Arbeidsnotater fra Høgskolen i Buskerud nr. 62

ARBEIDSNOTAT

ARBEIDSNOTAT

Grunnleggende statistikk

Jon Reinertsen

HØGSKOLEN i Buskerud

Arbeidsnotater fra Høgskolen i Buskerud

Nr. 62

Grunnleggende statistikk

Av

Jon Reinertsen

Hønefoss 2005

ISSN 0807-447X

INNHOLD

1.Forord.

Et poeng med dette lille heftet er å dekke behovet for repetisjon av noen viktige statistiske begrep fra grunnutdanningen. Spesielt gjelder det beskrivende statistikk, enkel regresjon og korrelasjon, estimering og hypoteseprøving. Utgangspunktet er kurset grunnleggende statistikk (6 studiepoeng) på ØK.ADM. på HIBU avd. Hønefoss. Selv om ØK.ADM.-studentene også har tatt et kurs i kvantitative metoder og et matematikkurs før statistikkurset så kreves det ikke mye matematikk for å lese dette heftet. La meg si det slik at hvis man har kunnskaper i matematikk svarende til to år med matematikk fra videregående skole er man godt rustet. Har man noe i tillegg til dette er det selvfølgelig ingen ulempe, men det viktigste er at man har den rette innstillingen, godt humør, er full av innsats, er sta i forhold til å ville lære og ikke blir "svimmel" om det skulle dukke opp noen symboler og formler som man ikke kan huske å ha sett før. Det er kanskje nettopp symbolbruken og en del definisjoner som man må holde styr på som er vanskeligheten i statistikkfaget, og ikke matematikken.

Jeg ønsker med dette heftet er å gå litt videre med en del av temaene fra grunnutdanningen, men også ta opp noen nye som man sannsynligvis kommer til å støte på et eller annet sted i sitt hovedfag. Uansett om man er i repetisjonsdelen eller i delen med nytt stoff er målet at man skal få en grundig forståelse av de temaene som presenteres. Forutsetningen er at man regner igjennom eksemplene på flere måter: i) Den "tunge veien" med papir, blyant og enkel kalkulatorbruk; ii) Bruk av kalkulatorens statistikkprogrammer (TI-83 vil bli brukt her, men også Casio og andre kan brukes) og iii) ved hjelp av det som først og fremst sannsynligvis blir verktøyet under hovedfaget: Statistikkprogrammet SPSS. Det at man skal bruke papir og blyant er ikke et forsøk på å sysselsette deg som student med masse regning, men baseres på erfaringen om at man forstår begrepene mye bedre i etterhånd ved å være mer "oppe i regningen" enn bare trykke på noen taster, og så skjer det et eller annet inne i et "mørkt rom". Det vil dessuten gjennomgående bli brukt små talleksempler slik at arbeidsmengden blir relativt liten.

Innholdet i hele heftet svarer til ca. 3 studiepoeng, dvs. at alt bør kunne foreleses på ca 20 timer. Det har vært vanlig å sette av ca. 10 timer til denne delen på hovedfag, dvs. at vi rekker å gjennomgå ca. halvparten, og den andre halvdelen man sørge for å dekke selv. Det er fornuftig å ikke jukse her, idet det er kun en selv som blir lurt. Ved en skikkelig innsats på mellom 30 timer (kan en del fra før av) og 60 timer (kan nesten ingen ting fra før av) så vil man greie å sette seg grundig inn i dette stoffet. Mye av den statistikken som dere vil støte på i løpet av hovedfaget vil kun være av den typen hvor man trenger en "fornuftig overfladisk" forståelse, dvs. tilstrekkelig forståelse til å kjenne forutsetningene for å kunne bruke metodene og kunne tolke beregningene som er gjort i SPSS. Men ved å ha en dypere forståelse av noen viktige begreper vil man mye lettere forstå de øvrige temaene som man vil støte på.

Fordi en del av stoffet er repetisjon, vil formen dette presenteres i være noe mer kortfattet enn hva som er naturlig i en vanlig lærebok.

2.Bokstavbruk i statistikk.

I statistikk bruker en konsekvent bokstaver fra det norske (engelske) alfabetet til å betegne begreper i utvalget, og greske bokstaver til å betegne begreper i populasjonen.
For eksempel betegnes det aritmetiske gjennomsnittet i utvalget med $\bar{x}$ ("x strek"), mens gjennomsnittet i populasjonen betegnes med den greske bokstaven μ ("my"). Standardavviket i utvalget betegnes med s, mens standardavviket i populasjonen betegnes med den greske bokstaven σ ("sigma") .....osv. Mange av de greske bokstavene vil bli brukt på forskjellige temaer i dette heftet, og derfor følger her en presentasjon av **det greske alfabetet** (med store og små bokstaver og uttale)

A	α	*alfa*	N	ν	*ny*
B	β	*beta*	Ξ	ξ	*ksi*
Γ	γ	*gamma*	O	o	*omikron*
Δ	δ	*delta*	Π	π	*pi*
E	ε	*epsilon*	P	ρ	*rho*
Z	ς	*zeta*	Σ	σ	*sigma*
H	η	*eta*	T	τ	*tau*
Θ	θ	*teta*	Y	υ	*ypsilon*
I	ι	*iota*	Φ	ϕ	*fi*
K	κ	*kappa*	X	χ	*kji*
Λ	λ	*lambda*	Ψ	ψ	*psi*
M	μ	*my*	Ω	ω	*omega*

Noen tilleggskommentarer :
Du kjenner sikkert utrykket: Hun var alfa og omega (f.o.m. alfa (første bokstav) t.o.m. omega (siste bokstav), dvs. hele alfabetet, dvs. hun betydde alt.
Hvis du en gang i framtiden kommer til Hellas er det greit å kunne det greske ordet for apotek: ΦΑΡΜΑΣΙΑ (hvordan vil du uttale dette?). En, to, tre på gresk er ENA ("*ena*"), ΔΥΟ ("*dyo*"),TPIA ("*tria*")

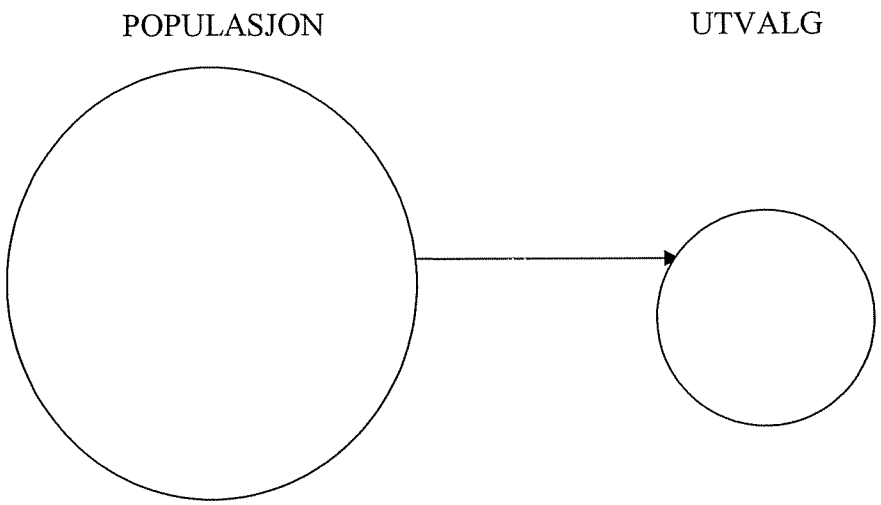
De mest brukte bokstavene i statistikk er: $\alpha, \beta, \varepsilon, \theta, \lambda, \mu, \pi, \rho, \sigma, \chi, \Theta$ *og* Σ

3.Innledning.

Opprinnelig så definerte man statistikk som data som hadde med status (tilstand) å gjøre(ofte til et land). Det var derfor vanlig å si at statistikk var data knyttet til staten. Etter hvert som faget utviklet seg kom det til å brukes på langt flere områder. Det finnes i dag praktisk talt ikke et eneste fagområde som ikke anvender statistiske metoder i en eller annen sammenheng. **Statistikk** blir i dag ofte definert som en **samling metoder** som brukes til **å ta fornuftige avgjørelser** under **usikkerhet**. Vi skal gjennom dette heftet se på en del forskjellige statistiske metoder og analyser, og se på hvorledes disse kan brukes til å trekke fornuftige konklusjoner under usikkerhet.

To meget viktige begrep i statistikk er **populasjon** og **utvalg**. Med populasjon skal en forstå samlingen av alle elementer (objekter, individer) en i øyeblikket er interessert i. Med utvalg skal en forstå en mindre og representativ del trukket tilfeldig fra populasjonen.

Eks. Hvis man er opptatt av oppslutningen om EU så er populasjonen alle stemmeberettigede personer i Norge. Et utvalg fra denne vil da for eksempel være 1000 tilfeldig utrukne personer. Det finnes mange forskjellige måter å trekke på (noen bedre enn andre), men vi skal ikke komme inn på det her. Det viktigste er at utvalget er tilfelgig og representativt.



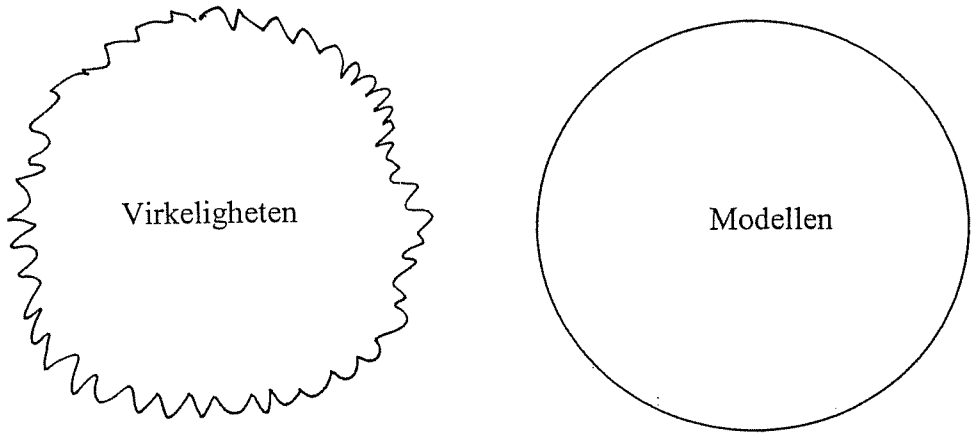
Noen viktige spørsmål som man må ha klart for seg når man skal drive med statistikk er:

- Hva er problemstillingen?
- Hva er populasjonen?
- Hvordan skal utvalget tas?
- Hvordan formulere hypotesene?
- Hvordan skal forsøkene planlegges?
- Hvordan skal dataene bearbeides?

- Hvilke statistiske metoder skal brukes/hvordan skal dataene analyseres?
- Hvilke konklusjoner kan trekkes og hvor pålitelige er metodene?

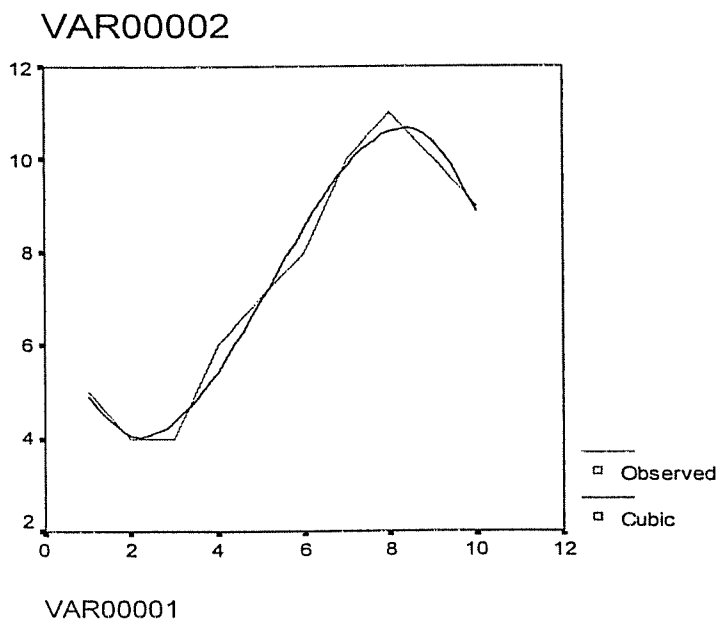
I dette heftet skal vi først og fremst konsentrere oss om de 3 siste punktene, men også (sporadisk) komme innom mange av de andre.

En **modell** er en forenkling og etterlikning av virkeligheten. Det betyr at situasjonen vi skal beskrive er renset for små detaljer og forenklet på en slik måte at vi kan gjøre tilnærmede og tilstrekkelig pålitelige beregninger i virkeligheten ved hjelp av modellen.



Ofte brukes det matematiske "språket" når man skal beskrive slike. Dette skjer ofte ved at man angir modellen i form av noen likninger, eller at man angir et funksjonsutrykk, eller en kombinasjon av disse.

Virkelighet og modell:



Dette er et eksempel på hvorledes virkeligheten som er uregelmessig blir etterlignet tilnærmet med en glatt kurve av tredje grad (se minste kvadraters metode s.....). Modellen er her beskrevet ved en matematisk funksjon, som gjør en i stand til å gjøre beregninger basert på denne. Ofte vil det være å lage prognoser framover i tid, dvs. å si noe om framtida ved hjelp av den beregnede modellen.

Statistisk inferens/ analytisk statistikk er uten tvil den viktigste delen av statistikkfaget. Den består bl.a. av:

- Estimering dvs. metoder for å anslå parametere enten
  - i) ved å angi et ett tall (punktestimering) eller
  - ii) ved å angi et intervall (intervallestimering)

- Modellkonstruksjon og prognoser, dvs. konstruere en modell ved hjelp av en gitt datamengde, og så bruke denne til å lage prognoser.

- Hypotesetesting som går ut på å avgjøre om en framsatt påstand skal forkastes eller ikke.

Hele formålet med denne typen statistikk er å få bedre kunnskap om den verden vi lever i, dvs. mer presist: bedre kunnskap om den populasjonen vi har definert.

Før vi kommer til den analytiske statistikken trenger vi imidlertid en del "verktøy". Det er forskjellige mål som kan beregnes ved hjelp av det observerte tallmaterialet, som for eksempel mål på sentral tendens, mål på spredning, mål på skjevhet, mål på spisshet, mål på sammenhenger osv.
Denne delen av statistikken kalles for beskrivende statistikk, eller også ofte deskriptiv statistikk.

4. Beskrivende statistikk.

4.1 Innledning.

Tilfeldige (random) forsøk er forsøk hvor resultatet ikke er kjent før etter at forsøket er gjennomført (dette i motsetning til deterministiske forsøk hvor resultatet er kjent på forhånd)

Eks. Kaster en terning for å studere antall øyne som terningen viser.
Eks. Stille et spørsmål til en tilfeldig valgt person med hensyn til om hun/han er for eller mot norsk medlemskap i EU.

Tilfeldig (random) variabel: Hvis vi gjør et tilfeldig forsøk så vil dette resultere i at en størrelse X antar en verdi eller et kjennetegn. Denne X-en kaller man dermed en tilfeldig (random) variabel.

Eks. Antall øyne som en terning viser (etter at den har blitt kastet) er en tilfeldig variabel.

Eks. Øyenfargen hos en tilfeldig valgt person er en tilfeldig variabel.

Tilfeldige variable kan deles i to hovedgrupper:

i) kvantitative variable. og ii) kvalitative variable

En kvantitativ variabel er en variabel som har en tallverdi knyttet til seg.

Eks. Antall øyne som en terning viser er en kvantitativ variabel.
Eks. En persons høyde og vekt er begge kvantitative variable.

En kvalitativ variabel ((kategorivariabel) er en variabel som ikke har noen tallverdi. Mao. en kvalitativ variabel er en variabel som ikke er kvantitativ.

Eks. Øyenfargen hos en person er en kvalitativ variabel
Eks. Nasjonalitet er en kvalitativ variabel.

Variabler kan også klassifiseres etter hvilket skalanivå de er målt på. Det finnes 4 nivåer:

i) Nominal skala
ii) Ordinal/rangskala
iii) Intervallskala
iv) Forholdstall

Kvalitative variable er på nominal- eller ordinalnivå, og kvantitative variable er enten på ordinal-, intervall-, eller forholdstallnivå.

Kvantitative variable kan deles i to hovedgrupper:

i) Diskrete variable
ii) Kontinuerlige variable

Diskrete variable er variable som kun antar bestemte verdier innenfor sitt variasjonsområde. Disse er oftest resultater av telleprosesser.

Eks. Antall aviser en tilfeldig person leser daglig
Eks. Hvor mange søsken en tilfeldig person har.

Kontinuerlige variable er variable som antar en hvilken som helst verdi innenfor sitt variasjonsområde. Disse er oftest resultater av måleprosesser.

Eks. Vekten til en tilfeldig valgt person.
Eks. Den tiden en tilfeldig person bruker til å se TV pr. uke.

4.2 Litt om summer:

Når man skal summere gitte tall gjør man det selvfølgelig direkte. Gitt tallene 2, 7, 5, 8, 5, 3. Summen av disse = 2+7+5+8+5+3 = 30. Dette kan man jo gjøre i hodet eller med papir og blyant eller med kalkulatoren (direkte) eller med kalkulatorens statistikkprogrammer eller med SPSS.(kommer tilbake til IKT-metodene litt senere)

I statistikk er det i mange sammenhenger interessant "å snakke om tallene" før de er kjente. Anta at en skal beregne (finne) n tall i en undersøkelse og man er interessert i å beregne summen av disse. Man trenger da n symboler for disse tallene, og det er vanlig å bruke $x_1, x_2, \cdots\cdots x_n$. Summen av disse blir da = $x_1 + x_2 + \cdots\cdots + x_n$, eller mye mer praktisk(les kortere) $\sum_{i=1}^{n} x_i$. Det betyr m.a.o. at

$$x_1 + x_2 + \cdots\cdots + x_n = \sum_{i=1}^{n} x_i$$

der $\sum$ er den store greske bokstaven sigma (m.a.o. "s-lyden" som i sum) som brukes til å symbolisere en sum av tall. Når det ikke er noen tvil skriver en ofte kortere $\sum_i x_i$, eller $\sum x_i$, eller $\sum x$, eller SUM(x) (les "summen av x-ene") Det siste brukes ofte der det nærmest er livsnødvendig å unngå bruk av symboler.

i-en i x_i kalles en indeks og er den mest vanlige, men man bruker også ofte j og k som indekser. Det betyr m.a.o. at

$$x_1 + x_2 + \cdots\cdots + x_n = \sum_{i=1}^{n} x_i = \sum_{j=1}^{n} x_j = \sum_{k=1}^{n} x_k$$

En sier derfor at indeksen er en dummy variabel. Dvs. den kan egentlig være hva som helst, men i, j og k er altså de vanligste.

5. Beliggenhetsmål

Anta at vi har gjennomført forsøket og har de n resultatene av en kvantitativ variabel : $x_1, x_2, \cdots\cdots x_n$. Dette skrives også ofte x_i, i = 1, 2, …,n og kalles for råmaterialet, fordi det er det ubehandlede tallmaterialet.

I mange sammenhenger er det nyttig å angi ett tall som representant for alle tallene. For å si noe om hvor tallene ligger plassert på tallinja (eller er lokalisert) så er det vanlig å angi et såkalt beliggenhetsmål (også kalt mål på sentral tendens). Dvs. det er et tall som sier noe om hvor tallmassen er ligger (eller er lokalisert).

