6/60

519.2

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

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

over

MATEMATIKK

og

STATISTIKK

ved

NORGES LANDBRUKSHØGSKOLE

I. Elementære matematiske emner

II. Statistikk

MEIERIINSTITUTTETS AVDELING
FOR KJEMI, BAKTERIOLOGI OG
KONSUMMJØLK

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

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

over

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

ved

NORGES LANDBRUKSHØGSKOLE

I. Elementære matematiske emner

Skrivemaskinstua
Oslo sept. 1946

Innhold :

I. Elementære matematiske emner.

1. Om funksjoner.

Fra geometrien kjenner vi til hvordan vi skal beregne en sirkels periferi når vi har oppgitt radien. Betegner vi radien med x og periferien med y, er sammenhengen mellom disse to størrelser uttrykt i formelen

$$y = 2\pi x \qquad (\pi = 3{,}14159265...)$$

Radien og periferien er to egenskaper ved den forestilling som vi kaller sirkelen. Verdien av disse to egenskaper er bundet sammen med hverandre på en bestemt måte. Når radien forandres, vil også periferien forandre seg og det på en helt bestemt måte. En uttrykker dette ved å si at radien og periferien står i funksjonsforhold til hverandre.

La oss med x betegne en størrelse som vi etter forgodtbefinnende kan tildele verdier mellom en nedre grense (a) og en øvre grense (b). La videre y være en annen størrelse som er bundet sammen med x på en slik måte at verdien av y er gitt når verdien av x er valt. Til enhver verdi som vi kan tildele x innen intervallet fra a til b, skal svare en bestemt verdi av y. En sier da at y er en funksjon av x.

x og y er variable størrelser og x som vi etter forgodtbefinnende kan velge verdier for mellom grensene a og b, kalles den uavhengig variable eller argumentet. y hvis verdi bestemmes av verdien for x, kalles den avhengig variable eller funksjonen. Slik som vi ovenfor har oppfattet funksjonsforholdet mellom radien og periferien, er radien den uavhengig variable og periferien den avhengig variable.

Det er tre forskjellige elementære måter å angi et funksjonsforhold på, nemlig ved 1) formler, 2) tabeller og 3) grafiske bilder. I den rene matematikk er et funksjonsforhold alltid primært gitt ved en formel mens tabeller og grafiske bilder brukes som deskriptive hjelpemidler. I empiriske vitenskaper derimot er det i regelen tabellen som er det primære, mens det grafiske bilde og formelen er avledete uttrykk for funksjonsforhold.

1. Formler.

At funksjonsforholdet mellom x og y er gitt ved en formel vil si at vi har en ligning mellom x og y, en ligning som foruten x og y i regelen også inneholder flere eller færre andre størrelser som er uavhengige av både x og y og som er å betrakte som ikke-variable eller konstante. En ordner ligningen helst slik at y blir isolert på den ene siden av likhetstegnet. En sier i dette tilfelle at y er en eksplisitt funksjon av x. Et eksempel på en eksplisitt funksjon har vi i formelen

$$y = 2\pi x$$

Et annet eksempel har vi i det matematiske pendels svingetid (y) som en

funksjon av pendellengden (x):

$$y = \pi\sqrt{\frac{x}{g}}$$

hvor g er tyngdens akselleråsjon.

Disse to eksempler viser at det er to slags konstanter. Vi har konstanter som er uavhengig av tid og sted og som vi kan kalle universelle konstanter. Tallene 2 og π er slike universelle konstanter. Om g vet vi at den forandrer seg med den geografiske bredde og med høyden over havflaten. De aller fleste av de formler vi har med å gjøre i praksis, inneholder som regel begge slags konstanter.

Formelen er det enkleste middel en har til å angi et funksjonsforhold på. Når funksjonsforholdet er gitt på denne måten, kan vi i mange tilfelle, men ikke alltid, ved hjelp av elementære midler beregne verdien av den avhengig variable når verdien av den uavhengig variable er oppgitt. Når en sirkels radius er oppgitt, kan vi lett beregne periferien og når et matematisk pendels lengde er oppgitt, kan vi beregne svingetiden forutsatt at også tyngdens akselleråsjon er kjent på det sted der pendlet tenkes opphengt.

Selv om funksjonsforholdet er gitt ved en formel, er det meget ofte hensiktsmessig å bruke andre midler til å beskrive det på. Formlene er i regelen så innviklet at en må bruke andre framstillingsmåter for å kunne danne seg et inntrykk av funksjonens karakter. En tabulerer da funksjonsforholdet eller framstiller det grafisk. Men også i mange tilfelle der formelen er enkel nok, er det praktisk og nødvendig å bruke tabellen og den grafiske metode.

Vi har ovenfor holdt oss til det enkle tilfelle at der er bare to variable størrelser. I praksis har vi imidlertid meget ofte å gjøre med en variabel størrelse som er en funksjon ikke bare av en, men av to eller flere andre variable størrelser som er uavhengige av hverandre innbyrdes. Volumet (v) av en kasse er en funksjon av kassens lengde (l), bredde (b) og høyde (h). Funksjonsforholdet er gitt ved formelen:

$$v = l.b.h$$

Den arbeidsmengde som utføres er avhengig av arbeidets art, av arbeidsstyrken (antall mann), arbeidsdagens lengde og av antall arbeidsdager som brukes på arbeidet. Betegner a arbeidsmengden, m antall mann, t arbeidsdagens lengde i timer, d antall arbeidsdager og er c en størrelse som karakteriserer arbeidet, kan en som en første tilnærmelse for funksjonsforholdet bruke formelen

$$a = c.m.t.d$$

Denne formel er den som brukes i praksis. Den gir et tilnærmet riktig uttrykk for funksjonsforholdet innenfor rimelige variasjonsgrenser for de uavhengig variable. Funksjonsforholdet er imidlertid i virkeligheten langt mer komplisert enn denne formel.

Renten (r) av en kapital er en funksjon av kapitalens størrelse, (k), av rentefoten (p % p.a.) og av forrentningstiden (t). Dette funksjonsforhold er gitt ved formelen:

$$r = k.t.p/100$$

2. Tabeller.

Er funksjonsforholdet primært gitt ved en formel, bruker en som allerede nevnt ofte tabellen når en skal undersøke funksjonsforholdet nærmere. Tabellen er således bindeleddet mellom formelen og det grafiske bilde. Når det gjelder empiriske funksjonsforhold er tabellen som regel det primære, formelen og det grafiske bilde avledete uttrykk for funksjonsforholdet.

Når en skal stille opp en tabell, gjelder det å stille den opp slik at den blir oversiktlig og lett å finne fram i. Grunnlaget for tabellen er en rekke sammenhørende verdier av de to eller flere variable som skal tabuleres hva nå enten disse først må beregnes av en formel eller de er gitt primært. En skriver opp disse sammenhørende verdier ved siden av hverandre etter at verdiene av den (eller de) uavhengig variable er ordet i en rekkefølge fra den minste verdi til den største, eller omvendt.

Det er to prinsipper som går igjen i de tabellene vi har bruk for. Det er tabellen med en inngang og tabellen med to innganger. Vi skal først se på et eksempel på den første slags tabell som er den som oftest brukes når en skal tabulere funksjonsforholdet mellom to variable. Tabell 1 er et eksempel på en tabell med en inngang. Den angir funksjonsforholdet mellom vanndampens maksimumstrykk (e gitt i mm Hg.) og temperaturen (t gitt i C^{o}), for variasjonsområdet $t = 0^{o}$ til $t = 19^{o}$.

Tabell 1

t	e	t	e
0	4,6	10	9,1
1	4,9	11	9,8
2	5,3	12	10,4
3	5,7	13	11,1
4	6,1	14	11,9
5	6,5	15	12,7
6	7,0	16	13,5
7	7,5	17	14,4
8	8,0	18	15,3
9	8,5	19	16,3

For å spare plass bruker en imidlertid også i dette tilfelle tabeller med to innganger. Skulle Tabell 1 føres videre til f.eks. å omfatte alle hele gradtall fra 0° til 99°, ville den oppta fem ganger så meget plass som den refererte tabell. Vi skal se at vi kan nøye oss med meget mindre plass når vi bruker en tabell med to innganger. Først skal vi imidlertid bruke en tabell med to innganger til å tabulere funksjonsforholdet mellom tre variable. Tabell 2 er et eksempel på en slik tabell. Her er tabulert funksjonsforholdet mellom det barometriske høydetrin som avhengig variabel (tallene inne i tabellen) og lufttrykk (P) og temperatur (t) som uavhengig variable.

Tabell 2

P \ t	16°	14°	12°	10°	8°	6°	4°	2°	0°
780	10,89	10,82	10,74	10,66	10,58	10,50	10,42	10,33	10,26
770	11,04	10,96	10,88	10,80	10,72	10,64	10,56	10,47	10,39
760	11,19	11,11	11,02	10,94	10,86	10,78	10,70	10,61	10,53
750	11,34	11,25	11,17	11,08	11,01	10,92	10,84	10,75	10,67
740	11,49	11,41	11,32	11,23	11,16	11,07	10,98	10,90	10,82
730	11,64	11,55	11,47	11,38	11,30	11,21	11,13	11,04	10,97
720	11,81	11,72	11,64	11,55	11,45	11,38	11,29	11,20	11,12
710	11,98	11,89	11,80	11,71	11,63	11,54	11,45	11,35	11,28
700	12,15	12,06	11,97	11,87	11,79	11,70	11,61	11,52	11,44
690	12,33	12,23	12,14	12,05	11,97	11,87	11,78	11,70	11,60
680	12,51	12,41	12,32	12,22	12,14	12,05	11,95	11,87	11,77
670	12,70	12,60	12,51	12,41	12,33	12,23	12,14	12,05	11,95
660	12,89	12,79	12,70	12,60	12,51	12,41	12,32	12,23	12,13

Ved det barometriske høydetrin forstår en høyden i meter av en luftsøyle som med hensyn til trykk ekvivalerer 1 mm. Hg. Forutsetningen ~~for denne tabell over~~ funksjonsforholdet er ~~at luftens relative fuktighet~~ er 75 %. Når en skal avlese det barometriske høydetrin av denne tabellen, går en inn i tabellen to veier, nemlig over P og over t. Det barometriske høydetrin svarende til P = 720 og t = 8 er altså 11,45 meter.

Som nevnt bruker en også for å økonomisere med plassen en tabell med to innganger til å tabulere funksjonsforholdet mellom to variable. Vi skal som eksempel utvide Tabell 1 til å omfatte alle hele gradtall fra $t=0^{\circ}$ til $t=99^{\circ}$. Vi setter da t = a+10.b hvor a er antall enere i tallet og b er antallet av tiere i tallet. Til eksempel er t=53 = 3 + 10.5 (a=3 og b=5). Vi tabulerer nå enerne (a) i tabellens hode, tierne i venstre kollonne og vanndampens maksimumstrykk (e) i rekker og kolonner inne i tabellen slik som vist i Tabell 3.

Tabell 3.

a / 10.b	0	1	2	3	4	5	6	7	8	9
0	4,6	4,9	5,3	5,7	6,1	6,6	7,0	7,5	8,0	8,5
10	9,1	9,8	10,4	11,1	11,9	12,7	13,5	14,4	15,3	16,3
20	17,4	18,5	19,6	20,9	22,2	23,5	25,0	26,5	28,1	29,7
30	31,5	33,4	35,3	37,4	39,5	41,8	44,2	46,6	49,3	52,0
40	54,9	57,9	61,0	64,3	67,8	71,4	75,1	79,1	83,2	87,5
50	92,0	96,7	101,6	106,7	112,0	117,5	123,3	129,3	135,6	142,1
60	148,9	156,0	163,3	171,0	178,9	187,1	195,7	204,6	213,8	223,4
70	233,3	243,6	254,3	265,4	276,9	288,8	301,1	313,8	327,1	340,7
80	354,9	369,5	384,6	400,3	416,5	433,2	450,5	468,3	486,8	505,8
90	525,5	545,8	566,7	588,3	610,6	633,7	657,4	681,9	707,1	733,2

Skal en tabulere en størrelse som funksjon av tre uavhengig variable, kan en ikke klare seg med en enkelt tabell. En må ty til et helt sett av tabeller, en tabell over funksjonsforholdet mellom den avhengig variable og to av de uavhengig variable for hver enkelt valt verdi av den tredje uavhengig variable. Funksjonsforholdet mellom det barometriske høydetrin, lufttrykk, temperatur og relativ fuktighet må således framstilles i en hel bunke trykk-temperaturtabeller av samme slag som Tabell 2, en tabell for relativ fuktighet lik 10 %, en tabell for relativ fuktighet lik 20 % osv.

3. Grafisk framstilling.

Formler gir bare i de aller færreste tilfelle en klar forestilling om funksjonsforholdet. Det samme kan sies om tabellene. Vår evne til å oppfatte lange tallrekker i sin helhet er nemlig sterkt begrenset. Ved å framstille funksjonsforholdet grafisk oppnår vi imidlertid å skaffe oss et helhetsbilde av sammenhengen.

Vi har forskjellige slags grafiske framstillingsmåter. De viktigste og mest brukte er kurvefremstillingene. Vi minner om at vi som grunnlag for disse bruker et aksesystem som består av en horisontal akse (abscisseaksen) og en vertikal akse (ordinataksen). På den horisontale akse danner vi oss en målestokk for den uavhengig variable og på den vertikale akse en målestokk for den avhengig variable. Skjæringspunktet mellom aksene brukes som oftest, men ikke alltid, som nullpunkt for begge variable.

Er funksjonsforholdet primært gitt ved en formel, må det først tabuleres. Dette gjøres ved at vi velger bestemte verdier for den uavhengig variable og beregner av formelen de tilsvarende verdier av den avhengig variable. Er funksjonsforholdet primært gitt ved en tabell, brukes denne uten videre som grunnlag for den grafiske framstilling. Vi avsetter verdiene av den uavhengig

variable fra nullpunktet - positive verdier utover til høyre og negative verdier utover til venstre - langs abscisseaksen, oppreiser perpendikulærer i de punktene som representerer verdiene av den uavhengig variable og avsetter langs disse perpendikulærer verdiene av den avhengig variable. Når dette er utført for alle sammenhørende verdier av de to variable, har vi funksjonsforholdet mellom størrelsene framstilt i planet ved en rekke punkter (endepunktene av alle perpendikulærene). Gjennom disse punkter kan vi nå i de fleste tilfelle legge en sammenhengende linje (kurve) hvis form vi prøver å slutte oss til så godt vi kan på basis av den måte punktene ligger plasert i forhold til hverandre. Denne kurven er da det grafiske bilde av funksjonsforholdet.

Å legge en slik ubrutt kurve gjennom punktene på en riktig og forsvarlig måte, er ikke alltid noen enkel oppgave. Og det må naturligvis ikke gjøres for andre funksjoner enn de som kan framstilles av en sammenhengende kurve, de såkalte kontinuerlige funksjoner. Jo mindre avstand det er mellom de valte eller på forhånd gitte verdier av den uavhengig variable, jo lettere er det å trekke opp kurven. I regelen vil en stå seg på - iallfall til en begynnelse - å nøye seg med å trekke rette linjer fra punkt til punkt og får derved funksjonsforholdet framstilt ved et rettlinjet linjediagram. Hva enten en bruker den ene eller den annen framstillingsmåte, må en vise stor forsiktighet fordi det er meget lett å forvanske inntrykket av funksjonsforholdet. Et eksempel på kurveframstilling er gitt i Figur 1. Det er et bilde av funksjonen

$$y = \frac{x^3 - 12x + 3}{10}$$

Det er imidlertid noen funksjoner som ikke lar seg framstille av en sammenhengende kurve, de såkalte diskontinuerlige funksjoner. Vi kan her skjelne mellom to slags. Vi har funksjoner som er kontinuerlige stykkevis innen det variasjonsområde for den uavhengig variable som vi er interessert i. Disse funksjoner lar seg framstille grafisk av en sammenhengende kurve som i et eller flere punkter plutselig blir avbrutt og etterfølges av en ny sammenhengende kurve som begynner i et annet nivå enn der hvor den første kurven slutter. Vi skal ikke komme nærmere inn på disse funksjoner da vi ikke senere vil ha noe bruk for dem. Viktigere er den slags diskontinuerlige funksjoner som er diskontinuerlige fordi det bare er visse bestemte verdier av den uavhengig variable - som regel ekvidistante verdier - som det er noen mening i å bruke. Ofte er det slik at funksjonen ikke er definert for andre verdier av den uavhengig variable enn de som er hele tall (f.eks. 0,1,2,3,...). Vi skal nevne et eksempel som vi senere kommer tilbake til i Sannsynlighetsregningen.

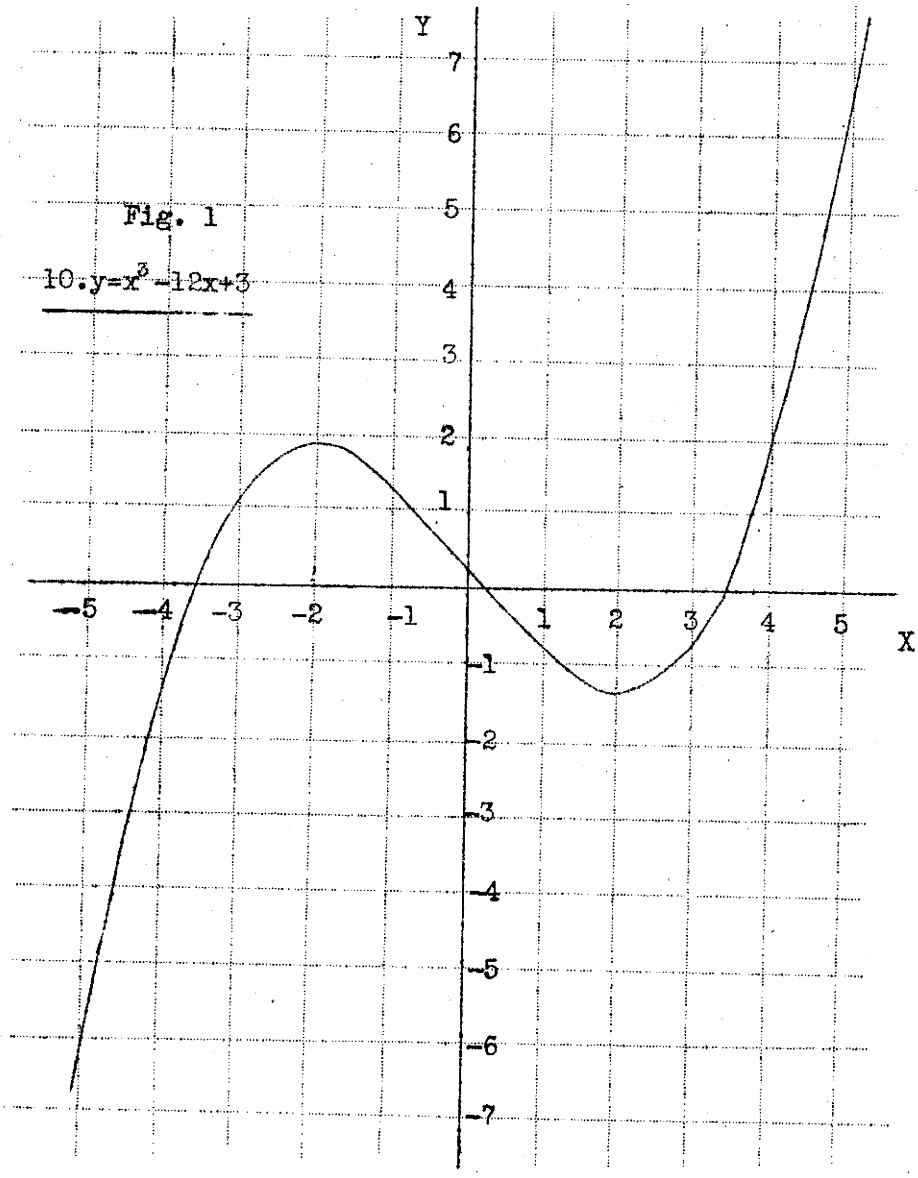


Fig. 1

10. $y=x^3-12x+3$

Funksjonen

$$y = 1024.\frac{4.3\ldots(4-x)}{1.2\ldots. x}\ (\tfrac{1}{2})^4$$

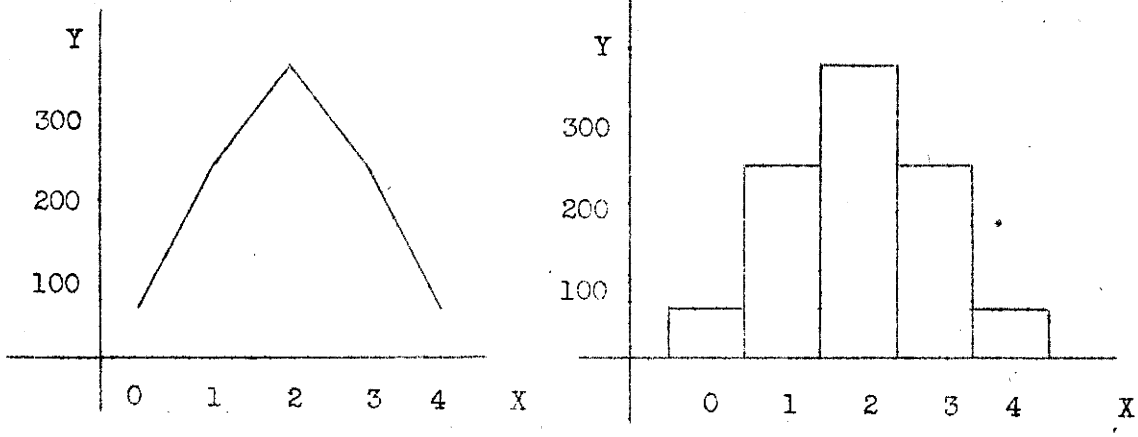
er bare definert for verdiene x=0, x=1, x=2, x=3 og x=4. Vi skal senere komme inn på fortolkningen av funksjonen. Her må vi nøye oss med å referere en tabell over sammenhørende verdier av x og y. Disse er gitt i Tabell 4.

Tabell 4.

x	y
0	64
1	256
2	384
3	256
4	64

Vi kan framstille slike funksjoner ved et rettlinjet linjediagram slik som beskrevet ovenfor. Vi kan også la verdiene av y representeres av søyler som har like stor bredde og hvis høyder er lik verdiene av y. Funksjonsforholdet gitt i Tabell 4 er framstilt på begge måter i Figur 2.

Fig. 2.



Vi har en rekke andre grafiske framstillingsmåter enn de som er nevnt her og som i mange tilfelle er å foretrekke framfor de metodene som baseres på et rettvinklet koordinatsystem. Vi har imidlertid ikke tid til å komme inn på disse andre metoder her. I flere lærebøker i statistikk finner en ofte en oversikt over forskjellige slags grafiske framstillingsmetoder, f.eks. i Gunnar Jahn: "Statistikkens teknikk og metode".

Vi har ovenfor holdt oss til det tilfelle at det er bare to variable, en uavhengig og en avhengig variabel. Er det flere uavhengig variable, f.eks. to, er det ikke mulig å gi en alminnelig grafisk framstilling i et plan. Er

det to uavhengig variable, skulle en bruke tre akser, en for den avhengig variable og en for hver av de to uavhengig variable. Et slikt system forutsetter rommet med dets tre utstrekninger. Å framstille et funksjonsforhold i i rommet er meget upraktisk. Vi kan imidlertid også framstille et slikt funksjonsforhold i planet på følgende måte.

La den avhengig variable være y, de to uavhengig variable være x og z. Velger vi nå en fast verdi for z, har vi bare et funksjonsforhold mellom x og y og dette funksjonsforhold kan framstilles i planet på alminnelig måte. Velger vi en rekke verdier for z, vil vi for hver enkelt z-verdi kunne framstille funksjonsforholdet mellom x og y i planet ved en kurve eller et rettlinjet linjediagram. La oss tenke oss at dette er gjort for k forskjellig valte verdier av z og i samme rettvinklete koordinatsystem. Funksjonsforholdet mellom x,y og z er da framstilt ikke av en enkelt kurve, men av k forskjellige kurver og hver enkelt av disse kurvene refererer seg til en bestemt verdi av z. Funksjonsforholdet mellom tre variable kan altså framstilles i planet ved en kurveskare. Har en mer enn tre variable, kan en gå fram på tilsvarende måte og framstille funksjonsforholdet ved flere sett av kurveskarer.

Oppgave 1.

Gi en grafisk framstilling av følgende funksjoner.

a) $y = \frac{1 - 3x}{1 + 2x}$, b) $y = \frac{2x}{1+x^2}$.

Oppgave 2.

Ved å velge forskjellige pos. og neg. verdier for z, skal man framstille funksjonsforholdet

$$y = 2x.z$$

grafisk i et plant rettvinklet koordinatsystem.

Oppgave 3.

Framstill grafisk de funksjonsforhold som er gitt i Tabell 1 og Tabell 2 (Tabell 2 framstilles av en kurveskare).

2. Proporsjonalitet.

La x være den uavhengig variable og y den avhengig variable. Er funksjonsforholdet gitt ved formelen

$$y = c.x$$

hvor c er en konstant, sier en at y er proporsjonal med x. For å få nærmere rede på hva dette betyr kan vi velge to forskjellige verdier for x, $x=x_1$ og $x=x_2$, og beregne de til disse svarende verdier av y (y_1 og y_2). Sett at $x_2=a.x_1$ hvor a er et vilkårlig valt tall. Vi har da:

$$y_1 = c.x_1$$

$$y_2 = c.x_2 = c.a.x_1$$

Ved divisjon får vi:

$$\frac{y_2}{y_1} = \frac{c.a.x_1}{c.x_1} = a$$

eller:

$$y_2 = a.y_1$$

Vi ser av dette at når vi øker ($a > 1$) eller minsker ($a < 1$) x i et bestemt forhold (a), vil y også økes eller minskes i samme forhold.

Faktoren (konstanten) c kalles proporsjonalitetsfaktoren. Setter vi $x = 1$, har vi $y = c$. c er derfor verdien av den avhengig variable når vi setter den uavhengig variable lik enheten. Det grafiske bilde av funksjonsforholdet $y = c.x$ er en rett linje som går gjennom origo (nullpunktet) og som med den positive X-aksen danner en vinkel v hvis størrelse er bestemt ved at $tg.v = c$.

Mer alminnelig kan vi skrive ligningen eller formelen for proporsjonalitet slik:

$$y + b = c.x$$

eller

$$y = c.x - b$$

I dette tilfelle er det ikke y selv som er proporsjonal med x, men y plus en konstant størrelse (b) som kan være positiv eller negativ etter forholdene. Det grafiske bilde av funksjonsforholdet $y = c.x - b$ er en rett linje som skjærer ordinataksen (Y-aksen) i en avstand fra origo lik (-b) og som med den positive X-aksen danner en vinkel bestemt ved at $tg.v = c$.

Er funksjonsforholdet mellom x og y gitt ved formelen

$$y = \frac{c}{x} \quad \text{eller} \quad x.y = c$$

er y omvendt proporsjonal med x. For å få nærmere rede på karakteren av dette funksjonsforholdet kan vi som ovenfor velge to forskjellige verdier for x, $x=x_1$ og $x=x_2=a.x_1$ hvor a er et vilkårlig valt tall. De tilsvarende verdier av y er:

$$y_1 = \frac{c}{x_1}$$

$$y_2 = \frac{c}{x_2} = \frac{c}{a.x_1}$$

Ved divisjon får vi:

$$\frac{y_2}{y_1} = \frac{1}{a}$$

eller:

$$y_2 = \frac{y_1}{a}$$

Vi ser av dette at dersom vi øker x i et visst forhold ($a > 1$), vil y minskes i samme forhold og dersom vi minsker x i et visst forhold ($a < 1$), vil y økes i samme forhold.

Konstanten c er også i dette tilfelle verdien av den avhengig variable når vi for den uavhengig variable velger verdien 1. Det grafiske bilde av funksjonsforholdet $x.y=c$ kalles en hyperbel. I Figur 3 er vist det grafiske bilde av funksjonsforholdet $x.y=5$.

Mer alminnelig kan vi skrive formelen for omvendt proporsjonalitet på følgende måte:

$$y = \frac{c}{x+b} \qquad \text{eller} \qquad y.(x+b) = c$$

Her er y omvendt proporsjonal med (x+b) hvor b er en konstant som kan være positiv eller negativ.

Eksempel 1.

Sirkelens omkrets er proporsjonal med radien : $y = 2\pi x$. Faktoren 2π er omkretsen av en sirkel hvis radius er lik 1.

