6
2.7	11.5	20.6	42.7
2.8	11.0	19.4	39.4
2.9	10.5	18.4	36.6
3.0	10.1	17.4	34.1
3.1	9.7	16.6	32.0
3.2	9.4	15.9	30.2
3.3	9.1	15.3	28.6
3.4	8.9	14.7	27.2
3.5	8.7	14.2	25.9
3.6	8.5	13.7	24.8
3.7	8.3	13.3	23.7
3.8	8.1	12.9	22.8
3.9	7.9	12.5	22.0
4.0	7.7	12.2	21.2
4.1	7.5	11.9	20.5
4.2	7.4	11.6	19.9
4.3	7.3	11.3	19.3
4.4	7.2	11.0	18.8
4.5	7.1	10.8	18.3
4.6	7.0	10.6	17.8
4.7	6.9	10.4	17.4
4.8	6.8	10.3	17.0
4.9	6.7	10.2	16.6