Det aritmetiske middeltall (the arithmetic mean) er det mest brukte beliggenhetsmålet. Det er definert ved:

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

M.a.o. finn summen av x-ene og divider så på antall observasjoner, dvs. $\bar{x} = \frac{SUM(x)}{n}$ for de

med "$\sum$ - fobi"

Eks. Anta at tallmaterialet x_i, i = 1, 2, ...,10 er gitt ved: 1, 2, 1, 3, 4, 3, 2, 2, 3, 2. Da blir

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+2+1+3+4+3+2+2+3+2}{10} = \frac{23}{10} = 2{,}3 \ (*)$$

Tallet 2,3 er nå et tall som representerer de 10 tallene, og forteller hvor disse tallene er lokalisert (eller ligger)
Ser en litt nærmere på tallene som inngår i telleren ser en at en del av tallene er innbyrdes like. Det medfører at en kan skrive

$$\bar{x} = \frac{2 \cdot 1 + 4 \cdot 2 + 3 \cdot 3 + 1 \cdot 4}{10} = \frac{23}{10} = 2{,}3 \ (**)$$

Om man her regner ut $\bar{x}$ ved hjelp av (*) eller (**) spiller ikke noen særlig rolle, men hvis tallmaterialet hadde vært stort, og mange av observasjonene var like, så ville det vært svært besparende å bruke (**). Man sier her at frekvensen (hyppigheten (the frequency)) av tallet 1 er 2, frekvensen av tallet 2 er 4, frekvensen av tallet 3 er 3, og frekvensen av tallet 4 er 1. Dette skrives

$$f_1 = 2,\ f_2 = 4,\ f_3 = 3 \text{ og } f_4 = 1$$

Den generelle formelen for beregning av $\bar{x}$ når flere av observasjonene er innbyrdes like (dvs. verdien x_k har frekvensen f_k, k=1,2,...m, der m er antall forskjellige verdier av x. I eks. foran er m=4. (m.a.o. verdien x_1 forekommer f_1 ganger, verdien x_2 forekommer f_2 ganger, ..., verdien x_m forekommer f_m ganger,) blir dermed:

$$\bar{x} = \frac{\sum_{k=1}^{m} f_k x_k}{n} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + \cdots + f_m \cdot x_m}{n}$$

eller bare kortere $\bar{x} = \frac{\sum_{k} f_k x_k}{n}$ eller $\bar{x} = \frac{\sum f \cdot x}{n}$ eller $\bar{x} = \frac{SUM(f \cdot x)}{n}$

Et annet men ikke så mye brukt mål på sentral tendens er **typetallet** (eng.: the mode) $\tilde{x}$ (=T) som ganske enkelt er den observerte verdi med størst frekvens.

Eks. I eks. over er typetallet $\tilde{x} = 2$ fordi verdien 2 forekommer hyppigst, nemlig 4 ganger.

Noen ganger inneholder våre tallmaterialer enkelte ekstreme verdier i forhold til de fleste andre. (disse kalles av noen "uteliggere" etter det engelske outlier. Se definisjonen s. 12.) I slike tallmaterialet blir det aritmetiske gjennomsnittet lett påvirket i retning av de(n) ekstremt store/små verdiene.
Eks. Anta at man har observert alderen x på 5 personer og funnet: x_i: 1, 2, 3, 4, og 60.
Beregner en her gjennomsnittsalderen ved hjelp av det aritmetiske gjennomsnittet finner en

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{1+2+3+4+60}{5} = \frac{70}{5} = 14 \text{ (år)}$$

som neppe kan sies å være et representativt tall for tall for dataene. I slike sammenhenger er det man bruker et mål på sentral tendens som ikke så lett lar seg påvirke av ekstreme verdier. Det finnes flere slike mål. Et mye brukt mål er den såkalte medianen.

Medianen M er det tallet som deler det ordnede tallmaterialet (ordnet i stigende eller avtagende rekkefølge) i to like store deler. Medianen sies derfor ofte å være den midterste observasjonen i det ordnede tallmaterialet hvis det er et odde antall observasjoner, og gjennomsnittet av de to midterste hvis det er et like antall observasjoner.

Eks. La x_i være: 2, 5, 3, 4, 16. Ordner man tallmaterialet har en: 2, 3, 4, 5, 16 og da ser en at medianen blir 4.

Eks. Sløyfer en nå for eksempel observasjonen 2 ser en at det ikke lenger er noen observasjon i midten, og medianen er dermed gjennomsnittet av de to midterste, d.v.s.

$$M = \frac{4+5}{2} = 4{,}5$$

Medianen behøver m.a.o. ikke være en observasjon. Man sier ofte at medianen M er den verdien som er slik at 50% av tallmaterialet(det ordnede) ligger under denne og 50% ligger over denne.

Andre nyttige beliggenhetsmål er de såkalte **kvartilene** Q_1, Q_2 *og* Q_3. De deler også det ordnede tallmaterialet i to deler:

Q_1 slik at 25% av observasjonene ligger under og 75% ligger over denne.
Q_2 slik at 50% av observasjonene ligger under og 50% ligger over denne.
Q_3 slik at 75% av observasjonene ligger under og 25% ligger over denne.

Det betyr m.a.o. at medianen M og 2.kvartil Q_2 er den samme.

Et tallmateriale kan deles inn i 2 deler på mange måter. Noen andre mye brukte er :

Densilene $D_1, D_2, \ldots\ldots, D_{10}$ deler tallmaterialet inn 10-deler analogt til over. Det betyr at D_1 deler tallmaterialet i to slik at 10% ligger under D_1 og 90% ligger over denne verdien, D_2 deler tallmaterialet i to slik at 20% ligger under D_2 og 80% ligger over denne verdien, osv.

Prosentilene $P_1, P_2, \ldots\ldots, P_{100}$ deler tallmaterialet inn i 100-deler analogt til over. Det betyr at P_1 deler tallmaterialet i to slik at 1% ligger under P_1 og 99% ligger over denne verdien. osv....

6. Spredningsmål

To forskjellige tallmaterialer kan ha samme beliggenhetsmål. Bl.a. for å kunne skille mellom disse så innføres såkalte spredningsmål, som gir et mål på hvor stor spredning det er i observasjonene.
Eks. Tallmaterialene x_i: 1, 4, 5, 9, 11 og y_i: 3, 5, 7, 9 er forskjellige, men har allikevel samme aritmetiske gjennomsnitt (kontroller selv). Er medianene like? Spredningen i de to tallmaterialene er imidlertid forskjellig.

Variasjonsbredden (the range) er et enkelt, men ikke så mye brukt variasjonsmål. Det er definert som differansen mellom den største og den minste observasjonen, dvs.

$$V = x_{maks} - x_{\min}$$

Eks. I tallmaterialene over finner en $V_x = 11 - 1 = 10$ og $V_y = 9 - 3 = 6$

Kvartibredden (the interquartilrange) er et annet variasjonsmål, som er noe mer brukt en variasjonsbredden. Det er differansen mellom 3. og 1. kvartil, dvs.

$$\text{Kv.br.} = Q_3 - Q_1 = \text{IQR}$$

Det betyr at kvartilbredden er avstanden mellom de to verdiene (Q_1 *og* Q_3) som er slik at 50% av observasjonene i det ordnede tallmaterialet ligger mellom disse (75% ligger på nedsiden av 3.kvartil og 25% ligger på nedsiden av 1.kvartil)

IQR brukes ofte til å definere hva en **outlier** (ekstremverdi) er for noe. En observasjon kalles en outlier den er

$$< Q_1 - 1{,}5IQR \text{ eller } > Q_3 + 1{,}5IQR$$

Hvis observasjonen er

$$< Q_1 - 3IQR \text{ eller } > Q_3 + 3IQR$$

kalles den ofte for en **ekstrem outlier**

Variansen er det klart mest brukte spredningsmålet. Dette målet forteller hvor mye observasjonene avviker fra sitt gjennomsnitt med. Variansen er definert ved

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2)$$

En ser m.a.o. mer presist at variansen først regner ut hvor mye x_1 avviker fra $\bar{x}$ med, deretter kvadreres dette, så gjøres det tilsvarende for x_2, osv...., tilslutt gjøres det for x_n. Etter dette deles alle disse kvadrerte avvikene med n, dvs. m.a.o. si at variansen er gjennomsnittlig kvadrert avvik fra gjennomsnittet for alle observasjonene. Grunnen til at man kvadrerer er at man ellers ville få 0 hver eneste gang, fordi det kan vises generelt at man alltid har at

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) \ + \cdots + (x_n - \bar{x}) = 0$$

Forklaring for de som måtte ønske det (de andre kan hoppe over):

$$(x_1 - \bar{x}) + (x_2 - \bar{x}) \ + \cdots + (x_n - \bar{x}) = (x_1 + x_2 + \cdots + x_n) - \bar{x} - \bar{x} - \cdots - \bar{x} = n \cdot \bar{x} - n \cdot \bar{x} = 0$$

Eks. Betrakter nå tallmaterialet på side 9 der x_i, i = 1, 2, ...,10 var gitt ved: 1, 2, 1, 3, 4, 3, 2, 2, 3, 2. Her fant vi $\bar{x} = 2{,}3$. Dermed blir variansen

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}((x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2) =$$
$$= \frac{1}{10}((1 - 2{,}3)^2 + (2 - 2{,}3)^2 + \cdots + (2 - 2{,}3)^2) = 0{,}81$$

En ser m.a.o. at først så beregnes avvikene fra gjennomsnittet for hver eneste observasjon:

$$(1 - 2{,}3),\ (2 - 2{,}3), \cdots,\ (2 - 2{,}3) \qquad \text{(Vis at summen av disse avvikene = 0)}$$

Deretter kvadreres disse avvikene før de så adderes. Summen av de kvadrerte avvikene blir 8,1 (kontroller selv). Tilslutt deles summen av disse 10 kvadrerte avvikene på 10, en regner m.a.o. ut gjennomsnittlig kvadrert avvik for de 10 tallene.
Som vist over må man altså gjøre noe med avvikene før man deler på 10 ellers vil man kun få 0 i gjennomsnittlig avvik hver eneste gang. Den ene muligheten er altså som her å kvadrere avvikene (da blir de negative avvikene kvadrert positive). Den andre muligheten er å beregne absoluttverdiene av avvikene, og så addere disse og tilslutt dividere med 10. Grunnen til at man har valgt kvadreringen er at dette i den generelle teorien som er utviklet i forbindelse med dette gir mye bedre "matematiske arbeidsforhold". En ulempe med kvadreringen er imidlertid at variansen får en annen benevning enn de opprinnelige data. Tenk for eksempel at de 10 tallene er beløp i kroner. Da vil gjennomsnittlig beløp være 2,3 kroner, mens variansen blir 0,81 kroner^2 (m.a.o. 0,81 kvadratkroner, hva nå det måtte være for noe?). For å korrigere for dette (m.a.o. ha et spredningsmål med samme benevning som dataene) så innføres det såkalte standardavviket som er kvadratroten av variansen. M.a.o.:

$$\text{Standardavviket} = \sigma = \sqrt{\textit{Variansen}}$$

Det betyr at standardavviket i tallmaterialet over er $\sigma = \sqrt{0{,}81} = 0{,}9$ (kroner).

Man kan her analogt til overgangen fra $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ til $\bar{x} = \frac{\sum_{k} f_k x_k}{n}$ sette opp en tilsvarende

kortere beregningsformel for variansen ved å slå sammen de like leddene:

$$= \frac{1}{10}((1-2{,}3)^2 + (2-2{,}3)^2 + (1-2{,}3)^2 + (3-2{,}3)^2 + (4-2{,}3)^2 + (3-2{,}3)^2 + (2-2{,}3)^2 +$$
$$+ (2-2{,}3)^2 + (3-2{,}3)^2 + (2-2{,}3)^2)) =$$

$$= \frac{1}{10}(2 \cdot (1-2{,}3)^2 + 4 \cdot (2-2{,}3)^2 + 3 \cdot (1-2{,}3)^2 + 1 \cdot (3-2{,}3)^2) = 0{,}81$$

Dette leder dermed til følgende formel:

$$\sigma^2 = \frac{1}{n}\sum_{k=1}^{m} f_k (x_k - \bar{x})^2 = \frac{1}{n}(f_1 \cdot (x_1 - \bar{x})^2 + f_2 \cdot (x_2 - \bar{x})^2 + \cdots + f_m \cdot (x_m - \bar{x})^2)$$

der m er antall forskjellige x-verdier.
Dette er jo praktisk (forenklende) når man har mange like tall å arbeide med og skal gjøre beregningene "for hånd", men så fort en overlater beregningene til TI-83 `s statistikkprogrammer eller SPSS er det helt uvesentlig hvilken beregningsformel som ligger bak.

Nå er det kanskje noen som husker at man skal dele på (n-1) og ikke n når man beregner variansen. Når skal man gjøre hva? Det er vanlig å kalle σ^2 for populasjonsvariansen, dvs. variansen til alle elementene en i øyeblikket interesser seg for. Nå er det vanlig at ikke hele populasjonen er kjent, men at man tar et tilfeldig utvalg for å få kunnskap om populasjonen. I dette utvalget kan man så beregne variansen som dermed kalles for utvalgsvariansen, og betegnes med s^2. Denne utvalgsvariansen vil jo måtte være et tall i nærheten av σ^2 siden utvalget vårt er representativt. Det kan i den matematiske statistikken vises at s^2 ligger nærmere σ^2 (treffer bedre) når man deler på (n-1) enn hvis man deler på n. Mer presist for de som har vært borte i forventningsverdier og estimering: Det kan vises at $E(S^2) = \sigma^2$, m.a.o. S^2 (=variabelen knyttet til s^2) er en forventningsrett estimator for σ^2.
Det betyr da at

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{m} f_k (x_k - \bar{x})^2 = \frac{1}{n-1}(f_1 \cdot (x_1 - \bar{x})^2 + f_2 \cdot (x_2 - \bar{x})^2 + \cdots + f_m \cdot (x_m - \bar{x})^2)$$

er et bra estimat for σ^2. Det er vanlig å bruke s^2 når man opererer med et utvalg av data. Kjenner man hele populasjonen bruker man σ^2. Når tallmaterialene blir store spiller det liten rolle om man deler på (n-1) eller n.

Eks. Anta at summen av de kvadrerte avvikene er 2250 og at n=500. Da blir

$$\sigma^2 = \frac{2250}{500} = 4{,}500\text{, mens} \qquad s^2 = \frac{2250}{500-1} = 4{,}509$$

som resulterer i følgende standardavviker:

$$\sigma = \sqrt{4{,}500} = 2{,}121 \text{ og} \qquad s = \sqrt{4{,}509} = 2{,}123$$

M.a.o. det blir helt ubetydelige forskjeller. For å kunne skjelne litt bedre mellom σ^2 og s^2 bruker en i noen bøker N på antallet i populasjonen, og n på antallet i utvalget. Det betyr at populasjonsvariansen blir gitt ved

$$\sigma^2 = \frac{1}{N}\sum_{k=1}^{m} f_k (x_k - \bar{x})^2 = \frac{1}{N}(f_1 \cdot (x_1 - \bar{x})^2 + f_2 \cdot (x_2 - \bar{x})^2 + \cdots + f_m \cdot (x_m - \bar{x})^2) \text{ og}$$

utvalgsvariansen blir gitt ved

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{m} f_k (x_k - \bar{x})^2 = \frac{1}{n-1}(f_1 \cdot (x_1 - \bar{x})^2 + f_2 \cdot (x_2 - \bar{x})^2 + \cdots + f_m \cdot (x_m - \bar{x})^2)$$

Grupperte tallmaterialer. Det er ofte slik at en del store tallmaterialer er ordnet i tabellform, for å skape mer oversikt (se for eksempel statistisk årbok) enn det råmaterialet gjør. Dette vil da være en tilnærmet angivelse av de opprinnelige dataene.
Eks. Anta at et tilfeldig utvalg på n=20 observasjoner er gitt ved:

$$x_i : 2, 3, 6, 5, 7{,}11, 8, 9{,}14{,}12{,}10, 5, 3, 6, 6, 14, 9, 8, 7{,}13$$

Først skal vi regne eksakt på dette tallmaterialet, for deretter å organisere tallene i en tabell og så sammenlikne resultatene. Det ordnede tallmaterialet gjør det litt lettere mht. beregningene.

$$x_{(i)} : 2, 3, 3, 5, 5, 6, 6, 6, 7, 7, 8, 8, 9, 9, 10, 11, 12, 13{,}14{,}14$$

En finner nå det aritmetiske gjennomsnittet

$$\bar{x} = \frac{\sum_{k=1}^{m} f_k x_k}{n} = \frac{f_1 \cdot x_1 + f_2 \cdot x_2 + \cdots + f_m \cdot x_m}{n} = \frac{1 \cdot 2 + 2 \cdot 3 + 2 \cdot 5 + 3 \cdot 6 + \cdots + 2 \cdot 14}{20} = 7{,}9$$

og utvalgsvariansen

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{m} f_k (x_k - \bar{x})^2 = \frac{1}{n-1}(f_1 \cdot (x_1 - \bar{x})^2 + f_2 \cdot (x_2 - \bar{x})^2 + \cdots + f_m \cdot (x_m - \bar{x})^2) =$$

$$= \frac{1}{20-1}(1\cdot(2-7{,}9)^2 + 2\cdot(3-7{,}9)^2 + \cdots + 2\cdot(14-7{,}9)^2) = 12{,}9 \ (12{,}9368\ldots)$$

Dermed blir standardavviket $s = \sqrt{12{,}9} = 3{,}6$ (= 3.59678..)

I mange sammenhenger er et slikt tallmateriale gitt i tabellform som følger:

Klassegrenser	Frekvens f_k	Klassemidtpkt. x_k
$[0,5\rangle$	3	2,5
$[5,10\rangle$	11	7,5
$[10,15\rangle$	6	12,5

og da er ikke råmaterialet kjent slik som her. Det betyr at en nå kun vet at det er 3 observasjoner mellom fra og med 0 og til 5, 11 observasjoner mellom 5 (f.o.m.) og 10 (til), osv.. Man velger nå punktet midt i klassen som representant for de ukjente verdiene. M.a.o. det er 3 observasjoner som er 2,5 (eksakt er de 2, 3 og 3 hvis man ser på råmaterialet), 11 observasjoner som er 7,5, osv.
Med denne tilnærmingen finner en nå

$$\bar{x} = \frac{\sum_{k=1}^{m} f_k x_k}{n} = \frac{f_1\cdot x_1 + f_2\cdot x_2 + \cdots + f_m\cdot x_m}{n} = \frac{3\cdot 2{,}5 + 11\cdot 7{,}5 + 6\cdot 12{,}5}{20} = 8{,}3 \ (8{,}25)$$

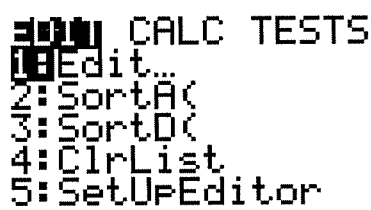
som avviker litt fra den eksakte verdien på 7,9. Nå skal det bemerkes at ved større tallmaterialer så blir forskjellene gjennomgående mye mindre.
Tilsvarende finner man variansen i tabellen:

$$s^2 = \frac{1}{n-1}\sum_{k=1}^{m} f_k(x_k - \bar{x})^2 = \frac{1}{20-1}(3\cdot(2{,}5-8{,}3)^2 + 11\cdot(7{,}5-8{,}3)^2 + 6\cdot(12{,}5-8{,}3)^2) = 11{,}3$$

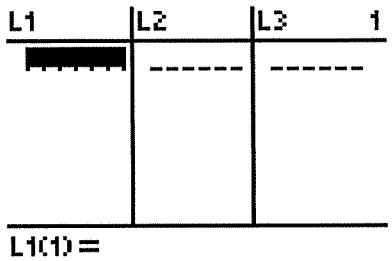
Herav finner en da standardavviket $s = \sqrt{11{,}3} = 3{,}4$

Ønsker man å legge disse tallene inn i listene i TI 84 går en fram som følger:

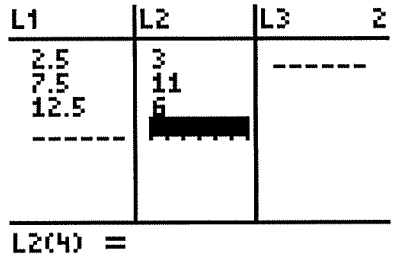
Trykk først på STAT-tasten. Da får du opp følgende bilde:



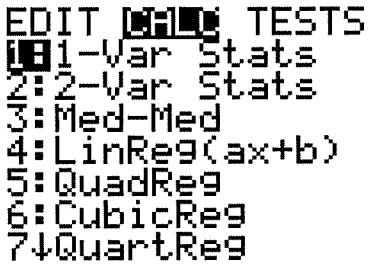
Trykk så på ENTER-tasten og du får opp følgende bilde:



Kalkulatoren er nå klar til å ta imot tall i de forskjellige listene. Legger så klassemidtpunktene inn i liste 1, L_1, og frekvensene inn i liste 2 , L_2 .
Dette gir da følgende bilde:



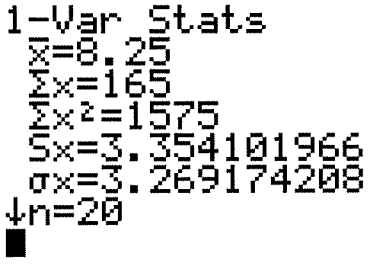
Nå trykker en så på STAT-tasten igjen, men velger nå isteden alternativet CALC (calculations). Dette gir følgende bilde:

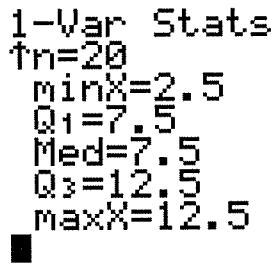


En bruker nå 1: 1-Var Stats (envariabelstatistikk) på følgende måte:

Trykk først på ENTER og deretter på 2ND 1, så på kommatasten, og tilslutt på 2ND 2. Du vil da få opp følgende bilde:

Trykker en nå på ENTER-tasten får en følgende bilde:





Her får en nå bekreftet beregningene over, og i tillegg beregnet de tre kvartilene.