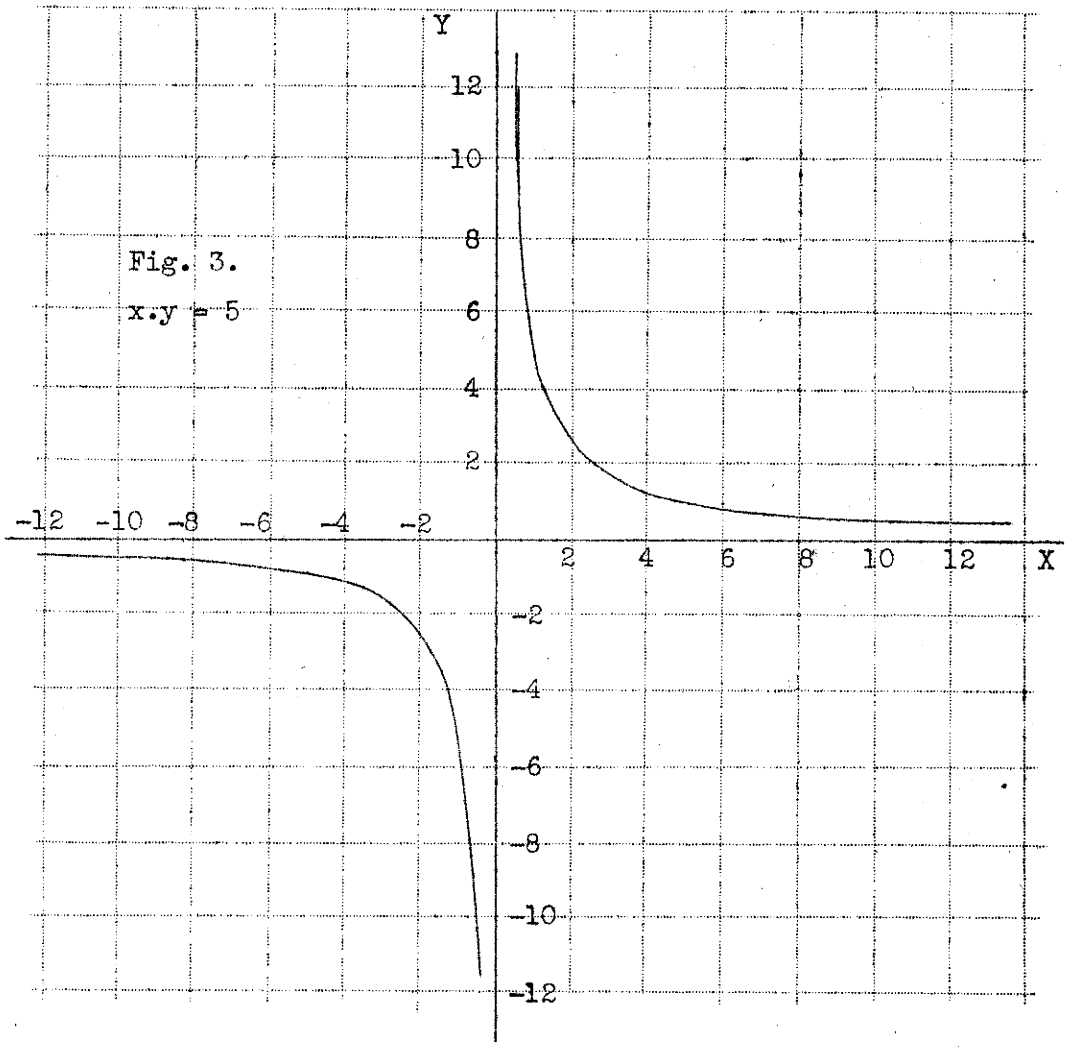


Fig. 3.
$x.y = 5$

Eksempel 2.

Arbeidsmengden (a) er proporsjonal med arbeidsstyrken (m = antall mann), med arbeidsdagens lengde (t) og med antall arbeidsdager (d)

$$a = c.m.t.d$$

Proporsjonalitetsfaktoren er lik det arbeid som utføres av en mann i en time. Setter vi arbeidsmengden konstant og antall timer pr. dag konstant og skriver formelen slik:

$$m.d = \frac{a}{ct}$$

ser vi at antall mann (m) og antall arbeidsdager (d) er omvendt proporsjonale med hverandre.

Eksempel 3.

Som kjent gjør en om temperaturen angitt i Fahrenheitgrader til Celsiusgrader ved fra Fahrenheitgradtallet å trekke 32 og multiplisere det utkomne med $\frac{5}{9}$.

Betegner vi C-gradene med y og F-gradene med x, kan vi uttrykke sammenhengen ved formelen:

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

Det grafiske bilde av funksjonsforholdet er en rett linje som skjærer ordinataksen i en avstand fra origo lik $-\frac{160}{9}$ og som med abcsisseaksen danner en vinkel (v) bestemt ved at tg $v = \frac{5}{9}$, dvs. $v = 29^{o},05$.

Eksempel 4.

Volumet (V) av en avsperret gassmasse, hvis temperatur holdes konstant, er omvendt proporsjonalt med trykket (P):

$$V = \frac{C}{P} \quad \text{eller.} \quad P.V = C$$

hvor C er en konstant hvis størrelse avhenger av hvor stor gassmasse som er under behandling og av temperaturen. Funksjonsforholdet kan uttrykkes i setningen: Produktet av volum og trykk er konstant.

Alle de oppgaver som en løser i den praktiske regning ved den framgangsmåte som går under navn av reguladetri, er basert på den forutsetning at de størrelser en opererer med, er proporsjonale eller omvendt proporsjonale med hverandre. En størrelse kan naturligvis samtidig være proporsjonal med en eller flere andre størrelser og omvendt proporsjonal med en eller flere størrelser.

Eksempel 5.

Den forlengelse som en tråd får ved strekning, er proporsjonal med trådens lengde og med belastningen og omvendt proporsjonal med trådens tverrsnitt. Betegner vi forlengelsen med Δl, trådens lengde med l, belastningen med P og tverrsnittet med q, har vi:

$$\Delta l = \frac{C.l.P}{q} \qquad \text{eller} \qquad \Delta l.q = C.l.P$$

Her er C materialkonstant som kalles strekningskoeffisienten. Den er lik forlengelsen av en tråd av samme materiale av lengde 1 meter og tverrsnitt 1 cm^2 når belastningen er lik 1 kg.

Oppgave 4.

Framstill grafisk funksjonsforholdet mellom x,y og z gitt ved formelen:

$$y = \frac{10x - 4}{2z}$$

Dette skal gjøres på to forskjellige måter: 1) ved at en velger noen faste verdier av z og 2) ved at en velger noen faste verdier av x.

3. Den inverse funksjon.

Vi har hittil forestilt oss at funksjonsforholdet er gitt på eksplisitt form, dvs. at den avhengig variable er gitt ved en formel som inneholder de uavhengig variable og konstantene. I mange tilfelle har vi imidlertid interesse av å bytte ut den avhengig variable med en av de uavhengig variable og betrakte denne som avhengig variabel. La oss til eksempel betrakte sammenhengen mellom Celsiusgrader og Fahrenheitgrader. Når vi skal gjøre om Fahrenheitgrader til Celsiusgrader, skjer dette ved hjelp av den eksplisitte funksjon

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

hvor y er Celsiusgrader og x er Fahrenheitgrader. Det er imidlertid like ofte spørsmål om å gjøre om Celsiusgrader til Fahrenheitgrader som omvendt. Da må vi betrakte Celsiusgradene (y) som uavhengig variabel og Fahrenheitgradene (x) som avhengig variabel. For å finne den formel som vi da skal regne etter, har vi bare å finne x uttrykt som en eksplisitt funksjon av y, og dette gjøres ved at vi løser den ligning som eksisterer mellom x og y med hensyn på x. Vi finner lett at

$$x = \frac{9}{5} \cdot y + 32$$

Vi gjør altså om Celsiusgrader til Fahrenheitgrader ved å multiplisere antallet Celsiusgrader med $\frac{9}{5}$ og til resultatet å addere 32 grader. Skal vi nå framstille dette funksjonsforholdet grafisk, må vi huske på at nå er det x som er avhengig variabel og skal avsettes langs ordinataksen og y som er uavhengig variabel og skal avsettes langs abscisseaksen.

Dersom y er en eksplisitt funksjon av x, kan vi uttrykke dette ved symbolet

$$y = f(x)$$

som uttales "y er lik f-funksjonen av x". I stedet for f kan vi bruke andre bokstaver (g,h,t,G,F,....). f(x) er i hvert enkelt tilfelle en formel som inneholder x som eneste variabel og dessuten de konstanter som karakteriserer funksjonsforholdet. Vi har til eksempel:

$$y = f(x) = 2\pi x \quad \text{og} \quad y = h(x) = \frac{5}{9} \cdot (x-32)$$

Av ligningen $y = f(x)$ kan vi nå i mange tilfelle finne x uttrykt som en eksplisitt funksjon av y. Resultatet kan skrives slik:

$$x = g(y)$$

hvor g(y) i hvert gitt tilfelle er en formel som inneholder y som eneste variabel, og som er direkte utledet av ligningen $y = f(x)$. For funksjonsforholdet mellom

Celsiusgrader og Fahrenheitgrader har vi

$$x = g(y) = \frac{9}{5} \cdot y + 32$$

Når funksjonsforholdet mellom x og y er skrevet på formen $x = g(y)$, er x å oppfatte som avhengig variabel og y som uavhengig variabel. Ved grafisk framstilling av funksjonsforholdet skal da x avsettes langs ordinataksen og y langs abscisseaksen. Er $x = g(y)$ utledet av $y = f(x)$, kalles $x = g(y)$ den <u>inverse</u> funksjon til $y = f(x)$.

Framstiller vi funksjonen $y = f(x)$ og dens inverse funksjon $x = g(y)$ i samme koordinatsystem, vil vi få to kurver av samme utseende som ligger forskjellig til i forhold til koordinataksene. Vi innser også lett at vi kan få kurven for $x = g(y)$ av kurven for $y = f(x)$ ved en enkel dreining av koordinatsystemet. Som akse for dreiningen kan vi bruke en rett linje gjennom origo som halverer vinkelen mellom den positive abscisseaksen og den positive ordinataksen (første kvadrant). Er nå kurven for $y = f(x)$ inntegnet, vil vi få kurven for $x = g(y)$ ved å dreie hele systemet med kurven 180° omkring den nevnte aksen. Etter denne dreining vil nemlig den positive ordinataksen falle på den plass den positive abscisseaksen hadde før dreiningen, og omvendt. Den negative ordinataksen vil innta den plass den negative abscisseaksen hadde før dreiningen, og omvendt.

La oss som eksempel ta for oss funksjonen

$$y = \sqrt{x} + 5$$

Herav finner vi at

$$x = g(y) = (y - 5)^2$$

Kurvene for disse to funksjoner er inntegnet i Figur 4.

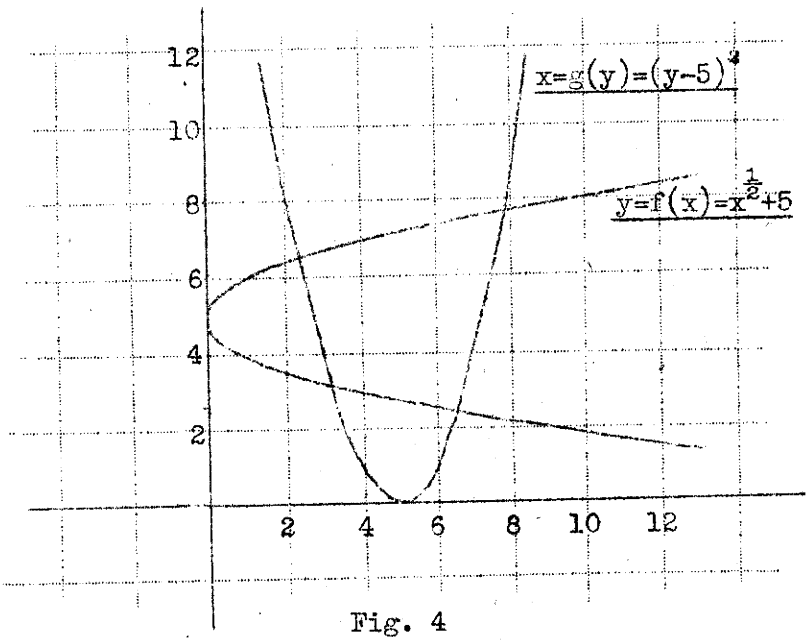


Fig. 4

Har vi flere enn to variable, f.eks. tre (x,y og z) og y er den avhengig variable, kan vi uttrykke dette ved symbolet:

$$y = f(x,z)$$

hvor f(x,z) i hvert enkelt tilfelle er en formel som inneholder x og z og konstantene. Denne funksjon har da to inverse funksjoner som vi får ved å uttrykke x som en eksplisitt funksjon av y og z, og ved å uttrykke z som en eksplisitt funksjon av x og y. Resultatene kan skrives slik:

$$x = g(y,z) \qquad \text{og} \qquad z = h(x,y)$$

Eksempel 6.

$$y = f(x,z) = \frac{10x - 4}{2z}$$

Herav finner vi lett at:

$$z = h(x,y) = \frac{10x - 4}{2y}$$

og

$$x = g(y,z) = \frac{2yz + 4}{10}$$

I mange tilfelle er funksjonsforholdet mellom de variable gitt på såkalt implisitt form, dvs. funksjonsforholdet er gitt ved en alminnelig ligning som inneholder de variable og de konstanter som er med på å karakterisere funksjonsforholdet. Til eksempel kan funksjonsforholdet

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

skrives på formen:

$$5x - 9y - 160 = 0$$

Den implisitte skrivemåte kan vi uttrykke ved symbolet

$$f(x,y) = 0$$

når det er to variable (x og y). f(x,y) er da i hvert enkelt tilfelle en formel som inneholder x og y og konstantene. Er det tre variable, kan vi uttrykke dette ved symbolet: $h(x,y,z) = 0$.

I eksemplet ovenfor har vi:

$$f(x,y) = 5x - 9y - 160 = 0$$

Når funksjonsforholdet er gitt på implisitt form, kan vi komme over til den eksplisitte form ved å løse ligningen med hensyn på en av de variable hvis dette er mulig.

Eksempel 7.

$$F(x,y,z) = 10x - 2yz - 4 = 0$$

Til denne implisitte funksjon svarer følgende tre eksplisitte funksjonsuttrykk:

$$x = \frac{2yz + 4}{10}, \quad y = \frac{10x - 4}{2z} \quad \text{og } z = \frac{10x - 4}{2y}$$

Oppgave 5.

$$f(x,y) = x.y - 2x + 3y - 4 = 0$$

Uttrykk først y som eksplisitt funksjon av x og dernest x som eksplisitt funksjon av y. Framstill grafisk begge funksjoner i samme rettvinklete koordinatsystem.

4. Eksponensialfunksjonen og logaritmefunksjonen.

Ved en potens a^n forstår en når n er et helt positivt tall et produkt av n a'er:

$$a^n = a.a.a.a.....a$$

Når a er et positivt tall er imidlertid potensen a^x definert for alle reelle verdier av eksponenten (x). Oppfatter en nå potensen som en funksjon av eksponenten, har en den såkalte eksponensialfunksjon:

$$y = a^x$$

Funksjonen er kontinuerlig. Dens grafiske bilde er en sammenhengende kurve som i hele sin utstrekning ligger på oversiden av abscisseaksen.

Oppgave 6.

Framstill grafisk funksjonene: a) $y = (\frac{1}{4})^x$, b) $y = (\frac{1}{2})^x$, c) $y = 2^x$ og d) $y = 3^x$ i samme koordinatsystem.

Vi skal repetere hva en forstår ved visse potensuttrykk.

1) Potens med brøkeksponent:

$$a^{\frac{t}{n}} = \sqrt[n]{a^t}$$

2) Potens med negativ eksponent:

$$a^{-m} = \frac{1}{a^m}$$

3) Potens med eksponenten 0:

$$a^0 = 1$$

For å finne eksponensialfunksjonens inverse funksjon måtte vi løse ligningen $y = a^x$ med hensyn på x. Vi ser at x er den eksponent som et gitt tall (a) skal opphøyes i for å gi y som resultat. Og vi husker da at dette er logaritmen til y i det logaritmesystem som har a til grunntall. Logaritmen til y er eksponenten i den potens som en gitt rot (a) skal opphøyes i for at potensen skal bli lik y. Den inverse funksjon til $y = f(x) = a^x$ er derfor

$$x = g(y) = \log_a y$$

Den siste funksjon kalles logaritmefunksjonen.

I Figur 5 er gitt en grafisk framstilling av funksjonen $y = 2^x$ og dens inverse funksjon $x = \log_2 y$.

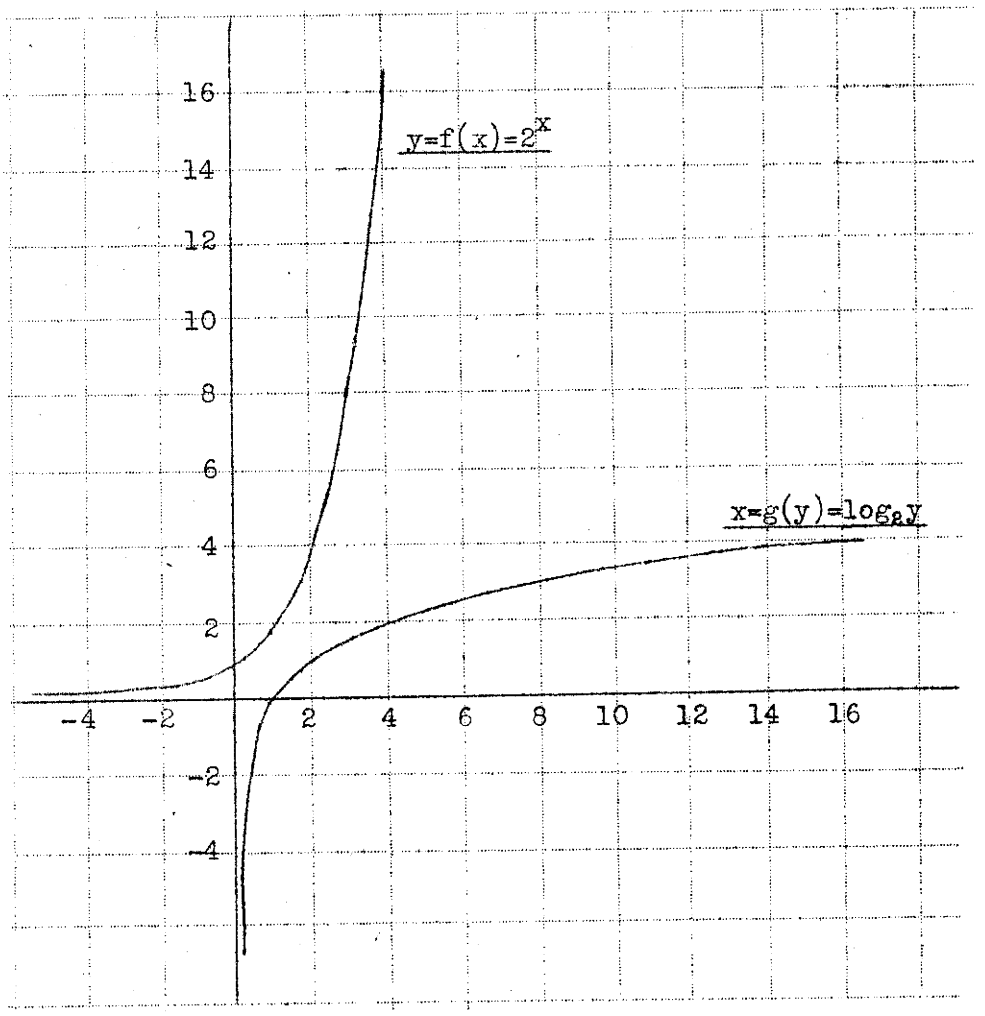


Fig. 5

Verdiene av logaritmefunksjonen for forskjellige verdier av den uavhengig variable kan ikke beregnes ved hjelp av elementære midler. Det samme gjelder eksponensialfunksjonen med unntagelse av hele verdier av den uavhengig variable. Men begge funksjoner er beregnet og tabulert i logaritmetabeller. Den egentlige logaritmetabell gir x=logy når y er gitt og antilogaritmetabellen gir $y = a^x$ når x er gitt. Begge tabeller er ordnet med to innganger. De logaritmetabeller som brukes hos oss, er bygget på grunntallet a = 10. Dette er det Briggske system. Vi har et annet system som bygger på det irrasjonale tall: e = 2,71828.... som grunntall. Dette system kalles det naturlige logaritmesystem.

Den tekniske fordel ved logaritmeregningen er den at regneoperasjonene multiplikasjon og divisjon reduseres til de enklere regneoperasjoner addisjon

og subtraksjon. Potensering og rotutdraing reduseres til multiplikasjon og divisjon. Dette vil framgå av følgende operasjonssetninger for logaritmeregning.

1) Logaritmen til et produkt er lik summen av faktorenes logaritmer. La grunntallet være a. Da er:

$$b = a^{\log b} \qquad \text{og} \qquad c = a^{\log c}$$

Altså er:

$$b.c = a^{\log b} \cdot a^{\log c} = a^{\log b + \log c}$$

Følgelig er

$$\log (b.c) = \log b + \log c$$

2) Logaritmen til en brøk er lik differensen mellom tellerens logaritme og nevnerens logaritme, eller:

$$\log \frac{b}{c} = \log b - \log c \qquad \text{(Bevis dette).}$$

3) Logaritmen til en potens er lik logaritmen til roten multiplisert med eksponenten. Vi har at

$$b^n = (a^{\log b})^n = a^{n.\log b}$$

og derfor er:

$$\log b^n = n.\log b$$

4) Logaritmen til en rot er lik logaritmen til radikanden dividert med roteksponenten, eller:

$$\log \sqrt[n]{b} = \frac{\log b}{n} \qquad \text{(Bevis dette)}$$

5. Binomialformelen.

Binomialformelen er en formel som har fått utstrakt anvendelse under løsningen av en rekke forskjellige oppgaver både av rent teoretisk og praktisk art. Formelen gir det alminnelige uttrykk for hvordan $(q+p)^n$ lar seg skrive som en sum av ledd av formen $A_x.q^{n-x}.p^x$.

For n = 0,1,2,3,4,.... får vi direkte ved multiplikasjon:

$(q+p)^0 = 1$

$(q+p)^1 = q + p$

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

$(q+p)^3 = q^3 + 3q^2p + 3qp^2 + p^3$

$(q+p)^4 = q^4 + 4q^3p + 6q^2p^2 + 4qp^3 + p^4$

osv.

I alminnelighet har vi at

$$(q+p)^n = q^n + nq^{n-1}p + \frac{n(n-1)}{1.2}q^{n-2}p^2 + \frac{n(n-1)(n-2)}{1.2.3}q^{n-3}p^3$$

$$+ \ldots\ldots\ldots\ldots + \frac{n(n-1)}{1.2}q^2p^{n-2} + nqp^{n-1} + p^n$$

Denne formelen kan utvikles ved hjelp av elementære hjelpemidler, men denne utvikling skal vi ikke ta med her. At formelen er riktig kan bevises ennå enklere, ved det såkalte induksjonsbevis. Vi forutsetter at alle studentene har gjennomgått dette bevis og skal derfor ikke ta det med. Vi skal heller nytte tiden til å se nærmere på leddene i formelen og særlig da på koeffisientene, de såkalte binomiske koeffisienter.

Vi ser av formelen at i alle ledd unntatt det første og det siste, forekommer både en potens av q og en potens av p som faktorer. Vi legger videre merke til at summen av eksponentene i disse potenser er den samme i alle ledd, nemlig lik n. Dette gjelder forresten også første og siste ledd. Første ledd kan nemlig skrives slik: $q^n = q^n.p^o$, og siste ledd kan skrives slik: $p^n = q^o.p^n$. Eksponenten til den potens som q er opphøyet i, avtar fra n med en enhet fra ledd til ledd når rekken leses fra venstre mot høyre, mens eksponenten til den potens som p er opphøyet i, øker med en enhet fra ledd til ledd.

Når det gjelder koeffisientene, skal vi først merke oss at de to ledd som har samme avstand fra rekkens to ender, har samme koeffisient. Formelen er altså symmetrisk i koeffisientene.

Det alminnelige, det (x+1)'te ledd, i formelen er:

$$\frac{n(n-1)(n-2)\ldots\ldots(n-x+1)}{1.2.3.4\ldots\ldots x}\ q^{n-x}p^x$$

Koeffisienten er dannet på den måte at telleren er produktet av alle hele positive tall fra n nedover til (n-x+1) og nevneren er produktet av alle hele positive tall fra 1 oppover til og med x. La oss multiplisere i teller og nevner med produktet av alle hele positive tall fra 1 oppover til (n-x), altså med

$$1.2.3.4.5\ldots\ldots\ldots\ (n-x)$$

For koeffisienten har vi da:

$$\frac{n(n-1)(n-2)\ldots(n-x+1)}{1.2.3\ldots\ldots\ x} = \frac{n(n-1)\ldots\ldots(n-x+1).(n-x)\ldots3.2.1}{1.2.3\ldots\ldots x.\ 1.2.3\ldots\ldots\ (n-x)}$$

I telleren har vi nå produktet av alle hele tall fra 1 oppover til n, i nevneren har vi produktet av alle hele tall fra 1 til x og produktet av alle hele tall fra 1 til (n-x).

For produktet av alle hele pos. tall fra 1 oppover til et bestemt tall, f.eks. til r, er innført en bestemt betegnelse. Vi skriver:

$$1.2.3.4.5.....(r-1).r = r!$$

og uttaler dette "r fakultet".

Eksempelvis har vi:

$1 = 1!$, $1.2 = 2!$, $1.2.3 = 3!$, $1.2.3.4 = 4!$ osv.

Et spesïaltilfelle har vi når r=0. Dette tilfelle faller helt utenfor den definisjon vi ovenfor har gitt av r! Vi kan ikke forklare sammenhengen uten å komme inn på vanskeligere matematiske emner og må derfor her nøye oss med å slå fast at en ved "0 fakultet" skal forstå enheten, eller $0! = 1$

Ved å bruke fakultetsbetegnelsen kan vi skrive det alminnelige ledd i binomialformelen på kortere form. Vi kan skrive det slik:

$$\frac{n(n-1)(n-2)..... (n-x+1)}{1.2.3.4.....x} q^{n-x}p^x = \frac{n!}{x!\,(n-x)!} q^{n-x}p^x$$

Vi innser nå lett at hvis vi i dette alminnelige uttrykk innsetter x=0, får vi formelens første ledd. Vi har nemlig

$$\frac{n!}{0!\,(n-0)!} q^{n-0}p^0 = \frac{n!}{0!\,n!} q^n.1 = q^n$$

Annet ledd i formelen får vi ved å innsette x=1. Vi har:

$$\frac{n!}{1!\,(n-1)!} q^{n-1}p^1 = \frac{n.\,(n-1)!}{1.(n-1)!} q^{n-1}p = nq^{n-1}p$$

Tredje ledd får vi ved å innsette x= 2, fjerde ledd ved å innsette x=3 osv. Siste ledd (det n+1'te ledd) får vi ved å innsette x=n. Vi har nemlig:

$$\frac{n!}{n!\,(n-n)!} q^{n-n}p^n = \frac{n!}{n!\,0!} q^0p^n = p^n$$

Vi får altså samtlige ledd i binomialformelen ved i det alminnelige ledd

$$\frac{n!}{x!\,(n-x)!} q^{n-x}p^x$$

etter hvert å innsette for x de hele tall fra x=0 til x=n. Dette uttrykker en ved foran formelen for det alminnelige ledd å sette Σ (summetegnet, et stort gresk sigma) og ved over og under dette tegn å sette grenseverdiene for x. Vi kan derfor skrive binomialformelen slik:

$$(q+p)^n = \sum_0^n \frac{n!}{x!\,(n-x)!} q^{n-x}p^x$$

Eksempel:

$$(q+p)^3 = \sum_0^3 \frac{3!}{x!\,(3-x)!}\, q^{3-x} p^x$$

$$= \frac{3!}{0!\,3!}\, q^{3-0} p^0 + \frac{3!}{1!\,2!}\, q^{3-1} p^1 + \frac{3!}{2!\,1!}\, q^{3-2} p^2 + \frac{3!}{3!\,0!}\, q^{3-3} p^3$$

$$= q^3 + 3q^2 p + 3qp^2 + p^3$$

Dersom p er negativ, $p = -c$, kan vi bruke samme formel ved å sette $(q+p)^n = (q-c)^n = (q+(-c))^n$.

Et spesialtilfelle av binomialformelen kommer vi til senere i sannsynlighetsregningen. Er $q+p = 1$, er naturligvis også $(q+p)^n = 1$. I dette tilfelle har vi da at

$$\sum_0^n \frac{n!}{x!\,(n-x)!}\, q^{n-x} p^x = 1$$

Oppgave 7.

Bruk binomialformelen på følgende uttrykk:

1) $(1 + x)^4$
2) $(2x + 3)^3$
3) $(3x + 2y)^3$
4) $(3 + y)^5$
5) $(x^3 + y)^6$
6) $(x - y^3)^6$
7) $(x^3 - y^2)^7$
8) $(x - y)^5$

6. Eksempel på anvendelse av binomialformelen.

Binomialformelen har vist seg å være et meget nyttig redskap til løsning av mange oppgaver. Vi skal i det følgende vise hvordan den kan brukes til å finne summen av de k første naturlige tall $(1+2+3+ \ldots. +k)$, summen av de k første kvadrattall $(1^2+2^2+3^2+ \ldots. +k^2)$, summen av de k første kubikktall $(1^3+2^3+3^3+ \ldots. +k^3)$ osv.

Vi vil da benytte oss av formelen

$$(1 + p)^n = \sum_0^n \frac{n!}{x!\,(n-x)!}\, p^x \qquad (q=1)$$

I denne formel skal vi nå etter hvert innsette n=2, n=3, n=4 .... og for hver verdi av n etter hvert sette p=0, p=1, p=2, p=3, ... p=k.

1. n = 2

Vi har i dette tilfelle: $(1+p)^2 = 1 + 2p + p^2$ og etter hvert for $p = 0, 1, 2, 3, \ldots. k$:

$$(1+0)^2 = 1^2 = 1$$
$$(1+1)^2 = 2^2 = 1 + 2.1 + 1^2$$
$$(1+2)^2 = 3^2 = 1 + 2.2 + 2^2$$
$$(1+3)^2 = 4^2 = 1 + 2.3 + 3^2$$
$$(1+4)^2 = 5^2 = 1 + 2.4 + 4^2$$
$$\cdots\cdots\cdots\cdots$$
$$\cdots\cdots\cdots\cdots$$
$$\cdots\cdots\cdots\cdots$$
$$(1+k)^2 = (1+k)^2 = 1+2.k + k^2$$

La oss nå tenke oss at hele den kolonne av tall som står lengst til venstre, er sløyfet. Vi har da igjen k+1 ligninger. Vi summerer nå disse ligninger ved å summere hver for seg de to sider av likhetstegnene. Vi innser lett at alle kvadrattallene fra og med 1^2 til og med k^2 vil finnes som addender på begge sider av likhetstegnet i summeligningen. Disse tall vil derfor falle bort. På venstre side har vi da igjen bare kvadrattallet $(1+k)^2$. På høyre side av likhetstegnet har vi først summen av k+1 ettall, og denne summen er lik k+1. Dessuten har vi igjen $2(1+2+3+4+ \ldots. +k)$. Summeligningen blir derfor:

$$(1+k)^2 = (1+k) + 2(1+2+3+4+\ldots.. +k)$$

Vi finner herav at

$$1+2+3+4+ \ldots. +k = \tfrac{1}{2}\left[(1+k)^2 - (1+k)\right] = \tfrac{1}{2}(1+k)(1+k-1)$$

eller:

$$1+2+3+4+ \ldots\ldots\ldots +k = \frac{(1+k).k}{2}$$

Dette resultatet kjenner vi igjen fra før. De naturlige tall danner jo en aritmetisk rekke, og summen av de k første ledd av en slik rekke er lik middeltallet av første og siste ledd multiplisert med antall ledd.

2. n = 3.

Vi har da: $(1+p)^3 = 1 + 3p + 3p^2 + p^3$

og etter hvert for p=0, 1, 2, 3, 4, ... k:

$$(1+0)^3 = 1^3 = 1$$
$$(1+1)^3 = 2^3 = 1 + 3.1 + 3.1^2 + 1^3$$
$$(1+2)^3 = 3^3 = 1 + 3.2 + 3.2^2 + 2^3$$
$$(1+3)^3 = 4^3 = 1 + 3.3 + 3.3^2 + 3^3$$
$$(1+4)^3 = 5^3 = 1 + 3.4 + 3.4^2 + 4^3$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\cdots\cdots\cdots\cdots\cdots\cdots$$
$$(1+k)^3 = (1+k)^3 = 1 + 3.k + 3.k^2 + k^3$$

Vi summerer nå på nøyaktig samme måte som i første tilfelle og finner:

$$(1+k)^3 = (1+k) + 3(1+2+3+ \ldots + k) + 3(1^2+2^2+3^2+ \ldots + k^2)$$

$$= (1+k) + 3.\frac{(1+k).k}{2} + 3.(1^2+2^2+3^2+ \ldots + k^2)$$

Herav finner vi at

$$1^2+2^2+3^2+4^2+ \ldots + k^2 = \frac{k.(k+1)(2k+1)}{6}$$

3. n = 4.

Vi har da: $(1+p)^4 = 1 + 4p + 6p^2 + 4p^3 + p^4$

og etter hvert for p=0,1,2,3,4, ... k:

$$(1+0)^4 = 1^4 = 1$$
$$(1+1)^4 = 2^4 = 1 + 4.1 + 6.1^2 + 4.1^3 + 1^4$$
$$(1+2)^4 = 3^4 = 1 + 4.2 + 6.2^2 + 4.2^3 + 2^4$$
$$(1+3)^4 = 4^4 = 1 + 4.3 + 6.3^2 + 4.3^3 + 3^4$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$(1+k)^4 = (1+k)^4 = 1 + 4.k + 6.k^2 + 4.k^3 + k^4$$

Vi summerer nå på samme måte som i de to foregående tilfelle og finner:

$$(1+k)^4 = (1+k) + 4(1+2+3+\ldots+k) + 6(1^2+2^2+3^2+ \ldots + k^2)$$
$$+ 4.(1^3+2^3+3^3+ \ldots\ldots\ldots k^3)$$

eller $$(1+k)^4 = (1+k) + 4\,\frac{(1+k).k}{2} + 6\,\frac{k.(k+1)(2k+1)}{6}$$
$$+ 4.(1^3+2^3+3^3+ \ldots\ldots\ldots + k^3)$$

Herav finner vi at

$$1^3+2^3+3^3+4^3+ \ldots\ldots + k^3 = \left[\frac{(1+k).k}{2}\right]^2$$

Summen av de k første kubikktall er altså lik kvadratet på summen av de k første naturlige tall.

Vi kunne lett gå videre, velge n=5 og derved finne $1^4+2^4+3^4+ \ldots + k^4$ osv. Vi har imidlertid ikke her bruk for flere summer enn de som er funne, og skal derfor ikke gå videre.

7. T r e g h e t s m o m e n t e r .

Det forutsettes kjent at treghetsmomentet av en flate med hensyn til en fast akse er lik summen av produktene av flateelementene og kvadratet på deres avstander fra aksen. En forlanger av disse flateelementer at de skal være så små at alle punkter innen elementet kan ansees for å ha samme avstand fra aksen. Vi kan finne treghetsmomentet av en flate med hensyn til en akse ved hjelp

av elementære midler i visse enkle tilfelle. Vi deler da flaten opp i et endelig antall elementer av liten størrelse og søker å finne et tilnærmet riktig uttrykk for treghetsmomentet. Deretter undersøker vi om dette uttrykket nærmer seg en grense når antallet av elementer økes over alle grenser, dvs. når elementenes størrelse avtar mot 0.

Treghetsmomentet av et rektangel.

Vi skal bestemme treghetsmomentet av et homogent rektangel med hensyn til en akse i planet gjennom rektanglets tyngdepunkt og parallell med den ene av rektangletš sider. La rektanglets høyde være H og dets bredde være B. Aksen er parallell med breddesiden (se figur 6.)

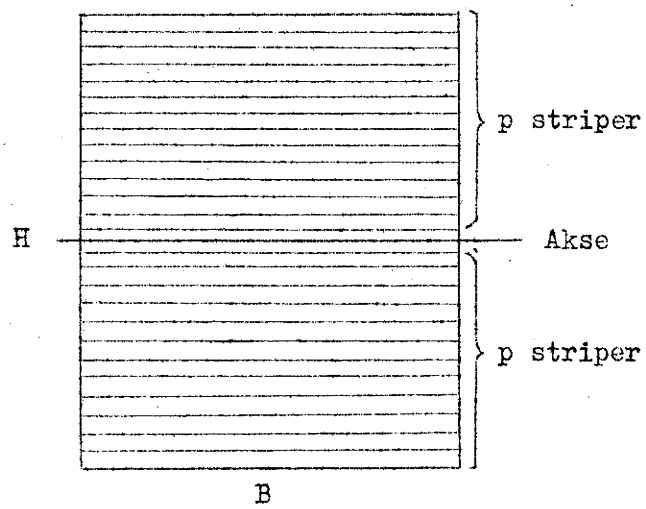


Fig. 6

Vi deler høyden i n like store deler, hver del lik $d=\frac{H}{n}$, og velger for n et ulike tall, $n=2p+1$. Ved å trekke paralleller med aksen gjennom delingspunktene får vi delt opp rektanglet i n kongruente striper. Vi har da p striper på hver side av aksen og en stripe i midten (den som inneholder aksen). Dersom nå stripene er tilstrekkelig smale, vil alle punkter innen den enkelte stripe kunne ansees å ha tilnærmet samme avstand fra aksen. Arealet av hver stripe er B.d. Treghetsmomentet av den midterste stripen kan vi sette ut av betraktning. Avstanden fra aksen til de andre stripers tyngdepunkter (stripenes middelavstander) er d, 2d, 3d, 4d, ... pd. Treghetsmomentet blir derfor (tilnærmet) lik:

$$T = 2\left[Bd.d^2 + Bd.(2d)^2 + Bd.(3d)^2 + ... + Bd.(pd)^2\right]$$
$$= 2Bd^3.(1^2+2^2+3^2+ ..... +p^2) = 2Bd^3.\frac{p(p+1)(2p+1)}{6}$$

Fører vi nå inn H og n, får vi etter noe forenkling:

$$T = \frac{B.H^3}{12}\left(1 - \frac{1}{n^2}\right)$$

Lar vi nå n vokse over alle grenser, vil $\frac{1}{n^2}$ avta mot 0, og det eksakt riktige uttrykk for treghetsmomentet blir

$$T = \frac{B.H^3}{12}$$

Treghetsmomentet av en sirkelflate.

Vi skal først bestemme treghetsmomentet av en sirkelflate med hensyn til en akse loddrett på sirkelflatens plan gjennom dens sentrum, det såkalte polære treghetsmoment.

La sirkelens diameter være D (se figur 7). Vi deler denne i n like store deler (n er et stort tall) og lar n være et ulike tall (n=2p+1). Hver del er lik $d = \frac{D}{n}$. Vi slår så sirkler om sentrum gjennom diameterens delingspunkter og får derved sirkelflaten delt opp i p ringformige elementer av bredde d. Dessuten får vi en sirkelflate innerst med radius $\frac{d}{2}$. Treghetsmomentet av denne sirkelflate kan settes ut av betraktning.

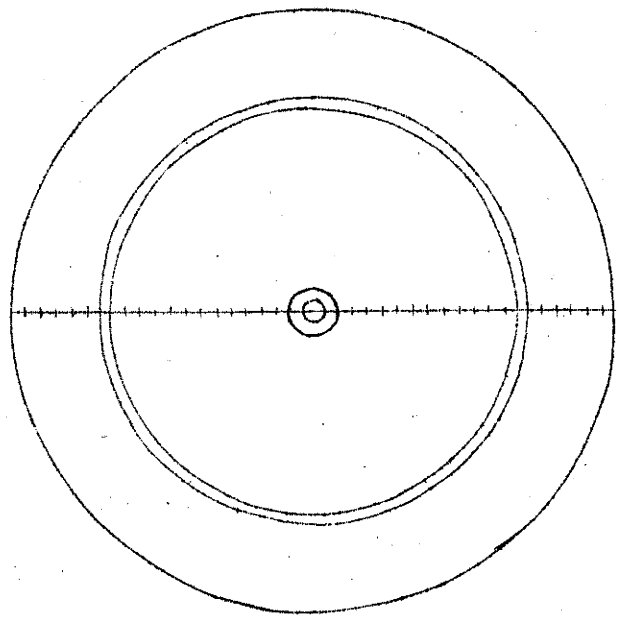


Fig. 7

Arealet av en ringformig stripe er $2\pi rd$, hvor r er avstanden fra sentrum (eller aksen) til vedkommende stripes midtlinje, dvs. stripens middelavstand fra aksen. Disse middelavstander fra aksen er for de forskjellige striper:

$$d, 2d, 3d, 4d, \ldots\ldots pd$$

Treghetsmomentet er derfor (tilnærmet):

$$T = 2\pi d.d.d^2 + 2\pi d.2d.(2d)^2 + 2\pi d.3d.(3d)^2 + \ldots\ldots$$
$$\ldots\ldots\ldots\ldots + 2\pi d.pd.(pd)^2$$
$$= 2\pi d^4.(1^3 + 2^3 + 3^3 + \ldots\ldots\ldots + p^3)$$
$$= 2\pi d^4.\left(\frac{1+p}{2}.p\right)^2 = \tfrac{1}{2}\pi d^4.(1+p)^2 p^2$$

Fører vi her inn D og n, får vi etter noe forenkling at

$$T = \frac{\pi D^4}{32}\left(1 - \frac{1}{n^2}\right)^2$$

Lar vi nå n vokse over alle grenser, vil $\frac{1}{n^2}$ avta mot 0. Det eksakt riktige uttrykk for treghetsmomentet er derfor:

$$T = \frac{\pi D^4}{32}$$

Vi kan lett vise at treghetsmomentet av en sirkelflate med hensyn til en diameter som akse er lik halvdelen av det polære treghetsmoment. Vi tegner inn i sirkelen to diametre I og II (se figur 8) som står loddrett på hverandre. Betegner vi nå elementenes avstand fra diameter I med x og fra diameter II med y, har vi

$$x^2 + y^2 = r^2$$

hvor r er elementenes avstand fra den polære akse. Vi har derfor også at

$$fx^2 + fy^2 = fr^2$$

hvor f er elementet (se fig.). Stiller vi opp denne ligning for alle elementer og summerer, har vi

$$\Sigma fx^2 + \Sigma fy^2 = \Sigma fr^2$$

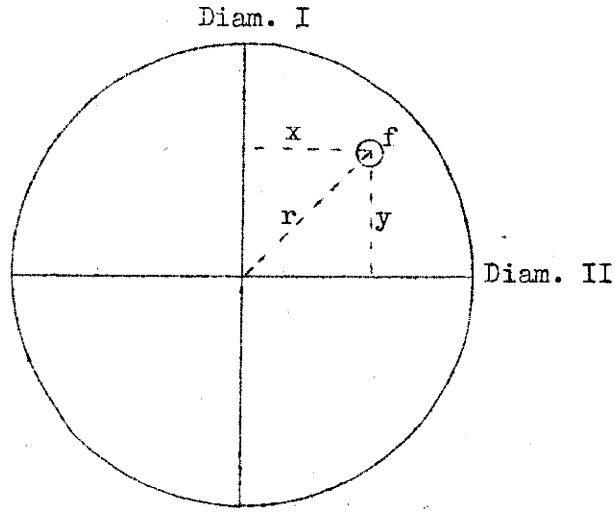


Fig. 8

Her er nå Σfx^2 treghetsmomentet med hensyn til diameter I, Σfy^2 er treghetsmomentet med hensyn til diameter II og Σfr^2 er det polære treghetsmoment. Treghetsmomentet med hensyn til en diameter er selvsagt det samme for alle diametre. Betegner vi treghetsmomentet med hensyn til en diameter med T', har vi at

$$T' = \Sigma fx^2 = \Sigma fy^2$$

Vi har derfor at

$$2T' = \Sigma fr^2 = \frac{\pi D^4}{32}$$

eller

$$T' = \frac{\pi D^4}{64}$$

8. Noen regler for regning med summetegn.

La a_1, a_2, a_3, ... a_n være en rekke bestående av n tall. Vi vil ikke forutsette noe om disse tallene. De kan være positive, negative, hele eller brudne, og rekken behøver ikke å være dannet etter noen bestemt regel. Som en kortere skrivemåte for en slik rekke bruker en ofte betegnelsen

$$a_i \quad (i=1,2,3,....n)$$

Summen av disse tallene, altså $a_1+a_2+a_3+ .....+a_n$, skrives ofte slik:

$$a_1+a_2+a_3+....+a_n = \sum_{i=1}^{n} a_i$$

På samme måte er

$$\sum_{i=m}^{n} a_i = a_m+a_{m+1}+a_{m+2}+ ..... + a_n$$

Fotskriften i som også kalles summeringsindeks, angir altså bare hvilken plass tallet har i rekken når rekken begynner med a_1. Begynner rekken med a_m, har tallet a_i den (i-m+1)'te plass fra venstre. Er første tall a_m og siste tall a_n, omfatter rekken ialt n-m+1 tall.

Et uttrykk som nybegynnere ofte stusser ved er

$$S = \sum_{i=1}^{n} a$$

hvor leddet under summetegnet ikke har noen fotskrift. Dette betyr simpelten tallet a addert til seg selv n ganger. Altså er

$$\sum_{i=1}^{n} a = n.a$$

På samme måte er

$$\sum_{i=m}^{n} a = (n-m+1).a$$

(En bør være oppmerksom på at fotskriften i mange tilfelle sløyfes. Ofte sløyfes også grensebetegnelsen over og under summetegnet. Det vil imidlertid da fremgå av sammenhengen hvordan summeringen er å forstå).

I den teoretiske statistikk er summetegnet et så alminnelig brukt symbol at det ikke er mulig å følge med i selv en forholdsvis elementær fremstilling uten at en er fortrolig med det og kan benytte noen enkle regneregler. Vi skal derfor allerede nå gjennomgå noen eksempler på regning med summetegn.

Eksempel 1.

La oss tenke oss at det under summetegnet står flere ledd bundet sammen med pluss eller minus, f. eks.

$$S = \sum_{i=1}^{n} (a_i - b_i + c_i)$$

Denne summen er lik:

$$S = (a_1 - b_1 + c_1) + (a_2 - b_2 + c_2) + (a_3 - b_3 + c_3) + \ldots\ldots + (a_n - b_n + c_n)$$

$$= (a_1 + a_2 + a_3 + \ldots + a_n) - (b_1 + b_2 + b_3 + \ldots\ldots + b_n) + (c_1 + c_2 + c_3 + \ldots + c_n)$$

eller:

$$S = \sum_{i=1}^{n} (a_i - b_i + c_i) = \sum_{i=1}^{n} a_i - \sum_{i=1}^{n} b_i + \sum_{i=1}^{n} c_i$$

Hvis vi altså under summetegnet har flere ledd forbundet med plus eller minus, kan vi summere hvert ledd for seg og forbinde disse summene med de samme fortegn som vi hadde foran leddene under det opprinnelige summetegn. Beviset for dette er ovenfor gitt for det tilfelle at det under summetegnet er 3 ledd, men det innsees lett at regelen må gjelde i sin alminnelighet.

Denne regelen kan også brukes den motsatte vei. Hvis vi har en rekke summetegn forbundet med plus eller minus, kan vi samle alle leddene under et felles summetegn hvis summeringsområdet er det samme for alle summetegnene. Men en kan ikke samle

$$\sum_{i=1}^{n} a_i + \sum_{i=1}^{m} b_i - \sum_{i=1}^{k} c_i$$

under ett summetegn fordi summeringsområdene i dette tilfelle er ulike.

Oppgave 8.

Vis hvordan en skal summere

$$\sum_{i=1}^{n} (a_i - b)$$

Eksempel 2.

Hvis det under summetegnet finnes en faktor uten fotskrift, kan denne faktoren settes utenfor summetegnet, altså:

$$\sum_{i=1}^{n} p.a_i = p.\sum_{i=1}^{n} a_i$$

Vi har nemlig at

$$\sum_{i=1}^{n} p.a_i = p.a_1+p.a_2+p.a_3+\ldots\ldots + p.a_n$$
$$= p.(a_1+a_2+a_3+ \ldots\ldots +a_n) = p.\sum_{i=1}^{n} a_i$$

Disse to reglene (eks. 1 og 2) kan vi naturligvis bruke samtidig. Vi har f. eks. at

$$\sum_{i=1}^{n} (p_1.a_i+p_2.b_i-p_3.c_i) = p_1.\sum_{i=1}^{n} a_i +p_2 \sum_{i=1}^{n} b_i - p_3.\sum_{i=1}^{n} c_i$$

Eksempel 3.

La oss nå tenke oss at vi har to rekker tall:

$a_1, a_2, a_3, \ldots a_n$ eller a_i $(i = 1, 2, 3, \ldots n)$

og $b_1, b_2, b_3, \ldots b_m$ eller b_j $(j = 1, 2, 3, \ldots m)$

Av disse to rekkene skal vi danne en ny rekke på den måten at hvert tall i den nye rekken er summen av ett tall i a-rekken og ett tall i b-rekken og slik at hvert tall i a-rekken skal kombineres med hvert tall i b-rekken.

Hvis n=1 og m=1, vil den nye rekken inneholde bare ett tall, nemlig a_1+b_1. Er n=2 og m=1, vil den nye rekken inneholde to tall, nemlig a_1+b_1 og a_2+b_1. Er n=2 og m=2, vil den nye rekken inneholde 4 tall, nemlig

$$a_1+b_1,\ a_2+b_1,\ a_1+b_2,\ a_2+b_2$$

I alminnelighet vil den nye rekken inneholde n.m tall, og vi skal stille oss som oppgave å finne summen av disse, uttrykt ved summene av a-rekken og b-rekken. Vi må da begynne med å stille opp tallene i den nye rekken. Først kan vi danne alle de tallene som inneholder a_1. Dette er a_1+b_1, a_1+b_2, a_1+b_3, ... a_1+b_m. Deretter danner vi de tallene som inneholder a_2. Dette er $a_2+b_1, a_2+b_2, a_2+b_3$, ... a_2+b_m. Dernest danner vi alle de tallene som inneholder a_3, de som inneholder a_4 osv. Tallene i den nye rekker kan stilles opp i følgende rektangulære skjema:

$$a_1+b_1,\ a_1+b_2,\ a_1+b_3,\ \ldots\ldots\ a_1+b_m$$
$$a_2+b_1,\ a_2+b_2,\ a_2+b_3,\ \ldots\ldots\ a_2+b_m$$
$$a_3+b_1,\ a_3+b_2,\ a_3+b_3,\ \ldots\ldots\ a_3+b_m$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$a_n+b_1,\ a_n+b_2,\ a_n+b_3,\ \ldots\ldots\ a_n+b_m$$

Summen av disse n.m tallene kan vi finne ved først å summere tallene i hver rekke for seg og dernest summere alle rekkesummene. La oss betegne disse rekkesummene med $S_1, S_2, S_3, \ldots S_n$. Summen av tallene i første rekke (S_1) er:

$$S_1 = (a_1+b_1) + (a_1+b_2) + (a_1+b_3) + \ldots + (a_1+b_m)$$
$$= m.a_1 + (b_1+b_2+b_3+ \ldots +b_m) = m.a_1 + \sum_{j=1}^{m} b_j$$

På samme måte finnes

$$S_2 = m.a_2 + \sum_{j=1}^{m} b_j$$
$$S_3 = m.a_3 + \sum_{j=1}^{m} b_j$$
$$\ldots\ldots\ldots\ldots\ldots$$
$$S_n = m.a_n + \sum_{j=1}^{m} b_j$$

Summen av alle n.m tallene blir derfor:

$$S = S_1+S_2+S_3+ \ldots +S_n = m.(a_1+a_2+a_3+ \ldots +a_n) + n.\sum_{j=1}^{m} b_j$$

Eller:

$$S = m.\sum_{i=1}^{n} a_i + n.\sum_{j=1}^{m} b_j$$

Som betegnelse for denne summen bruker en

$$S = \sum_{i=1}^{n}\sum_{j=1}^{m} (a_i+b_j)$$

og dette er altså å oppfatte som summen av de n.m tallene som fremkommer når en lar fotskriftene i og j uavhengig av hverandre gjennomløpe tallene i=1,2,3, ... n og j=1,2,3, ... m. Denne dobbeltsummen kan altså uttrykkes ved summen av de to addendene under summetegnet ved formelen

$$\sum_{i=1}^{n}\sum_{j=1}^{m} (a_i+b_j) = m.\sum_{i=1}^{n} a_i + n.\sum_{j=1}^{m} b_j$$

Et spesialtilfelle av dette resultat har vi når alle tallene i b-rekken er like, $b_j=b$; dvs. når b mangler fotskrift. Da blir alle de tallene som befinner seg i samme rekke i det ovenfor nevnte rektangulære skjema like. Tallene i første rekke blir alle lik a_1+b, tallene i annen rekke a_2+b osv. Som vi allerede har vist er

$$\sum_{j=1}^{m} b = m.b$$

og følgelig er

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(a_i+b) = m.\sum_{i=1}^{n} a_i + n.\sum_{j=1}^{m} b = m.\sum_{i=1}^{n} a_i + n.m.b$$

Det er intet til hinder for at vi i den siste ligningen kan sette b=0. Vi finner da:

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(a_i+0) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i = m.\sum_{i=1}^{n} a_i$$

Oppgave 9.

Utled den siste setningen på annen måte.

Oppgave 10.

Vis at

$$\sum_{i=1}^{n}\sum_{j=1}^{m}(p.a_i - q.b_j) = mp\sum_{i=1}^{n} a_i - nq\sum_{j=1}^{m} b_j$$

Oppgave 11.

Vis at

$$\sum_{i=1}^{n}\sum_{j=1}^{m} a_i.b_j = \sum_{i=1}^{n} a_i * \sum_{j=1}^{m} b_j$$

Eksempel 4.

La oss tenke oss at vi har en rekke tall c_{ij} (i=1,2,3....n, j=1,2,3,....m) og det forutsettes at hver verdi av fotskriften i skal kombineres med hver verdi av fotskriften j. Rekken omfatter da alt i alt n·m tall. Vi kan skrive opp alle disse tallene i et rektangulært skjema hvor fotskriften i har en konstant verdi i hver rekke og den andre fotskriften j har en konstant verdi i hver kollonne. Tallene er

$$c_{11}\ c_{12}\ c_{13} \ldots\ldots\ldots\ldots\ c_{1m}$$
$$c_{21}\ c_{22}\ c_{23} \ldots\ldots\ldots\ldots\ c_{2m}$$
$$c_{31}\ c_{32}\ c_{33} \ldots\ldots\ldots\ldots\ c_{3m}$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$c_{n1}\ c_{n2}\ c_{n3} \ldots\ldots\ldots\ldots\ c_{nm}$$

Vi kan summere disse tallene ved først å danne rekkesummene og deretter summere disse. Summen av første rekke er

$$S_1 = c_{11}+c_{12}+c_{13}+ \ldots\ldots +c_{1m} = \sum_{j=1}^{m} c_{1j}$$

Summen av annen rekke, tredje rekke osv. er

$$S_2 = \sum_{j=1}^{m} c_{2j}$$

$$S_3 = \sum_{j=1}^{m} c_{3j}$$

osv.

Summen av alle de n.m tallene er da

$$S = S_1+S_2+S_3+ \ldots\ldots +S_n = \sum_{j=1}^{m} c_{1j} + \sum_{j=1}^{m} c_{2j} + \sum_{j=1}^{m} c_{3j} + \ldots\ldots\ldots + \sum_{j=1}^{m} c_{nj}$$

Da summeringsområdet for alle disse n summetegn er det samme, kan vi etter regelen i eks. 1 samle det hele under ett summetegn. Vi kan altså sette

$$S = \sum_{j=1}^{m} (c_{1j}+c_{2j}+c_{3j}+ \ldots\ldots +c_{nj})$$

Under summetegnet står nå en sum av ledd hvor fotskriften j finnes i alle ledd, mens fotskriften i gjennomløper verdiene i=1,2,3, ... n.
Vi kan derfor sette

$$c_{1j}+c_{2j}+c_{3j}+ \ldots\ldots + c_{nj} = \sum_{i=1}^{n} c_{ij}$$

Følgelig har vi at

$$S = \sum_{j=1}^{m} \sum_{i=1}^{n} c_{ij}$$

Vi kan imidlertid også summere de n.m tallene ved først å danne kollonnesummene og deretter summere disse. Ved nøyaktig samme fremgangsmåte kommer vi da til

$$S = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij}$$

Følgelig har vi at

$$\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij} = \sum_{j=1}^{m}\sum_{i=1}^{n} c_{ij}$$

Vi ser herav at det er likgyldig (i dette tilfelle) hvilken rekkefølge summetegnene har. Forutsetningen for gyldigheten av denne regelen er imidlertid at summeringsområdet for den ene fotskriften er uavhengig av den annen fotskriften. Hvis f. eks. summeringsområdet for j på en eller annen måte er avhengig av i, må summetegnet for i skrives først. At det må være slik vil vi lett innse når vi løser følgende oppgave.