7. Mål på skjevhet og spisshet.

Vi har til nå sett på mål på sentral tendens og mål på spredning. Disse kalles ofte henholdsvis første- og andre-ordens mål. I en del sammenhenger er det også nyttig å se på høyere ordens mål. Anta at vi har n observasjoner $x_1, x_2, \cdots\cdots x_n$ Vi definerer derfor nå det såkalte **r.te-ordens momentet omkring** $\bar{x}$ ved

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

hvis alle observasjonene er forskjellige, eller ved

$$m_r = \frac{\sum_{k} f_k (x_k - \bar{x})^r}{n} \qquad r = 1, 2, 3, \ldots..$$

hvis en del av observasjonene er like, eller dataene er gruppert. Det er ikke egentlig noen forskjell på de to formlene (jfr. de to formlene for varians) idet hvis alle frekvensene var lik 1 så er alle x-ene forskjellige og formel 1 fremkommer. En annen ting er at en godt kan bruke formel 1 i alle tilfellene, men en blir da sittende å addere mange like ledd der en har like observasjoner istedenfor å multiplisere (m.a.o. 5+5+5+5+5+5+5+5+5+5 er tyngre å regne ut enn 9*5) Det betyr m.a.o. at formel 2 er en kortere (og greiere) formel å bruke enn formel 1 når det er mange like data.

Det kan vises at $m_1 = 0$ (beviset under kun for de spesielt interesserte) uansett tallmateriale (se varians side...)

$$m_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^1}{n} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} = \frac{\sum_{i=1}^{n} x_i - \sum_{i=1}^{n}\bar{x}}{n} = \frac{n\bar{x} - n\bar{x}}{n} = 0$$

Dessuten har en at

$$m_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \text{variansen i et tallmateriale} = \sigma^2 \text{ (egentlig populasjonsvariansen)}$$

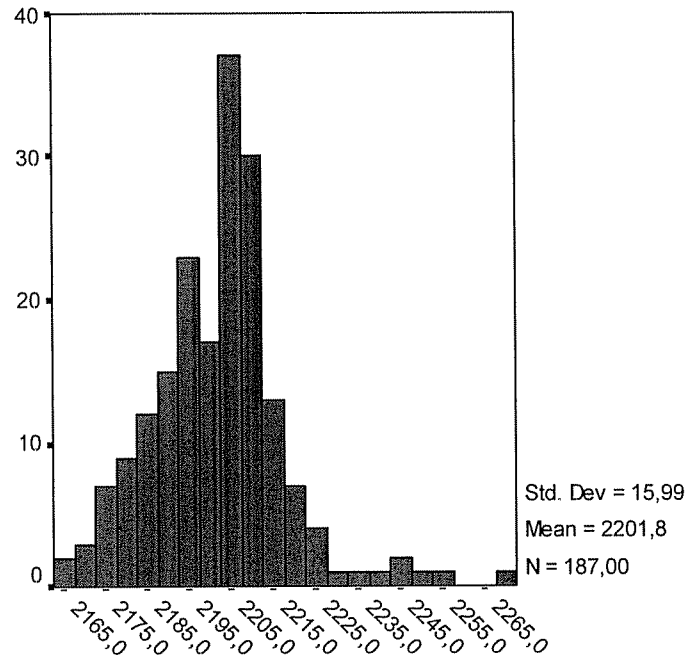
Nå skal vi også betrakte m_3 og m_4. Disse har betydning for en del av analysene som skal gjøres senere.

Tredjeordensmomentet omkring $\bar{x}$ definert ved

$$m_3 = \frac{\sum_k f_k (x_k - \bar{x})^3}{n}$$

brukes til å beregne **skjevheten (the skewness)** i en fordeling.
Hvis en fordeling har enkelte små verdier som skiller seg fra de øvrige (fordelingen vil da ha en hale mot venstre) så sier man at skjevheten er negativ. Hvis fordelingen er symmetrisk så er skjevheten 0. Har fordelingen enkelte store verdier som skiller seg fra de øvrige (fordelingen har da en hale mot høyre) så er skjevheten positiv.

Henter en noen data fra SPSS (kap.6) får en for de 187 scoredataene følgende histogram:



I følge dette histogrammet så skal man ha en positiv skjevhet. Dette får man bekreftet av kommandoene:

ANALYSE

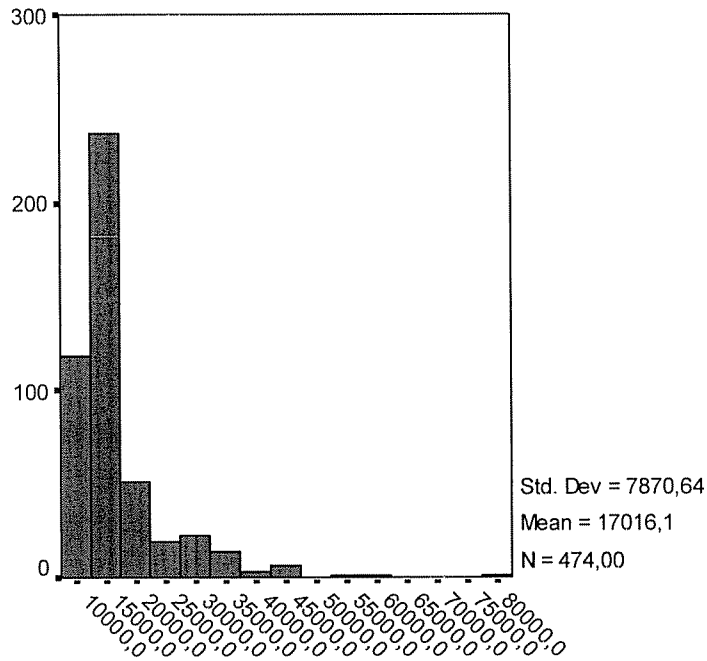
DESCRIPTIVE STATISTICS

DESCRIPTIVES

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
SCORE	187	2165	2270	2201,84	15,986	,616	,178	2,169	,354
Valid N	187								

Hvor en ser at Skewness er +0,616. Ser en på noen andre data fra SPSS (Employee-data) ser man en større skjevhet enn over:



Her er skjevheten på hele +2,853, m.a.o. en mye skjevere fordeling enn foran

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Beginning Salary	474	$9,000	$79,980	$17,016.09	$7,870.638	2,853	,112	12,390	,224
Valid N (listwise)	474								

Ifølge Jøreskog (Formulas for Skewness and Kurtosis 1999) så beregnes skjevheten ved først å regne ut

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{m_3}{s^3}$$

der s er standardavviket.

g_1 vil være negativ hvis m_3 er negativ, og positiv hvis m_3 er positiv.
Deretter beregnes (justert g_1)

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} \cdot g_1$$

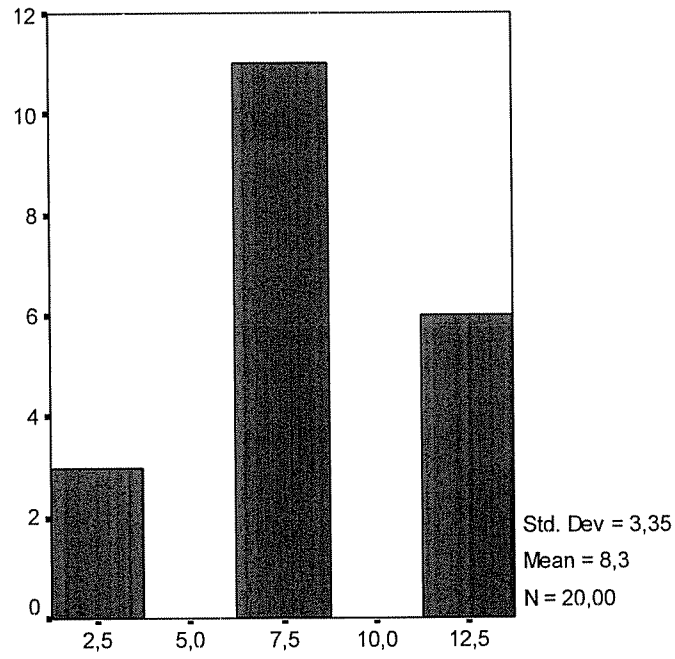
(justert g_1 som er forventningsrett (normalitetsforuts.))

Legger en nå inn tallene fra side 15 i SPSS får en:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
VAR00002	20	2,50	12,50	8,2500	3,35410	-,177	,512	-,548	,992
Valid N (listwise)	20								

Og dermed finner en skjevhet på -0.177, m.a.o. en svak venstreskjevhet, hvilket en kan se av grafen:



Nå skal vi prøve å kontrollregne denne verdien, og vi trenger altså både m_2 og m_3.

Vi har tidligere funnet $s^2 = 11{,}25 \Rightarrow m_2 = \frac{19}{20} \cdot 11{,}25 = 10{,}69$ (se side.....) I tillegg finner en nå m_3 ved

$$m_3 = \frac{\sum_k f_k (x_k - \bar{x})^3}{n} = \frac{3 \cdot (2{,}5 - 8{,}25)^3 + 11 \cdot (7{,}5 - 8{,}25)^3 + 6 \cdot (12{,}5 - 8{,}25)^3}{20} = -5{,}72$$

Dermed blir

$$g_1 = \frac{m_3}{m_2^{3/2}} = \frac{-5{,}72}{(10{,}69)^{3/2}} = -0{,}164$$

og dermed finner en

$$G_1 = \frac{\sqrt{n(n-1)}}{n-2} \cdot g_1 = \frac{\sqrt{20 \cdot 19}}{18} \cdot (-0{,}164) = -0{,}178$$

som avviker svakt fra SPSS sitt tall på -0,171 (differanse på -0,171-(-0,178) = 0,007. Dette tilskrives forskjellen i antall desimaler som det regnes med.
Iflg. SPSS sin hjemmeside så er utrykket for standardfeilen til g_1 (the standard error of g_1 ,dvs standardavviket til g_1) gitt ved

$$se(g_1) = \sqrt{\frac{6W_N(W_N-1)}{(W_N-2)(W_N+1)(W_N+3)}}$$

der $W_N = \sum_{i=1}^{N} w_i = \sum_{i=1}^{N} (vektene\ for\ observasjon\ i) = 1+1+\cdots+1 = N$, der N er antall observasjoner. Dvs. at

$$se(g_1) = \sqrt{\frac{6N(N-1)}{(N-2)(N+1)(N+3)}} = \sqrt{\frac{6 \cdot 20 \cdot 19}{18 \cdot 21 \cdot 23}} = 0{,}512$$

som stemmer helt med SPSS-utskriften. Dette er et tall som kan brukes til hypotesetesting og estimering (konfidensintervaller).

Et annet viktig mål i en fordeling baserer seg på fjerdeordensmomentet omkring $\bar{x}$, og dette måler graden av spisshet (kurtosis) i fordelingen. Nå er iflg. def. s.17

$$m_4 = \frac{\sum_k f_k (x_k - \bar{x})^4}{n}$$

Definer så g_2 ved

$$g_2 = \frac{m_4}{m_2^2} - 3$$

Grunnen til at 3-tallet kommer inn er at i normalfordelingen er spissheten akkurat lik 3,0. Det betyr dermed at hvis en fordeling er spissere enn normalfordelingen (spisshet > 3,0) så er $g_2 > 0$, og hvis den er mindre spiss enn normalfordelingen så blir $g_2 < 0$. Tilsvarende til definisjonen av G_1 defineres nå G_2 ved

$$G_2 = \frac{n-1}{(n-2)(n-3)}\left[(n+1)g_2 + 6\right]$$

(som også er en forventningsrett estimator under normalitetsforutsetningen).

Prøver nå å sjekke beregningene i SPSS-utskriften. Må da først finne m_4 (m_2 er kjent fra før).

$$m_4 = \frac{\sum_k f_k (x_k - \bar{x})^4}{n} = \frac{3 \cdot (2{,}5 - 8{,}25)^4 + 11 \cdot (7{,}5 - 8{,}25)^4 + 6 \cdot (12{,}5 - 8{,}25)^4}{20} = 262{,}0195\ldots$$

Dermed blir

$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{262{,}02}{10{,}69^2} - 3 = -0.707$$

og det forventningsrette estimatet G_2

$$G_2 = \frac{n-1}{(n-2)(n-3)}\left[(n+1)g_2 + 6\right] = \frac{19}{18 \cdot 17}\left[21 \cdot (-0.707) + 6\right] = -0{,}5495 = -0{,}55$$

som stemmer svært så bra med SPSS sin verdi som er -0,548. Fordelingen er m.a.o. litt mindre spiss enn normalfordelingen.

På SPSS sin hjemmeside finner man også formelen til standardfeilen (les standardavviket) til g_2:

$$se(g_2) = \sqrt{\frac{4(N^2 - 1)(se(g_1))^2}{(N-3)(N+5)}}$$

som innsatt N=20 og $se(g_1) = 0{,}512$ gir

$$se(g_2) = \sqrt{\frac{4(20^2 - 1)(0{,}512)^2}{(20-3)(20+5)}} = 0{,}992$$

som stemmer eksakt med datautskriften på side 20
Dette kan da igjen brukes til å gjennomføre hypotesetesting og etimering.

8. Enkel regresjon

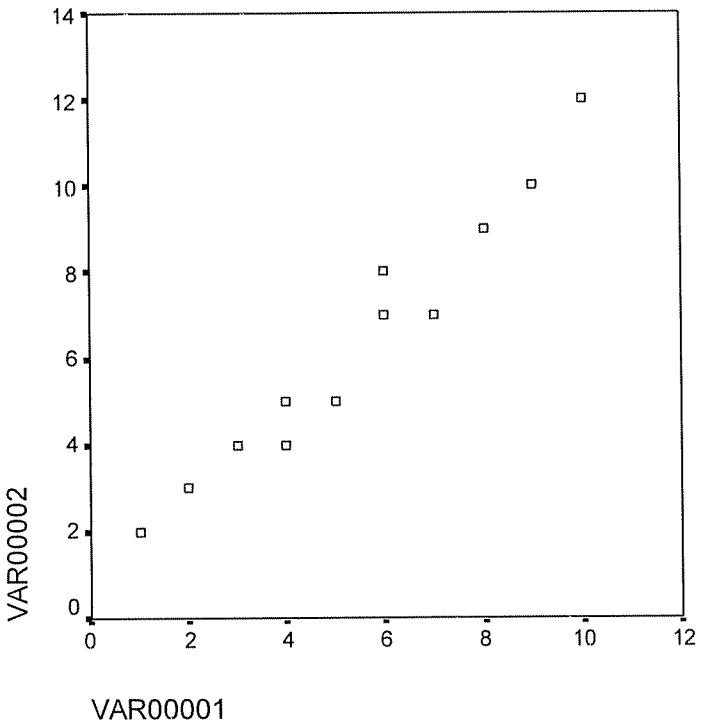
Anta man har n parobservasjoner (x_i, y_i) der x_i er gitte verdier av en tilfeldig variabel X og y_i er verdien av en tilfeldig variabel Y.

x_1	x_2	x_3	.....	x_n
y_1	y_2	y_3	.....	y_n

Avsetter man punktene (x_i, y_i), i = 1,2,3,....,n i et xy-koordinatsystem fremkommer det såkalte spredningsdiagrammet (the scatterplot) :

Eks. Anta man har observert følgende sammenheng mellom X og Y.

x	1	2	3	4	4	5	6	6	7	8	9	10
y	2	3	4	4	5	5	7	8	7	9	10	12



Ser en på spredningsdiagrammet observerer man at det er en positiv rettlinjet trend i sammenhengen mellom x og y. Dette kan da beskrives ved følgende modell (husk at en modell er en etterlikning og forenkling av virkeligheten (som her er representert ved de 12 parobservasjonene)):

$$y = \alpha + \beta x + \varepsilon \quad \text{der } \varepsilon \text{ er } N(0, \sigma^2) \qquad (*)$$

ε kalles ofte støyen (eller feilleddet, eng.:the error) og antas å være normalfordelt med forventning 0 og med en varians σ^2 (se normalfordelingen s 40)
$\alpha + \beta x$ kalles ofte for regresjonslikningen (den teoretiske (eller sanne)) for y med hensyn på x, eller av og til for signalet. Det betyr at man kan si at y = "signal" + "støy". I statistikk er det vanligst å angi likningen for en rett linje med a+bx istedenfor ax+b som er vanligst norske matematikkbøker. Modellen (*) over gjelder selvfølgelig for alle n observasjonsparene. Ofte beskrives modellen derfor noe mer presist som følger:

De tilfeldige variablene $Y_1, Y_2, ..., Y_n$ (gitt de tilsvarende x-ene) er uavhengige med

$$\text{forventning} = \mu_{Y|x} = \alpha + \beta x \qquad \text{og}$$

$$\text{varians} = \sigma^2$$

eller ekvivalent

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, 2, ..., n$$

der $e_1, e_2, ..., e_n$ er n uavhengige feilledd som har

$$\text{forventning} = 0 \text{ og varians} = \sigma^2$$

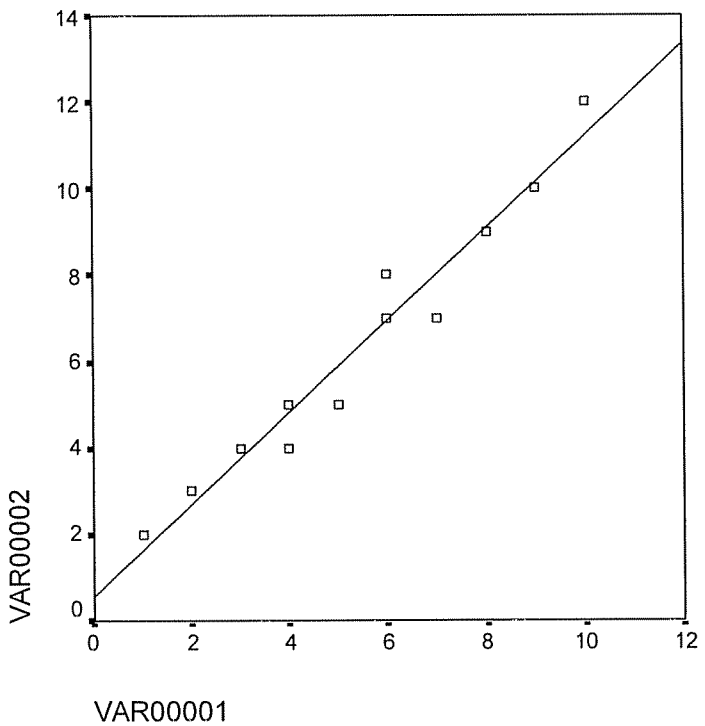
Likningen $\mu_{Y|x} = \alpha + \beta x$ kalles ofte for **populasjonsregresjonslikningen** for Y m.h.t. x. Denne skal vi prøve å estimere ved hjelp av et utvalg av n observasjonspar. Vi kan da finne en såkalt estimert regresjonslikning eller en såkalt **utvalgsregresjonslikning** som betegnes ved

$$\hat{y} = a + b x$$

Denne vil da kunne brukes til å estimere fremtidige verdier av Y, dvs. å lage prognoser. a og b er da estimater for henholdsvis α *og* β . Disse finner en ved hjelp av den såkalte **minste kvadraters metode**, som går ut først å beregne avvikene

$$e_i = observert\ y\ verdi - estimert\ y\ verdi = y_i - \hat{y}_i \text{ for all de n punktene}$$

(se grafisk billede)



Det betyr at i hvert eneste punkt så beregnes avviket mellom den observerte y-verdien og den y-verdien den ukjente linja $\hat{y} = a + b x$ (det som er ukjent er a og b; det som er kjent er at det vi skal finne er en rett linje). Man beregner først

$$e_i = y_i - \hat{y}_i = y_i - (a + b x_i) \text{ for i = 1, 2, 3,.....,n}$$

Nå kvadreres alle disse avvikene og deretter adderes de. Man beregner m.a.o.

$$f(a,b) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

for de n = 12 leddene. Dette vil dermed være en funksjon av 2 variable (a og b), mer presist en annengradsfunksjon som betegnes med $f(a,b)$. Hva tror du er grunnen til at avvikene a og b kvadreres?. Nå skjer det et lite stykke matematikk som krever at man har minst 6 studiepoeng med matematikk på høgskole/universitetsnivå, for eksempel 6 studiepoengskurset i matematikk på Øk.Adm. . Vi skal ikke utlede det som skjer her, men henvise til (for eksempel) Sydsæters Matematisk analyse bind I s 484 for den spesielt interesserte. Imidlertid skal vi påpeke at det som skjer er:
Funksjonen f som altså er en funksjon av 2 variable deriveres nå (partielt) med hensyn på a og på b. Dette gjøres for å bestemme minimum av f. Man beregner mao.

$$\frac{\partial f(a,b)}{\partial a} \quad og \quad \frac{\partial f(a,b)}{\partial b}$$

Deretter settes de deriverte lik 0, dvs. man løser likningene

$$\frac{\partial f(a,b)}{\partial a} = 0 \quad og \quad \frac{\partial f(a,b)}{\partial b} = 0$$

Dette utgjør to likninger med to ukjente (fortsatt a og b) . Løses disse to likningene m.h.p. a og b finner en:

$$\begin{aligned} an \quad & +b(\sum_i x_i) = \sum_i y_i \\ a(\sum_i x_i) + b(\sum_i x_i^2) &= \sum_i x_i y_i \end{aligned}$$

Disse to likningene kalles for **normallikningene** i regresjonsanalyse utledet ved minste kvadraters metode. Grunnen til at metoden kalles **minste kvadraters metode** er man først finner summen av de **kvadrerte** avvikene, og deretter **minimum** av dette. Dette betyr m.a.o.

Eks. Går vi tilbake til vårt talleksempel på s.20 finner en

$$\begin{aligned} & n = 12, \ \sum_i x_i = 1+2+3+.....+9+10 = 65 \\ & \sum_i y_i = 2+3+4+.....+10+12 = 76 \\ & \sum_i x_i^2 = 1^2 + 2^2 + 3^2 + ... + 9^2 + 10^2 = 437 \\ & \sum_i x_i y_i = 1 \cdot 2 + 2 \cdot 3 + 3 \cdot 4 + ... + 9 \cdot 10 + 10 \cdot 12 = 502 \end{aligned}$$

dermed blir normallikningene:

$$\begin{aligned} 12a + 65b &= 76 \\ 65a + 437b &= 502 \end{aligned}$$

Løser en disse m.h.p. a og b (ved en eller annen metode) finner en a = 0,571 og b = 1,064 (med 3 desimalers nøyaktighet)

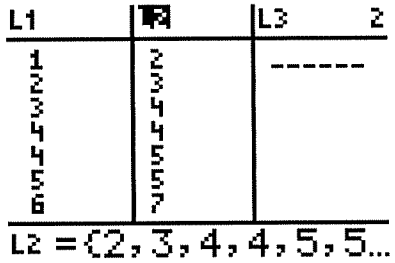
Nå skal vi kontrollere disse beregningene ved hjelp av kalkulatorens statistikkprogram og SPSS.

Regresjonsanalyse ved hjelp av kalkulator:

Med TI-83 gjør man følgende :

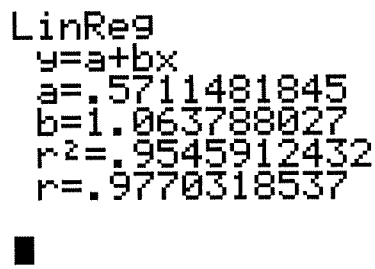
i) Trykk på STAT-tasten.
ii) Trykk på ENTER.

Du er nå klar til å legge inn tallene i liste 1(x-ene) og i liste 2 (y-ene). Legg så inn tallene. Kalkulatoren viser nå:



iii) Trykk så på STAT- tasten på nytt.
iv) Gå så bort til CALC med piltastene.
v) Gå så ned til 8:LinReg(a+bx)
vi) Trykk så ENTER.
vii) Skriv så inn L_1, L_2 rett etter LinReg(a+bx) (ved å trykke 2nd 1 deretter , (kommatasten) og tilslutt 2nd 2
viii) Trykk så ENTER

Du vil nå se at kalkulatoren viser.



M.a.o. vi får bekreftet våre beregninger over og i tillegg noen beregninger knyttet til begrepet korrelasjon som vi kommer til litt senere.

Regresjonsanalyse ved hjelp av SPSS:

i) Trykk på SPSS-ikonet
ii) Skriv inn tallenene kolonne 1 og 2 (spiller ingen rolle hvor x-ene legges, og hvor y-ene legges)
iii) Gå deretter til analyse (se verktøylinjen øverst) og klikk på denne
iv) Gå så ned til regression og velg linear (du kan også velge curve estimation (kommer imidlertid tilbake til denne senere)

Du vil nå få en utskrift som inneholder flere momenter som ennå ikke er omtalt. De fleste av disse skal vi komme tilbake til senere. Den siste av tabellene er den vi skal bruke nå, og den ser ut som følger (selv her er det en del verdier som for øyeblikket vi ikke skal kommentere)

Model		Coefficients Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	,571	,443		1,290	,226
	VAR00001	1,064	,073	,977	14,499	,000

a Dependent Variable: VAR00002

Her ser en at de estimerte koeffisientene er 0,571 og 1,064 for h.h.v. α *og* β , hvilket stemmer med våre tidligere beregninger. I tillegg angis standardfeilen til estimatorene til α *og* β til h.h.v. 0,443 og 0,073. Kolonnen som angir de standardiserte koeffisientene baserer seg på variablene angitt på såkalte standardform (dvs z-scorene som framkommer ved å beregne

$$z = \frac{x - \bar{x}}{s}$$

og tilsvarende for y-ene før analysen gjennomføres. $\bar{x}$ er det aritmetiske gjennomsnittet og s er standardavviket. Det betyr at benevningen i teller og nevner blir like, og dermed blir variablene "dimensjonsløse", dvs de blir uavhengige av de enhetene som brukes.)

Hvis man ser på det generelle likningssystemet på s. 24, og løser dette (generelt) mhp. a og b, så kan det vises at

$$b = \frac{s_{xy}}{s_x^2} \quad og \quad a = \bar{y} - b\bar{x}$$

der en har innført

$$s_{xy} = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y})$$

(som ofte kalles for kovariansen mellom x og y) og

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

som er utvalgsvariansen i x-dataene (se side 14)

Eks. Nå kan det vises (ved å multiplisere ut $(x_i - \bar{x})(y_i - \bar{y})$ og ved å summere leddvis) at

$$s_{xy} = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left[\sum_i x_i y_i - n\bar{x}\,\bar{y}\right]$$

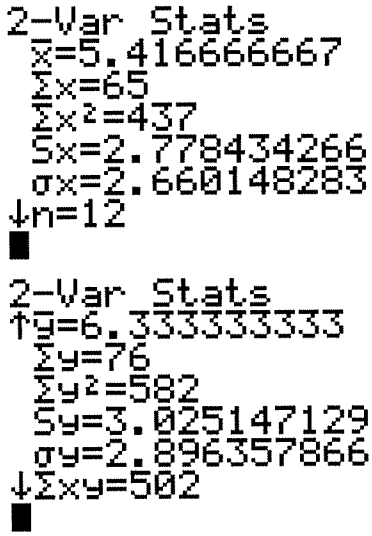
og at

$$s_x^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 = \frac{1}{n-1}\left[\sum_i x_i^2 - n\bar{x}^2\right]$$

Disse reduserte utrykkene gjør det lettere å beregne a og b ved formlene over. Man får ved hjelp av TI-83 og kommandoene

STAT
CALC
2: 2-VAR STAT ENTER
2ND 1, 2ND 2 ENTER

følgende bilde



Herav finner en da greitt

$$s_{xy} = \frac{1}{n-1}\left[\sum_i x_i y_i - n\bar{x}\,\bar{y}\right] = \frac{1}{12-1}\left[502 - 12\cdot\frac{65}{12}\cdot\frac{76}{12}\right] = 8{,}2121\ldots$$

og

$$s_x^2 = \frac{1}{n-1}\left[\sum_i x_i^2 - n\bar{x}^2\right] = \frac{1}{12-1}\left[437 - 12\cdot(\frac{65}{12})^2\right] = 7{,}7196\ldots$$

Dermed har man:

$$b = \frac{s_{xy}}{s_x^2} = \frac{8{,}2121}{7{,}7196} = 1{,}0638.. = 1{,}064$$

og herav:

$$a = \bar{y} - b\bar{x} = \frac{76}{12} - 1{,}064\cdot\frac{65}{12} = 0{,}57$$

9. Enkel korrelasjon

Anta man har n parobservasjoner (x_i, y_i) der x_i er verdien av en tilfeldig variabel X og y_i er verdien av en tilfeldig variabel Y.

x_1	x_2	x_3	.....	x_n
y_1	y_2	y_3	.....	y_n

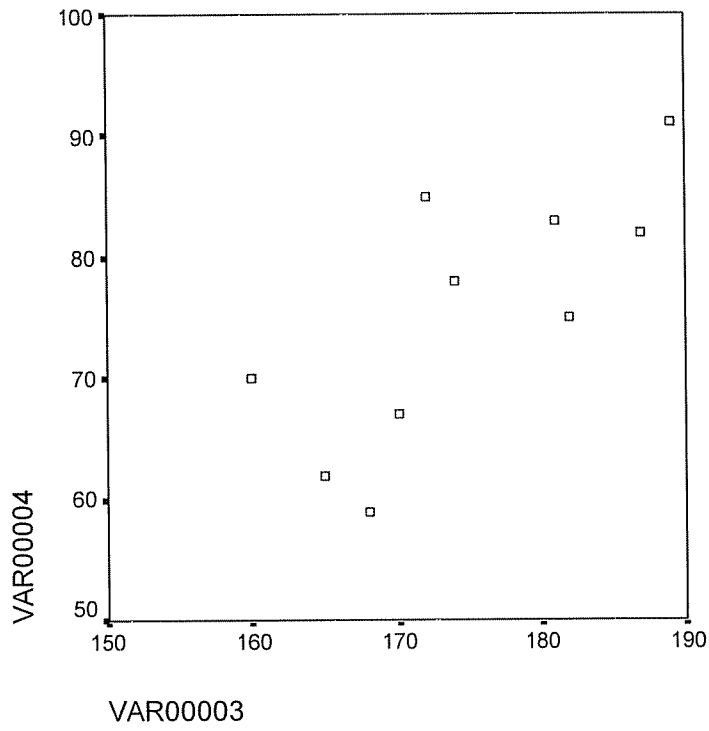
Merk nå at x_i ikke er gitte verdier av X som under regresjon, men verdier av en tilfeldig variabel, og det betyr at de ikke kan bestemmes på forhånd. M.a.o. er nå både x_i *og* y_i verdier av tilfeldige variable.

En sier (litt forenklet) at regresjonslikningen forteller hvordan sammenhengen mellom X og Y er.
Hvor god sammenhengen er, dvs. hvor godt punktene er knyttet til linja måles ved den såkalte **korrelasjonskoeffisienten** r_{xy} som er et tall mellom -1 og 1.
Hvis det gjennomgående er slik at små x-verdier hører sammen med små y-verdier, og store x-verdier hører sammen med store y-verdier, så sier vi at X og Y er **positivt korrelerte** (dvs. at $0 < r_{xy} < 1$

Eks. Vekt og høyde er to variable som er positivt korrelerte. Anta man har målt høyde og vekt hos en tilfeldig valgt gruppe på 10 mennesker og funnet:

x	160	165	189	187	182	168	181	170	174	172
y	70	62	91	82	75	59	83	67	78	85

Spredningsdiagrammet vil da kunne se ut som følger:



Hvis man nå ber SPSS å regne ut korrelasjonskoeffisienten ved kommandoene

ANALYZE
REGRESSION
PEARSON

får man følgende utskrift:

Correlations

		VAR00003	VAR00004
VAR00003	Pearson Correlation	1	,748
	Sig. (2-tailed)	,	,013
	N	10	10
VAR00004	Pearson Correlation	,748	1
	Sig. (2-tailed)	,013	,
	N	10	10

* Correlation is significant at the 0.05 level (2-tailed).

Herav ser en at korrelasjonskoeffisienten er 0,748 (m.a.o. positiv)

Nå skal vi kontrollregne denne utskriften. Det kan vises at korrelasjonskoeffisienten r_{xy} er gitt ved:

$$r_{xy} = \frac{s_{xy}}{s_x s_y} \qquad (*)$$

der s_{xy} *og* s_x er definert som foran på side 28, og s_y er gitt ved

$$s_y^2 = \frac{1}{n-1}\sum_i (y_i - \bar{y})^2$$

Hva tror du at denne størrelsen representerer? (Vink: Sammenlikn med s_x^2)
Utrykket for r_{xy} gitt ved (*) over er dividert med s_x *og* s_y for at r_{xy} skal være et tall mellom -1 og 1. Husk at s_{xy} (=kovariansen mellom X og Y) også måler graden av lineær sammenheng mellom X og Y . Dermed er m.a.o. r_{xy} et standardisert mål på graden av lineær sammenheng.

Nå tilbake til talleksempelet:
En finner her:

$$\sum_i x_i = 160 + 165 + \cdots + 172 = 1748$$

$$\sum_i y_i = 70 + 62 + \cdots + 85 = 75$$

$$s_x^2 = \frac{1}{n-1}\left[\sum_i x_i^2 - n\bar{x}^2\right] = \frac{1}{10-1}\left[306384 - 10\cdot(\frac{1748}{10})^2\right] = 92{,}622..$$

$$s_y^2 = \frac{1}{n-1}\left[\sum_i y_i^2 - n\bar{y}^2\right] = \frac{1}{10-1}\left[57542 - 10\cdot(\frac{752}{10})^2\right] = 110{,}177...$$

$$s_{xy} = \frac{1}{n-1}\sum_i (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n-1}\left[\sum_i x_i y_i - n\bar{x}\,\bar{y}\right] =$$

$$= \frac{1}{10-1}((160\cdot 70 + 165\cdot 62 + \cdots + 172\cdot 85) - 10\cdot\frac{1748}{10}\frac{752}{10}) =$$

$$= \frac{1}{9}\left[132130 - 10\cdot 174{,}8\cdot 75{,}2\right] = 75{,}6$$

Dermed finner en til slutt :

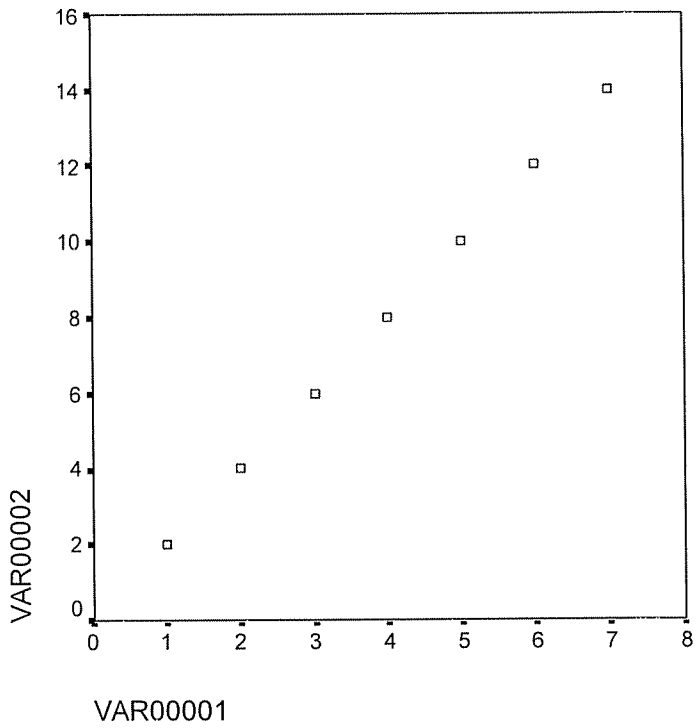
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{75{,}6}{\sqrt{92{,}622}\sqrt{110{,}177}} = 0{,}748$$

som stemmer overens med SPSS-utskriften.

Anta at sammenhengen mellom X og Y er som følger:

x	1	2	3	4	5	6	7
y	2	4	6	8	10	12	14

Spredningsdiagrammet blir dermed som følger:



og man ser at sammenhengen er perfekt hvilket betyr at korrelasjonskoeffisieneten =1. Dette bekrefter også SPSS:

Correlations

		VAR00001	VAR00002
VAR00001	Pearson Correlation	1	1,000
	Sig. (2-tailed)	,	,
	N	7	7
VAR00002	Pearson Correlation	1,000	1
	Sig. (2-tailed)	,	,
	N	7	7

** Correlation is significant at the 0.01 level (2-tailed).

Hvis det gjennomgående er slik at små x-verdier hører sammen med store y-verdier, og store x-verdier hører sammen med små y-verdier så sier vi at det er **negativ korrelasjon** mellom X og Y.

Et eksempel på dette er følgende observerte sammenheng mellom etterspørselen (=y) og prisen (=x) på en vare:

x	86	81	75	90	95	99
y	125	142	150	120	118	115

Correlations

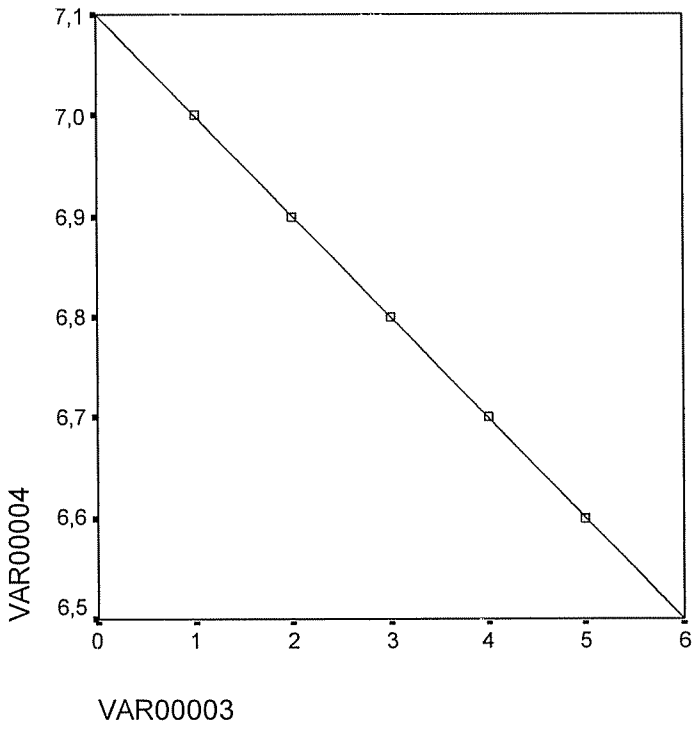
		VAR00003	VAR00004
VAR00003	Pearson Correlation	1	-,954
	Sig. (2-tailed)	,	,003
	N	6	6
VAR00004	Pearson Correlation	-,954	1
	Sig. (2-tailed)	,003	,
	N	6	6

** Correlation is significant at the 0.01 level (2-tailed).

Kontrollregn selv at tallene over stemmer.
En ser m.a.o. at korrelasjonen mellom X og Y er negativ, og nesten lik -1.