Oppgave 12.

Skriv opp i skjemaform de tallene som skal tas med i summen

$$S = \sum_{i=1}^{n}\sum_{j=1}^{i} c_{ij}$$

I det foregående er det ikke gjort noen forutsetning om hvordan tallene c_{ij} er fremkommet. Setter vi f. eks.

$$c_{ij} = a_i + b_j$$

vil vi ha (eks. 3):

$$\sum_{i=1}^{n}\sum_{j=1}^{m} (a_i+b_j) = \sum_{j=1}^{m}\sum_{i=1}^{n} (a_i+b_j) = m.\sum_{i=1}^{n} a_i + n.\sum_{j=1}^{m} b_j$$

Setter vi $c_{ij} = a_i.b_j$, vil vi ha (oppgave 11):

$$\sum_{i=1}^{n}\sum_{j=1}^{m} a_i.b_j = \sum_{j=1}^{m}\sum_{i=1}^{n} a_i.b_j = \sum_{i=1}^{n} a_i \cdot \sum_{j=1}^{m} b_j$$

Eksempel 5.

Til slutt skal vi ta for oss det tilfelle at det under et dobbelt summetegn står et produkt av 3 faktorer hvorav to har forskjellig fotskrift og den tredje har begge de to andre faktorers fotskrifter. Vi skal undersøke hvordan vi skal utføre summeringen

$$S = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i.b_j.c_{ij}$$

når det forutsettes at hver i-verdi, i = 1,2,3, ... n, skal kombineres med hver j-verdi, j = 1,2,3, ... m.

I dette tilfelle har vi for oss tre rekker tall. Den ene rekken, a-rekken, omfatter n tall. Den andre rekken, b-rekken, omfatter m tall. Og

den tredje rekken, c-rekken, omfatter n.m tall. Av disse tre rekkene skal vi danne rekken a.b.c og summere denne. Vi kan da først tenke oss at vi danner produktrekken $(a_i.b_j)$ på samme måte som i oppgave 11. Denne rekken omfatter n.m tall. Av denne og av c-rekken som også omfatter n.m tall, skal vi så danne produktrekken a.b.c. Men når vi skal danne produktene $(a_i.b_j)c_{ij}$, kan vi ikke kombinere tallene fra disse to rekkene fritt. Har vi nemlig valt i og j, er det dermed gitt hvilket tall i c-rekken som skal brukes. Tallet $a_1.b_2$ i produktrekken (a.b) kan bare kombineres med det tall i c-rekken som har fotskriftene 1 og 2, dvs. med tallet c_{12}. Produktet $a_1.b_2.c_{34}$ f. eks. har ikke noen plass i produktrekken (a.b.c). Det følger av dette at ethvert tall i produktrekken (a.b) kan kombineres med bare ett tall i c-rekken nemlig med det tallet som har både a's og b's fotskrift. Av følgende skjema vil det lett kunne sees hvilke produkter det blir tale om (tabell med to innganger):

b_j \ a_i	a_1	a_2	a_3	.....	a_n
b_1	c_{11}	c_{21}	c_{31}	.....	c_{n1}
b_2	c_{12}	c_{22}	c_{32}	.....	c_{n2}
b_3	c_{13}	c_{23}	c_{33}	.....	c_{n3}
.....	.....				
b_m	c_{1m}	c_{2m}	c_{3m}	.....	c_{nm}

En ser herav at summen av rekken (a.b.c.) er:

$$
\begin{aligned}
S = {} & a_1b_1c_{11} + a_1b_2c_{12} + \ldots\ldots\ldots\ldots + a_1b_mc_{1m} \\
+ {} & a_2b_1c_{21} + a_2b_2c_{22} + \ldots\ldots\ldots\ldots + a_2b_mc_{2m} \\
+ {} & \ldots\ldots\ldots\ldots \\
+ {} & a_nb_1c_{n1} + a_nb_2c_{n2} + \ldots\ldots\ldots\ldots + a_nb_mc_{nm} \\
= {} & a_1(b_1c_{11} + b_2c_{12} + \ldots\ldots\ldots\ldots + b_mc_{1m}) \\
+ {} & a_2(b_1c_{21} + b_2c_{22} + \ldots\ldots\ldots\ldots + b_mc_{2m}) \\
+ {} & \ldots\ldots\ldots\ldots \\
+ {} & a_n(b_1c_{n1} + b_2c_{n2} + \ldots\ldots\ldots\ldots + b_mc_{nm})
\end{aligned}
$$

Summene i disse parentesene er summene av tallene i de kollonnene som fremkommer når vi etter hvert multipliserer tallene i b-kollonnen med tallene i c-kollonnene. La oss betegne disse summene med $S_1, S_2 \ldots\ldots S_n$. Vi har da:

$$S_1 = b_1c_{11} + b_2c_{12} + \ldots\ldots + b_mc_{1m} = \sum_{j=1}^{m} b_jc_{1j}$$

og videre:

$$S_2 = \sum_{j=1}^{m} b_jc_{2j}$$

$$S_3 = \sum_{j=1}^{m} b_jc_{3j}$$

osv.

og i sin alminnelighet

$$S_i = \sum_{j=1}^{m} b_jc_{ij}$$

Følgelig blir

$$S = a_1S_1 + a_2S_2 + a_3S_3 + \ldots\ldots + a_nS_n = \sum_{i=1}^{n} a_iS_i$$

$$= \sum_{i=1}^{n} a_i \sum_{j=1}^{m} b_jc_{ij}$$

På samme måte kan en vise at

$$S = \sum_{j=1}^{m} b_j \sum_{i=1}^{n} a_ic_{ij}$$

Følgelig har vi at

$$\sum_{i=1}^{n} \sum_{j=1}^{m} a_ib_jc_{ij} = \sum_{i=1}^{n} a_i \sum_{j=1}^{m} b_jc_{ij} = \sum_{j=1}^{m} b_j \sum_{i=1}^{n} a_ic_{ij}$$

Oppgave 13.

La a_i, b_j og c_{ij} være tallene:

b_j \ a_i	-2	-1	0	+1	+2
5	10	7	4	0	1
6	8	5	4	2	5
7	5	2	4	5	8
8	1	0	4	7	10

Beregn

$$\Sigma\,\Sigma\; a_ib_jc_{ij}$$

Oppgave 14.

Vis hvordan en skal finne summen

$$\sum_{i=1}^{n} \sum_{j=1}^{m} a_ib_j(a_i-b_j)$$

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

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

over

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

ved

NORGES LANDBRUKSHØGSKOLE

II. Statistikk

Innhold:

II. Statistikk.

1. Innledning.

Behovet for statistiske metoder melder seg hver gang en skal orientere seg i, beskrive eller analysere en rekke observasjoner. En rekke observasjoner er en rekke tall som er funnet ad empirisk vei ; det kan være gjentatte målinger av lengden av en linje i terrenget, kjemiske analyseresultater av melkeprøver, veieresultater av kløveravlingen på et antall like store forsøksflater, målinger av høyden av et antall grantrær osv. Til en begynnelse vil slike observasjonsrekker foreligge som uordnete rekker noteringer. Den første oppgaven blir derfor å ordne observasjonene og gi en kort karakteristikk av dem. I denne delen av forelesningene skal det bli gitt en beskrivelse av noen av de metodene vi har til rådighet til dette formål.

Til å begynne med må vi da ta for oss noen grunnleggende begreper som det er viktig å kjenne til, nemlig enhet, kjennetegn og statistisk masse.

La oss sette at det er nødvendig for et eller annet formål å undersøke og karakterisere høyden av grantrærne i en skog. Vi kan da etter undersøkelsens formål og omfang, måle høyden av hvert enkelt tre eller bare av et mindre antall trær tatt ut på en eller annen nærmere definert måte. I statistisk forstand er da det enkelte grantre en enhet. En samling av enheter med visse felles karakteristikker kaller vi en statistisk masse. Alle grantrærne i en skog er en statistisk masse. Enhetene i denne statistiske massen har "gran" som felles karakteristikk. Alle grantrærne innen et skogområde som er behandlet etter et bestemt skjøtselsprinsipp, er også en statistisk masse. Det samme gjelder alle 10 år gamle grantrær innen et nærmere avgrenset skogområde.

Enheten har mange forskjellige kjennetegn som skiller det ut fra de andre enhetene innen samme statistiske masse. Et grantre har slike kjennetegn som f. eks. høyde, alder, brysthøydediameter, høyde av kvistfri stamme osv. Et kjennetegn er altså en karakteristikk av en statistisk enhet, hva nå enten den er gitt ved et enkelt ord eller ved en lengre beskrivelse. Resultatene av målingene av et bestemt kjennetegn hos enheter som tilhører en nærmere avgrenset statistisk masse, er en observasjonsrekke eller observasjonssamling.

Hvilke enheter en skal regne med i en statistisk masse vil avhenge av hva hensikten med undersøkelsen er, eller m.a.o. av hva en vil bruke observasjonene til. Prøver av melken fra hver enkelt ku i Akershus fylke på en bestemt dag er en statistisk masse. Prøver av melken fra en bestemt ku på Landbrukshøgskolens fjøs for hver dag i et helt år er også en statistisk masse. Men observasjonene av kvelstoffinnholdet i disse melkeprøvene kan

naturligvis ikke benyttes for samme formål i de to tilfelle. Det forarbeid for enhver statistisk undersøkelse, som gjelder bedømmelse og valg av statistisk masse, er derfor uhyre viktig.

I alminnelighet ordner en kjennetegnene i to grupper. Først har vi de kjennetegnene som observeres ved måling (veiing) eller opptelling. Eksempler er høyden av gran, kvelstoffinnholdet i melk, antall kronblader hos Soleihov osv. Disse kjennetegnene kalles variable eller kvantitative. Andre kjennetegn karakteriseres ved en kvalitetsbetegnelse. Disse kalles konstante kjennetegn. Til eksempel er kjennetegnet for kjønn (han, hun) og for øyefarge (brun, blå) konstante kjennetegn. Når vi imidlertid i det følgende taler om kjennetegn, vil vi alltid forstå variable kjennetegn. Observasjonene av konstante kjennetegn fører nemlig alltid til observasjoner av variable kjennetegn, og det er disse siste som har størst interesse. Egenskapen kjønn hos nyfødte barn har de to konstante kjennetegn jente og gutt. La oss tenke oss at det innen en bestemt befolkningsgruppe er notert fra år til år antallet av guttefødsler og antallet av jentefødsler. Disse observasjonene kan da brukes til å vise hvordan fødselsantallet har variert fra år til år for begge kjønn under ett eller hver for seg. I alle tre tilfelle er det årene som er enhetene og kjennetegnet er antall fødsler. Dette kjennetegnet er variabelt. Eller, det kan være forholdet mellom antall jentefødsler og antall guttefødsler en er interessert i. Da er også årene enhetene og kjennetegnet er f. eks. den såkalte kjønnsproporsjon = prosentisk antall gutter av det totale antall fødte. Også i dette tilfelle er kjennetegnet variabelt.

De variable kjennetegnene inndeles gjerne i to grupper. Noen kjennetegn kan bare uttrykkes i hele tall og kalles diskrete. Eksempler er antall kronblader, antall grisunger pr. kull, antall ord pr. side i en bok osv. Slike kjennetegn observeres ved opptelling. Andre kjennetegn kan ikke uttrykkes i hele tall (uten ved avrunding) og kalles kontinuerlige. Eksempler er kjennetegnet for høyavling (antall kg), kjennetegnet for kvelstoffinnhold (% kvelstoff), kjennetegnet for legemshøyde (antall cm) osv. Slike kjennetegn observeres ved måling eller veiing. Unntagelsesvis vil også observasjonene av diskrete kjennetegn bli uttrykt i brudne tall. La oss tenke oss at en skal undersøke spireevnen hos et parti gulerotfrø. Dette kan gjøres ved at en sår ut et antall like store prøver, hver prøve på k frø, og teller antallet av spirte frø. Observasjonene er da selvsagt hele tall. Men hvis vi angir resultatene i forhold til k, vil observasjonene bli brudne tall, men likevel diskrete fordi det er bare bestemte tall det kan være tale om, nemlig 0/k, 1/k, 2/k, 3/k, ... k/k.

Observasjoner av et diskret kjennetegn kaller vi diskrete observasjoner og observasjoner av et kontinuerlig kjennetegn kontinuerlige observasjoner.

2. Fordelingsrekken.

Før vi kan gi en beskrivelse eller kort karakteristikk av en lang rekke observasjoner, må vi ordne dem i en oversiktlig tabell. En slik tabell kalles en fordelingsrekke. Hvordan en slik fordelingsrekke skal stilles opp, læres best av et par eksempler.

Eksempel 1.

På et forsøksfelt i granskog på høy bonitet, alder 70 år, ble den dobbelte barktykkelse (i brysthøyde) målt på n=29 trær. Observasjonene (i mm) er:

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

Sum = 584

I dette eksemplet omfatter rekken bare 29 observasjoner. Rekken tar derfor liten plass som den er, og en viss oversikt over rekken som helhet har vi jo også. Vi skal imidlertid bruke disse observasjonene til å vise hvordan fordelingsrekken stilles opp.

Vi ser at observasjonene er hele tall. Det er ikke diskrete observasjoner, men kontinuerlige som er avrundet til hele tall. Vi ser at flere av observasjonene finnes igjen i rekken flere ganger. Både første og annen observasjon er lik 14, observasjon nr. 5 og nr. 6 er lik 18 osv. Fordelingsrekken for disse observasjonene er en tabell (med en inngang) over observasjonsverdiene og det antall ganger hver verdi finnes i observasjonsrekken. Fordelingsrekken er:

x_i	h_i
14	2
15	0
16	2
17	3
18	2
19	1
20	5
21	2
22	7
23	0
24	3
25	2
	29

La i sin alminnelighet

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

være betegnelsene for observasjonene i en rekke på n diskrete observasjoner. La verdiene av disse observasjonene være $x_1, x_2, x_3, \ldots x_c$ og la antallet av disse verdier være $h_1, h_2, h_3, \ldots h_c$. h_1 er altså antallet av observasjoner med verdien x_1, og i sin alminnelighet: h_i er antallet av observasjoner med verdien x_i. Fordelingsrekken for disse n observasjonene er da:

x_i	h_i
x_1	h_1
x_2	h_2
x_3	h_3
.....	
.....	
x_c	h_c

Her er naturligvis

$$\Sigma\ h_i = n = \text{antall observasjoner.}$$

(Her og i det følgende hvor det ikke kan oppstå misforståelser, sløyfer vi grenseverdiene for summeringsområdet).

Når o'ene er hele tall, vil også x'ene være hele tall. Vi ordner alltid fordelingsrekken slik at

$$x_1 < x_2 < x_3 < \ldots\ldots < x_c$$

I eks. 2 er gjengitt en annen fordelingsrekke for diskrete observasjoner.

Eksempel 2.

Observasjonene er antallet av arrstråler i arret hos en valmue (fra England). Enheten er det enkelte arr.

x_i	h_i	$\% h_i$
6	3	0,16
7	11	0,58
8	38	1,99
9	106	5,56
10	152	7,98
11	238	12,49
12	305	16,01
13	315	16,53
14	302	15,85
15	234	12,28
16	128	6,72
17	50	2,62
18	19	1,00
19	3	0,16
20	1	0,05
	1905	99,98

Det er i dette tilfelle ialt $n = 1905$ observasjoner, og vi ser at fordelingsrekken gir en meget god oversikt over hele observasjonsrekken som helhet. Vi ser at alle observasjonene ligger mellom grensene 6 og 20. Det er få observasjoner nær den nedre og nær den øvre grensen og en tydelig opphopning av observasjoner omtrent på midten av variasjonsområdet.

En slik fordelingsrekke kan fremstilles grafisk på flere måter. Mest alminnelig er det å bruke et søylediagram i et rettvinklet koordinatsystem. En bruker da x som uavhengig variabel og h som avhengig variabel. (h er altså høyden av søylene). Det er videre alminnelig at en omregner h-verdiene til prosent av n (sml. eks. 2).

Oppgave 1.

Fremstill rekken i eks. 2 ved hjelp av et søylediagram.

Når observasjonene er kontinuerlige, må vi først ordne observasjonsverdiene i klasser. Fordelingsrekken er da en tabell over klassene og antallet av observasjoner som faller innen de enkelte klasser.

Eksempel 3.

En tilsynelatende ensartet tredje års timoteivoll ble inndelt i 240 kvadratiske ruter på 25 m^2 hver. Avlingen ble veiet for hver rute. En fikk på den måten 240 observasjoner av timoteiavlingen pr. 25 m^2. Ordnes disse observasjonene i klasser på ett kg, får en følgende fordelingsrekke:

Klasser	x_i	h_i	% h_i
11-12	11,5	2	0,83
12-13	12,5	6	2,50
13-14	13,5	9	3,75
14-15	14,5	18	7,50
15-16	15,5	30	12,50
16-17	16,5	40	16,67
17-18	17,5	33	13,75
18-19	18,5	30	12,50
19-20	19,5	21	8,75
20-21	20,5	16	6,67
21-22	21,5	12	5,00
22-23	22,5	12	5,00
23-24	23,5	3	1,25
24-25	24,5	2	0,83
25-26	25,5	4	1,67
26-27	26,5	2	0,83
		240	100,00

Fordelingsrekken viser at 2 ruter hadde gitt avlinger på mellom 11 kg og 12 kg, 6 ruter hadde gitt avlinger på mellom 12 kg og 13 kg osv. På samme måte som i eks. 2 er det en sterk opphopning av observasjoner omtrent på midten av variasjonsområdet og antallet av observasjoner innen klassene blir stort sett mindre og mindre utover mot begge variasjonsgrensene.

Når en skal stille opp fordelingsrekken for kontinuerlige observasjoner, må en først bestemme seg for hvilken klassevidde en vil bruke og for klassegrensene. Da en under senere beregninger benytter klassenes midtverdier (se eks. 3) som representanter for klassene, er det klart at en innfører en feil ved å bruke slike klasser. Det er også klart at jo større klassevidde en bruker, jo større må denne feilen bli. På den annen side vil en også gjerne ha få klasser. Det ville derfor vært greit om en hadde en regel for arbeidet, men noen slik regel har det ikke vært mulig å stille opp. I enkelte lærebøker anbefales det å bruke ikke flere enn 30 og ikke færre enn 10 klasser. Bestemmer en seg først for hvor mange klasser en vil bruke, kan

en finne den omtrentlige klassevidde ved å dividere hele variasjonsområdet (dvs. differensen mellom den største og den minste observasjon) med antallet av klasser. For observasjonene i eks. 3 er variasjonsområdet omtrent (27-11) kg = 16 kg, og en har derfor valt å bruke 16 klasser med en klassevidde på ett kg. En må imidlertid også passe på å velge klassene slik at en får greie klassegrenser og helst også enkle tall for klassenes midtverdier.

Observasjonsantallet h_i kalles frekvensene. h_i er altså frekvensen for observasjonsverdien x_i når observasjonene er diskrete. Er observasjonene kontinuerlige, er h_i frekvensen for den klassen som har midtverdien x_i.

3. M i d d e l t a l l.

Når en har ordnet observasjonene i en fordelingsrekke, har en skaffet seg en oversikt over observasjonsrekken som helhet. Dette er imidlertid ikke nok. En må på en eller annen måte søke å finne uttrykk for det vesentlige ved observasjonene. Dette gjør en ved å beregne visse størrelser som hver for seg gir uttrykk for en bestemt egenskap ved observasjonsrekken som helhet. De viktigste av disse størrelsene er middeltallene.

Middeltall blir stadig brukt i det daglige liv og brukes oftest helt ubevisst som uttrykk for det alminnelige eller karakteristiske. Når det sies at en mann er middels høy, at høsten er som vanlig eller at en gutt har middels evner, er det i virkeligheten middeltall som blir brukt. Det er imidlertid flere slags middeltall, og disse forskjellige middeltall tilfredsstiller ikke alltid samme formål. Vi må her nøye oss med å gjøre rede for de to viktigste, nemlig det aritmetiske gjennomsnitt (eller bare gjennomsnittet) og typetallet.

La observasjonene være

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

Det aritmetiske gjennomsnitt er lik summen av observasjonene dividert med antallet (n). La oss betegne gjennomsnittet med m. Da er

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

For observasjonene i eks. 1 er summen lik 584 og antallet er n=29. Gjennomsnittet er derfor lik

$$m = \frac{584}{29} = 20{,}14 \text{ mm} \quad (20{,}1379)$$

Eksempel 4.

En rekke melkeprøver fra samme ku, tatt over et lengre tidsrom (fra 19/2 1931 til 13/6 1932), ble undersøkt bl.a. med hensyn på kvelstoffinnhold. Kvelstoffmengden ble omregnet til prosent av melkeprøvens vekt. Observasjonene er:

0,510	0,500	0,518	0,497	0,514	0,509
0,500	0,512	0,515	0,561	0,514	0,549
0,500	0,518	0,476	0,569	0,509	0,536
0,500	0,514	0,511	0,543	0,514	
0,525	0,503	0,508	0,541	0,503	
				Sum =	14,469

Her er n=28. Det gjennomsnittlige kvelstoffinnhold i melken er derfor

$$m = \frac{14,469}{28} = 0,5168\ \%$$

Eksempel 5.

Lengden av en linje i terrenget ble målt med stålbånd n=20 ganger. Observasjonene (i meter) er:

901,375	901,465	901,500	901,460
901,405	901,430	901,460	901,480
901,395	901,430	901,450	901,505
901,435	901,395	901,450	901,495
901,430	901,500	901,490	901,510
		Sum =	18029,060

Gjennomsnittet er

$$m = \frac{18029,060}{20} = 901,453 \text{ meter}$$

Gjennomsnittet av observasjoner ordnet i en fordelingsrekke er også lik summen av observasjonene dividert med antallet. Under beregningene må vi bare huske at hver enkelt observasjonsverdi forekommer i observasjonsrekken så mange ganger som frekvensen angir. Dessuten innfører vi den forenkling for kontinuerlige observasjoner at alle observasjonene som tilhører samme klasse, settes lik klassens midtverdi. Summen av observasjonene er altså lik summen av produktene $h_i . x_i$. Gjennomsnittet er derfor:

$$m = \frac{\Sigma\, h_i . x_i}{n}$$

Eksempel 6.

En har tellet antallet av grisunger i n=334 kull. I følgende tabell (fordelingsrekke) er altså h_i antallet av kull med x_i unger.

x_i	h_i	$h_i x_i$
2	1	2
3	1	3
4	4	16
5	6	30
6	17	102
7	20	140
8	30	240
9	35	315
10	51	510
11	52	572
12	39	468
13	45	585
14	21	294
15	7	105
16	5	80
	334	3462

Summen av observasjonene er altså 3462, og gjennomsnittet er

$$m = \frac{3462}{334} = 10{,}37 \quad (10{,}3653)$$

Oppgave 2.

Beregn gjennomsnittet for observasjonene i eks. 3.

Det aritmetiske gjennomsnitt er en så enkel størrelse at det ikke skulle være noen grunn til å gi en nærmere begrunnelse for berettigelsen av å bruke det som en karakteristikk av observasjonsrekken som helhet. Det er imidlertid grunn til å advare mot å bruke det uten støtte enten i observasjonsrekken selv eller i andre karakteristiske størrelser. Meget forskjellige observasjonsrekker kan nemlig ha samme eller omtrent samme gjennomsnitt. Sett f. eks. at vi får oppgitt at den gjennomsnittlige årsinntekt for voksne menn i et herred er 4000 kr. Vi må da ikke uten videre bruke dette gjennomsnittet som uttrykk for inntektsforholdene i vedkommende herred. Gjennomsnittet = 4000 kr kan nemlig være gjennomsnittet av inntekter mellom 2000 kr og 6000 kr. Det kan imidlertid også være gjennomsnittet av inntekter mellom 1000 kr og 5000 kr og en enkelt inntekt på en million kr.

Gjennomsnittet er derfor en størrelse som en bør være noe forsiktig med. Det bør enten oppgis sammen med observasjonene selv eller sammen med størrelser som gir en karakteristikk av andre egenskaper ved observasjonsrekken.

De to fordelingsrekkene som er gjengitt i eks. 2 og eks. 3, har det til felles at frekvensene vokser fra den nedre variasjonsgrensen til et maksimum omtrent på midten av variasjonsområdet og avtar igjen fra dette maksimum mot den øvre variasjonsgrensen. Fordelingsrekkene har bare ett frekvens-

maksimum og er nokså symmetriske omkring dette maksimum. Rekken i eks. 3 er noe usymmetrisk.

Det er denne typen av fordelingsrekker vi som oftest støter på i empirisk materiale. Det er imidlertid også mange avvikelser fra denne alminneligste formen. Vi har eksempler på fordelingsrekker hvor frekvensene avtar fra den ene variasjonsgrensen til den annen. Det finnes også eksempler på fordelingsrekker med mer enn ett frekvensmaksimum. Eksempel 7 er et eksempel på en helt skjev fordelingsrekke.

Eksempel 7.

En fordelingsrekke for antallet av kronblader hos Soleihov.

x_i	h_i
5	223
6	45
7	6
8	4
9	3
	n = 381

Når vi skal karakterisere skjeve fordelingsrekker, har vi bruk for et annet middeltall ved siden av gjennomsnittet, nemlig typetallet. Typetallet er et middeltall og er den verdi av x_i som har den største frekvens. Typetallet i eks. 2 er 13 og typetallet i eks. 7 er 5. Typetallet eller typeklassen i eks. 3 er (16-17) kg. For skjeve fordelingsrekker er typetallet i alminnelighet noe forskjellig fra gjennomsnittet, men det behøver ikke å være det. For rekken i eks. 7 er typetallet 5 og gjennomsnittet m=5,29.

Typetallet er et middeltall som kan brukes til å karakterisere en fordelingsrekke fordi det angir den verdi av kjennetegnet som er den alminneligste. Det gir i mange tilfelle en mer verdifull karakteristikk enn gjennomsnittet. Det er således i mange tilfelle viktigere å kjenne den alminneligste inntekt enn gjennomsnittsinntekten, og det er viktigere for en skofabrikant å kjenne det alminneligste skonummer enn gjennomsnittsnummeret.

Typetallet må naturligvis ikke brukes til karakteristikk av en fordelingsrekke med mindre rekken har et tydelig frekvensmaksimum. Det er derfor også klart at en ikke må oppgi et tall som typetall uten at rekken er basert på et meget stort antall observasjoner. Det er selvsagt også klart at rekken ikke må ha mer enn ett frekvensmaksimum.

4. Varians og middelavvik (middelfeil).

Middelavviket er den størrelse som oftest blir brukt som uttrykk for variasjonens størrelse. Vi skal foreløpig nøye oss med å definere det og vise hvordan det skal beregnes.

La observasjonene være o_i (i=1,2,3, ... n), og la gjennomsnittet være m. Differensene mellom de enkelte observasjonene og gjennomsnittet er

$$(o_1-m),(o_2-m),(o_3-m), \ldots\ldots\ldots\ldots (o_n-m)$$

Ved å kvadrere disse differensene, summere kvadratene og dividere denne summen med et tall som er en ener mindre enn antallet av observasjoner, med n-1, får vi en størrelse som kalles variansen. Vi vil betegne denne med V og har da at

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

Middelavviket er lik kvadratroten av variansen. Betegner vi middelavviket med s, har vi altså at

$$s = \sqrt{V}$$

Eksempel 8.