Hvis det er perfekt lineær sammenheng mellom to variable og de er negativt korrelerte vil korrelasjonskoeffisienten være nøyaktig lik -1.
Eks.

x	1	2	3	4	5
y	7,0	6,9	6,8	6,7	6,6



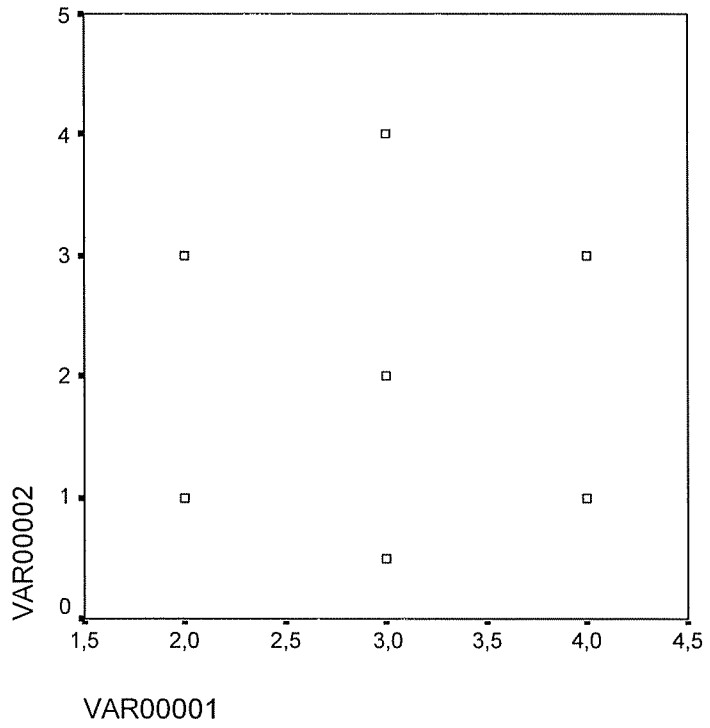
Her har man en perfekt lineær negativ sammenheng, og finner dermed en korrelasjonskoeffisienten lik -1 (Kontroller selv at på etter eller annet vis at dette stemmer)

I de situasjonene vi har sett på her har r vært i nærheten av 1 eller -1, men i mange situasjoner er r i nærheten av 0. Det betyr at utvalget viser at det tyder på at det ikke er noen grad av lineær sammenheng mellom de to variablene som er involvert, og det er en viktig konklusjon og eventuelt komme fram til. Dette skal vi komme nærmere tilbake til senere under avsnittet

om hypotesetestingen. Imidlertid skal vi kort bemerke her at når SPSS-utskriften nederst på side 31 angir en sig.(=signifikanssannsynlighet) på 0.013 betyr det at vi forkaster påstanden om at populasjonskorrelasjonskoeffisienten er 0, og påstår at den er forskjellig fra 0. Risikoen (= sannsynligheten) for at vi tar feil er 0,013.

Eks. Hvis det ikke ser ut til å være noen lineær sammen heng mellom de to variablene, som for eksempel i den observerte sammenhengen:

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



Correlations

		VAR00001	VAR00002
VAR00001	Pearson Correlation	1	,000
	Sig. (2-tailed)	,	1,000
	N	7	7
VAR00002	Pearson Correlation	,000	1
	Sig. (2-tailed)	1,000	,
	N	7	7

En ser av tabellen at det ikke er noen tendens verken i den ene eller andre retningen, og dette bekreftes av tabellen som angir r = 0,000. Et klassisk eksempel på en slik situasjon er sammenhengen mellom skonummer og inntekt.

Man kan imidlertid **ikke konkludere** at det er **årssaksammenheng** mellom to variable selv om man finner en korrelasjonskoeffisient som er signifikant forskjellig fra 0. Det finnes mange eksempler på såkalt nonsenskorrelasjon hvor man kan sette sammen data fra to variable i en tabell og så få regnet ut en korrelasjonskoeffisient. Et par klassiske eksempler på dette er sammenhengen mellom antall barnefødsler og antall registrerte storker i Danmark, eller sammenhengen mellom lærerlønningene i Norge og antall prester på Jamaica.

Hvis man kun har observerte y-verdier og ingen kjennskap til de tilsvarende x-verdiene, og man ønsker å lage en prognose så vil det være naturlig å bruke $\bar{y}$. Har man derimot også de tilhørende x-verdiene, og det er en viss grunn til å tro at det er en sammenheng mellom x og y så kan man bruke denne tilleggskunnskapen til å lage en mye bedre prognose. Anta at (x_i, y_i) er et av de n observasjonsparene, og at sammenhengen mellom x og y er beskrevet ved den estimerte regresjonslinjen for x mhp. y ($\hat{y} = a + b\,x$). Vi definerer nå

det såkalte **totalavviket** ved $y_i - \bar{y}$, (eng.: total deviation)

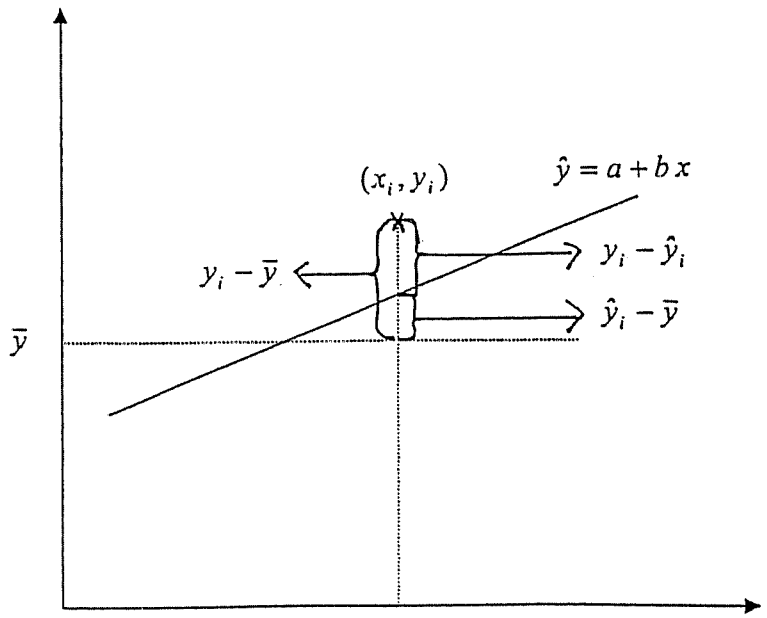
det såkalte **forklarte avviket** ved $\hat{y}_i - \bar{y}$ (eng.: explained deviation) og

det såkalte **uforklarte avviket** $y_i - \hat{y}_i$. (eng.: unexplained deviation)

En ser at

totalavviket = forklart avvik + uforklart avvik (vis dette)

Geometrisk ser dette ut som følger:



Nå kan det vises hvis man i hvert eneste punkt kvadrerer og summerer avvikene over at

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y})^2$$

(hopp gjerne over denne "forklaringen", men det krever ikke så mye matematikk ut over det at $(a-b)^2 = a^2 - 2ab + b^2$ og at summen av flere ledd kan summeres leddvis)

Dette uttrykker en ofte som følger:

Total variasjon = Forklart variasjon + uforklart variasjon

Legg merke til at **kvadrerte avvik** (eng.: squared deviation) betegnes med begrepet **variasjon** (eng.: variation). Den uforklarte variasjonen kalles også ofte **restvariasjon**, og er den delen av totalvariasjonen som ikke blir forklart av regresjonsanalysen. Det er også vanlig å kalle *X*-en for en forklaringsvariabel idet verdien av denne er med på å forklare verdien av *Y*.

En annen måte å tolke korrelasjonskoeffisienten r på er ved følgende sammenheng:
Det kan vises at

$$r^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{Forklart\ variasjon\ i\ y}{Total\ variasjon\ i\ y}$$

Det betyr at den kvadrerte korrelasjonskoeffisienten kommer nærmere og nærmere 1 ettersom den forklarte variasjonen kommer nærmere og nærmere den totale variasjonen, (husk at: Total variasjon = Forklart variasjon + Uforklart variasjon) dvs at den uforklarte variasjonen nærmer seg 0.
r^2 angir dermed et mål på hvor god regresjonsanalysen er, og er et forklaringsmål. På engelsk kalles den ofte for " the coeffesient of determination".

Eks. Vi går nå tilbake til eksempelet på side 28 hvor den observerte sammenhengen mellom x og y var som følger

x	160	165	189	187	182	168	181	170	174	172
y	70	62	91	82	75	59	83	67	78	85

Vi fant her r = 0,748 hvilket gir $r^2 = 0{,}56$. Dette betyr at i denne regresjonsanalysen forklarer X 56% av totalvariasjonen i Y, hvilket igjen betyr at 44% av totalvariasjonen i Y forblir uforklart (skyldes andre faktorer).

Legger man nå inn disse dataene på nytt i SPSS så får man bl.a. følgende utskrifter ved å gjøre en regresjonsanalyse

Coefficients

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-67,475	44,769		-1,507	,170
	VAR00001	,816	,256	,748	3,191	,013

a Dependent Variable: VAR00002

og

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,748[a]	,560	,505	7,38448

a. Predictors: (Constant), VAR00001

Herav ser en bl.a. angivelsen av korrelasjonskoeffisienten R til 0,748 og $R^2 = R\ Square = 0{,}56$ som stemmer med det vi antydet over. I tillegg til dette ser man av utskriften på side 34 at

$$\hat{y} = -67{,}475 + 0{,}816x$$

I tillegg til dette er

$$\bar{y} = \frac{70 + 62 + \cdots + 85}{10} = 75{,}2$$

Dermed er det mulig å beregne forklart variasjon og totalvariasjon ved å sette opp følgende tabell:

x	160	165	189	187	182	168	181	170	174	172
y	70	62	91	82	75	59	83	67	78	85
$y - \bar{y}$	-5,2	-13,2	15,8	6,8	-0,2	-16,2	7,8	-8,2	2,8	9,8
$\hat{y} - \bar{y}$	-12,1	-8,0	11,5	9,9	5,8	-5,6	5,0	-4,0	-0,7	-2,3

Herav finner en

$$\sum_i (y_i - \bar{y})^2 = (-5{,}2)^2 + (-13{,}2)^2 + \cdots + 9{,}8^2 = 991{,}6$$

og

$$\sum_i (\hat{y}_i - \bar{y})^2 = (-12{,}1)^2 + (-8{,}0)^2 + \cdots + (-2{,}3)^2 = 555{,}1$$

(Kontroller beregningene over.)

Dermed blir

$$r^2 = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} = \frac{\textit{Forklart variasjon i } y}{\textit{Total variasjon i } y} = \frac{555{,}1}{991{,}6} = 0{,}56$$

som stemmer med beregningene over.

10. Noen viktige diskrete fordelinger.

Forutsetningen for å kunne forstå dette avsnittet er man har vært innom fordelingene i videregående skoles pensum i sannsynlighetsregning, eller at man har lest sannsynlighetsregning på en høgskole (for eksempel 6 studiepoengskurset i statistikk på Øk.Adm.)

Binomisk fordeling brukes når man gjør n forsøk som oppfyller følgende krav:

i) De n enkeltforsøkene er **u**avhengige.
ii) Det er to mulige utfall (suksess og fiasko) i hvert enkeltforsøk.
iii) p = P(Suksess) er konstant gjennom alle forsøkene.

Mange forsøk i virkelighetens verden oppfyller disse kravene eksakt eller tilnærmet (husk at en modell er en etterlikning og forenkling av virkeligheten)

Hvis variabelen X er definert ved

X = antall suksesser på de n forsøkene

så kan det vises at sannsynlighetsfordelingen til X er gitt ved

$$f_{BIN}(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} \quad , x = 0,1,2,\ldots,n$$

Det betyr at sannsynligheten for at det inntreffer x suksesser på de n forsøkene er gitt ved utrykket over. X betegner variabelen, mens x er en verdi av denne.

Det kan vises at

$$E(X) = np$$

(dvs. forventet antall suksesser = antall forsøk* s.h for suksess i hvert forsøk)
Forventningen til en variabel er det samme som gjennomsnittet i modellen, eller det teoretiske gjennomsnittet.

$$Var(X) = np(1-p)$$

Variansen til en variabel er det samme som spredningen i modellen, eller den teoretiske variansen. Denne måler hvor mye variabelen X i gjennomsnitt avviker fra sitt gjennomsnitt

(= E(X)) med. Mer presist: Var(X) = E(X-E(X))^2 . Grunnen til at man kvadrerer er analogt til tidligere bemerkninger knyttet til den empiriske variansen at E(X-E(X)) alltid blir lik 0.

Eks. Anta at X er binomisk fordelt med n = 100 og p = P(suksess) = 0,4 i hvert enkelt forsøk. Da blir forventet antall suksesser = E(X) = $100 \cdot 0{,}4 = 40$. M.a.o. 40% av antall forsøk er suksesser. Videre blir Var(X) = $100 \cdot 0{,}4 \cdot (1-0{,}4) = 24$.
Sannsynligheten for at det blir akkurat 45 suksesser er da gitt ved

$$P(X=45) = f_{BIN}(45) = \binom{100}{45} 0{,}4^{45}(1-0{,}4)^{100-45} = 0{,}0478$$

som beregnes greit på TI-83 ved kommandoene 2nd VARS 0:binompdf(ENTER100,0.4,45) og sannsynligheten for at det blir høyst 45 suksesser =

$$P(X \le 45) = \binom{100}{0} \cdot 0{,}4^0 \cdot (1-0{,}4)^{100-0} + \cdots + \binom{100}{45} 0{,}4^{45}(1-0{,}4)^{100-45} = 0{,}8689$$

som beregnes ved kommandoene

2nd VARS
A: binomcdf (
ENTER
100,0.4,45)

Hypergeometrisk fordeling brukes når man gjør n forsøk som oppfyller følgende krav:

i) De n enkeltforsøkene er avhengige.
ii) Det er to mulige utfall (suksess og fiasko) i hvert enkeltforsøk.
iv) p = P(Suksess) er konstant gjennom alle forsøkene.

Legg merke til at den eneste forskjellen på binomisk og hypergeometrisk fordelingen er at i binomisk fordeling så er enkeltforsøkene **u**avhengige, mens i hypergeometrisk fordeling så er enkeltforsøkene avhengige. Hypergeometrisk fordeling brukes når man tar tilfeldige utvalg fra populasjon bestående av spesielle elementer (de som har en egenskap/et kjennetegn) og alminnelige elementer (de som ikke har egenskapen/kjennetegnet). Trekker man et spesielt element inntreffer "en suksess", og trekker man et alminnelig element inntreffer "en fiasko". p = P(trekke spesielt element) er konstant i alle forsøkene (skal ikke vises her). Grunnen til at enkeltforsøkene er avhengige er at resultatet i en trekning er avhengig av resultatet i en annen trekning (antall spesielle/alminnelige elementer endres fra trekning til trekning.)

Anta at populasjonen består av N elementer og at a av disse er spesielle. Anta videre at vi tar et tilfeldig ikkeordnet utvalg uten tilbakelegging på n. La X være antall spesielle elementer i utvalget. Da kan det vises at sannsynlighetsfordelingen til X er gitt ved

$$f_{HYP}(x) = P(X = x) = \frac{\binom{a}{x} \cdot \binom{N-a}{n-x}}{\binom{N}{n}} \qquad , x = 0,1,2,\ldots,\min(a,n)$$

Det kan vises at

$$E(X) = n \cdot \frac{a}{N} = n \cdot p \quad der\ p = \frac{a}{N} = P(Trekke\ spesielt\ element) = P(Suksess)$$

Det er m.a.o. ingen forskjell på forventningen i den binomiske og den hypergeometriske fordeling.
Videre kan det vises at

$$Var(X) = \frac{N-1}{N-n} n \frac{a}{N}(1 - \frac{a}{N}) = \frac{N-1}{N-n} np(1-p) \quad der\ p = \frac{a}{N} = P(Trek\ \text{ker}\ spesielt\ element)$$

Denne skiller seg noe fra variansen i den binomiske fordeling hvor Var(X) = np(1-p).
Forskjellen er m.a.o. det såkalte korreksjonsleddet (p.g.a. avhengigheten mellom forsøkene)

$$\frac{N-1}{N-n}$$

Nå har en at

$$\frac{N-1}{N-n} = \frac{1-\frac{1}{N}}{1-\frac{n}{N}} \approx \frac{1-0}{1-0} = 1$$

når N er stor (også i forhold til n). M.a.o. i en slik situasjon er det (tilnærmet) ingen forskjell på binomisk og hypergeometrisk fordeling idet de har samme forventning (teoretiske mål på sentral tendens), og samme varians (dvs. samme teoretiske mål på spredning). En annen måte å se dette på er at hvis man tar et lite utvalg relativt til en stor populasjon, så er det liten sjanse for at man kommer til å trekke det samme elementet to ganger hvis man trekker med tilbakelegging. M.a.o.hvis man har et lite utvalg relativt til en stor populasjon, så kan man nettopp av denne grunn si at man har uavhengighet (tilnærmet), og m.a.o. en binomisk situasjon (tilnærmet) istedenfor den eksakte hypergeometriske.

Hvis n er liten i forhold til N (som er stor) så har en m.a.o. at

$$f_{HYP}(x) \approx f_{BIN}(x)$$

(En vanlig tommelfingerregel er at man bør ha $n \leq \frac{1}{10} N$ der N er stor)
Vi skal se senere i våre anvendelser at man bruker den tilnærmede binomiske fordeling eller vel så ofte normalfordelingen som vi skal se på i neste avsnitt.

Eks. Anta at vi har et vareparti bestående av N = 500 elementer og at a = 10 av disse er defekte. Det blir tatt et tilfeldig ikkeordnet utvalg uten tilbakelegging på n = 25 elementer. Hvis 1 eller flere av disse er defekte så forkastes varepartiet. La nå X = antall defekte elementer i utvalget. Vi ønsker nå å finne $P(X \geq 1)$
Regner vi nå først eksakt (hypergeometrisk) finner man:

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{\binom{10}{0}\binom{490}{25}}{\binom{500}{25}} = 1 - 0{,}596 = 0{,}404$$

Hvis man nå isteden bruker den binomiske fordelingen, dvs. antar at X er binomisk fordelt med n = 25 og $p = P(Suksess) = \frac{a}{N} = \frac{10}{500} = 0{,}02$
Da har en (tilnærmet):

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \binom{25}{0} 0{,}02^0 (1 - 0{,}02)^{25-0} = 1 - 0{,}98^{25} = 0{,}397$$

M.a.o. man har en feil av størrelsesorden 0,404-0,397 = 0,007 som må sies å være bra.
I mange sammenhenger er populasjonen mange ganger større og utvalget av en størrelsesorden på ca 1000 (jfr. galluper) Da blir feilene betydelig mindre enn dette.

Poissonfordelingen er en tredje og siste av de diskrete fordelingene vi skal se litt på. Vi skal ikke bruke den i dette heftet, men det kan tenkes du kan komme bort i den i en artikkel eller få bruk for den senere i en eller annen analyse.
En måte (ofte brukt) å presentere Poissonfordelingen på er å si at den er "i slekt med" den binomiske fordelingen der n er stor og p = P(Suksess) er liten. For eksempel hvis man tenker seg et flyselskap som gjør n = 10000 flygninger årlig, og hvor P(Nestenulykke i en tilfeldig flygning) = 0,0005. Da er forventet antall nestenulykker = np = $10000 \cdot 0{,}0005 = 5$ pr år. Ofte er kun denne intensiteten som betegnes med λ kjent (m.a.o. her er $\lambda = np = 5$).

Anta at X er Poissonfordelt med parameter λ. Da kan det vises at sannsynlighetsfordelingen til X er gitt ved:

$$f_{POI}(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad ,x = 0, 1, 2, 3,\ldots$$

Det kan vises at

$$f_{BIN}(x) \xrightarrow[np=\lambda]{n\to\infty \,\&\, p\to 0} f_{POI}(x)$$

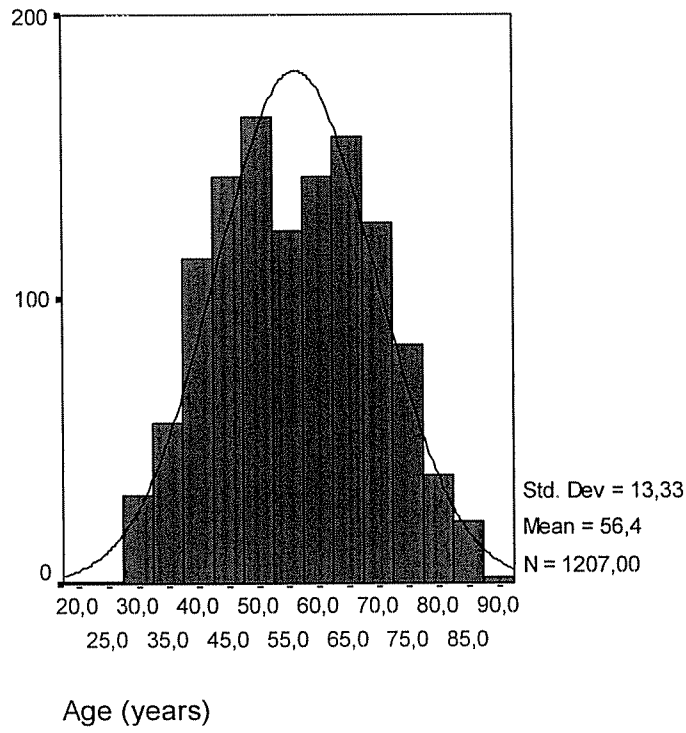
Det betyr at binomiske sannsynligheter kan regnes ut ved hjelp av Poissonfordelingen. Bruker vi eksempelet over (hvor riktignok ikke forutsetningene med en stor n er oppfylt) hvor altså $\lambda = np = 25 \cdot \frac{10}{500} = 0{,}5$ finner man

$$P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{0{,}5^0}{0!} e^{-0{,}5} = 1 - e^{-0{,}5} = 0{,}393$$

som sies å være en brukbar tilnærmelse til den eksakte binomiske beregningen som altså ga svaret 0,397. M.a.o. en har en feil på 0,004.