Samme observasjonene som i eks. 1. n=29 og m=20,14.

o_i	o_i-m	$(o_i-m)^2$
14	- 6,14	37,6996
14	- 6,14	37,6996
17	- 3,14	9,8596
22	+ 1,86	3,4596
18	- 2,14	4,5796
18	- 2,14	4,5796
20	- 0,14	0,0196
16	- 4,14	17,1396
17	- 3,14	9,8596
17	- 3,14	9,8596
16	- 4,14	17,1396
22	+ 1,86	3,4596
20	- 0,14	0,0196
20	- 0,14	0,0196
22	+ 1,86	3,4596
25	+ 4,86	23,6196
24	+ 3,86	14,8996
21	+ 0,86	0,7396
22	+ 1,86	3,4596
19	- 1,14	1,2996
21	+ 0,86	0,7396
20	- 0,14	0,0196
22	+ 1,86	3,4596
20	- 0,14	0,0196
25	+ 4,86	23,6196
22	+ 1,86	3,4596
22	+ 1,86	3,4596
24	+ 3,86	14,8996
24	+ 3,86	14,8996
584	- 0,06	267,4484

Herav finnes

$$V = \frac{267,4484}{28} = 9,5517$$

$$s = \sqrt{9,5517} = 3,09 \text{ (mm)}$$

En bør merke seg at

$$\Sigma (o_i - m) = \Sigma o_i - \Sigma m = \Sigma o_i - nm = nm - nm = 0$$

Når denne summen ikke er blitt nøyaktig lik 0 i eks. 8, kommer dette av at vi har brukt en avrundet verdi for m, nemlig m = 20,14.

Er observasjonene ordnet i en fordelingsrekke, er formelen for variansen lik

$$V = \frac{\Sigma h_i (x_i - m)^2}{n-1}$$

Nå blir det jo differensene $(x_i - m)$ som skal kvadreres og summeres, men under summeringen må vi naturligvis ta med hvert kvadrat så mange ganger som den til x_i svarende frekvens (h_i) angir.

Oppgave 3.

Beregn middelavviket for observasjonene i eks. 5 og eks. 7.

5. Praktisk metode til beregning av gjennomsnittet og middelavviket.

Den metoden vi har brukt ovenfor til beregning av gjennomsnittet og middelavviket kan ikke anbefales til bruk i praksis. Den er tungvint. Gjennomsnittet er jo i regelen et tall med flere desimaler selv om observasjonene er hele tall. Det vil derfor som oftest være meget arbeidskrevende å beregne differensene $(o_i - m)$, og dessuten skal jo disse differensene også kvadreres og kvadratene summeres. I praksis bruker en derfor helst en indirekte metode som vi nå skal forklare.

La a være et bestemt valt tall. Vi danner alle differensene $(o_i - a)$ og summerer disse. Vi finner da:

$$\Sigma (o_i - a) = \Sigma o_i - na = nm - na$$

Følgelig er

$$m = \frac{\Sigma (o_i - a)}{n} + a$$

For å finne gjennomsnittet kan vi altså beregne alle differensene $(o_i - a)$, summere disse, dividere summen med antallet av observasjonene og til dette resultat addere det valte tall a.

Disse differensene kan også brukes til beregning av variansen.

For kvadratet $(o_i-a)^2$ har vi nemlig:

$$(o_i-a)^2 = [(o_i-m) + (m-a)]^2 = (o_i-m)^2 + 2(m-a)(o_i-m) + (m-a)^2$$

Følgelig er:

$$\Sigma (o_i-a)^2 = \Sigma(o_i-m)^2 + 2(m-a) \Sigma(o_i-m) + n.(m-a)^2$$

og siden nå

$$\Sigma(o_i-m) = 0$$

$$\Sigma(o_i-m)^2 = (n-1).V$$

finner vi at

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

Når vi bruker denne fremgangsmåten, står det oss fritt å velge a slik det passer best i hvert enkelt tilfelle. Vi må da naturligvis velge a slik at differensene (o_i-a) blir enklest mulige tall. Hensikten med å bruke denne indirekte metoden er jo nettopp den å få så enkle tall å regne på som overhodet mulig. I mange lærebøker anbefales det å velge for a en verdi som antas å ligge i nærheten av gjennomsnittet. Men det er bedre i mange tilfelle å sette a lik den minste observasjonsverdi eller lik et tall som er noe mindre enn denne. Derved oppnår en nemlig det regneteknisk sett heldige at alle differensene (o_i-a) blir positive. I andre tilfelle kan det lønne seg å sette a=0. Det er ikke mulig å gi noen alminnelig regel for valget av a. En får velge den verdi som passer best i hvert enkelt tilfelle, og hvilken verdi det er som passer best, avhenger både av hvordan observasjonene er og av hvilke tekniske hjelpemidler en har til rådighet.

Velger en a=0, har en følgende formler:

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

og

$$V = \frac{\Sigma o_i^2 - nm^2}{n-1}$$

Oppgave 4.

Beregn gjennomsnittet og middelavviket for observasjonene i eks. 1 og eks. 4. I eks. 1 velges a=14 og i eks. 4 a=0,51.

Er observasjonene ordnet i en fordelingsrekke, skal vi bruke følgende formler til beregning av m og V:

$$m = \frac{\Sigma h_i(x_i-a)}{n} + a$$

$$V = \frac{\Sigma h_i(x_i-a)^2 - n(m-a)^2}{n-1}$$

Det regneskjemaet en skal bruke er vist i eks. 9.

Eksempel 9.

Samme observasjonene som i eks. 6. a = 10.

x_i	h_i	x_i-a	$h_i(x_i-a)$	$h_i(x_i-a)^2$
2	1	- 8	- 8	64
3	1	- 7	- 7	49
4	4	- 6	- 24	144
5	6	- 5	- 30	150
6	17	- 4	- 68	272
7	20	- 3	- 60	180
8	30	- 2	- 60	120
9	35	- 1	- 35	35
10	51	0	0	0
11	52	1	52	52
12	39	2	78	156
13	45	3	135	405
14	21	4	84	336
15	7	5	35	175
16	5	6	30	180
	334		+ 122	2318

Herav:

$$m = \frac{122}{334} + 10 = 10,37 \qquad (10,3653)$$

$$V = \frac{2318 - 334.(10,3653-10)^2}{333} = 6,8271$$

$$s = \sqrt{6,8271} = 2,61$$

Oppgave 5.

Beregn gjennomsnittet og middelavviket for observasjonene i eks. 2 og eks. 3. I eks. 2 velges a=12 og i eks. 3 a=16,5.

6. Middelavviket som karakteristikk av observasjonsrekken.

Vi har allerede nevnt at de aller fleste fordelingsrekkene viser en mer eller mindre markert opphopning av observasjoner omtrent på midten av variasjonsområdet. Dette er iallfall regelen når det gjelder biologiske observasjoner. Fordelingsrekken i eks. 7 er en unntagelse.

Når fordelingsrekken har dette karakteristiske utseende, vil gjennomsnittet ha en verdi som ligger omtrent på midten av variasjonsområdet. Gjennomsnittet for rekken i eks. 6 er 10,37 og for rekken i eks. 3 17,83. For beskrivelsen av observasjonsrekken som helhet er det naturligvis av stor interesse å kunne gi et kvantitativt uttrykk for hvor sterk denne opphopningen er, eller m.a.o. hvor tett samlet omkring gjennomsnittet observasjonene

er. En har forskjellige størrelser til bruk for dette formål. I biologisk statistikk brukes mest middelavviket. Vi skal senere komme inn på spørsmålet om hvordan middelavviket kan brukes som karakteristikk eller kvantitativt mål for opphopningen. Her skal vi nøye oss med å konstatere at en i regelen vil finne praktisk talt alle observasjonene innen området fra (m-3s) til (m+3s). For rekken i eks. 9 er m=10,37 og s=2,61. Herav finnes:

$$m - 3s = 2{,}54 \quad \text{og} \quad m + 3s = 18{,}20$$

Vi ser at den minste observasjon er 2 og den største 16. Unntatt den ene observasjon på x = 2, faller altså alle observasjonene innenfor det nevnte område.

For rekken i eks. 3 har vi m = 17,83 og s = 2,91. Altså er

$$m - 3s = 9{,}10 \quad \text{og} \quad m + 3s = 26{,}56$$

I dette tilfelle faller altså alle observasjonene innenfor området mellom (m-3s) og (m+3s).

I sin alminnelighet kan vi si at observasjonene vil falle innenfor et område mellom (m-ts) og (m+ts) hvor t har en verdi på 3-4. Det kan også bevises, noe vi skal komme tilbake til senere, at for t > 1 vil minst

$$H_t = 100 \cdot \left(1 - \frac{1}{t^2}\right) \%$$

av observasjonene falle innenfor området (m-ts) til (m+ts). Setter vi her t=3, finnes $H_3 = 88{,}9$ %, og settes t=4, finnes $H_4 = 93{,}8$ %.

Gjennomsnittet og middelavviket gir derfor sammen en ganske god totalbeskrivelse av observasjonsrekken som helhet. Ved siden av disse to størrelsene bruker en meget ofte et eller annet mål for fordelingsrekkens skjevhet, men vi har ikke anledning til å komme noe inn på dette spørsmålet.

7. Universet.

Vi har hittil utelukkende beskjeftiget oss med spørsmålet om hvordan en observasjonsrekke skal beskrives. Denne oppgaven skal vi også fortsatt beskjeftige oss med. Men det vil ha en viss betydning allerede nå å søke å utvide synsfeltet noe. Denne rent beskrivende teknikk er selvsagt nødvendig i enhver statistisk undersøkelse. Men en slik beskrivelse kan ikke gi oss umiddelbart noe grunnlag for mer alminnelige slutninger som må til dersom det skal vinnes ny erkjennelse.

La oss ta for oss et eksempel.

Praktiske arbeidshensyn som meldte seg under en større kjemisk undersøkelse av gjødsel (sauegjødsel), gjorde det ønskelig å få bragt på det rene om gjødslas kjemiske sammensetning endret seg når den ble konser-

vert i kloroform og lagret i isskap noen døgn. Spesielt var en interessert i å få rede på om kvelstoffinnholdet endret seg.

I den hensikt å få bragt dette på det rene ble det tatt ut to mindre prøver av en større gjødselprøve. Den ene av disse prøvene ble analysert med det samme, mens gjødsla var frisk. Den andre prøven ble konservert i kloroform og lagret i isskap i 5 døgn og deretter analysert. Resultatet ble at en fant litt større relativ kvelstoffmengde i den konserverte enn i den friske prøven.

Vedkommende som sto for undersøkelsen, var på forhånd klar over at han ville finne en viss forskjell mellom de to prøvene selv om det ikke skjer noen kjemisk forandring under lagringen. For det første måtte han regne med en forskjell på grunn av at det ikke er mulig å utføre en slik kvantitativ kjemisk analyse helt feilfritt. Og for det annet hadde han ingen garanti for at de to prøvene fra begynnelsen av var identisk like m.h.p. kvelstoffinnholdet. Den store prøven behøvde jo ikke være helt ensartet. Han var derfor klar over at han ikke kunne nøye seg med dette ene forsøket og gjentok det derfor ialt 16 ganger. Resultatene av disse forsøkene er gjengitt i eks. 10. Tallene som er betegnet med o_i i tabellens hode, er analyseresultatene av de friske prøvene. I annen kollonne (under o_i') er gitt analyseresultatene av de tilsvarende lagrete prøvene. $O_i = o_i - o_i'$ er differensene. Observasjonene o_i og o_i' er gitt i prosent.

Eksempel 10.

o_i	o_i'	O_i
0,9328	0,9376	- 0,0048
0,9456	0,9120	+ 0,0336
0,6800	0,6768	+ 0,0032
0,9088	0,8736	+ 0,0352
0,8736	0,8528	+ 0,0208
0,9904	0,9472	+ 0,0432
0,5504	0,5408	+ 0,0096
0,6672	0,6640	+ 0,0032
0,9360	0,9120	+ 0,0240
0,8560	0,8192	+ 0,0368
0,5584	0,5360	+ 0,0224
0,6272	0,5904	+ 0,0368
0,9184	0,8928	+ 0,0256
0,7600	0,7136	+ 0,0464
0,5264	0,5056	+ 0,0208
0,6080	0,5904	+ 0,0176

Differensene, $O_i = o_i - o_i'$, kan vi oppfatte som en observasjonsrekke. Vi ser at disse observasjonene varierer fra forsøk til forsøk, vi ser at 15

av observasjonene er positive og bare en negativ. Vi kunne videre beskrive rekken mer inngående ved gjennomsnittet, middelavviket osv. Dette ville imidlertid ikke kunne gi oss noe svar på det spørsmålet som ble stilt, om kvelstoffinnholdet forandrer seg under lagringen i sin alminnelighet. Når vi stiller spørsmålet slik, tenker vi ikke lenger på de faktiske observasjonene, men på de uendelig mange observasjonene vi ville få ved å gjenta forsøket utover i all framtid.

Vi skal senere vise at svaret på dette spørsmålet er at det under lagringen foregår en minskning av kvelstoffinnholdet. En vurderende undersøkelse av de 16 forsøksresultatene gir altså et resultat som ikke bare gjelder de faktiske observasjonene, men hele den tenkte rekken av uendelig mange observasjoner som ville kunne skaffes til veie ved forsøk utover i framtiden. I logikken kalles en slik slutning en induksjon. Det er altså en slutning (eller dom) om noe alminnelig på grunnlag av noe spesielt.

Denne tenkte uendelige rekken av observasjoner kaller vi et univers. De faktiske observasjonene betrakter en som et utvalg av universet.

Praktisk talt alle observasjonsrekker må betraktes som utvalg av uendelige universer. Dette er kanskje særlig lett forståelig når observasjonene skaffes til veie ved forsøk. Vi kan gjenta målingen av lengden av en linje i terrenget så ofte vi vil, og vi kan tenke oss en bestemt kjemisk analyse gjentatt i det uendelige. I praksis er naturligvis muligheten for gjentagelse begrenset. Det tar tid å måle en linje, og materialet for en kjemisk analyse foreligger i begrenset mengde. Men også i andre tilfelle må vi som oftest betrakte universet som uendelig. Teller en f. eks. antallet av kronblader i blomsten hos Engsoleie, er mulighetene for stadig nye observasjoner langt flere enn en i praksis kan benytte seg av. Den eksisterende bestand av planter må imidlertid oppfattes som en representant for en meget større bestand (teoretisk sett uendelig stor) som også omfatter planter som ville ha vært eksisterende hvis de frø de skulle vokset opp fra, hadde fått levelige vilkår.

Vi skal senere komme tilbake til spørsmålet om hvordan vi kan slutte noe om universet på grunnlag av et endelig utvalg av faktiske observasjoner. Det vi skal merke oss her, er at de faktiske observasjonene alltid må betraktes som et utvalg av et univers som i regelen er ubegrenset. Til utvalget må en i alminnelighet stille det krav at det skal være tilfeldig, men hva som ligger i dette kravet, kan vi først forklare etter at vi har lært litt sannsynlighetsregning.

8. Om årsakene til variasjonen.

Observasjonene vil alltid variere i større eller mindre grad. Årsakene til denne variasjonen er mange og av mange forskjellige slags. I de fleste tilfelle kan vi ved logisk resonnement eller ved undersøkelse skille ut noen av disse årsaker eller årsakskomplekser. La oss som eksempel tenke oss at det foreligger observasjoner over kløveravlingen på en rekke like store forsøksruter. (Sml. eks. 3). Vi kan da først skille ut en gruppe årsaker som vi kan kalle feilårsaker. Det er ikke til å unngå at hver enkelt observasjon blir beheftet med observasjonsfeil. Det gjøres feil under utstikkingen av rutene, under slåtten og bergingen av høyet og under veiingen. Når en betrakter rekken i eks. 3, vil en imidlertid snart bli klar over at selv om en ved å arbeide nøyaktig reduserer feilårsakenes virkning til et minimum, vil variasjonen ikke bli synderlig minsket. Det er andre årsaker som i dette tilfelle er dominerende. Er forsøksrutene ens bearbeidet og gjødslet, vil variasjonen i observasjonene i hovedsaken skyldes jordvariasjonen.

Variasjonen skyldes at det årsakskomplekset som er bestemmende for vedkommende kjennetegn, ikke er helt uforandret fra enhet til enhet, det er ikke helt stabilt. Vi skulle derfor vente at hvis vi deler opp et univers i flere subuniverser etter et slikt prinsipp at det årsakskomplekset som er bestemmende for kjennetegnet, er mer stabilt innen subuniversene enn i det uoppdelte univers, vil variasjonen i subuniversene bli mindre enn i det uoppdelte univers. Erfaringsmessig er også dette tilfelle. Vi skal ta for oss et eksempel.

Eksempel 11.

Avlingsforsøk med beter. En veiet avlingene for hver forsøksenhet og omregnet vektene til avling (rötter) pr. dekar. I forsøket var det tatt med tre sorter: forsukkerbeter (Fsb), forbeter (Fb) og sukkerbeter (Sb). For Fsb hadde en n=19 paralleller (stammer), for Fb n=12 og for Sb n=6 paralleller. Observasjonene for hver forsøksenhet er gitt i tonn pr. dekar.

Fsb (n=19)			Fb (n=12)		Sb (n=6)
3,8	4,6	3,3	6,1	5,9	3,6
4,3	5,7	3,4	4,9	5,3	2,2
4,7	5,1	4,5	5,0	5,9	3,4
4,1	4,1	5,2	4,9	5,4	2,3
3,8	4,3	4,5	5,1	5,5	3,5
4,6	3,7	4,5	5,8	6,5	2,5
5,0					

Gjennomsnittet og middelavviket for hver av disse observasjonsrekker og for alle tre rekker under ett er:

Sort	m	s
Alle sorter under ett	4,51	1,05
Fsb	4,38	0,62
Fb	5,53	0,52
Sb	2,92	0,65

Vi ser at gjennomsnittsavlingene er forskjellige for de tre sortene. Dette kan skyldes årsaker som ikke har noe med sortsegenskapene å gjøre (f. eks. jordvariasjonen). Vi vil her tenke oss at forsøket er planlagt og utført på en slik måte at sortsegenskapene har kunnet gjøre seg gjeldende, og vi vil videre gå ut fra at resultatet av forsøket gir et riktig uttrykk for forskjellene mellom sortsuniversene. Hvis det er tilfelle, er det klart at sortsegenskapene er årsakskomplekser som har en dominerende innflytelse på kjennetegnet i det univers som omfatter alle sortene. Ved en oppdeling av dette universet etter sortene, har vi fått fram subuniverser med ulike gjennomsnitter og sterkt reduserte middelavvik. Variasjonen i det uoppdelte univers skyldes for en stor del at sortsuniversene har ulike gjennomsnitter.

Dette eksemplet er et eksempel på en helt alminnelig regel. Hvis oppdelingen av en observasjonsrekke foretas på en slik måte at gjennomsnittene i delrekkene blir ulike, vil middelavviket bli redusert. Og hvis dette empiriske resultat kan overføres på universene (de uoppdelte og subuniversene), vil det bety at vi ved oppdelingen har dannet universer hvor det årsakskompleks som er bestemmende for kjennetegnet, er mer stabilt enn i det uoppdelte univers. Oppdelingen har m.a.o. ført til renere eller mer enhetlige universer.

Hvordan en i praksis skal kunne avgjøre om de empiriske resultatene kan overføres til universet, skal vi ikke beskjeftige oss med her. Vi kommer tilbake til dette spørsmål senere.

9. K o r r e l a s j o n.

La oss nå tenke oss at to variable kjennetegn er observert hos n enheter. Vi har da for oss en rekke på n parobservasjoner. Eks. 12 er et eksempel på en slik rekke. Enhetene er i dette tilfelle byggkorn, og de to observerte kjennetegn er vekten og kvelstoffinnholdet. Observasjonene av vekten (o_i) er gitt i mgr og observasjonene av kvelstoffinnholdet (o_i') er gitt i prosent. Korn nr. 1 veier altså 66 mgr og har et kvelstoffinnhold på 1,71 %.

Eksempel 12.

Korn nr.	o_i	o_i'
1	66,0	1,71
2	62,4	1,57
3	58,8	1,66
4	53,4	1,52
5	51,1	1,36
6	51,2	1,41
7	49,0	1,29
8	51,2	1,31
9	55,2	1,45
10	55,3	1,42
11	48,5	1,31
12	52,4	1,44
13	54,8	1,31
14	51,8	1,33
15	59,6	1,74
16	56,8	1,51
17	53,4	1,67
18	54,8	1,39
19	51,8	1,49
20	51,8	1,45
21	55,4	1,53
22	51,0	1,24
23	54,6	1,41
24	50,2	1,45
25	61,4	1,87

Sammenlikner vi nå observasjonene parvis, ser vi at stort sett er store o-verdier kombinert med store o'-verdier og små o-verdier med små o'-verdier. Det er m.a.o. en tendens til at korn med stor vekt har større kvelstoffinnhold pr. vektenhet enn mindre korn. Hvis denne regelen også gjelder det univers av parobservasjoner som disse observasjonene er et utvalg av, sier vi at det er positiv korrelasjon mellom de to kjennetegnene.

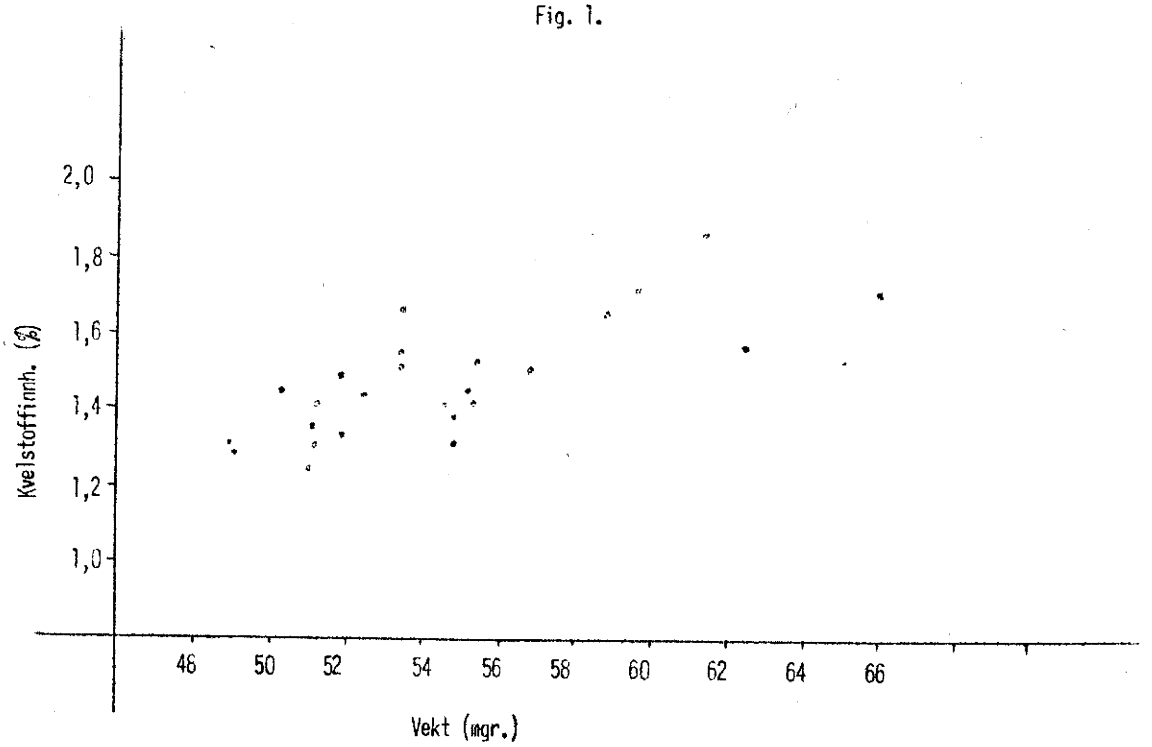
I andre tilfelle kan regelen være den at store o-verdier er fortrinsvis kombinert med små o'-verdier og små o-verdier med store o'-verdier. Hvis denne regelen også gjelder universet, sier vi at det er negativ korrelasjons mellom de to kjennetegnene.

For å skaffe oss bedre oversikt kan vi fremstille observasjonsparene i et rettvinklet koordinatsystem. Til hvert observasjonspar svarer et punkt hvis anscisse er o_i og hvis ordinat er o_i'. Til n observasjonspar svarer altså n punkter, en punktsverm. Figur 1 viser et slikt punktdiagram for observasjonene i eks. 12.

Oppgave 6.

Fremstill observasjonene i eks. 10 (o_i, o_i') ved et punktdiagram.

Fig. 1.



Vi har tidligere pekt på at verdien av ett kjennetegn hos en enhet er påvirket av mange årsaker. To eller flere kjennetegn hos samme enhet vil derfor i mange tilfelle være avhengig av årsaker som i større eller mindre grad er felles for dem. La oss f. eks. tenke oss at kjennetegnene er lengden av høyre og venstre lårben hos voksne menn. Det er umiddelbart innlysende at disse to kjennetegnene må være påvirket av et stort antall årsaker som virker noenlunde likt på de to lårbens vekst. Det er årsaker av arvelig natur og ernæringsforholdene under oppveksten. Måler vi derfor lengden av høyre og venstre lårben hos et antall voksne menn, må de observasjonene vi får, være positivt korrelerte. De to lårbens vekst er i hovedsaken bestemt av de samme årsakene slik at det blant de årsakene som er bestemmende for høyre lårbens vekst, er det mange som også er bestemmende for veksten av venstre lårben.

På grisens forben finnes noen kjerteldannelser som kalles Müllerske kjertler. Vi skal senere vise at det er positiv korrelasjon mellom antallet av disse kjertlene på høyre og venstre forben. Dette skyldes at det blant de årsaker som er bestemmende for antallet på høyre forben, er noen som samtidig også er bestemmende for antallet på venstre forben.

Korrelasjon mellom to kjennetegn kan imidlertid også skyldes at det ene kjennetegn er en av de årsaker som er bestemmende for det annet. Vi vet f. eks. at avlingens størrelse er avhengig av hvor sterkt det gjødsles.

Innenfor rimelige grenser for de brukte gjødselmengder vil avlingen stort sett vokse når en bruker større og større mengder gjødsel pr. arealenhet. Mellom avlingens størrelse og gjødselmengden pr. arealenhet må det derfor være positiv korrelasjon. Dette kommer av at gjødselmengden er en av de årsaker som er medbestemmende for størrelsen av avlingen. Men en bør være oppmerksom på at størrelsen av avlingen også er avhengig av andre faktorer. Utføres forsøket som feltforsøk, vil jordvariasjonen være en av disse andre faktorer.

Korrelasjonsundersøkelser spiller nå en meget stor rolle i naturvitenskapelige undersøkelser. I de fleste tilfelle kommer en ikke lenger enn til en påvisning av at to eller flere kjennetegn er korrelerte, mens oppklaringen av det eller de årsaksforhold som betinger denne korrelasjon, må oppgis. Dette er kanskje særlig alminnelig innen biologiske undersøkelser hvor årsaksforholdene er særlig kompliserte. Men selv om en må stoppe opp med dette, er selvsagt påvisningen av korrelasjon av meget stor interesse.

10. K o r r e l a s j o n s t a b e l l e n.

Vi har tidligere lært hvordan en kan skaffe seg en oversikt over observasjonene av ett kjennetegn ved å ordne dem i en fordelingsrekke. Vi benytter en tilsvarende metode når vi har observasjoner av to kjennetegn. Det skjema som brukes til dette, kalles en korrelasjonstabell. Vi skal referere et par eksempler.

Vi skal først ta for oss et eksempel på en korrelasjonstabell der begge kjennetegnene er diskrete (eks. 13). I denne tabellen er x antallet av Müllerske kjertler på grisens høyre forben og y antallet av Müllerske kjertler på venstre forben. Vi ser at antallet varierer fra 0 til 10 på begge ben. Frekvensene (tallene inne i tabellen) viser hvor mange av de n=2000 undersøkte grisene det er som har x=0,1,2,....10 kjertler på høyre ben og y=0,1,2,3....10 kjertler på venstre ben. Vi ser at det er 8 griser som mangler kjertler på begge ben. Det er 5 griser som har 1 kjertel på venstre og ingen på høyre. Det er 151 griser som har 1 kjertel på høyre og 1 kjertel på venstre ben. Det er 119 griser som har 4 kjertler på høyre og 3 kjertler på venstre ben osv.

Eksempel 13.

y \ x	0	1	2	3	4	5	6	7	8	9	10	h
0	8	4	2									14
1	5	151	65	14	5	1						241
2	2	58	154	38	27	7						336
3		9	96	173	119	24	8	1				430
4		3	28	128	153	92	16	8	1			429
5			7	28	77	101	58	20	3	1		295
6			1	6	26	52	48	18	5	3		159
7					3	11	16	17	3	3		53
8					1	9	7	9	2	2		30
9								5	2	2	1	10
10							2			1		3
h	15	225	353	437	411	297	155	78	16	12	1	2000

Tallene i nederste rekke i tabellen er summene av frekvensene i kollonnene. Disse tallene er frekvensene i fordelingsrekken for antall kjertler på høyre ben (for x) når en ikke tar noe hensyn til hvor mange kjertler det er på venstre ben. På samme måte er tallene i kollonnen lengst til høyre i tabellen frekvensene i fordelingsrekken for antall kjertler på venstre forben (for y) når en ikke tar noe hensyn til antallet på høyre ben. Disse to fordelingsrekkene kalles korrelasjonstabellens marginale fordelingsrekker.

Tar vi nå ut av den statistiske massen de enhetene som har f. eks. x=5 kjertler på høyre ben, det er 297 slike, finnes frekvensene i fordelingsrekken for antall kjertler på venstre ben (for y) for denne gruppen under x=5 i tabellen. Denne fordelingsrekken er altså:

y	h
1	1
2	7
3	24
4	92
5	101
6	52
7	11
8	9
	297

Tar vi ut de enhetene som har f. eks. y=4 kjertler på venstre ben, finnes frekvensene i fordelingsrekken for antallet på høyre ben (for x) for denne gruppen i rekken ut for y=4. Denne fordelingsrekken er:

x	h
1	3
2	28
3	128
4	153
5	92
6	16
7	8
8	1
	429

Slike fordelingsrekker som dette kalles betingete fordelingsrekker. I denne korrelasjonstabellen har vi altså 11 betingete fordelingsrekker for x og 11 for y.

Både for de to marginale fordelingsrekkene og for de betingete kan vi selvsagt beregne gjennomsnittet og middelavviket på vanlig måte. Vi har to marginale gjennomsnitter og to marginale middelavvik. Antallet av betingete gjennomsnitter og middelavvik er lik antallet av betingete fordelingsrekker.

Vi ser at korrelasjonstabellen gir en meget god oversikt over observasjonene som helhet. Videre ser vi at frekvensene er ordnet nokså regelmessig omkring diagonalen fra tabellens øvre venstre til nedre høyre hjørne, dvs. at stort sett er store x-verdier kombinert med store y-verdier og små x-verdier med små y-verdier. Dette tyder på at det er positiv korrelasjon mellom de to observerte kjennetegn i det univers som disse observasjonene er et utvalg av.

Eks. 13 viser hvordan korrelasjonstabellen er bygd opp når begge kjennetegnene er diskrete. Er enten begge eller bare det ene kjennetegnet kontinuerlig, må en først foreta en klasseinndeling av observasjonene på samme måte som når en skal stille opp fordelingsrekken for observasjonene av ett kontinuerlig kjennetegn. Deretter samordnes observasjonsparene i korrelasjonstabellen på samme måte som for diskrete kjennetegn.

I eks. 14 er x verdiene av observasjonene av fettinnholdet (i %) i havrekorn og y verdiene av observasjonene av kornvekten (i mgr.). I tabellen er ført opp både de klasser som er brukt til klasseinndelingen av observasjonene og klassenes midtverdier (x og y). Vi ser at av de 224 undersøkte korn er det f.eks. 48 som veier mellom 40 og 45 mgr. og hvis fettinnhold ligger mellom 6 og $6\frac{1}{2}$ %.

Vi ser at små kornvekter fortrinnsvis er kombinert med store fettprosenter og store kornvekter med små fettprosenter. Frekvensene er ordnet omkring diagonalen fra tabellens nedre venstre til øvre høyre hjørne. Dette tyder på at det er negativ korrelasjon mellom kjennetegnene i det univers som disse observasjonene er et utvalg av.

Eksempel 14.

Vekt-klasser	x / y	Fettprosent-klasser								h
		4½-5	5-5½	5½-6	6-6½	6½-7	7-7½	7½-8	8-8½	
		4,75	5,25	5,75	6,25	6,75	7,25	7,75	8,25	
30-35	32,5					8	2	1		11
35-40	37,5		1	6	22	33	10	2	1	75
40-45	42,5	1	2	10	48	37	8	1		107
45-50	47,5		1	12	11	2				26
50-55	52,5		2	1	1					4
55-60	57,5			1						1
h		1	6	30	82	80	20	4	1	224

Når ett eller begge kjennetegn er kontinuerlig, bør en ikke samordne observasjonene i korrelasjonstabell uten i de tilfelle der antallet er så stort at de forskjellige beregningene som skal utføres blir for meget arbeidskrevende på grunnlag av primærlisten. Det at en grupperer observasjonene i klasser og bruker klassenes midtverdier gjør jo at det kommer inn feil som helst bør unngås. I det følgende vil vi forutsette at observasjonene er ordnet i korrelasjonstabell. Senere skal vi så vise hvordan beregningene skal utføres i de tilfelle der antallet av observasjoner er så lite at en ikke bør bruke korrelasjonstabell.

11. Regressjonslinjene.

Vi skal nå se noe nærmere på de betingete gjennomsnittene. Disse beregnes på helt ordinær måte av de betingete fordelingsrekkene i korrelasjonstabellen. Vi vil betegne de betingete gjennomsnittene for y med m(y/x) og de betingete gjennomsnittene for x med m(x/y). I følgende tabell er disse gjennomsnittene stilt sammen med verdiene av x og y for eks. 13.

x	m(y/x)	y	m(x/y)
0	0,60	0	0,57
1	1,36	1	1,44
2	2,31	2	2,30
3	3,20	3	3,18
4	3,89	4	3,90
5	4,78	5	4,85
6	5,51	6	5,44
7	6,14	7	6,28
8	6,50	8	6,27
9	7,33	9	7,90
10	9,00	10	7,00

Vi ser av denne tabellen at m(y/x) er nær proporsjonal med x og m(x/y) nær proporsjonal med y. Vi får et meget bedre inntrykk av dette når vi fremstiller sammenhengen mellom x og m(y/x) og mellom y og m(x/y) grafisk. Dette gjøres på følgende måte.

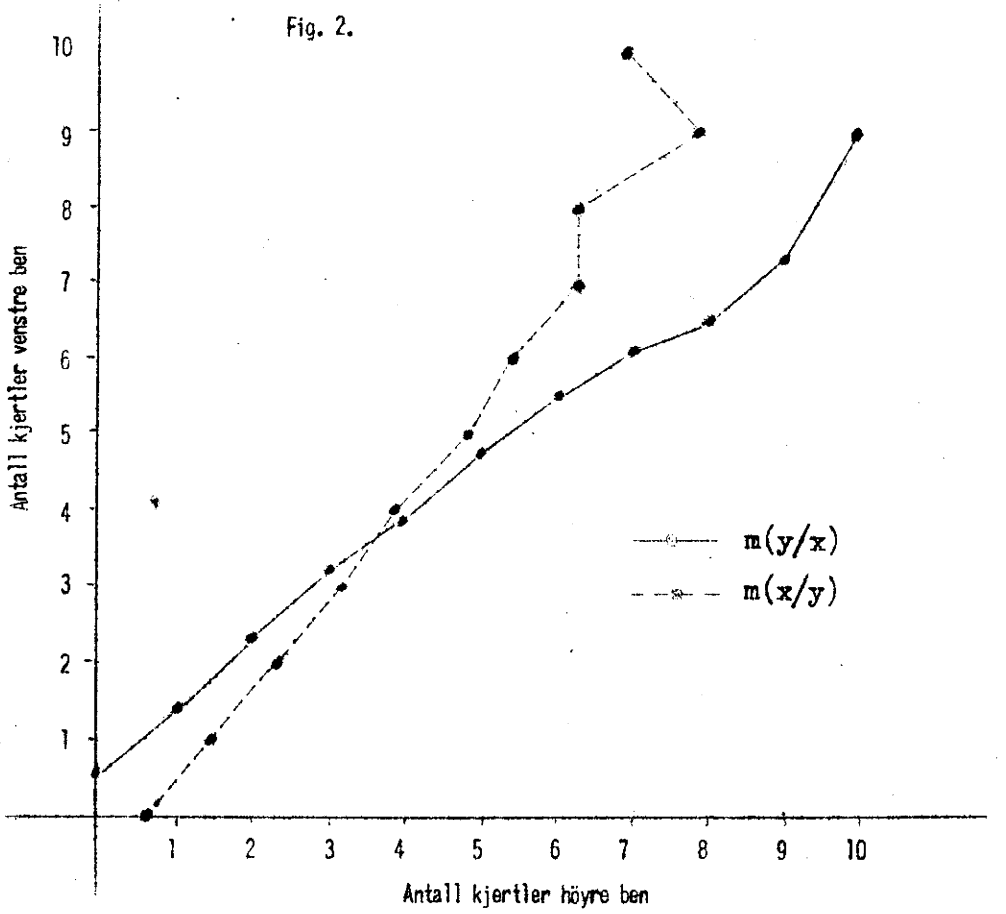
I et rettvinklet koordinatsystem avsettes punkter hvis abscisser er x og hvis ordinater er m(y/x). Disse punktene - det er like mange punkter som det er verdier av x - forbindes deretter med rette linjestykker. I samme koordinatsystemet avsettes punkter hvis ordinater er y og hvis abscisser er m(x/y). Disse punktene forbindes også med rette linjestykker. I figur 2 er de betingete gjennomsnittene for eks. 13 fremstilt på denne måten. Vi ser da at de to noe uregelmessige linjene har rettlinjet tendens. (Fig. 2 neste side).

Oppgave 7.

Beregn de betingete gjennomsnittene i eks. 14. Tegn figur.

Hvis universet var kjent - det vil i alminnelighet si at vi hadde uendelig mange observasjoner - ville vi kunne gi en tilsvarende grafisk fremstilling av sammenhengen mellom de betingete gjennomsnittene for det ene

Fig. 2.



kjennetegnet og verdiene av det andre kjennetegnet. De betingete gjennomsnittene for y i universet vil vi betegne med Y og de betingete gjennomsnittene for x med X. Det er klart at Y er en funksjon av x og X en funksjon av y. De to kurvene som er det grafiske bilde av disse to funksjonene, kaller vi regressjonslinjene. I svært mange tilfelle er disse kurvene rette linjer, og vi sier da at regressjonen er rettlinjet. Ellers taler en om krumlinjet regressjon.

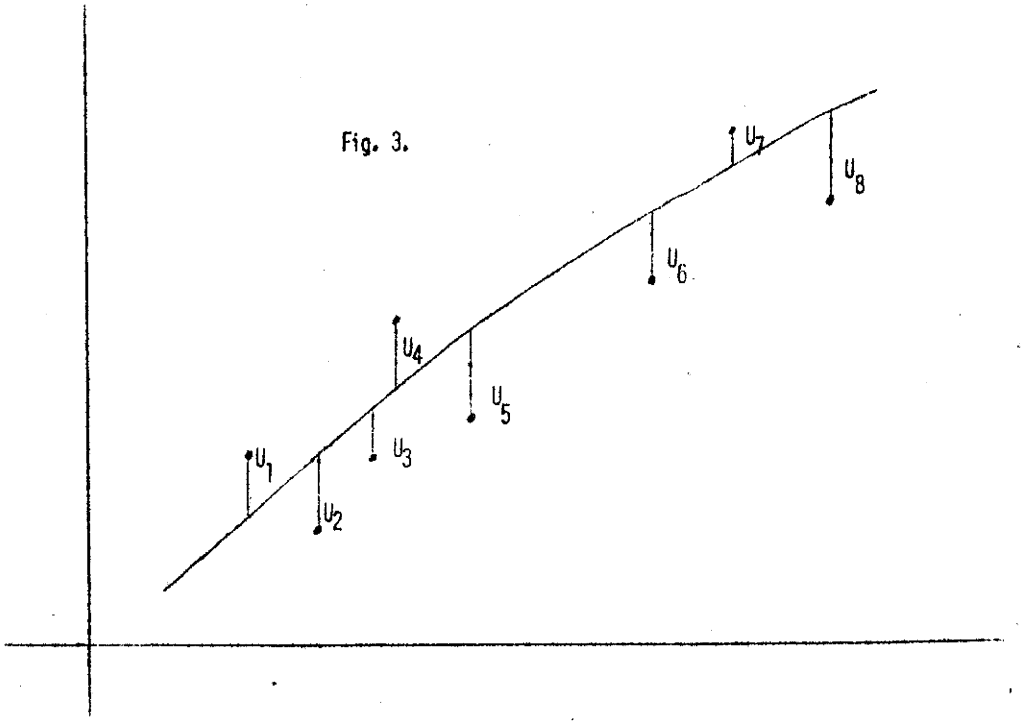
Universet er imidlertid ikke kjent, og av den grunn har vi ikke noe middel til å bestemme regressjonslinjene eksakt. Men når vi vet eller kan forutsette at regressjonslinjene tilhører en bestemt kurvetype, f. eks. den rette linjen, kan vi bestemme dem tilnærmet riktig. Da Y er en funksjon av x og X en funksjon av y, kan vi sette

$$Y = f(x) \qquad \text{og} \qquad X = \varphi(y)$$

Disse formlene, $f(x)$ og $\varphi(y)$, inneholder naturligvis ved siden av de uavhengig variable flere eller færre konstanter (parametrer). Det er disse konstantene som må bestemmes. Og den metoden som brukes til dette, kalles minste kvadraters metode.

La oss tenke oss at det i et rettvinklet koordinatsystem er avsatt

en rekke punkter: $(x_1,y_1),(x_2,y_2),(x_3,y_3)$ ..... (se fig. 3). Vi tenker da ikke på x og y som observasjonsverdier, men kort og godt som abscisser og ordinater til en rekke punkter. Videre tenker vi oss at det i samme



koordinatsystem er tegnet inn en kurve hvis ligning er $y=f(x)$. For punktet (x_1,y_1) er differensen mellom punktets ordinat og ordinaten til kurven for $x=x_1$ lik $u_1=y_1-f(x_1)$. Denne differensen er avhengig av hvilke verdier konstantene i $f(x)$ har. Ved å variere verdiene av disse konstantene kan vi flytte kurven i forhold til punktene omtrent etter forgodtbefinnende. Vi vil nå tenke oss at alle disse ordinatdifferensene er dannet. Vi kvadrerer dem og summerer kvadratene. Dersom da kurven ligger slik til i forhold til punktene at summen

$$S = u_1^2 + u_2^2 + u_3^2 + ...$$

er mindre når differensene $(u_1,u_2,u_3, ...)$ regnes ut med utgangspunkt i denne kurven, enn når de regnes ut med utgangspunkt i en hvilket som helst annen kurve med samme formel (men andre konstanter), er denne kurven bestemt etter minste kvadraters metode. Metoden går altså ut på å bestemme konstantene i formelen $f(x)$ slik at denne kvadratsummen blir minst mulig.

Hvis $y=ax+b$, dvs. at kurven er en rett linje, og a og b er bestemt etter minste kvadraters metode, er kvadratsummen av differensene $y_i-f(x_i) = y_i-ax_i-b$ mindre når differensene regnes ut fra denne rette linjen enn når de regnes ut fra en hvilken som helst annen rett linje. Vi skal ta for oss et eksempel. La punktenes koordinater være:

x	y
1	0,5
1	1,2
2	1,5
3	3,5
3	2,8
4	4,3

Disse punktene er avmerket på figur 4. Benytter vi en rett linje, y=ax+b, og bestemmer a og b etter minste kvadraters metode, finner vi a=1,17 og b= - 0,44. Regner vi nå ut differensene mellom punktenes ordinater og or-

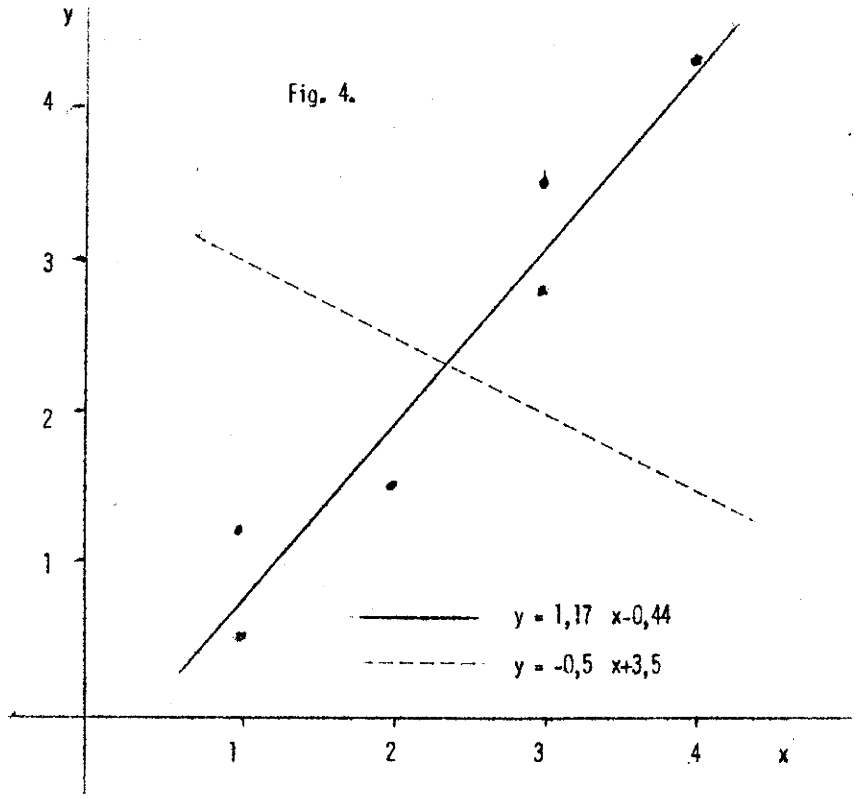


Fig. 4.

dinatene til denne rette linjen, kvadreres disse differensene og summerer, kommer vi til en kvadratsum på S=0,69. Velger vi en annen rett linje, f. eks. y = - 0,5x + 3,5, og regner ut differensene kommer vi til en kvadratsum på S = 21,22.

Disse to rette linjene er tegnet inn på figur 4. Vi ser da at den linjen som er bestemt etter minste kvadraters metode, skjærer rett gjennom punktsvermen. Den andre rette linjen som er valt på slump, skjærer punktsvermen over omtrent på midten.

Vi ser at ingen av punktene ligger nøyaktig på den rette linjen som er bestemt etter minste kvadraters metode. Men linjen ligger slik til i forhold til punktene at den gir en beskrivelse av punktsvermen som helhet.

Det fremgår av det foregående at vi i hvert enkelt tilfelle har to regressjonslinjer. Vi har regressjonslinjen for y med hensyn på x (forkor-

tet m.h.p. x), og vi har regressjonslinjen for x m.h.p. y. Dessuten må vi bruke betegnelsen regressjonslinje i to forskjellige betydninger som vi må holde skarpt atskilt:

1) i betydningen universets regressjonslinje
2) i betydningen regressjonslinje når konstantene i dens ligning er bestemt etter minste kvadraters metode.

For universets regressjonslinje vil vi i det følgende bruke betegnelsen regressjonslinje. Men når betegnelsen brukes i den andre betydningen, vil vi, når det kan oppstå forveksling, bruke betegnelsen empirisk regressjonslinje.

12. Rettlinjet regressjon.

Vi vil nå forutsette at universets regressjonslinjer er rette. Da er

$$Y = ax + b$$

og

$$X = cy + d$$

Oppgaven er å bestemme a,b,c og d etter minste kvadraters metode.

La oss anta at observasjonene (antall = n) er ordnet i en korrelasjonstabell. Den frekvens som svarer til observasjonsverdiene x_i og y_j betegner vi med h_{ij}. Altså er $\Sigma\Sigma h_{ij} = n$.

Den til x_i svarende ordinat til regressjonslinjen $Y = ax+b$ er $Y_i = ax_i+b$. For hver enkelt y-verdi svarende til $x=x_i$, skal vi nå danne differensene (y_j-Y_i), kvadrere disse og summere kvadratene. Vi må imidlertid også ta hensyn til at hver y-verdi forekommer så ofte som frekvensene angir. Holder vi oss til det geometriske bilde, vil det si det samme som at ethvert punkt med abscissen x_i og ordinaten y_j er et h_{ij}-dobbelt punkt. I eks. 13 er f. eks. punktet (0,0) et 8-dobbelt punkt, punktet (0,1) et 5-dobbelt punkt osv. Det bidrag som gruppen med $x=x_i$ gir til summen av alle kvadratene $(y-Y)^2$, er derfor

$$\Sigma h_{ij}(y_j-Y_i)^2 = \Sigma h_{ij}(y_j-ax_i-b)^2$$

hvor summeringen skal utstrekkes over hele den betingete fordelingsrekken for y svarende til $x=x_i$. Denne summen skal nå dannes for alle x-verdiene, og summen av alle disse gruppesummene er den totale sum av alle $(y-Y)^2$. Denne totalsum er derfor

$$S = \Sigma\Sigma h_{ij}(y_j-Y_i)^2 = \Sigma\Sigma h_{ij}(y_j-ax_i-b)^2$$

Oppgaven går ut på å bestemme a og b slik at S blir minst mulig. Den tek-

niske løsningen av denne oppgaven skal vi ikke komme inn på. Vi skal nøye oss med å referere de formlene som skal brukes til beregningen av a og b når S er et minimum. Disse er:

$$a = \frac{\Sigma\Sigma h_{ij} x_i y_j - n.m(x).m(y)}{(n-1).s(x)^2}$$

$$b = m(y) - a.m(x)$$

Konstantene i ligningen for den andre regressjonslinjen, $X = cy+d$, skal beregnes på tilsvarende måte. Formlene er:

$$c = \frac{\Sigma\Sigma h_{ij} x_i y_j - n.m(x).m(y)}{(n-1).s(y)^2}$$

$$d = m(x) - c.m(y)$$

I disse formlene er:

n = antall observasjoner, $n = \Sigma\Sigma h_{ij}$

$m(x)$ = det marginale gjennomsnitt for x

$m(y)$ = det marginale gjennomsnitt for y

$s(x)$ = det marginale middelavvik for x

$s(y)$ = det marginale middelavvik for y

Disse størrelsene beregnes på grunnlag av de to marginale fordelingsrekkene etter de metodene vi har lært tidligere (sml. eks. 9). For eks. 13 har vi $n = 2000$ og

$m(x) = 3,5465$ $\qquad$ $s(x) = 1,7195$

$m(y) = 3,5395$ $\qquad$ $s(y) = 1,7304$

Det står da igjen å beregne $\Sigma\Sigma h_{ij} x_i y_j$. Under et dobbelt summetegn har vi her tre faktorer hvorav to (x og y) har ulike fotskrifter og den tredje faktor (h) har begge de andre faktorenes fotskrifter. Summen beregnes derfor etter den metoden vi har beskrevet tidligere (I,8, eks. 5). For eks. 13 finnes $\Sigma\Sigma h_{ij} x_i y_j = 29\ 813$.

For eks. 13 har vi nå alle de tallene vi har bruk for til beregningen av a,b,c og d. Ved innsetting i formlene finner vi:

$a = 0,80$, $b = 0,72$, $c = 0,79$ og $d = 0,76$.

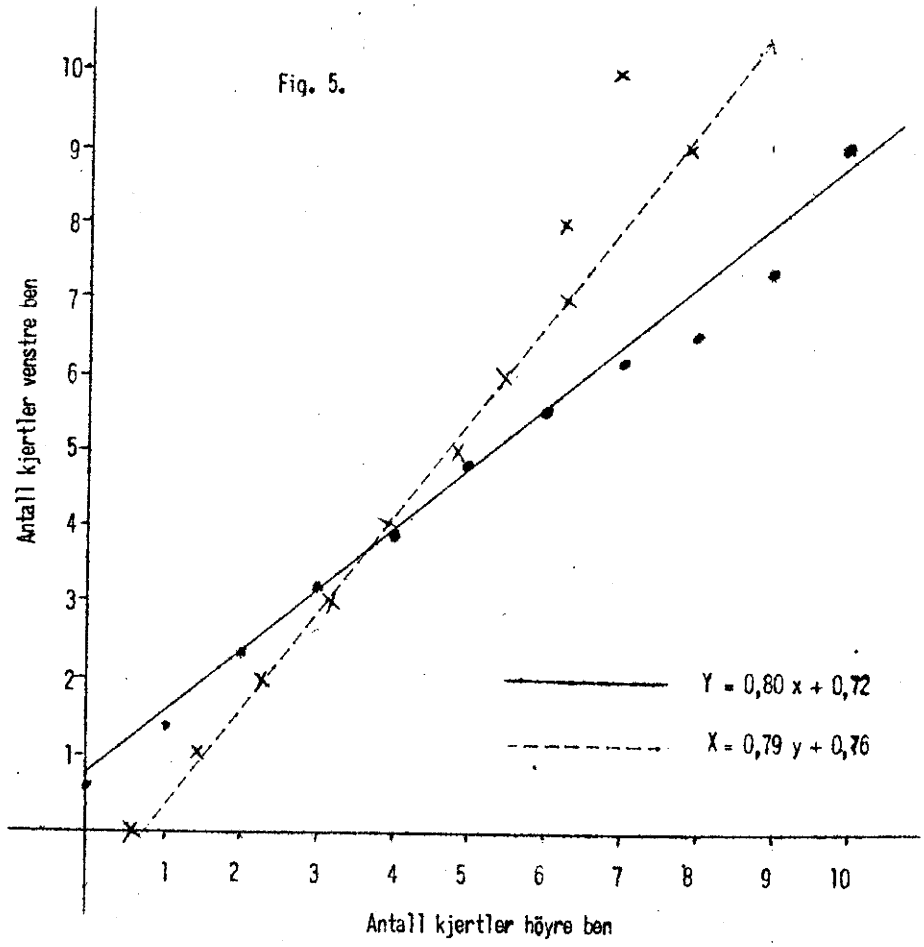
Ligningene for de to empiriske regressjonslinjene er derfor

$$Y = 0,80.x + 0,72$$

$$X = 0,79.y + 0,76$$

Disse to rette linjene er fremstillet grafisk i figur 5. Når en skal tegne inn disse to linjene i samme koordinatsystem, må en være oppmerk-

som på at når det gjelder den første linjen, Y = ax+b, er det den horisontale aksen (x-aksen) som er abscisseaksen. Når det gjelder den andre linjen, X = cy+d, er det den vertikale aksen (y-aksen) som er abscisseaksen.



I samme figur er også avmerket punkter hvis abscisser er x og ordinater m(y/x) og punkter hvis abscisser er y og ordinater m(x/y) (sml. fig. 2). Vi ser at disse punktene avviker lite fra de empiriske regressjonslinjene.

En bør merke seg at de to empiriske regressjonslinjene skjærer hverandre i det punkt som har m(x) til abscisse og m(y) til ordinat. Vi har nemlig at

$$m(y) = a.m(x) + b$$

og

$$m(x) = c.m(y) + d$$

Oppgave 8.

Finn ligningene for de empiriske regressjonslinjene for eks. 14. (Tegn figur). Det forutsettes at regressjonslinjene er rette.

Dersom observasjonsverdiene (x og y) er store tall, minsker en regnearbeidet betydelig ved å benytte avvikelsene (x-p) og (y-q) hvor p og q er valte tall. En kan lett bevise at

$$\Sigma\Sigma h_{ij}x_iy_j = \Sigma\Sigma h_{ij}(x_i-p)(y_j-q) + nqm(x) + npm(y) - npq$$

Oppgave 9.

Bevis dette.

Til beregningen av $\Sigma\Sigma h_{ij}(x_i-p)(y_j-q)$ benyttes samme regneskjema som til beregningen av $\Sigma\Sigma h_{ij}x_iy_j$. En skriver bare $(x-p)$ i stedet for x og $(y-q)$ i stedet for y. Setter vi spesielt $p = m(x)$ og $q = m(y)$, får vi

$$\Sigma\Sigma h_{ij}x_iy_j = \Sigma\Sigma h_{ij}[x_i-m(x)]\,[y_j-m(y)] + nm(x)m(y)$$

Oppgave 10.

Beregn $\Sigma\Sigma h_{ij}(x_i-p)(y_j-q)$ for eks. 13 og eks. 14. I eks. 13 settes $p=q=3$ og i eks. 14 $p=6{,}25$ og $q=42{,}5$.

Når antallet av observasjoner er så lite at vi ikke kan bruke korrelasjonstabellen, må vi beregne konstantene i regressjonslinjenes ligninger ved hjelp av den uordnete listen over parobservasjonene, primærlisten (sml. eks. 12). Vi bruker imidlertid akkurat samme metoden som før. Forskjellen er bare den at nå er alle frekvensene lik enheten ($h_{ij}=1$). I stedet for $\Sigma\Sigma h_{ij}x_iy_j$ skal vi nå beregne $\Sigma o_i o_i'$. Forresten vil fremgangsmåten læres best av et eksempel.