11. Noen viktige kontinuerlige fordelinger. Sentralgrense-teoremet

Normalfordelingen er uten tvil den viktigste av de kontinuerlige fordelingene, for ikke å si den viktigste av alle fordelinger. Det skyldes primært at mange variable i praksis viser seg å være normalfordelt eller tilnærmet normalfordelt. Mange variable kan dessuten gjennom transformasjoner bli normalfordelte. Normalfordelingen kalles også ofte Gaussfordelingen etter den tyske matematiker og filosof Carl Friedrich Gauss (1777-1855). Mange diskrete og kontinuerlige fordelinger kan tilnærmes med normalfordelingen (under gitte betingelser). Mange av de testene og estimeringsteknikkene som vi skal se på senere forutsetter at populasjonen er normalfordelt. Av denne grunn blir også normalfordelingen meget sentral i dette heftet.



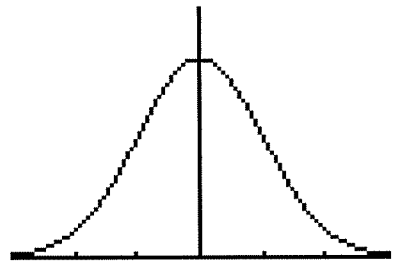
tillegg er det lagt inn en normalfordeling med et gjennomsnitt på 56,4 år og et standardavvik på 13,33 år. Om dataene kan sies å være et utvalg fra en normalfordelt populasjon eller ikke skal vi senere teste.
Vi skal nå bare innledningsvis se på noen egenskaper knyttet til normalfordelingen.

- Normalfordelingskurven er en entoppet symmetrisk glatt kurve. Toppunktet er ved μ som er gjennomsnittet i den teoretiske fordelingen (dvs. målet på sentraltendens i fordelingen)

  Ved hjelp av kalkulatoren kan en tegne normalfordelingen ved hjelp av kommandoene

  $Y=$
  2ND VARS
  1:normalpdf(
  X
  GRAPH

  Det er her viktig å passe på å la x gå mellom -3 og +3, mens y går mellom 0 og 0,5. En vil da få følgende bilde:



- X er normalfordelt med parametre $\mu\ (= E(X))$ *og* $\sigma(= \sqrt{Var(X)}\ )$

  $\Updownarrow$

  Sannsynlighetstettheten f til X er gitt ved $f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ der $-\infty < x < \infty$

  Dette skrives ofte:

  $$X \sim N(\mu, \sigma)$$

  I grafen over er $\mu = 0$ og $\sigma = 1$

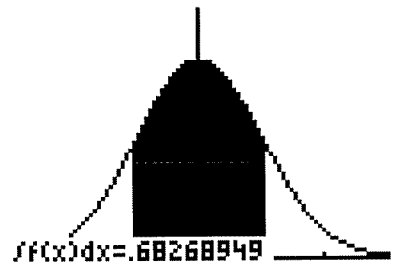
- Sannsynligheten for at utvalg av normalfordelte data skal ligge mellom det teoretiske gjennomsnittet minus et standardavvik og det teoretiske gjennomsnittet pluss et standardavvik er tilnærmet 0,68. Det vil m.a.o. si at når man har normalfordelte data så

vil ca. 68% av de som er med i undersøkelsen falle innefor det nevnte intervallet over. Mer presist :

Hvis X er normalfordelt med parametre $\mu\ (= E(X))$ *og* $\sigma(= \sqrt{Var(X)}\ )$ så er

$P(\mu - \sigma < X < \mu + \sigma) = 0{,}6827$

TI 83 viser nå

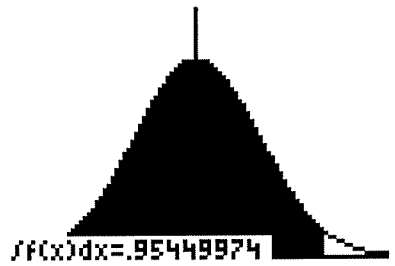


- Sannsynligheten for at utvalg av normalfordelte data skal ligge mellom det teoretiske gjennomsnittet minus to standardavvik og det teoretiske gjennomsnittet pluss to standardavvik er tilnærmet 0,95. Det vil m.a.o. si at når man har normalfordelte data så vil ca. 95% av de som er med i undersøkelsen falle innefor det nevnte intervallet over. Mer presist :

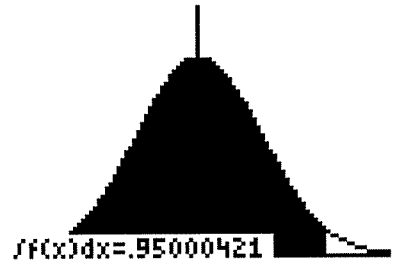
  Hvis X er normalfordelt med parametre $\mu\ (= E(X))$ *og* $\sigma(= \sqrt{Var(X)}\ )$ så er

  $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0{,}9545$.

  TI-83 viser nå.



Hvis man stedet for å gå 2 standardavvik går 1,96 standardavvik vil sannsynligheten bli presis 0,9500. Denne kunnskapen vil man få bruk for senere både under estimeringen og hypotesetestingen. TI-83 viser nå



- Hvis man spesielt har $\mu = 0$ *og* $\sigma = 1$ så fremkommer den såkalte standardnormalfordeling som er en normalfordeling som ligger symmetrisk omkring 2.-aksen. Det er vanlig å betegne en standardnormalt fordelt variabel med Z, og man skriver ofte

$$Z \sim N(0,1)$$

Tabeller over denne fordelingen finner man i de fleste statistikkbøker. Sammenhengen mellom en variabel $X \sim N(\mu, \sigma)$ og en variabel $Z \sim N(0,1)$ er gitt som følger:

$$Z = \frac{X - \mu}{\sigma}$$

Det betyr m.a.o. at $Z \sim N(0,1)$ fremkommer ved at en vilkårlig normalfordelt X **standardiseres** (reduseres med μ *og divideres med* σ. Av resultatene over må man dermed ha at

$$P(-1 < Z < 1) = 0{,}6827$$

og

$$P(-2 < Z < 2) = 0{,}9545$$

Dette kontrollerer man lett på kalkulatoren ved kommandoene:

2nd VARS
2:ENTER
normalcdf(-1,1)
ENTER

som gir 0,682689...

og helt analogt gir normalcdf(-2,2) = 0,954499...

Eks. Normalfordelingen kan også med stor grad av nøyaktighet brukes til å beregne diskrete sannsynligheter. Anta at X er binomisk fordelt med n = 100 og p = 0,4 (se side 40)
Vi beregnet her eksakt $P(X \leq 45) = 0{,}8689$. Tilnærmet har vi nå ved hjelp av normalfordelingen:

$$P(X \leq 45) \approx P(Z \leq \frac{45 - \mu}{\sigma}) = P(Z \leq \frac{45 - 100 \cdot 0{,}4}{\sqrt{100 \cdot 0{,}4 \cdot (1 - 0{,}4)}}) = P(Z \leq 1{,}02) = 0{,}8461$$

Bruker man i tillegg den såkalte 0,5-korreksjonen finner man

$$P(X \leq 45) \approx P(Z \leq \frac{45 + 0{,}5 - 100 \cdot 0{,}4}{\sqrt{100 \cdot 0{,}4 \cdot (1 - 0{,}4)}}) = P(Z \leq 1{,}12) = 0{,}8686$$

som kun har en feil på 0,8689-0,8686 = 0,0003.

Det går også an å angripe tilnærmingsberegningene direkte uten "å gå veien om Z" (den standardiserte variable). Ved hjelp av kalkulatoren har man da (med 0,5-korreksjonen):

$$P(X \leq 45) \approx Normalcdf(-10^{99},\, 45.5, 100 \cdot 0.4, \sqrt{100 \cdot 0.4 \cdot (1-0{,}4)}) = 0{,}8686$$

der -10^{99} er det største negative tallet kalkulatoren kan greie. Det skulle egentlig være $-\infty$.

Helt tilsvarende kan normalfordelingen brukes til å beregne tilnærmede verdier for den hypergeometriske fordeling, i Poissonfordelingen, og egentlig i alle fordelinger som har en form som likner på normalfordelingen.

Den omvendte situasjonen er også viktig. Dvs. Hvilken z-verdi svarer til gitt en sannsynlighet på 0,05? Dette og liknende problemer (sannsynligheten er risikoen for å gjøre feil) blir sentrale i estimeringsteorien og hypoteseprøving.
En bruker nå enten en "omvendt " normalfordelingstabell eller kalkulatorens invNorm:

2nd VARS
3: ENTER
invNorm(0.05)
ENTER

som gir verdien -1,6449.

Tilsvarende finner invNorm(0,95) = 1,6449 og invNorm(0,99) = 2,3263.

En svært viktig setning som legger grunnlaget for en rekke anvendelser av normalfordelingen er den såkalte

Sentralgrensesetningen: Anta at $X_1, X_2, \ldots, X_n$ er uavhengige stokastiske variable med samme sannsynlighetsfordeling. Anta dessuten at forventningen μ og standardavviket σ i populasjonen eksisterer. La nå

$$S_n = X_1 + X_2 + \cdots X_n$$

Da kan det vises at :

i) S_n er tilnærmet normalfordelt med forventning $n\mu$ og standardavvik $\sqrt{n}\sigma$ når n er stor. Dermed at den standardiserte variable

$$\frac{S_n - n\mu}{\sqrt{n}\sigma} \sim N(0,1) \text{ (tilnærmet)}$$

En har også at :

ii) $\overline{X} = \frac{X_1 + X_2 + \cdots X_n}{n} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ (tilnærmet) når n er stor. Dermed vil den standardiserte variable

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \text{ (tilnærmet)}$$

M.a.o. En sum av stokastiske variable (alle med samme sannsynlighetsfordeling) vil (uansett hva slags fordeling de følger) være tilnærmet normalfordelt bare n er tilstrekkelig stor. Det samme gjelder dermed også for gjennomsnittet. Matematisk er dette et ganske tungt bevis som vi selvfølgelig ikke skal ta her. Det vises også at når $n \to \infty$ så vil resultatene over være eksakte.
Dette er en setning som har stor nytteverdi idet man i mange sammenhenger ikke kjenner populasjonsfordelingen, men bruker ofte metoder hvor normalfordeling er en forutsetning. Det finnes også mer generelle versjoner av setningen.

t-fordelingen:

Hvis $X \sim N(\mu, \sigma)$ så er

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

I de fleste situasjoner så er imidlertid σ ukjent og må dermed estimeres. Til dette formålet brukes utvalgsstandardavviket s gitt ved

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

Betrakter en nå

$$\frac{\overline{X} - \mu}{s/\sqrt{n}}$$

så kan det vises at denne er tilnærmet normalfordelt når n er stor, men den er såkalt t-fordelt med v = n-1 frihetsgrader når n er liten. Den betegnes med t. En har m.a.o. at

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

Det kan vises at t-fordelingen er en en-toppet symmetrisk kurve som har samme symmetripunkt (=0) som normalfordelingen, og at t-fordelingen nærmer seg normalfordelingen når n vokser og blir stor. Sannsynlighetstettheten til t er gitt ved

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\,\Gamma(\frac{\nu}{2})}(1+\frac{t^2}{\nu})^{-\frac{\nu+1}{2}}$$

der Γ er den såkalte gammafunksjonen som er gitt ved et integralutrykk. Det kan imidlertid vises at

$$\Gamma(x+1) = x\Gamma(x) \quad \forall x > 0$$

og at

$$\Gamma(n) = (n-1)! \quad \forall n \in N$$

Dessuten er

$$\Gamma(\frac{1}{2}) = \sqrt{\pi} \quad og \quad \Gamma(2) = \Gamma(1) = 1$$

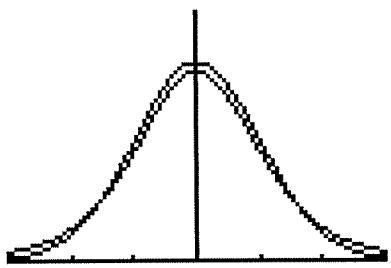
Dette gjør det lettere å finne utrykket for t-fordelingen når v er gitt. Anta f..eks. at antall frihetsgrader v = 5. Da blir sannsynlighetstettheten til t

$$f(t) = \frac{\Gamma(\frac{5+1}{2})}{\sqrt{\pi 5}\,\Gamma(\frac{5}{2})}(1+\frac{t^2}{5})^{-\frac{5+1}{2}} = \frac{8}{3\sqrt{5\pi}}(1+\frac{t^2}{5})^{-3}$$

Hvis du tegner denne sammen med den standardnormale kurve så vil du se at t-fordelingen har litt tyngre "haler" enn normalfordelingen, men at de to kurvene for øvrig ikke er så veldig forskjellige selv med kun v=5 frihetsgrader, dvs. med kun 6 forsøk. Imidlertid kan man ved å overse den forskjellen som allikevel er der ved et lite antall forsøk fort komme til å trekke motsatt konklusjon når man lager konfidensintervaller eller gjennomfører hypoteseprøving.
På kalkulatoren ligger både normalfordelingen og t-fordelingen (se: 2nd VARS 1:normalpdf og 4:tpdf) . Hvis man ønsker å tegne normalfordelingen og for eksempel t-fordelingen med 5 frihetsgrader så går en inn på Y= og skriver:

$$Y_1 = normalpdf(X)$$
$$Y_2 = tpdf(X,5)$$

Trykker en nå på GRAPH vil en få følgende grafiske bilde:



Normalfordelingen er den øverste av de to kurvene nær 0, og den som fortest nærmer seg 1.- aksen.

Lar en nå antall frihetsgrader v øke så vil en se hvorledes t-fordelingen nærmer seg normalfordelingen. I mange tabeller så stopper v på 30. Dvs at når n>30 så kan en like gjerne bruke normalfordelingen istedenfor t-fordelingen. Tegner du

$$Y_1 = normalpdf(X)$$
$$Y_2 = tpdf(X,30)$$

så vil du skjønne hvorfor. En skal imidlertid være oppmerksom på at det selv opp mot v=100 er forskjeller på de to fordelingene. Vi kommer mer tilbake til dette senere under estimering og hypotesetesting.
Arealer under de to kurvene er viktige innenfor dette området. Bruker en nå kalkulatoren så har en for eksempel

i) Under normalfordelingen:
$$P(X \geq 1.96) = Normalcdf(1.96,10\text{^}99) = 0.0249978\ldots = 0{,}0250$$

ii) Under t-fordelingen med 30 frihetsgrader:
$$P(t \geq 1.96) = tcdf(1.96,10\text{^}99,30) = 0.0296711\ldots = 0.0297$$

Vi har m.a.o. kun en forskjell på 0.0297-0.0250 = 0,0047. Dette kan imidlertid ha en viss betydning hvis man skal multiplisere med et stort standardavvik.

χ^2 (kjikvadrat)-fordelingen.

Den tredje og siste av de kontinuerlige fordelingene som vi skal nevne er kjikvadratfordelingen.

i) Den er nyttig når man tar utvalg fra normalfordelte (eller tilnærmet normalfordelte) populasjoner.
ii) Den kan brukes til å teste om dataene kommer fra bestemte fordelinger. (for eksempel: Er våre data normalfordelte?)
iii) Den kan brukes til å teste eventuell uavhengighet mellom variable.

Det kan vises at hvis

iv) $X \sim N(0,1)$ så er $X^2 \sim \chi_1^2$ (kjivadratfordelt med 1 frihetsgrad), og hvis
v) $X_1, X_2, \cdots, X_n$ er uavhengige og standardnormalfordelte variable ($dvs.\ X_i \sim N(0,1)$, $i = 1, 2, \ldots, n$) så er
$X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi_n^2$ (kjikvadratfordelt med n frihetsgrader)
vi) Hvis $X \sim \chi_v^2$ så er sannsynlighetstetthetsfunksjonen til X gitt ved

$$f(x) = \begin{cases} \dfrac{1}{2^{v/2}\Gamma(\frac{v}{2})} x^{\frac{v-2}{2}} e^{-\frac{v}{2}} & ,x > 0 \\ 0 \quad ellers \end{cases}$$

der Γ er den såkalte gammafordelingen gitt på side 49.

Hvis for eksempel $\nu = 6$ (= antall frihetsgrader) så er

$$f(x) = \frac{1}{2^{\frac{6}{2}}\Gamma(\frac{6}{2})} x^{\frac{6-2}{2}} e^{-\frac{6}{2}} = \frac{1}{16} x^2 e^{-3} \quad , x > 0$$

som er en entoppet høyreskjev kurve som lett lar seg tegne ved hjelp av kalkulatoren. For øvrig ligger også kjikvadratfordelingen på kalkulatoren. Ønsker man å tegne kjikvadratfordelingen over med 6 frihetsgrader direkte så bruker en som før

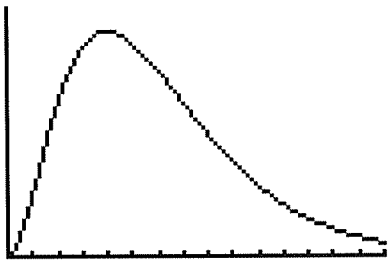
$Y =$

2nd VARS

6: $\chi^2 pdf(X,6)$

GRAPH

så vil en få følgende grafiske bilde på kalkulatoren:



Når n vokser vil formen på kjikvadratfordelingen bli mer og mer lik normalfordelingen. Det kan vises at hvis X er kjikvadratfordelt med ν frihetsgrader så vil en ha at for store ν så er tilnærmet

$$X \sim N(\nu, \sqrt{2\nu})$$

Prøv for eksempel med $\nu = 50$ å tegne χ^2 -fordelingen og normalfordelingen sammen. La

$$Y_1 = \chi^2 pdf(X,50)$$

og

$$Y_2 = Normalpdf(X,50,10)$$

En ser da at det er helt ubetydelige forskjeller på de to kurvene.

Arealer under kjikvadratfordelingen, og verdier på 1.aksen med gitte arealer (les sannsynligheter) er viktige innenfor områdene estimering og hypotesetesting. Anta at X er kjikvadratfordelt med 20 frihetsgrader. Da er for eksempel eksakt:

$$P(X \geq 30) = \chi^2 cdf(30,10\text{^}99,20) = 0,06998...$$

Bruker en isteden normalfordelingen finner en tilnærmet:

$$P(X \geq 30) = Normalcdf(30,10\text{^}99,20) = 0,05692...$$

12. Estimering.

Punktestimering.

Anta at θ er en ukjent størrelse i populasjonen (dvs en ukjent størrelse som inngår i den sannsynlighetsfordelingen som gjelder for populasjonen). En slik størrelse kalles **en parameter.** Det kan for eksempel være andelen i populasjonen som er for EU, andelen defekte i et vareparti, alkoholkonsentrasjonen i blodet (noen timer etter en fest) osv.... En slik størrelse er det ofte ønskelig å kunne anslå, dvs det vi i statistikken kaller å **estimere.** Når vi skal estimere en parameter så betrakter vi en stokastisk variabel, dvs en variabel viss verdier i det lange løp (hvis vi gjør flere forsøk) vil treffe det ukjente tallet θ som vi er på jakt etter. En slik variabel kalles for **en estimator**, og betegnes med $\hat{\Theta}$ (les "teta hatt". Se det greske alfabetet). Når forsøket er gjennomført kan vi beregne verdien av estimatoren $\hat{\Theta}$, som kalles for **estimatet** for θ, og betegnes med $\hat{\theta}$.

En god estimator $\hat{\Theta}$ for θ er slik at

$$E(\hat{\Theta}) = \theta \ .$$

Den kalles da for en forventningsrett estimator.Dvs. at gjennomsnittsverdien av $\hat{\Theta}$ er i det lange løp lik θ.
Og den skal dessuten være slik at

$Var(\hat{\Theta})$ er så liten som mulig.