La oss forutsette at regressjonslinjene i det univers som eks. 12 er et utvalg av, er rette. Eks. 15 viser da hvilke beregninger som må utføres.

Eksempel 15.

o_i	o_i'	o_i^2	$o_i'^2$	$o_i \cdot o_i'$
66,0	1,71	4356,00	2,9241	112,860
62,4	1,57	3893,76	2,4649	97,968
58,8	1,66	3457,44	2,7556	97,608
53,4	1,52	2851,56	2,3104	81,168
51,1	1,36	2611,21	1,8496	69,496
51,2	1,41	2621,44	1,9881	72,192
49,0	1,29	2401,00	1,6641	63,210
51,2	1,31	2621,44	1,7161	67,072
55,2	1,45	3047,04	2,1025	80,040
55,3	1,42	3058,09	2,0164	78,526
48,5	1,31	2352,25	1,7161	63,535
52,4	1,44	2745,76	2,0736	75,456
54,8	1,31	3003,04	1,7161	71,788
51,8	1,33	2683,24	1,7689	68,894
59,6	1,74	3552,16	3,0276	103,704
56,8	1,51	3226,24	2,2801	85,768
53,4	1,67	2851,56	2,7889	89,178
54,8	1,39	3003,04	1,9321	76,172

(tab. forts. neste side)

o_i	o_i'	o_i^2	$o_i'^2$	$o_i \cdot o_i'$
51,8	1,49	2683,24	2,2201	77,182
51,8	1,45	2683,24	2,1025	75,110
55,4	1,53	3069,16	2,3409	84,762
51,0	1,24	2601,00	1,5376	63,240
54,6	1,41	2981,16	1,9881	76,986
50,2	1,45	2520,04	2,1025	72,790
61,4	1,87	3769,96	3,4969	114,818
1361,9	36,84	74644,07	54,8838	2019,523

Herav finnes:

$$m(o) = \frac{\Sigma o_i}{n} = \frac{1361,9}{25} = 54,4760$$

$$m(o') = \frac{\Sigma o_i'}{n} = \frac{36,84}{25} = 1,4736$$

$$V(o) = \frac{\Sigma o_i^2 - n.m(o)^2}{n-1} = \frac{74644,07 - 25.54,4760^2}{24} = 18,8841$$

$$s(o) = \sqrt{V(o)} = \sqrt{18,8841} = 4,33$$

$$V(o') = \frac{\Sigma o_i'^2 - n.m(o')^2}{n-1} = \frac{54,8838 - 25.1,4736^2}{24} = 0,0249$$

$$s(o') = \sqrt{V(o')} = \sqrt{0,0249} = 0,16$$

Bruker en også i dette tilfelle betegnelsene Y (for o') og X (for o) for universets betingete gjennomsnitter og x og y som betegnelser for observasjonsverdiene, er regressjonslinjenes ligninger som før

$$Y = ax + b \qquad \text{og} \qquad X = cy + d$$

Konstantene beregnes slik:

$$a = \frac{\Sigma o_i.o_i' - n.m(o).m(o')}{(n-1).s(o)^2} = \frac{2019,523 - 25,54,476,1,4736}{24 \cdot 18,8841} = 0,0279$$

$$b = m(o') - a.m(o) = 1,4736 - 0,0279.54,476 = - 0,0441$$

På samme måte finnes

$$c = 21,13 \qquad \text{og} \qquad d = 23,34$$

De empiriske regressjonslinjenes ligninger er derfor:

$$Y = 0,0279.x - 0,0441$$

$$X = 21,13.y + 23,34$$

Oppgave 11.

Tegn inn disse to regressjonslinjene i samme rettvinklete koordinatsystem sammen med de punktene som representerer observasjonsparene.

Oppgave 12.

I følgende tabell er gitt observasjonene av den gjennomsnittlige årringbredde (o_i) og kvistmengde (o_i') for n=15 grantrær. Arringbredden er angitt i mm og kvistmengden i prosent (kvistsårflatenes samlete areal i prosent av stammens overflate).

Tre nr.	o_i	o_i'
1	2,3	0,19
2	3,0	0,26
3	2,2	0,18
4	2,9	0,53
5	3,5	0,44
6	3,4	0,45
7	3,9	0,71
8	4,1	0,79
9	3,9	0,74
10	3,6	0,69
11	3,6	0,60
12	3,5	0,67
13	4,6	0,96
14	5,6	1,05
15	5,1	0,88

Finn ligningene for de empiriske regressjonslinjene under forutsetning av at de er rette. Tegn figur.

Når observasjonene er store tall, minsker en - som allerede nevnt - regnearbeidet betydelig ved å bruke avvikelsene fra to valte tall p (for o_i) og q (for o_i'). Til beregning av $\Sigma o_i.o_i'$ bruker en ligningen:

$$\Sigma o_i.o_i' = \Sigma(o_i-p)(o_i'-q) + nqm(o) + npm(o') - npq$$

Oppgave 13.

I følgende tabell er o_i observasjoner av brystomfanget og o_i' observasjoner av vekten for n=25 dølaokser. Oksene var ved undersøkelsen mellom 18 og 20 måneder. Brystomfanget er gitt i cm og vekten i kg.

Okse nr.	o_i	o_i'	Okse nr.	o_i	o_i'
1	153	240	14	157	266
2	156	289	15	155	292
3	158	286	16	160	326
4	156	294	17	157	318
5	161	383	18	159	326
6	158	305	19	161	372

(tab. forts. neste side)

(Tab. forts. fra forrige side)

Okse nr.	o_i	o_i'	Okse nr.	o_i	o_i'
7	155	266	20	159	343
8	153	276	21	157	303
9	159	322	22	156	291
10	156	305	23	161	337
11	157	313	24	161	330
12	160	370	25	160	325
13	159	331			

Finn ligningene for de empiriske regressjonslinjene under forutsetning av at de er rette. Velg p=153 og q=240. Tegn figur.

13. K r u m l i n j e t r e g r e s s j o n.

Når regressjonslinjene er krumme, er det alltid meget vanskelig å bestemme seg for hvilken funksjonstype en skal bruke. Det er jo d a så mange funksjoner å velge mellom. I praksis bruker en helst hele rasjonale funksjoner. Men også andre typer blir benyttet. Eksempelvis kan nevnes funksjonen

$$Y = a.\log x + b$$

hvor Y altså er en lineær funksjon av log. x. I dette tilfelle skal a og b bestemmes slik som vi har lært ovenfor.

Vi skal her nøye oss med å vise hvordan konstantene i en hel funksjon av 2. grad skal bestemmes. Vi setter altså:

$$Y = ax^2 + bx + c$$

Omfatter observasjonsrekken så få parobservasjoner at en ikke samordner dem i en korrelasjonstabell, men beholder dem i primærlisten, skal a,b og c etter minste kvadraters metode beregnes ved hjelp av følgende ligninger:

$$(\Sigma o_i^4)\ a + (\Sigma o_i^3)\ b + (\Sigma o_i^2)\ c = \Sigma o_i^2 o_i'$$

$$(\Sigma o_i^3)\ a + (\Sigma o_i^2)\ b + (\Sigma o_i)\ c = \Sigma o_i o_i'$$

$$(\Sigma o_i^2)\ a + (\Sigma o_i)\ b + n.c = \Sigma o_i'$$

Velges også for den andre empiriske regressjonslinjen en hel funksjon av 2. grad, skal konstantene i denne bestemmes av det sett ligninger som en får ved i de tre refererte ligninger å bytte om o_i og o_i'.

Eksempel 16.

Dette eks. er hentet fra en undersøkelse av planteavstandens innflytelse på forskjellige egenskaper ved grantrær. o_i er de brukte planteav-

stander (i meter) og o_i' er prosent kvist (sml. oppgave 12).

o_i	o_i'	o_i^2	o_i^3	o_i^4	$o_i o_i'$	$o_i^2 o_i'$
1,25	0,19	1,5625	1,9531	2,4414	0,2375	0,2969
1,25	0,26	"	"	"	0,3250	0,4063
1,25	0,18	"	"	"	0,2250	0,2813
1,50	0,53	2,2500	3,3750	5,0625	0,7950	1,1925
1,50	0,44	"	"	"	0,6600	0,9900
1,50	0,45	"	"	"	0,6750	1,0125
2,00	0,71	4,0000	8,0000	16,0000	1,4200	2,8400
2,00	0,79	"	"	"	1,5800	3,1600
2,00	0,74	"	"	"	1,4800	2,9600
2,50	0,69	6,2500	15,6250	39,0625	1,7250	4,3125
2,50	0,60	"	"	"	1,5000	3,7500
2,50	0,67	"	"	"	1,6750	4,1875
3,50	0,96	12,2500	42,8750	150,0625	3,3600	11,7600
3,50	1,05	"	"	"	3,6750	12,8625
3,50	0,88	"	"	"	3,0800	10,7800
32,25	9,14	78,9375	215,4843	637,8867	22,4125	60,7920

Disse summene innsettes i de tre ligningene, og ved å løse disse finnes:

$$a = -0,1107, \qquad b = 0,8153 \qquad \text{og} \qquad c = -0,5610$$

Hvis nå ligningen for regressjonslinjen for % kvist m.h.p. planteavstanden i universet er en hel funksjon av 2. grad, er

$$Y = -0,1107\,x^2 + 0,8153\,x - 0,5610$$

et tilnærmet riktig uttrykk for denne.

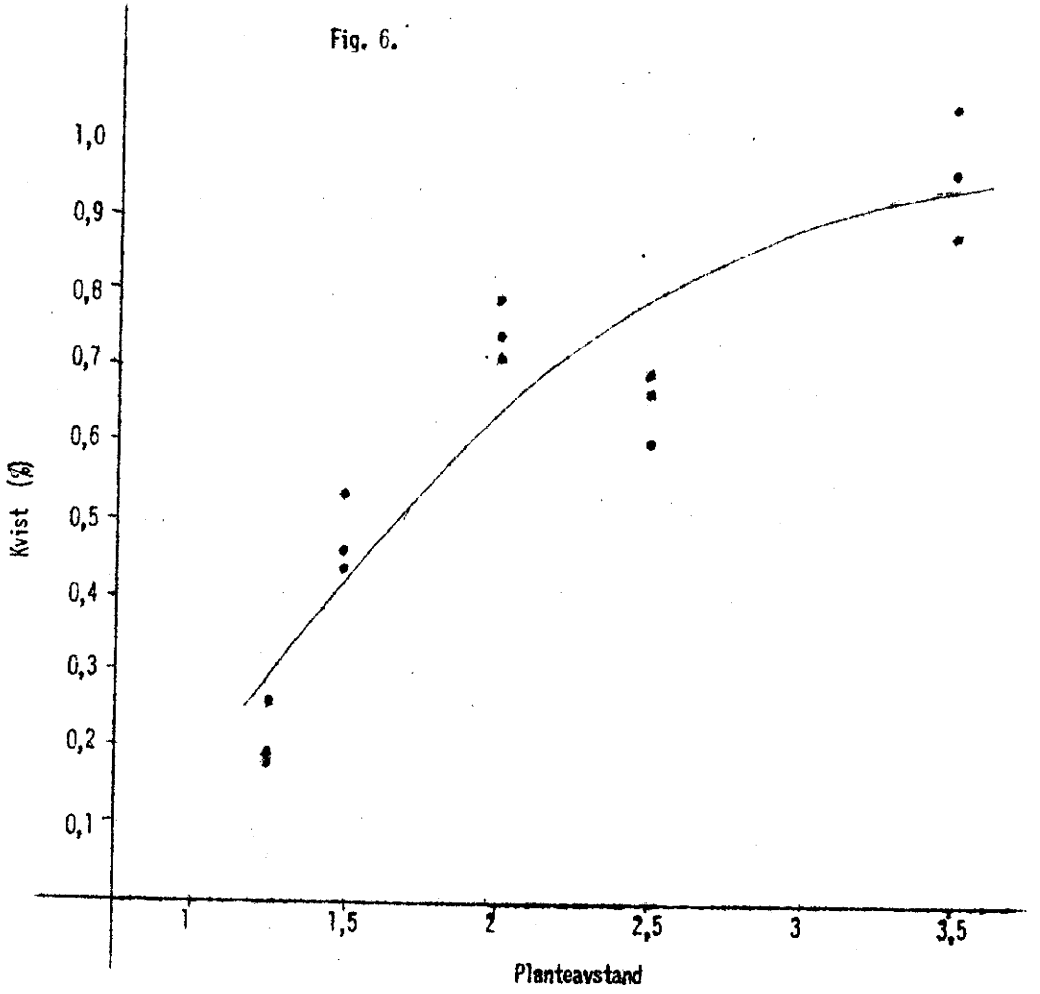
I figur 6 (neste side) er kurven for denne funksjonen inntegnet sammen med de punktene som representerer observasjonsparene, dvs. punkter hvis abscisser er o_i og hvis ordinater er o_i'. Vi ser at kurven gir en bra, men tilsynelatende ikke helt god, beskrivelse av observasjonene som helhet.

Bestemmelsen av konstantene i hele funksjoner av høyere grad enn 2 etter minste kvadraters metode byr ikke på noen prinsipielle vanskeligheter. Når en bruker en hel funksjon av 3. grad, må en for å bestemme konstantene løse en lineært ligningssett med 4 ukjente. For å løse et slikt ligningssett må en helst bruke en spesiell teknikk som vi ikke har anledning til å forklare her. Dessuten blir de tallene en må operere med, i regelen så pass store at en bør ha anledning til å bruke regnemaskiner.

Oppgave 14.

Karrforsøk med stigende mengder kvelstoffgjødsel til havre. For hvert trinn kvelstoffmengde 4 paralleller (karr). o_i = tilsatt kvelstoff, angitt i 25 mg N pr. karr. o_i' = tørrstoffavlingen (alle overjordiske plantedeler) angitt i gram pr. karr.

Fig. 6.



karr nr.	o_i	o_i'
1	1	5,9
2	1	5,4
3	1	6,1
4	1	5,7
5	5	15,2
6	5	16,4
7	5	16,0
8	5	16,9
9	10	30,6
10	10	29,8
11	10	29,6
12	10	29,2
13	20	45,1
14	20	42,9
15	20	43,0
16	20	42,8
17	40	45,4
18	40	46,9
19	40	44,6
20	40	47,8

Bestem ligningen for den empiriske regressjonslinjen for o' m.h.p. o. Bruk en hel funksjon av 2. grad. Tegn figur.

14. E s t i m e r i n g s l i n j e n.

Hvis det i universet til enhver verdi (x) av det ene kjennetegnet svarer bare en verdi (y) av det annet, er verdien av det ene kjennetegnet en funksjon av verdien av det annet, dvs. at y er en funksjon av x eller x en funksjon av y. Observerer en de to kjennetegnene i et utvalg av dette universet og observasjonene ikke er beheftet med observasjonsfeil, vil alle de punktene som i et rettvinklet koordinatsystem har observasjonene av det ene kjennetegn til abscisse og observasjonen av det annet kjennetegn til ordinat, ligge på den kurven som er det grafiske bilde av funksjonsforholdet mellom x og y. Kjenner en derfor verdien av det ene kjennetegnet, kan en beregne verdien av det annet.

Slike funksjonsforhold mellom to kjennetegns verdier eksisterer aldri når det gjelder erfaringsdata. I alle aktuelle tilfelle vil det til en verdi av det ene kjennetegnet svare mange verdier av det annet (sml. eks. 13 og 14).

Sett at en kjenner regressjonslinjen, $Y = f(x)$, i universet. Den til $x=x_i$ svarende funksjonsverdi, $Y_i = f(x_i)$, er da det aritmetiske gjennomsnitt for y i den del av universet (subuniverset) hvor $x=x_i$; eller, Y_i er det til x_i svarende betingete gjennomsnitt for y. Setter en nå y lik denne gjennomsnittsverdi, er dette aldri eksakt riktig, og den feilen en gjør ved å sette y lik gjennomsnittsverdien, er naturligvis desto større jo større variasjon det er innen vedkommende subunivers.

Vi har i praksis meget ofte bruk for å kjenne verdien av et kjennetegn som er slik at direkte observasjon er meget vanskelig. Kubikkmassen av et tre er jo - for å nevne et eksempel - ikke lett å observere direkte. En prøver i slike tilfelle å bestemme eller estimere dette kjennetegnet ved hjelp av andre lett observerbare kjennetegn. Kubikkmassen av et tre kan vi prøve å estimere ved hjelp av f. eks. trehøyden. Det er imidlertid klart at kubikkmassen ikke er en funksjon av høyden. Det ville den være bare hvis alle trærne innen det univers det er tale om, hadde akkurat samme stammeformen. Men hvis en kjenner regressjonslinjen for kubikkmassen m.h.p. høyden og variasjonen omkring denne linjen ikke er altfor stor, kan en estimere kubikkmassen for hver enkelt tre til gjennomsnittsverdien, Y_i, ved i ligningen for regressjonslinjen å sette inn observasjonen av trehøyden. Nøyaktigheten av en slik estimering beror selvsagt på hvor stor variasjon omkring regressjonslinjen (eller estimeringslinjen) det er. Det eksemplet vi har nevnt, er derfor sikkert ikke heldig valt (unntagen som illustrasjon). I regelen er stammeformen så pass varierende fra tre til tre at regressjonslinjen for kubikkmassen m.h.p. trehøyden i ke er et brukbart grunnlag for kuberingen.

Når regressjonslinjene brukes som estimeringslinjer, er det i hvert enkelt tilfelle bare den ene av de to regressjonslinjene som har interesse, og en må passe på at en velger den riktige. Skal en estimere kubikkmassen ved hjelp av høyden, er det naturligvis regressjonslinjen for kubikkmassen m.h.p. høyden (altså med høyden som uavhengig variabel) som skal brukes.

Da universets regressjonslinjer er ukjente i alle de tilfelle der det kan bli tale om å bruke dem til estimering, må en i praksis nøye seg med de empiriske regressjonslinjene. Har en imidlertid vært heldig med valget av funksjon, er den empiriske regressjonslinjen et tilnærmet riktig uttrykk for den tilsvarende regressjonslinjen i universet, men riktignok bare for regressjonslinjen i det universet som de benyttete observasjonene er et utvalg av. En må derfor ikke bruke den empiriske regressjonslinjen som estimeringslinje for enheter som ikke tilhører dette universet. Tenker vi oss f. eks. at regressjonslinjen for kubikkmassen m.h.p. trehøyden kan brukes til estimering av kubikkmassen og at den empiriske regressjonslinjen er bestemt på grunnlag av et utvalg av gran fra Østfold, må en være forsiktig med å bruke den utenfor dette område. Det kan jo meget godt tenkes at denne regressjonslinjen er forskjellig for gran fra Østfold og gran fra f. eks. Telemark.

15. K o r r e l a s j o n o g a v h e n g i g h e t.

Regressjonslinjen for y m.h.p. x gir en beskrivelse av hvordan det betingete gjennomsnitt for y endrer seg med x. Observasjonene har imidlertid som vi har sett, andre karakteristikker; middelavvik, typetall osv. I en undersøkelse av hvordan y avhenger av x burte vi derfor også undersøke om det betingete middelavvik for y forandrer seg med x og om det betingete typetallet for y forandrer seg med x og i tilfelle hvordan disse endringer kan beskrives. Det er umiddelbart innlysende at hvis bare en betinget karakteristikk for y endrer verdi med x, må det eksistere et eller annet slags avhengighetsforhold mellom x og y. Teoretisk sett kan en lett tenke seg at det betingete middelavvik for y er avhengig av x selv om det betingete gjennomsnitt for y er uavhengig av x. Men avhengighetsforhold av dette slag forekommer iallfall ytterst sjelden. Det viktigste i ethvert tilfelle er derfor regressjonsundersøkelsen. Men en skal derfor ikke se bort fra de andre karakteristikker hvor materialet er slik at en videregående undersøkelse er mulig.

Det mål en i praksis i første rekke stiller seg, er å få rede på regressjonen, og denne praktiske problemstillingen har vært så dominerende at den har gitt grunnlaget for selve definisjonen av korrelasjon. Hvis det be-

tingete gjennomsnitt for det ene kjennetegn er uavhengig av verdien av det annet, altså når det betingete gjennomsnitt for y er uavhengig av x og det betingete gjennomsnitt for x er uavhengig av y, sier en at de to kjennetegnene er ukorrelerte. Hvis det betingete gjennomsnitt for den ene kjennetegnet er avhengig av verdien av det annet, sier en at de to kjennetegnene er korrelerte.

Hvis en bygger korrelasjonslæren på denne definisjon, som synes å ha fått alminnelig tilslutning, må en altså være klar over at korrelasjon og avhengighet ikke er identiske begreper. To kjennetegn kan teoretisk sett være avhengige av hverandre selv om de ikke'er korrelerte.

16. Korrelasjonsforholdet.

I tillegg til beskrivelsen av regressjonen har vi bruk for et mål for korrelasjonsgraden. Det er kanskje lettest å forstå hva en mener med dette uttrykk når en tenker gjennom saken geometrisk. La oss tenke oss at parverdiene av to kjennetegn for universet er fremstilt ved en punktsverm og la oss tenke oss at regressjonslinjen for y m.h.p. x er trukket opp. Punktene i punktsvermen vil da ligge mer eller mindre spredt omkring denne regressjonslinjen. Og jo mindre spredning punktsvermen har, jo nærmere er korrelasjonen det idealtilfelle der y er en funksjon av x. I dette tilfelle vil hele punktsvermen være plasert på regressjonslinjen slik at det ikke er noen spredning eller variasjon omkring denne. Når vi taler om korrelasjonsgraden, mener vi graden av tilnærmelse til dette idealbilde.

Nå er det jo klart at spredningen omkring regressjonslinjen må være desto større jo større de betingete middelavvik for y er. Som mål for hvor stor spredningen er, kunne vi derfor bruke et gjennomsnitt av de betingete middelavvik for y. I det grensetilfelle at y er en funksjon av x, er alle disse betingete middelavvik lik 0 og gjennomsnittet av dem også lik 0.

Denne problemstilling har ledet til innføringen i praktisk statistikk av to størrelser som kalles korrelasjonsforholdet og korrelasjonsindeksen. Vi skal foreløpig beskjeftige oss med den første.

La oss tenke oss at det foreligger et utvalg av parobservasjoner av to kjennetegn, og la oss tenke oss at disse er ordnet i en korrelasjonstabell. Frekvensen svarende til $x=x_i$ og $y=y_j$ vil vi som før betegne med h_{ij}, og den til $y=y_j$ svarende frekvens i den marginale fordelingsrekken for y vil vi betegne med h_j, dvs. $h_j = \sum_i h_{ij}$. Videre vil vi bruke $m(y/x_i)$ som betegnelse for det til $x=x_i$ svarende betingete gjennomsnitt for y. Gjennomsnittet for y i den marginale fordelingsrekken vil vi betegne med $m(y)$. Korrela-

sjonsforholdet for y m.h.p. x (η = eta) er da:

$$\eta_{yx}^2 = 1 - \frac{\Sigma\Sigma h_{ij}\,[y_j - m(y/x_i)]^2}{\Sigma h_j\,[y_j - m(y)]^2}$$

Det er lett å innse at

$$0 \leq \Sigma\Sigma h_{ij}\,[y_j - m(y/x_i)]^2 \leq \Sigma\, h_j\,[y_j - m(y)]^2$$

Summene $\Sigma_j h_{ij}[y_j - m(y/x_i)]^2$ er jo de kvadratsummene vi bruker til beregning av de betingete middelavvik for y. I det grensetilfelle der y er en funksjon av x, er alle disse kvadratsummene lik 0 og summen av dem lik 0. Det annet grensetilfelle

$$\Sigma\,\Sigma h_{ij}[y_j - m(y/x_i)]^2 = \Sigma h_j[y_j - m(y)]^2$$

inntreffer når alle de betingete gjennomsnittene for y er like og lik det marginale gjennomsnittet. Dette vil inntreffe når regressjonslinjen for y m.h.p. x er parallell med x-aksen og det altså ikke er korrelasjon mellom kjennetegnene. I et endelig materiale (utvalg) blir disse betingete gjennomsnittene praktisk talt aldri nøyaktig like. Men tillater vi oss likevel å sette $m(y/x_i) = m(y)$, finner vi:

$$\Sigma\Sigma h_{ij}[y_j - m(y/x_i)]^2 = \Sigma\Sigma h_{ij}[y_j - m(y)]^2 =$$

$$= \sum_j [y_j - m(y)]^2 \sum_i h_{ij} = \Sigma[y_j - m(y)]^2 \cdot h_j$$

Vi innser derfor at korrelasjonsforholdet har en tallverdi mellom 0 og 1. Det er lik 1 når $\Sigma\Sigma h_{ij}[y_j - m(y/x_i)]^2 = 0$, dvs. når y er en funksjon av x. Dette svarer til korrelasjonsmaksimum. Det er lik 0 når $m(y/x_i) = m(y)$, dvs. når det ingen korrelasjon er. Korrelasjonsforholdet er derfor et rasjonelt mål for korrelasjonsgraden.

Å beregne korrelasjonsforholdet etter den formelen vi har gitt, er meget arbeidskrevende. Skulle vi bruke denne formelen, måtte vi beregne summen $\Sigma_j h_{ij}[y_j - m(y/x_i)]^2$ for hver enkelt betinget fordelingsrekke for y. Og når det er et stort antall av disse fordelingsrekkene, slik som i eks. 13 og 14, blir regnearbeidet nokså omfattende. Riktignok vil en ved å bruke denne fremgangsmåten, også få de kvadratsummene som eventuelt må brukes til beregningen av de betingete middelavvik for y og dermed disse middelavvik som omtrent gratis biprodukter. Men gjelder det bare å beregne korrelasjonsforhol-

det, er det en annen fremgangsmåte som fører fortere fram. Det kan nemlig bevises at

$$\eta^2_{yx} = \frac{\Sigma h_i[m(y/x_i) - m(y)]^2}{\Sigma h_j[y_j - m(y)]^2}$$

hvor h_i er frekvensen til $x=x_i$ i den marginale fordelingsrekken for x, altså $h_i = \Sigma_j h_{ij}$.

Nevneren i denne formelen er den kvadratsummen vi har bruk for til beregning av det marginale middelavvik for y. Denne regnes derfor ut etter en av de metodene vi har lært før (se II,4 og 5). For eks. 13 er denne summen lik 5988,8795. I telleren inngår det marginale gjennomsnitt for y. Dette er for eks. 13 lik 3,54. Hvordan utregningen av telleren skal gjøres er vist for det samme eksempel i følgende tabell. De betingete gjennomsnittene for y, m(y/x), er gitt tidligere (se II,11).

x_i	$m(y/x_i)$	$m(y/x_i)-m(y)$	$[m(y/x_i)-m(y)]^2$	h_i	$h_i[m(y/x_i)-m(y)]^2$
0	0,60	- 2,94	8,6436	15	129,6540
1	1,36	- 2,18	4,7524	225	1069,2900
2	2,31	- 1,23	1,5129	353	534,0537
3	3,20	- 0,34	0,1156	437	50,5172
4	3,89	0,35	0,1225	411	50,3475
5	4,78	1,24	1,5376	297	456,6672
6	5,51	1,97	3,8809	155	601,5395
7	6,14	2,60	6,5600	78	527,2800
8	6,50	2,96	8,7616	16	140,1856
9	7,33	3,79	14,3641	12	172,3692
10	9,00	5,46	29,8116	1	29,8116
				$\Sigma h_i[m(y/x_i)-m(y)]^2$	= 3761,7155

Følgelig er

$$\eta^2_{yx} = \frac{3761,7155}{5988,8795} = 0,6281$$

og

$$\eta_{yx} = \sqrt{0,6281} = \underline{0,79}$$

For å kunne beregne korrelasjonsforholdet må vi først ha beregnet de betingete gjennomsnittene. Disse kan i alminnelighet ikke beregnes med mindre observasjonene er ordnet i en korrelasjonstabell. Det er imidlertid en unntagelse fra denne regelen. Når observasjonene skaffes til veie ved forsøk, blir som oftest det kjennetegnet hvis virkning en vil undersøke, variert trinnvis. Og til hvert trin svarer det i regelen flere observasjoner av det annet kjennetegn. I eks. 16 er planteavstanden variert i 5 trin og til hvert trin svarer 3 observasjoner av kvistprosenten. I dette tilfelle

kan vi naturligvis beregne de betingete gjennomsnittene for % kvist og dermed også korrelasjonsforholdet for % kvist m.h.p. planteavstanden. Det marginale gjennomsnitt for % kvist er i dette tilfelle lik gjennomsnittet av alle n=15 observasjonene av % kvist. Det er $m(o') = 0,61$. Forresten skal beregningene utføres på samme måte som i forrige eksempel etter følgende skjema.

o_i	$m(o'/o_i)$	$[m(o'/o_i)-m(o')]$	$[m(o'/o_i)-m(o')]^2$	h_i	$h_i[m(o'/o_i)-m(o')]^2$
1,25	0,21	- 0,40	0,1600	3	0,4800
1,50	0,47	- 0,14	0,0196	3	0,0588
2,00	0,75	0,14	0,0196	3	0,0588
2,50	0,65	0,04	0,0016	3	0,0048
3,50	0,96	0,35	0,1225	3	0,3675
					0,9699

Summen $\Sigma(o'-m(o')^2$ regnes ut på vanlig måte. Den er 1,0031.