Dvs at spredningen/usikkerheten knyttet til $\hat{\Theta}$ er så liten som mulig. Det finnes forventningsrette estimatorer som har mindre varians enn alle andre forventningsrette estimatorer (i hele universet). Slike estimatorer kalles for "**minimum variance unbiased estimator"**

Anta at $X \sim bin(100, p)$ og at vi ønsker å estimere p som står for sannsynligheten for at det inntreffer en suksess i hvert av de 100 forsøkene. Under avsnittet om binomisk fordeling så nevnte vi at $E(X) = np$. Herav kan det vises at

$$E(\frac{X}{n}) = p$$

I tillegg kan det vises at

$$Var(\frac{X}{n}) = \frac{p(1-p)}{n}$$

og at denne variansen er mindre enn variansen til alle andre forventningsrette estimatorer for p .

$\hat{P} = \frac{X}{n}$ er m.a.o. den beste forventningsrette estimatoren for p som finnes.

Hvis man nå observerer $X = 38$ så er dermed et forventningsrett estimat for p

$$\hat{p} = \frac{38}{100} = 0{,}38.$$

Intervallestimering.

Når man angir ett tall som estimat for den ukjente parameteren så sier vi at vi punktestimerer. Nå er det imidlertid mye vanligere å intervallestimere. Dvs å angi et intervall $[a,b]$ på tall-linja som med en viss sannsynlighet* inneholder den ukjente parameteren θ. Denne sannsynligheten kalles for konfidenskoeffisienten. *Mer presist så representerer denne sannsynligheten metodens pålitelighet idet en parameter ikke er en variabel, og således ikke kan ha en sannsynlighet knyttet til seg. (dette gjøres imidlertid i såkalt Bayesiansk statistikk som vi ikke skal komme inn på her)

Hvis man ønsker å angi et intervallestimat i en binomisk situasjon så kan det vises at

$$\frac{x}{n} \pm 1{,}96\sqrt{\frac{\frac{x}{n}(1-\frac{x}{n})}{n}} = \hat{p} \pm 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \hat{p} \pm 1{,}96SE(\hat{p})$$

danner utgangspunktet for å lage et 95% konfidensintervall for p. Legg merke til at konfidensintervallet består av punktestimatet for p pluss/minus et såkalt feilledd som inneholder tallet 1,96 (arealet under den normalfordelte kurve mellom -1,96 og +1,96 er akkurat 0,95) og et estimat for standardavviket til estimatoren

$$\hat{P} = \frac{X}{n}$$

$SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ er et estimat for den såkalte standardfeilen (standarderror) eller standarsavviket til til $\hat{P} = \frac{X}{n}$.

I en gallup utført av MMI for Dagbladet 18.desember 2004 blant 862 stemmeberettigede er et estimat for Arbeiderpartiets oppslutning 24,8% og for venstres oppslutning 2,7%. 95% konfidensintervall for Arbeiderpartiets og Venstres oppslutning blir da henholdsvis

$$0{,}248 \pm 1{,}96\sqrt{\frac{0{,}248(1-0{,}248)}{862}} = 0{,}248 \pm 0{,}015$$

$$0{,}027 \pm 1{,}96\sqrt{\frac{0{,}027(1-0{,}027)}{862}} = 0{,}027 \pm 0{,}006$$

Legg merke til at pluss-minus-leddet er størst for Arbeiderpartiet. Det skyldes at $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ vokser med $\hat{p}$ så lenge $\hat{p}$ er mellom 0 og 0,5 (deretter avtar den når $\hat{p}$ er mellom 0,5 og 1,0)

Det kan m.a.o. vises at et 95% konfidensintervall for en vilkårlig parameter θ alltid er på formen

$$\hat{\theta} \pm 1{,}96\sigma(\hat{\theta}) = \hat{\theta} \pm 1{,}96SE(\hat{\theta})$$

m.a.o. estimatet for θ pluss-minus 1,96 multiplisert med standardavviket til estimatoren. Den siste skrivemåten den siste skrivemåten er den mest brukte av de to.
Forutsetningen er at $\hat{\Theta}$ kan tilnærmes med normalfordelingen. $SE(\hat{\theta})$ er verdien av $SE(\hat{\Theta})$, som er standardavviket til estimatoren $\hat{\Theta}$. I de fleste situasjoner er $SE(\hat{\Theta})$ ukjent, og må derfor estimeres. Det betyr at konfidensintervallet antar formen

$$\hat{\theta} \pm 1{,}96\hat{SE}(\hat{\theta})$$

der $\hat{SE}(\hat{\theta})$ er et estimat for $SE(\hat{\Theta})$. Hvis man ønsker en større konfidenskoeffisient (for eksempel 99%) så erstattes 1,96 med 2,58 (=invNorm(0,0005)) og intervallet blir selvfølgelig bredere.
I en del situasjoner følger ikke $\hat{\Theta}$ normalfordelingen, men en annen fordeling som for eksempel t-fordelingen eller kjikvadratfordelingen. Dermed må man finne fram den tilsvarende fraktilen under den gitte fordelingen (dette gjelder spesielt når n er liten). Konfidensintervallet blir dermed på formen

$$\hat{\theta} \pm f_{\alpha/2}\,\hat{SE}(\hat{\theta})$$

der $f_{\alpha/2}$ er $(1-\alpha/2)100\%$ – fraktilen i den aktuelle fordelingen, dvs. den verdien på 1.-aksen som gir et areal på $(1-\alpha/2)$ under kurven til venstre for denne (eller ekvivalent den verdien på 1.-aksen som gir et areal på $\alpha/2$ under kurven til høyre for denne). Det betyr at arealet mellom $-f_{\alpha/2}$ og $f_{\alpha/2}$ er presis $(1-\alpha)$ som er lik sannsynligheten som angir metodens pålitelighet.

P.g.a. sentralgrensesetningen så kan imidlertid normalfordelingen brukes når n er stor i de fleste situasjoner.

Hvis man ønsker å finne konfidensintervall for differansen mellom to andeler i populasjonen, $p_1 - p_2$, så har en nå ifølge resultatene over følgende 95% konfidensintervall for $p_1 - p_2$:

$$\hat{\theta} \pm 1{,}96\hat{\sigma}(\hat{\theta}) = \hat{p}_1 - \hat{p}_2 \pm 1{,}96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

der $\hat{p}_1$ og $\hat{p}_2$ er forventningsrette estimater for henholdsvis p_1 og p_2, n_1 og n_2 er antall elementer i de to uavhengige utvalgene.

Anta for eksempel at man er interessert i å finne et konfidensintervall for differansen mellom andelen menn (= p_1) og andelen kvinner (= p_2) i Norge for EU. Anta man har to uavhengige utvalg på henholdsvis $n_1 = 425$ menn og $n_2 = 450$ kvinner, og man fant at andelen menn for EU i utvalget var $\hat{p}_1$=0,53, og den tilsvarende andelen kvinner var $\hat{p}_2$ = 0,44. Da har man følgende utgangspunkt for å finne 95% konfidensintervall for $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \pm 1{,}96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 0{,}53 - 0{,}44 \pm 1{,}96\sqrt{\frac{0{,}53(1-0{,}53)}{425} + \frac{0{,}44(1-0{,}44)}{450}}$$

$=0{,}09 \pm 0{,}03.$

Dvs. at konfidensintervallett blir $[0{,}06;0{,}12]$. Mer presist: Vår metode påstår med en sikkerhet på 95% at differansen mellom andelen menn og andelen kvinner for EU ligger mellom 6% og 12%. Vi sier også dermed at det er en signifikant forskjell mellom andelen menn og andelen kvinner som er for EU. Se for øvrig avsnittet om hypoteseprøving.

Konfidensintervall i enkel regresjon.

På side 21 og 22 så vi på den enkle regresjonsmodellen

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, 2, ..., n$$

der $e_1, e_2, ..., e_n$ er n uavhengige feilledd som har

$$\text{forventning} = 0 \text{ og varians} = \sigma^2$$

Likningen $\mu_{Y|x} = \alpha + \beta x$ som vi kalte for populasjonsregresjonslikningen for Y m.h.t. x. estimerte vi ved hjelp av et utvalg av n observasjonspar. Vi fant da en såkalt estimert regresjonslikning eller en utvalgsregresjonslikning som vi betegnet ved

$$\hat{y} = a + b x$$

ved hjelp av den såkalte minste kvadraters metode.

Nå kan det vises at estimatoren $\hat{B} = \frac{S_{xY}}{S_x^2}$ (se side 28) som brukes til å estimere β (og dermed gir estimatet b) har følgende egenskaper :

$$E(\hat{B}) = \beta$$

og

$$SE(\hat{B}) = \frac{\sigma}{\sqrt{S_{xx}}}$$

Herav ser en altså at minste kvadraters estimatoren $\hat{B}$ er en forventningsrett estimator for β (m.a.o. verdien b av $\hat{B}$ treffer gjennomsnittlig i det lange løp den sanne ukjente verdi β). Dessuten ser en at jo større spredning det er på x – ene (m.a..o. jo større S_{xx} er) jo mindre blir standardfeilen til $\hat{B}$, og jo mer pålitelig blir dermed denne. Nå er også σ en ukjent verdi på lik linje med β og den må derfor også estimeres.
Nå kan det vises at den såkalte maksimum likelihood-estimatet $\hat{\sigma}$ for σ er gitt ved:

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2}$$

som altså måler gjennomsnittlig kvadrert avvik i den avhengige variable i alle de gitte punktene.
En ser at

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2$$

Dette er ikke en forventningsrett estimator for σ^2 . Det er derimot

$$\hat{\sigma}_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2 \ (*)$$

En har dermed at

$$\hat{\sigma}_e^2 = \frac{n}{(n-2)}\sigma^2$$

Grunnen til at man deler på (n-2) i utrykket (*) er at det er 2 størrelser som skal estimeres ved hjelp av tallmaterialet, nemlig α *og* β. Det medfører at tallmaterialet mister "kraft" slik at antall frihetsgrader nå bare er (n-2). (Jfr. utvalgsvariansen hvor antall frihetsgrader er (n-1). (Se side 14)
Nå kan det også vises at

$$t = \frac{\hat{\beta} - \beta}{\hat{\sigma}}\sqrt{\frac{(n-2)S_{xx}}{n}} = \frac{\hat{\beta} - \beta}{\hat{\sigma}_e}\sqrt{S_{xx}} = \frac{\hat{\beta} - \beta}{\frac{\hat{\sigma}_e}{\sqrt{S_{xx}}}} = \frac{\hat{\beta} - \beta}{SE(\hat{\beta})}$$

er verdien av en tilfeldig variabel som er t-fordelt med (n-2) frihetsgrader.

Ved hjelp av kunnskapen over kan det utledes et $(1-\alpha)100\%$ konfidensintervall for β. Det kan vises at dette blir:

$$\hat{\beta} - t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}} < \beta < \hat{\beta} + t_{\alpha/2,n-2}\hat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}}$$

eller tilsvarende ved at man bruker $\hat{\sigma}_e^2 = \frac{n}{(n-2)}\hat{\sigma}^2$

$$\hat{\beta} - t_{\alpha/2,n-2}\frac{\hat{\sigma}_e}{\sqrt{S_{xx}}} < \beta < \hat{\beta} + t_{\alpha/2,n-2}\frac{\hat{\sigma}_e}{\sqrt{S_{xx}}} \quad (*)$$

Vi går nå tilbake til eksempelet på side 28 og 35 hvor vi hadde følgende sammenhengende verdier mellom X og Y og hvor det var gjort en del beregninger:

x	160	165	189	187	182	168	181	170	174	172
y	70	62	91	82	75	59	83	67	78	85

Vi fant blant $s_x^2 = 92{,}622..$ og herav har en $S_{xx} = (n-1)s_x^2 = (10-1)92{,}622..=833{,}598..$
I tillegg fant vi den forklarte variasjonen i y til 555,1 og den totale variasjonen til 991,6. I og med at en har

Totalvariasjon = Forklart variasjon + Uforklart variasjon (Se side 34)

så finner en den uforklarte variasjonen = $\sum_{i=1}^{n}(y_i - \hat{y})^2 = 991{,}6\text{-}555{,}1 = 436{,}5$ og herav

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2} = \sqrt{\frac{1}{10}436{,}5} = 6{,}607$$

Av tabellen over t-fordelingen finner en $t_{\alpha/2,n-2} = t_{0.025,8} = 2{,}306 = 2{,}5\%$ - fraktilen i t-fordelingen med 10-2 = 8 frihetsgrader. Dermed blir 95% konfidensintervall for β idet $\hat{\beta} = 0{,}816$ (se side 35)

$$0{,}816 - 2{,}306 \cdot 6{,}61\sqrt{\frac{10}{(10-2)833{,}60}} < \beta < 0{,}816 + 2{,}306 \cdot 6{,}61\sqrt{\frac{10}{(10-2)833{,}60}}$$

$$0{,}226 < \beta < 1{,}406$$

Legger en nå inn tallene i tabellen i SPSS og bruker kommandoene

ANALYZE
REGRESSION
LINEAR (legger så tallene i kolonne 1 og 2)
STATISTICS
CONFIDENCE INTERVAL

får en bl.a. følgende utskrift

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-67,475	44,769		-1,507	,170	-170,712	35,762
	VAR00001	,816	,256	,748	3,191	,013	,226	1,406

a Dependent Variable: VAR00002

som stemmer helt overens med beregningene over.

Legg også merke til at

$$\hat{\sigma}_e = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y})^2} = \sqrt{\frac{1}{8}436{,}5} = 7{,}387$$

og dermed at

$$\frac{\hat{\sigma}_e}{\sqrt{S_{xx}}} = \frac{7{,}387}{\sqrt{833{,}6}} = 0{,}256$$

som man også finner igjen i tabellen (kolonne 4 siste rad). Dette er den såkalte standardfeilen til $\hat{B}$, Std.Error($\hat{B}$), eller kortere SE($\hat{B}$). Dermed vil konfidensintervallet (*) bli på formen:

$$\hat{\beta} - t_{\alpha/2,n-2}SE(\hat{\beta}) < \beta < \hat{\beta} + t_{\alpha/2,n-2}SE(\hat{\beta})$$

13. Hypotesetesting.

Hypotesetesting (eller hypoteseprøving) går ut på å avgjøre om en påstand skal forkastes eller ikke.
Påstanden som testes kalles for **nullhypotesen** og betegnes med H_0
H_0 testes alltid mot en **alternativ hypotese** H_A (eller H_1), dvs. disse to påstandene stilles alltid opp mot hverandre.
Som H_0 brukes nesten* alltid den påstanden som man ikke har tro på ("vil ha brent opp"), mens H_A er den påstanden som man har tro på. Hvis vi forkaster H_0 så påstår vi at H_A er riktig med en viss (liten) sannsynlighet for å ta feil.
* I visse typer modellkontroll, for eksempel at man har tro på at dataene er normalfordelte, så lar en H_0 være: dataene er normalfordelt og håper på at H_0 ikke skal forkastes, dvs. en håper på å finne en stor P-verdi.

Eks. I USA hadde man rundt 1985 framstilt en medisin, AZT,(bestående av bl.a. en bestemt type sopp) som man mente virket mot AIDS. N=285 pasienter med langt

fremskreden AIDS ble delt tilfeldig i to grupper. n = 143 pasienter fikk AZT, og de resterende N-n = 142 pasientene fikk en narremedisin. Ingen av pasientene og ingen av de som behandlet pasientene visste hvem som var Behandlingsobjekter og hvem som var Kontrollobjekter. Dette kaller man i statistikken å gjøre ***dobbelte blindforsøk***.
Før forsøket ble gjennomført formulerte man følgende H_0 og H_A:

H_0: AZT virker ikke mot AIDS og
H_A: AZT virker mot AIDS

Etter 6 mnd var 17 pasienter døde og koden ble brutt. Da viste det seg at man hadde fått følgende resultater (i dette meget brutale forsøket). Husk imidlertid at man på den tiden ikke hadde noen medisin mot AIDS, og at alle de 285 pasientene var så syke at de var oppgitt av helsevesenet):

Res. etter 6mnd Gruppe	Død	I live	SUM
Fikk AZT	1	142	143
Fikk narremed.	16	126	142
SUM	17	268	285

En ser altså at av de 17 som var døde var hele 16 i kontrollgruppa (de som fikk narremedisin).
Når dette ble oppdaget så lot man umiddelbart alle pasientene få AZT. Det viste seg imidlertid etter ytterligere noen måneder at AZT ikke hadde noen helbredende effekt mot AIDS. Medisinen hadde kun den effekt at den midlertidig bremset opp utviklingen av AIDS.
For å kunne gjennomføre hypoteseprøving trenger man data, og i den sammenheng observerer vi en stokastisk variabel X som vi kaller for **testobservatoren** . For å kunne gjøre beregninger må man kjenne fordelingen til X. I eksempelet over hvor dataene er gitt er det naturlig å la X = antall pasienter som er i live etter 6 mnd av de som fikk AZT (eller X = antall pasienter som er døde etter 6 mnd av de som fikk AZT. Dette kan en velge fritt. Beregningene videre blir imidlertid litt forskjellige)
Hvis det er rimelig å forkaste H_0 når $X \geq k$ så sier vi at **store verdier av X er signifikante**, hvis det er rimelig å forkaste H_0 når $X \leq k$ så sier vi at **små verdier av X er signifikante**. Valget av X vil avgjøre om store eller små verdier er signifikante. I begge tilfeller kaller vi k for **den kritiske verdien**. Hvis vi i eksempelet over velger X = antall pasienter som er i live etter 6 mnd av de som fikk AZT, så er store verdier av X signifikante (idet det er rimelig å forkaste påstanden om at AZT ikke har noen virkning og påstå at den har virkning når det er mange av behandlingsobjektene som er i live etter 6 mnd) Velger vi isteden X = antall pasienter som er døde etter 6 mnd av de som fikk AZT, så er små verdier av X signifikante. Det betyr lite beregningsmessig om man velger den ene eller den andre testobservatoren.

Kritisk verdi k bestemmes slik at sannsynligheten for å gjøre feil blir liten. Mer presist betyr det at sannsynligheten for å forkaste H_0 når H_0 er riktig er liten blir liten. Vanligst valg av denne sannsynligheten som betegnes med α er 0,05. Hvis testobservatoren er slik at H_0 forkastes når $X \geq k$ (store verdier av X er signifikante) så bestemmes altså k slik at

$$P_{H_0}(X \geq k) = \alpha$$

Indeksen H_0 signaliserer at sannsynligheten skal beregnes når H_0 er riktig. Noen skriver denne sannsynligheten $P(X \geq k|H_0)$ der $|$ er det vanlige symbolet som brukes når man skal regne ut betingede sannsynligheter(les "gitt H_0").

α kalles for **testens signifikansnivå** (eller bare kortere testens **nivå**). Noen ganger sier man også at α betegner sannsynligheten for å gjøre **feil av type I**.

Man kan også gjøre en annen feil ved hypoteseprøving : Akseptere (godta) H_0 når H_0 er gal (dvs når H_A er riktig). Denne feilen betegnes med β og kalles for **feil av type II.**

Hvis store verdier av X er signifikante, dvs. at H_0 forkastes hvis $X \geq k$, som igjen betyr at H_0 aksepteres hvis $X < k$ så er β gitt ved:

$$\beta = P_{H_A}(X < k)$$

En må vurdere i hvert enkelt tilfelle hvilken feil som er viktigst å unngå.

Eks. Hvis man har plukket sopp som man har stor tro på er spiselig vil det være naturlig å teste

$$H_0: \text{Soppen er giftig}$$

mot

$$H_A: \text{Soppen er spiselig}$$

I denne situasjonen vil feil av type I (forkaste H_0 når H_0 er riktig) medføre at man at påstår at soppen er spiselig når den er giftig. Feil av type II (akseptere H_0 når H_0 er gal (dvs når H_A er riktig)) vil medføre at man sier at soppen er giftig når den er spiselig. En ser her at det er viktigst å unngå feil av type I.