Altså er

$$\eta^2_{o'o} = \frac{0,9699}{1,0031} = 0,9669$$

og

$$\eta_{o'o} = \sqrt{0,9669} = 0,98$$

Oppgave 15.

Beregn korrelasjonsforholdet for fettinnholdet m.h.p. vekten for eks. 14.

Beregn også η for tørrstoffavlingen m.h.p. kvelstoffmengden for observasjonene i oppgave 14.

17. K o r r e l a s j o n s i n d e k s e n.

Korrelasjonsindeksen er dannet på samme måte som korrelasjonsforholdet. Vi betegner den med R. Korrelasjonsindeksen for y m.h.p. x er da:

$$R^2_{yx} = 1 - \frac{\Sigma\Sigma h_{ij}[y_j-Y_i]^2}{\Sigma h_j[y_j-m(y)]^2}$$

Sammenligner vi de to formlene for R og η, ser vi at forskjellen er den at vi i formelen for η har byttet ut $m(y/x_i)$ = de betingete gjennomsnitter for y med Y_i = ordinatverdiene til regressjonslinjen. Med regressjonslinjen forstår vi da den som er bestemt på grunnlag av observasjonene, altså den empiriske.

Er observasjonene ordnet i en korrelasjonstabell, er den gitte formelen lite egnet for beregning av R. Den er meget arbeidskrevende. Skulle vi

bruke den, måtte vi for hver betinget fordelingsrekke for y beregne summen $\Sigma_i h_{ij}[y_j - Y_i]^2$. I praksis bør en derfor bruke en annen fremgangsmåte. Det kan bevises at

$$R_{yx}^2 = \eta_{yx}^2 - \frac{\Sigma h_i[m(y/x_i) - Y_i]^2}{\Sigma h_j[y_j - m(y)]^2}$$

Er altså η beregnet, gjenstår det bare å beregne $\Sigma h_i[m(y/x_i) - Y_i]^2$.

Nevneren $\Sigma h_j[y_j - m(y)]^2$ er jo nødvendig også for beregningen av η.
For eks. 13 har vi

$$\eta_{yx}^2 = 0{,}6281$$

og

$$\Sigma\, h_j[y_j - m(y)]^2 = 5988{,}8795$$

Verdiene av Y_i finnes ved etter hvert å innsette x=0,1,2,3,....10 i regressjonsligningen

$$Y = 0{,}80x + 0{,}72$$

De videre beregningene vil fremgå av følgende tabell.

x_i	$m(y/x_i)$	Y_i	$m(y/x_i) - Y_i$	$[m(y/x_i) - Y_i]^2$	h_i	$h_i[m(y/x_i) - Y_i]^2$
0	0,60	0,72	- 0,12	0,0144	15	0,2160
1	1,36	1,52	- 0,16	0,0256	225	5,7600
2	2,31	2,32	- 0,01	0,0001	353	0,0353
3	3,20	3,12	0,08	0,0064	437	2,7968
4	3,89	3,92	0,03	0,0009	411	0,3699
5	4,78	4,72	0,06	0,0036	297	1,0692
6	5,51	5,52	- 0,01	0,0001	155	0,0155
7	6,14	6,32	- 0,18	0,0324	78	2,5272
8	6,50	7,12	- 0,62	0,3844	16	6,1504
9	7,33	7,92	- 0,59	0,3481	12	4,1772
10	9,00	8,72	0,28	0,0784	1	0,0784
					$\Sigma h_i[m(y/x_i) - Y_i]^2$	= 23,1959

Altså er:

$$R_{yx}^2 = 0{,}6281 - \frac{23{,}1959}{5988{,}8795} = 0{,}6281 - 0{,}0039 = 0{,}6242$$

og

$$R_{yx} = \sqrt{0{,}6242} = \underline{0{,}79}$$

Oppgave 16.

Beregn korrelasjonsindeksen for eks. 14 og 16. En bruker de ligningene for regressjonslinjene som allerede er funnet.

Korrelasjonsindeksen kan naturligvis beregnes også når observasjonene ikke er ordnet i korrelasjonstabell. Er observasjonsparene (o_i, o_i'), skal

korrelasjonsindeksen for o_i' m.h.p. o_i regnes ut etter formelen

$$R^2_{o'o} = 1 - \frac{\Sigma(o_i' - Y_i)^2}{\Sigma(o_i' - m(o'))^2}$$

For eks. 12 (se eks. 15) har vi bestemt en rettlinjet regressjonslinje for o_i' m.h.p. o_i. Ligningen for denne er

$$Y = 0{,}0278x - 0{,}0441$$

Y_i finnes ved etter hvert å innsette o_i for x i denne ligningen. Fra eks. 15 har vi at

$$\Sigma (o_i' - m(o'))^2 = V(o').(n-1) = 0{,}0249.24 = 0{,}5976$$

De videre beregninger er vist i følgende tabell:

o_i'	Y_i	$o_i' - Y_i$	$(o_i' - Y_i)^2$
1,71	1,79	- 0,08	0,0064
1,57	1,69	- 0,12	0,0144
1,66	1,59	0,07	0,0049
1,52	1,44	0,08	0,0064
1,36	1,38	- 0,02	0,0004
1,41	1,38	0,03	0,0009
1,29	1,32	- 0,03	0,0009
1,31	1,38	- 0,07	0,0049
1,45	1,49	- 0,04	0,0016
1,42	1,49	- 0,07	0,0049
1,31	1,30	0,01	0,0001
1,44	1,41	0,03	0,0009
1,31	1,48	- 0,17	0,0289
1,33	1,40	- 0,07	0,0049
1,74	1,61	0,13	0,0169
1,51	1,53	- 0,02	0,0004
1,67	1,44	0,23	0,0529
1,39	1,48	- 0,09	0,0081
1,49	1,40	0,09	0,0081
1,45	1,40	0,05	0,0025
1,53	1,50	0,03	0,0009
1,24	1,37	- 0,13	0,0169
1,41	1,47	- 0,06	0,0036
1,45	1,35	0,10	0,0100
1,87	1,66	0,21	0,0441
		$\Sigma (o_i' - Y_i)^2$	= 0,2449

Altså er $R^2_{o'o} = 1 - \frac{0{,}2449}{0{,}5976} = 1 - 0{,}4098 = 0{,}5902$

og $R_{o'o} = \sqrt{0{,}5902} = \underline{0{,}77}$

Oppgave 17.

Beregn korrelasjonsindeksen for de eksemplene som er gitt i oppgavene 12, 13 og 14. Bruk de ligningene for regressjonslinjene som allerede er funnet.

Beregn også korrelasjonsindeksen for vekten m.h.p. kvelstoffinnholdet for eks. 12 (se eks. 15).

Av ligningen

$$\eta^2_{yx} - R^2_{yx} = \frac{\Sigma h_i [m(y/x_i)-Y_i]^2}{\Sigma h_j [y_j - m(y)]^2}$$

vil det sees at R er mindre eller i høyden lik η. Likheten inntreffer når alle punktene som representerer de betingete gjennomsnittene for y ligger på den empiriske regressjonslinjen. Vi ser videre at forskjellen mellom korrelasjonsforholdet og korrelasjonsindeksen beror på hvor meget de betingete gjennomsnittene for y avviker fra regressjonsverdiene, altså på størrelsen av differensene $m(y/x_i)-Y_i$.

Uansett hvilken metode en bruker til bestemmelsen av regressjonslinjen, er målet å bestemme den slik at når antallet av observasjoner økes over alle grenser, vil den nærme seg til å falle sammen med universets tilsvarende regressjonslinje. I universet vil imidlertid regressjonslinjen for y m.h.p. x passere gjennom alle de punktene hvis abscisser er x og hvis ordinater er de tilsvarende betingete gjennomsnittene for y. For universet er derfor korrelasjonsindeksen og korrelasjonsforholdet like.

Det er derfor klart at forutsetningen for at η og R skal nærme seg til å bli like når antallet av observasjoner økes over alle grenser, er at den empiriske regressjonslinjen har fått en slik form at den vil nærme seg til å falle sammen med universets tilsvarende regressjonslinje. Hvis f. eks. universets regressjon er krumlinjet og en for utvalgets regressjon har valt en lineær funksjon, $Y = ax+b$, er det klart at R ikke vil nærme seg til å bli lik η når antallet av observasjoner økes.

Når antallet av observasjoner er stort, og korrelasjonsindeksen og korrelasjonsforholdet er tilnærmet like, ligger det nær å oppfatte dette som tegn på at en har vært heldig i valget av funksjon for den empiriske regressjonslinjen. For eks. 13 har vi funnet at når tallene avrundes til to desimaler, er både R og η lik 0,79. Siden vi i dette tilfelle har et meget stort antall observasjoner ($n=2000$), kan dette oppfattes som et tegn på at regressjonslinjen for y m.h.p. x er rett. En må imidlertid være ytterst forsiktig med å tillegge likheten mellom R og η for stor beviskraft. Blant annet må en være oppmerksom på at hvis en velger hele rasjonale funksjoner, vil en

ved å øke graden av funksjonen minske forskjellen mellom R og η uten at det ligger noen dypere betydning i det. Saken er nemlig den at en ved å øke graden av en hel rasjonal funksjon gjør den mer og mer elastisk. Er det f. eks. r betingete gjennomsnitter for y, kan en bestemme koeffisientene i en hel rasjonal funksjon av grad (r-1) slik at dens kurve passerer gjennom alle de r punktene som representerer gjennomsnittene. Da blir jo forskjellen mellom korrelasjonsindeksen og korrelasjonsforholdet automatisk lik 0 også for et endelig materiale (utvalg).

På den annen side må en også være forsiktig med å tillegge en forskjell mellom korrelasjonsforholdet og korrelasjonsindeksen for stor betydning. For eks. 16 er $\eta = 0{,}98$ og $R = 0{,}92$. Dette må imidlertid ikke oppfattes som noe bevis på at den empiriske regressjonslinjen ikke har fått en heldig form.

Vi kommer senere tilbake til spørsmål av denne art i en annen sammenheng.

18. Korrelasjonskoeffisienten.

Vi har allerede sett at hvis det ikke er korrelasjons mellom to kjennetegn, vil regressjonslinjene være parallelle med koordinataksene; regressjonslinjen for y m.h.p. x vil være parallell med x-aksen og regressjonslinjen for x m.h.p. y parallell med y-aksen. Er det funksjonell sammenheng mellom x og y, vil de to regressjonslinjene falle sammen.

Vi vil nå forutsette at regressjonslinjene (altså for universet) er rette. Hvis det altså ikke er noen korrelasjon, vil de to regressjonslinjene danne en vinkel med hverandre på 90°, og hvis det er funksjonell sammenheng, vil vinkelen mellom regressjonslinjene være lik 0. I alle andre tilfelle vil vinkelen mellom regressjonslinjene (vi tenker her på den minste vinkelen) ligge mellom 0 og 90° og den vil være desto mindre jo sterkere korrelasjonen er. Når regressjonslinjene er rette, skulle det derfor være rasjonelt å bruke vinkelen mellom dem eller en funksjon av denne vinkelen som mål for korrelasjonsgraden. Vi skal se at vi kan bruke en størrelse som kalles korrelasjonskoeffisienten. Vi vil betegne den med r.

Korrelasjonskoeffisienten kan defineres på flere forskjellige måter. Her er det naturlig å definere den ved formelen

$$r = \frac{\Sigma\Sigma h_{ij} x_i y_j - n \cdot m(x) m(y)}{(n-1)\, s(x) s(y)}$$

Når r defineres på denne måten, er den naturligvis en empirisk størrelse.

Når vi bruker r, kan konstantene a og c i ligningene for de to empiriske regressjonslinjene (se avsnitt 12) skrives på kortere og enklere måte. Vi finner lett at

$$a = r\,\frac{s(y)}{s(x)} \qquad \text{og} \qquad c = r\,\frac{s(x)}{s(y)}$$

Når vi kjenner vinkelkoeffisientene, k og k', for to rette linjer, kan vi beregne vinkelen mellom dem ved formelen

$$\operatorname{tg} v = \frac{k - k'}{1 + k.k'}$$

Her er k lik tangens til den vinkelen som den ene linjen danner med den positive absisseaksen, og k' tangens til den vinkelen den andre linjen danner med samme akse. I regressjonslinjen $Y = ax + b$ er a lik tangens til den vinkelen som denne regressjonslinjen danner med x-aksen, mens c i ligningen, $X = cy+d$, for den andre regressjonslinjen er lik tangens til den vinkelen denne danner med y-aksen. Skal vi derfor bruke a og c til beregning av vinkelen mellom regressjonslinjene, må vi sette enten $k=a$ og $k'=1/c$ eller $k=1/a$ og $k'=c$. Det er likegyldig hvilket av disse alternativene vi bruker. Velger vi det første, blir tangens til vinkelen mellom regressjonslinjene lik

$$\operatorname{tg} v = \frac{k-k'}{1+kk'} = \frac{a - \frac{1}{c}}{1 + a\frac{1}{c}} = \frac{ac - 1}{a + c}$$

Innsettes her $a = r\,\frac{s(y)}{s(x)}$ og $c = r\,\frac{s(x)}{s(y)}$, finner vi lett at

$$\operatorname{tg} v = \frac{1-r^2}{r} \qquad \frac{s(x).s(y)}{s(x)^2+s(y)^2}$$

Ved vinkelen v vil vi forstå den minste vinkelen (mellom 0 og 90^0) som regressjonslinjene danner med hverandre, og vi vil regne denne vinkelen som positiv. Det er derfor bare tallverdien av tg v vi behøver å diskutere.

Om r kan det bevises (beviset tas ikke med her) at den er en størrelse hvis tallverdi ikke kan bli større enn enheten. Den kan være positiv eller negativ, positiv når korrelasjonen er positiv og negativ når korrelasjonen er negativ.

Av formelen for tg v ser vi at når $r=+1$ eller $r=-1$, er $\operatorname{tg} v = 0$ og derfor $v=0$. I dette tilfelle faller altså regressjonslinjene sammen. Videre ser vi at når tallverdien av r nærmer seg 0, vil tg v vokse over alle grenser, og det betyr at v selv nærmer seg 90^0. Samtidig vil både a og c nærme seg 0. Regressjonslinjene blir altså parallelle med koordinataksene.

Når tallverdien av r avtar fra 1 til 0, vil tg v vokse fra 0 til et uendelig stort tall, og dette betyr igjen at v selv vil vokse fra 0 til 90°.

Vi kan på grunnlag av dette stille opp følgende regler:

1) Når r=0, er det ingen korrelasjon mellom kjennetegnene.
2) Når r=+1 eller r=-1, er det funksjonell sammenheng.
3) Når r har en tallverdi mellom -1 og +1 og forskjellig fra 0, er det positiv (når r positiv) eller negativ (når r negativ) korrelasjon mellom kjennetegnene og denne korrelasjonen er desto sterkere jo større tallverdi r har.

Strengt tatt gjelder disse reglene bare for universet, og de gjelder bare når regressjonslinjene er rette.

Korrelasjonskoeffisienten kan naturligvis beregnes ved hjelp av den formelen for den som er gitt ovenfor. Er imidlertid a og c beregnet, kan r beregnes ved hjelp av disse. Vi finner lett at

$$a.c = r^2$$

eller

$$r = \pm \sqrt{a.c}$$

Vi ser av formlene for a, c og r at disse tre størrelsene har samme fortegn. Produktet a.c er derfor alltid positivt. Og siden r har samme fortegn som a og c, skal vi bruke + foran rottegnet når a og c er positive og ÷ når a og c er negative.

For eks. 13 har vi funnet a = +0,80 og c = +0,79. Herav finnes

$$r = +\sqrt{0,80.0,79} = +0,79$$

Det kan bevises at når regressjonslinjene er rette, er korrelasjonskoeffisienten og korrelasjonsforholdet for universet like. En oppfatter derfor gjerne tilnærmet likhet mellom den empiriske korrelasjonskoeffisienten og det empiriske korrelasjonsforholdet som tegn på at regressjonen er rettlinjet. For eks. 13 er både korrelasjonskoeffisienten og korrelasjonsforholdet for y m.h.p. x lik 0,79 (avrundet til 2 desimaler). Samme verdi har forresten også korrelasjonsindeksen for y m.h.p. x.

Oppgave 18.

Beregn korrelasjonskoeffisienten for eks. 12, eks. 14, og for eksemplene i oppgavene 12 og 13.

19. Korrelasjon mellom flere kjennetegn.

Vi har hittil utelukkende beskjeftiget oss med korrelasjonen mellom to kjennetegn. Undersøkelser av korrelasjonen mellom flere kjennetegn samtidig har imidlertid i den senere tid kommet til å spille en større og større rolle i de fleste større empirisk statistiske arbeider, særlig kanskje i undersøkelser av økonomisk art. Men også i landbrukets forskjellige forsøksvirksomheter er undersøkelsene av korrelasjonen mellom flere kjennetegn blitt noe av en hovedsak. Teorien for slike undersøkelser er nokså innviklet, og det vi skal ta med her, er derfor å betrakte bare som en første elementære innføring og en antydning av problemene.

Vi har sett at hvis det i et univers er korrelasjon mellom to kjennetegn, vil variasjonen i verdiene av det ene i forhold til den totale variasjon bli redusert når en spalter opp universet i subuniverser etter verdiene av det annet kjennetegn. Dette gir seg utslag i at korrelasjonsforholdet er større enn 0. Den reduksjon i variasjonen i verdiene av den ene kjennetegnet som en finner ved en slik oppdeling av universet, er avhengig av hvor sterk korrelasjonen er. Bare i det teoretiske tilfelle at det består et funksjonsforhold mellom de to kjennetegns verdier, vil variasjonen innen subuniversene bli redusert til 0. Det vil alltid være en viss restvariasjon igjen.

Forutsetter vi for enkelhets skyld at det kjennetegnet som tas som uavhengig variabel, er en av de årsaker som er bestemmende for det annet kjennetegn, kan vi si at restvariasjonen skyldes virkningen av andre årsaker. Det er derfor klart at hvis vi trekker inn i undersøkelsen et tredje kjennetegn som også er korrelert med det som tas som avhengig variabel, vil derme restvariasjonen bli ytterligere redusert.

La oss tenke gjennom saken praktisk. Sett at vi skal estimere kubikkmassen av grantrær, og la oss tenke oss at vi først bruker observasjoner av trehøyden som grunnlag for estimeringen. Mellom kubikkmassen og trehøyden er det utvilsomt alltid et betydelig positiv korrelasjon. Men korrelasjonen er neppe i noe tilfelle så utpreget at det ikke blir en betydelig restvariasjon igjen slik at estimeringen på dette grunnlag blir temmelig usikker. En må derfor i alminnelighet ta med kjennetegn som enten alene eller sammen med trehøyden gir uttrykk for stammeformen. En kan f. eks. bruke brysthøydediameteren.

På samme måten innser vi lett at en neppe vil få noe godt resultat hvis en søker å estimere gårdsbruks nettoinntekt ved hjelp av areal dyrket mark alene. For å få en brukbar estimering må en også i dette tilfelle ta

med andre kjennetegn og da først og fremst kjennetegn som karakteriserer bruksmåten.

Det er klart at når en har observasjoner av flere enn to kjennetegn, er det ikke mulig å ordne dem i en korrelasjonstabell. En må beholde observasjonene i primærlisten. La oss tenke oss at vi har observert 3 kjennetegn hos n enheter. La observasjonene være o_i, o_i' og o_i'' (i=1,2,3,....n). La oss videre betegne observasjonsverdiene med x for o_i, med y for o_i' og med z for o_i''. Ved regressjonsfunksjonen for det siste kjennetegnet forstår vi da en funksjon av x og y:

$$Z = f(x,y)$$

Er regressjonen lineær, er

$$Z = ax + by + c$$

Konstantene a,b og c bestemmes også i dette tilfelle ved minste kvadraters metode. Etter denne metoden finnes verdiene av a,b og c ved følgende tre lineære ligninger:

$$(\Sigma o_i^2)a + (\Sigma o_i o_i')b + (\Sigma o_i)c = \Sigma o_i o_i''$$

$$(\Sigma o_i o_i')a + (\Sigma o_i'^2)b + (\Sigma o_i')c = \Sigma o_i' o_i''$$

$$(\Sigma o_i)a + (\Sigma o_i')b + nc = \Sigma o_i''$$

Eksempel 17.

Tre år etter en større snøbruddskade i Landbrukshøgskolens skog målte en lengden av råten i toppbruddene på de knekte trærne (gran). Samtidig ble en rekke andre kjennetegn observert hos de skadete trærne. Vi skal forsøke å estimere råtelengden ved 1) lengden av den levende kronen og 2) stammens diameter i bruddstedet. Vi vil bruke følgende betegnelser:

o_i = obs. (i m) av levende krone (tilsvarende variabel x)

o_i' = obs. (i cm) av diameteren i br.stedet (variabel y)

o_i'' = obs. (i m) av råtens lengde (variabel z)

Vi vil prøve med en lineær regressjonsfunksjon

$$Z = ax + by + c$$

Tabell neste side viser observasjonene og de beregningene som må utføres. Til bestemmelse av a,b og c har vi derfor følgende ligninger:

$$1245{,}81a + 1373{,}85b + 157{,}5c = 139{,}746$$

$$1373{,}85a + 2111{,}48b + 211{,}6c = 219{,}438$$

$$157{,}50a + 211{,}60b + 23c = 23{,}110$$

o_i	o_i'	o_i''	$o_i o_i''$	$o_i' o_i''$	$o_i o_i'$	o_i^2	$o_i'^2$	Z_i
3,8	15,3	1,40	5,320	21,420	58,14	14,44	234,09	1,29
9,0	10,0	0,79	7,110	7,900	90,00	81,00	100,00	0,75
5,8	11,6	1,03	5,974	11,948	67,28	33,64	134,56	1,10
8,1	8,4	0,65	5,265	5,460	68,04	65,61	70,56	0,86
5,0	8,2	1,10	5,500	9,020	41,00	25,00	67,24	1,23
12,2	7,1	0,85	10,370	6,035	86,62	148,84	50,41	0,41
9,0	7,3	0,55	4,950	4,015	65,70	81,00	53,29	0,78
5,2	9,8	1,10	5,720	10,780	50,96	27,04	96,04	1,19
4,6	12,0	1,00	4,600	12,000	55,20	21,16	144,00	1,23
5,7	5,2	1,20	6,840	6,240	29,64	32,49	27,04	1,18
4,0	9,0	2,10	8,400	18,900	36,00	16,00	81,00	1,34
8,0	6,3	0,65	5,200	4,095	50,40	64,00	39,69	0,90
8,2	8,3	1,25	10,250	10,375	68,06	67,24	68,89	0,86
5,3	11,0	1,00	5,300	11,000	58,30	28,09	121,00	1,16
11,5	5,5	0,30	3,450	1,650	63,25	132,25	30,25	0,51
6,8	10,1	0,70	4,760	7,070	68,68	46,24	102,01	0,99
13,3	6,0	0,39	5,187	2,340	79,80	176,89	36,00	0,29
6,3	7,4	1,10	6,930	8,140	46,62	39,69	54,76	1,09
5,5	8,4	1,20	6,600	10,080	46,20	30,25	70,56	1,16
5,0	10,8	1,35	6,750	14,580	54,00	25,00	116,64	1,20
3,7	7,8	1,10	4,070	8,580	28,86	13,69	60,84	1,38
3,5	10,6	1,60	5,600	16,960	37,10	12,25	112,36	1,38
8,0	15,5	0,70	5,600	10,850	124,00	64,00	240,25	0,79
157,5	211,6	23,11	139,746	219,438	1373,85	1245,81	2111,48	

Løsningene er: $a = -0,1157$

$b = -0,0114$

$c = +1,9020$

Regressjonsfunksjonen er altså

$$Z = 1,9020 - 0,1157x - 0,0114y$$

Innsetter vi nå i denne etter hvert de sammenhørende observasjonene av kronens lengde og bruddets diameter og regner ut, får vi de verdiene av Z som er gitt kollonnen lengst til høyre i tabellen. Dette er da estimeringsverdiene av råtelengden ved kronelengden og bruddets diameter etter en lineær regressjonsfunksjon.

Ved hjelp av korrelasjonsindeksen kan en videre måle hvor god estimeringen er. Korrelasjonsindeksen for o" m.h.p. o og o' er:

$$R^2_{o''oo'} = 1 - \frac{\Sigma(o_i'' - Z_i)^2}{\Sigma(o_i''-m(o''))^2}$$

I litteraturen går denne størrelsen som oftest under navn av "den multiple korrelasjonskoeffisient", men korrelasjonsindeks er en mer treffende betegnelse.

Følgende tabell viser hvordan beregningene utføres.

o_i''	$o_i''^2$	Z_i	$o_i'' - Z_i$	$(o_i'' - Z_i)^2$
1,40	1,9600	1,29	0,11	0,0121
0,79	0,6241	0,75	0,04	0,0016
1,03	1,0609	1,10	- 0,07	0,0049
0,65	0,4225	0,86	- 0,21	0,0441
1,10	1,2100	1,23	- 0,13	0,0169
0,85	0,7225	0,41	0,44	0,1936
0,55	0,3025	0,78	- 0,23	0,0529
1,10	1,2100	1,19	- 0,09	0,0081
1,00	1,000	1,23	- 0,23	0,0529
1,20	1,4400	1,18	0,02	0,0004
2,10	4,4100	1,34	0,76	0,5776
0,65	0,4225	0,90	- 0,25	0,0625
1,25	1,5625	0,86	0,39	0,1521
1,00	1,0000	1,16	- 0,16	0,0256
0,30	0,0900	0,51	- 0,21	0,0441
0,70	0,4900	0,99	- 0,29	0,0841
0,39	0,1521	0,29	0,10	0,0100
1,10	1,2100	1,09	0,01	0,0001
1,20	1,4400	1,16	0,04	0,0016
1,35	1,8225	1,20	0,15	0,0225
1,10	1,2100	1,38	- 0,28	0,0784
1,60	2,5600	1,38	0,22	0,0484
0,70	0,4900	0,79	- 0,09	0,0081
23,11	26,8121		+ 0,04	1,5026

Herav finnes:

$$m(o'') = \frac{23,11}{23} = 1,0048$$

$$\Sigma (o_i''-m(o''))^2 = \Sigma o_i''^2 - n.m(o'')^2$$

$$= 26,8121 - 23.1,0048^2 = 3,5916$$

Følgelig er

$$R^2_{o''oo'} = 1 - \frac{1,5026}{3,5916} = 0,5816$$

og

$$R_{o''oo'} = \sqrt{0,5816} = \underline{0,76}$$

Oppgave 19.

I følgende tabell er o_i smørprisen i kr. pr. kg, o_i' er kjøpernes inntekt i 1000 mill. kr og o_i'' er smørforbruket i mill. kg. Tallene gjelder for hele landet i årene fra 1926 til 1938. (Tabellen neste side).

Bestem den lineære empiriske regressjonsfunksjon for o" m.h.p. o og o'. Beregn korrelasjonsindeksen for o" m.h.p. o og o'.

Er det rimelig at en lineær funksjon kan brukes i dette tilfelle?

År	o_i	o'_i	o''_i
1926	4,6	2,5	3,8
1927	4,1	2,3	3,9
1928	4,0	2,2	3,5
1929	3,7	2,2	3,5
1930	3,5	2,2	4,2
1931	2,9	2,0	5,1
1932	2,8	1,9	5,7
1933	2,6	1,9	6,7
1934	2,8	2,0	5,8
1935	3,0	2,1	5,4
1936	3,0	2,3	5,5
1937	3,3	2,6	5,6
1938	3,6	2,9	5,1