En annen viktig sannsynlighet knyttet til hypoteseprøving er den såkalte **styrken**. Denne angir sannsynligheten for å forkaste H_0 når H_0 er gal (dvs H_A er riktig). Denne sannsynligheten betegnes med π og bør være så stor som mulig. En har hvis store verdier av X er signifikante at

$$\pi = P_{H_A}(X \geq k)$$

Nå er

$$\beta = P_{H_A}(X < k)$$

Dermed ser en at $\pi = 1 - \beta$

π vil (som β) være avhengig av forskjellige verdier av parameteren under alternativet. En betrakter derfor ofte den såkalte **styrkefunksjonen** $\pi(\theta)$. Denne kan framstilles grafisk med θ langs førsteaksen og styrken π langs andreaksen. Dette er da en kurve som kan brukes til å lese av styrken for enhver ønskelig verdi under alternativet. Jo brattere kurven går jo fortere stor blir dermed π og jo større er dermed sannsynligheten for å oppdage at H_0 ikke gjelder. Styrkefunksjonen kalles derfor for av og til for **oppdagelsesfunksjonen.** I noen situasjoner så brukes styrken i et gitt punkt sammen med nivået til å bestemme kritisk verdi k og antall forsøk n en trenger gjøre.

Eks. Går en nå til gallupeksempelet på side 48 hvor Venstre har en oppslutning på 2,7% i desembermålingen, mens de i valget 2001 hadde en oppslutning på 3,9% kan man stille spørsmålet om Venstre har hatt en signifikant tilbakegang fra populasjonsandelen på 0,039 på 5%-nivået.

Vi tester derfor

$$H_0 : p = 0{,}039$$

mot

$$H_A : p < 0{,}039$$

Både spørsmålet og H_0 og H_A bør formuleres før en ser dataene, ellers driver en fort såkalt **datafisking.**

La nå X = antall personer (av de n = 862) som har stemt Venstre. Små verdier av X er signifikante. Dvs. at H_0 forkastes hvis $X \le k$. Kritisk verdi k bestemmes slik at

$$P_{H_0}(X \le k) = 0{,}05$$

For så store tall er det vanligst å bruke normaltilnærmelsen (og regne tilnærmet)

$$P(X \le k) \approx P(Z \le \frac{k + 0{,}5 - 862 \cdot 0{,}039}{\sqrt{862 \cdot 0{,}039 \cdot (1 - 0{,}039)}}) = 0{,}05 \, (egentlig \le 0{,}05)$$

Herav får en da følgende likning

$$\frac{k + 0{,}5 - 862 \cdot 0{,}039}{\sqrt{862 \cdot 0{,}039 \cdot (1 - 0{,}039)}} = -1{,}645 (egentlig \le -1{,}645)$$

dvs. at

$$k \le 862 \cdot 0{,}039 - 0{,}5 - 1{,}645 \cdot \sqrt{862 \cdot 0{,}039 \cdot (1 - 0{,}039)} = 23{,}8$$

Det betyr mao. at nullhypotesen forkastes hvis

$$X = \text{antall personer (av de n} = 862) \text{ som har stemt Venstre} \le 23$$

(siden X må være et heltall og nivået ikke skal overstige 0,05). Det betyr at hvis $X \le 23$ så forkastes påstanden om at $p = 0{,}039$, og man påstår at $p < 0{,}039$ (venstre har fått en

redusert oppslutning i populasjonen). Sannsynligheten for at man tar feil er <0,05 (=nivået)
Nå er det faktisk akkurat 23 personer (av de 862) som stemmer Venstre. Det betyr at H_0 forkastes på 5%-nivået. Regner man isteden ut P-verdien har en i dette tilfellet at denne blir

$$P_{H_0}(X \le 23) = P_{H_0}(Z \le \frac{23 + 0{,}5 - 862 \cdot 0{,}039}{\sqrt{862 \cdot 0{,}039 \cdot (1 - 0{,}039)}}) = 0{,}0375 < 0{,}05$$

$H_0 : p = 0{,}039$ forkastes på 5%-nivået. Nivået på testen er da egentlig 0,0375.

Hvis man ikke vet noe om Venstre har fått en tilbakegang eller framgang før man ser tallene er det naturlig å teste

$$H_0 : p = 0{,}039$$

mot

$$H_A : p \neq 0{,}039$$

dvs. at man tester med tosidig alternativ. Da vil H_0 forkastes enten hvis k blir liten eller k blir stor dvs.

$$k \le 862 \cdot 0{,}039 - 0{,}5 - 1{,}96 \cdot \sqrt{862 \cdot 0{,}039 \cdot (1 - 0{,}039)} = 21{,}98$$

eller

$$k \ge 862 \cdot 0{,}039 + 0{,}5 + 1{,}96 \cdot \sqrt{862 \cdot 0{,}039 \cdot (1 - 0{,}039)} = 45{,}25$$

Det betyr at H_0 forkastes hvis $X \le 22$ eller $X \ge 46$.

Hvis vi nå går tilbake igjen til den første situasjonen med ensidig testing hvor en hadde $H_A : p < 0{,}039$ og antar (for eksempel) at $p = 0{,}03$ så vil sannsynligheten for å akseptere H_0 når H_0 er gal bli

$$\beta = P_{H_A}(X > 23) \approx P(Z > \frac{23 - 0{,}5 - 862 \cdot 0{,}03}{\sqrt{862 \cdot 0{,}03 \cdot (1 - 0{,}03)}}) = 0{,}7488... = 0{,}75$$

M.a.o en ganske stor sannsynlighet for å begå feil av type II. En ser også at styrken $\pi = 1 - \beta = 0{,}25$. Det er m.a.o. ikke så stor sannsynlighet for å oppdage at venstre har fått en tilbakegang når det i virkeligheten har skjedd.

Kjikvadrattester.

Vi skal nå se på en type tester som kan brukes til avgjøre om en statistisk modell (en forenkling og etterlikning av virkeligheten) er god/brukbar eller ikke ved hjelp av data.

Anta at en person skal kaste en mynt 100 ganger.

Vi tester da

H_0 : Modellen for myntkast (P(M) = P(K) = 0,5) er holdbar

mot

H_A : Modellen for myntkast (P(M) = P(K) = 0,5) er ikke holdbar.

Anta at man nå observerer

	Mynt	Kron	Sum
O_i	60	40	100

Testen går nå ut på å sammenlikne de observerte verdiene (O_i) med hva man kan forvente å få (E_i) hvis modellen er holdbar (brukbar). Idet antall suksesser (kron eller mynt) er binomisk fordelt med n=100 og P(suksess)=0,5 så har følgende forventede verdier:

	Mynt	Kron	Sum
$E_i = np_i$	50	50	100

Som testobservator skal vi i slike tester bruke

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

som måler avvikene mellom O_i og E_i. En ser at hvis forskjellene mellom O_i og E_i blir store så vil også χ^2 bli stor. Vi har derfor at store verdier χ^2 er signifikante. Ma.o. H_0 forkastes hvis $\chi^2 \geq k_{\alpha,(m-1)}$ = kritisk verdi på nivået α og med υ = (m-1) frihetsgrader (=antall mulige utfall i modellen -1). Velger en $\alpha = 0{,}05$ finner en av tabellen over kjikvadratfordelingen at $k_{\alpha,(m-1)} = k_{0.05,(2-1)} = 3{,}841$. M.a.o. H_0 forkastes hvis beregnet $\chi^2 \geq 3{,}841$.

En finner nå

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} = \frac{(60-50)^2}{50} + \frac{(40-50)^2}{50} = 2{,}0 + 2{,}0 = 4{,}0$$

Konklusjonen blir dermed at H_0 forkastes på 5%-nivået.

Velger en alternativt å beregne P-verdien i forsøket finner en av TI-83:

$$P(\chi^2 \geq 4{,}0) = \chi^2 cdf(4.0, 10^{99}, 1) = 0{,}0455$$

som er mindre enn 0,05 og dermed har en (selvfølgelig) samme konklusjon.

Kjikvadrattester kan også brukes til å teste uavhengighet mellom to variable X og Y som ofte presenteres i en $r \times c$-tabell.

Uavhengighetstesten baserer seg også på å sammenlikne observerte og forventede verdier, og dermed på

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i}$$

som testobservator.

Eks. Anta at man ønsker å teste om det er noen sammenheng mellom alderen på en person og meningen om et produkt. Vi ønsker derfor å teste

H_0 : Det er ingen sammenheng mellom alder og mening om produktet

mot

H_A : Det er sammenheng mellom alder og mening om produktet

120 tilfeldig valgte personer ble spurt om hva de synes om produktet. Anta at resultatet av undersøkelsen ble (de forventede verdiene står i parentes)

Mening→ Alder↓	Dårlig	Middels	Bra	SUM
20-30 år	10(16,9)	15(13,9)	20(14,3)	45
30-40 år	14(12,8)	9(10,5)	11(10,8)	34
40-50 år	21(15,4)	13(12,6)	7(13,0)	41
SUM	45	37	38	120

Her ser en for eksempel at i aldersgruppen 20 til 30 år så er det $O_1 = 10$ personer som mener at produktet er dårlig. I denne gruppen kan vi forvente $E_1 = 45 \cdot \frac{45}{120} = 16{,}9$ personer forutsatt at nullhypotesen er riktig.....osv. I aldersgruppen 40 til 50 år så er det $O_9 = 7$ personer som mener at produktet er bra. I denne gruppen kan vi forvente $E_9 = 438 \cdot \frac{41}{120} = 13{,}0$ personer forutsatt at nullhypotesen er riktig. Bregner en så kjikvadrattallet for hele matrisen finner en

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} = \frac{(10-16{,}9)^2}{16{,}9} + \cdots + \frac{(7-13{,}0)^2}{13{,}0} = 10{,}4$$

Herav finner en følgende P-verdi ved TI-83 (idet store verdier av testobservatoren χ^2 er signifikante)

$$P_{H_0}(\chi^2 \geq 10{,}4) = \chi^2 cdf(10.4,\ 10^{99},\ 4) = 0{,}034$$

idet en her har

d.f.=(ant.rader-1)(ant.kolonner -1) = (3-1)(3-1) = 4

Siden P-verdien = 0,034 < 0,05 ser vi at resultatet er signifikant på 5%-nivået. Konklusjonen blir m.a.o at H_0 forkastes til fordel for H_A.

Testing i den enkle regresjonsmodellen.

På side 24 og 25 så vi på den enkle regresjonsmodellen

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, 2, \ldots, n$$

der $e_1, e_2, \ldots, e_n$ er n uavhengige feilledd som har

$$\text{forventning} = 0 \text{ og varians} = \sigma^2$$

Likningen $\mu_{Y|x} = \alpha + \beta x$ som vi kalte for populasjonsregresjonslikningen for Y m.h.t. $X = x$. estimerte vi ved hjelp av et utvalg av n observasjonspar. Vi fant da en såkalt estimert regresjonslikning eller en utvalgsregresjonslikning som vi betegnet ved

$$\hat{y} = a + b x$$

ved hjelp av den såkalte minste kvadraters metode.

Ønsker en nå å teste

$$H_0 : \beta = \beta_0$$

mot

$$H_A : \begin{cases} \beta < \beta_0 \ eller \\ \beta > \beta_0 \ eller \\ \beta \neq \beta_0 \end{cases}$$

så bruker en testobservatoren T viss verdier er gitt ved

$$t = \frac{\hat{\beta} - \beta}{\hat{\sigma}} \sqrt{\frac{(n-2)S_{xx}}{n}} \quad (*)$$

Bruker en igjen eksempelet (fra side 30)med sammenhørende verdier mellom vekt (=Y) og høyde (=X), der vi blant annet under avsnittet om konfidensintervaller beregnet følgende størrelser:

$$S_{xx} = (n-1)s_x^2 = (10-1)92{,}622.. = 833{,}598..$$

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y})^2} = \sqrt{\frac{1}{10}436{,}5} = 6{,}607$$

Hvis man nå ønsker å teste

$$H_0 : \beta = 0 \text{ (Det er } \textbf{ingen sammenheng} \text{ mellom X og Y)}$$

mot

$$H_A : \beta > 0 \text{ (Det er en ("positiv") sammenheng mellom X og Y)}$$

så blir

$$t = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}} \sqrt{\frac{(n-2)S_{xx}}{n}} = \frac{0{,}816 - 0}{6{,}607} \sqrt{\frac{(10-2)833{,}598}{10}} = 3{,}189$$

Beregner en isteden t ved formelen

$$t = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_e} \sqrt{S_{xx}} = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} \text{ der } SE(\hat{\beta}) = \frac{\hat{\sigma}_e}{\sqrt{S_{xx}}}$$

(det siste uttrykket er kanskje det mest brukte av de to, og dette skal vi komme tilbake til senere)
En finner nå av det nye utrykket

$$t = \frac{\hat{\beta} - \beta_0}{SE(\hat{\beta})} = \frac{0{,}816 - 0}{0{,}256..} = 3{,}189$$

Dermed blir P-verdien

$$P_{H_0}(t \geq 3{,}189) = tcdf(3.189, 10^{99}, 8) = 0{,}0064$$

og man ser at resultatene er signifikante på 1%-nivået. Dvs at $H_0 : \beta = 0$ (Det er ingen sammenheng mellom X og Y) forkastes på 1%-nivået, og man påstår $H_A : \beta > 0$ (Det er en ("positiv") sammen heng mellom X og Y).

I SPSS finner en nå ved hjelp av kommandoene

ANALYZE
REGRESSION
LINEAR (henter så tallene i kolonne 1 og 2)
STATISTICS

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-67,475	44,769		-1,507	,170
	VAR00001	,816	,256	,748	3,191	,013

a Dependent Variable: VAR00002

Herav ser en at t-verdien blir 3,191 som stemmer godt overens med 3,189 over. Det som imidlertid ikke stemmer så godt overens er P-verdien = Sig.= 0,013. Dette forklares imidlertid greitt når man får vite at SPSS tester tosidig, dvs.

$$H_0 : \beta = 0 \text{ mot } H_A : \beta \neq 0 \text{ (det er en sammenheng mellom X og Y)}$$

og dermed blir P-verdien (pga. symmetri)

$$2 \cdot P_{H_0}(t \geq 3{,}57) = 2 \cdot tcdf(3.189, 10^{99}, 8) = 2 \cdot 0{,}0064 = 0{,}0128 = 0{,}013$$

Testing vedrørende korrelasjonskoeffisienten.

I mange sammenhenger er man interessert i å teste hypoteser knyttet til forskjellige verdier av korrelasjonskoeffisienten mellom to variable X og Y. Dvs. man tester

$$H_0 : \rho = 0 \text{ (Det er ingen sammenheng mellom X og Y)}$$

mot

$$H_A : \rho \begin{cases} > 0 \\ < 0 \\ \neq 0 \end{cases}$$

hvor de tre forskjellige alternativene representerer hhv. "Det er positiv korrelasjon mellom X og Y", "det er negativ korrelasjon mellom X og Y" og "det er ingen korrelasjon mellom X og Y".

Hvis vi igjen ser på tallmaterialet fra side 28, og legger dette inn i SPSS og bruker kommandoene

ANALYZE
CORRELATE
BIVARIATE

med alternativene Pearson Correlation Coefficient og 2-tailed (2-halet (les 2-sidig)) significance test. Dette betyr at man ønsker å teste

$$H_0 : \rho = 0 \text{ (Det er ingen korrelasjon mellom X og Y)}$$

mot

$$H_A : \rho \neq 0 \text{ (Det er korrelasjon mellom X og Y)}$$

Man får da følgende utskrift:

Correlations

		VAR00003	VAR00004
VAR00003	Pearson Correlation	1	,748
	Sig. (2-tailed)	,	,013

	N	10	10
VAR00004	Pearson Correlation	,748	1
	Sig. (2-tailed)	,013	,
	N	10	10

* Correlation is significant at the 0.05 level (2-tailed).

Herav ser man at korrelasjonskoeffisienten mellom X og Y er 0,748, og at den 2-sidige testen gir en signifikanssannsynlighet på 0,013 som er presis samme signifikanssannsynlighet som den man fant på side 66 hvor man testet

$$H_0 : \beta = 0 \text{ mot } H_A : \beta \neq 0 \text{ (det er en sammenheng mellom X og Y)}$$

Hvordan kan det ha seg slik? Jo, korrelasjonskoeffisienten måler jo nettopp graden av sammenheng (lineær) mellom to variable. Disse nullhypotesene er mao. helt ekvivalente.

At det virkelig er slik ser en av utrykkene for b og r_{xy} (se side 28 og 31)

$$b = \frac{s_{xy}}{s_x^2} \qquad r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Av det første uttrykket finner en nå

$$s_{xy} = b \cdot s_x^2$$

som så settes inn i det andre utrykket og gir

$$r_{xy} = b \cdot \frac{s_x}{s_y}$$

Mao.: Herav ser en at hvis $r_{xy} = 0$ så er $b = 0$ og omvendt. ($s_x > 0$ *og* $s_y > 0$)

14. Multippel regresjon.

2 forklaringsvariable.

Anta nå (i motsetning til enkel regresjon) at Y er avhengig av 2 forklaringsvariable (prediktorer) X_1 *og* X_2. Dvs . at vi nå har n observasjonstripler (x_{1i}, x_{2i}, y_i) der x_{1i} er en gitt verdi av en tilfeldig variabel X_1, x_{2i} er en gitt verdi av en tilfeldig variabel X_2 og y_i er verdien av en tilfeldig variabel Y. Tolket på en slik måte at når $X_1 = x_{11}$ og $X_2 = x_{21}$, så blir $Y = y_1, \ldots\ldots$, når $X_1 = x_{1n}$ og $X_2 = x_{2n}$, så blir $Y = y_n$

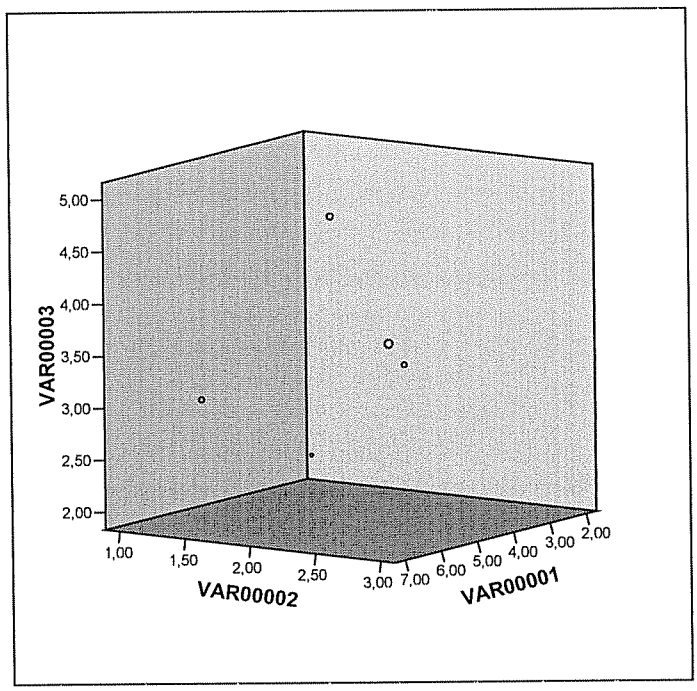
x_{11}	x_{12}	x_{13}	.....	x_{1n}
x_{21}	x_{22}	x_{23}	.....	x_{2n}
y_1	y_2	y_3	.....	y_n

Avsetter man punktene (x_{1i}, x_{2i}, y_i), i = 1,2,3,....,n i et xyz-koordinatsystem får en et spredningsdiagrammet (scatterplot) som ligger i rommet.

Anta at man har observert følgende sammenheng mellom X_1, X_2 og Y.

X_1	2	3	5	5	7
X_2	1	2	2	1	3
Y	2,1	3,2	4,8	2,9	3,9

Ber man nå SPSS om å tegne dette i et tredimensjonalt scatterplot får en følgende grafiske bilde:



Kommandoene som brukes er:

GRAPHICS
SCATTER
3-D
DEFINE

Velg så tallene i rad 1 som $X (= X_1)$, tallene i rad 2 som $Z (= X_2)$ og tallene i rad 3 som Y.

En antar nå (som i enkel regresjon) at det er en lineær sammenheng mellom X_1, X_2 og Y. Dette kan da beskrives ved følgende modell (husk at en modell er en etterlikning og forenkling av virkeligheten):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon \quad \text{der } \varepsilon \text{ er } N(0,\sigma^2) \quad (*)$$

ε kalles som tidligere støyen (eller feilleddet, eng.:the error) og som altså antas å være normalfordelt med forventning 0 og med en varians σ^2 (se normalfordelinge s.40). Modellen (*) over gjelder selvfølgelig for alle n observasjonsparene. Ofte beskrives modellen derfor noe mer presist som følger:

Den tilfeldige variable Y (gitt de tilsvarende x-ene) er uavhengige med

$$\text{forventning} = \mu_{Y|x} = E(Y|X = x) \;\; \beta_0 + \beta_1 x_1 + \beta_2 x_2 \text{ og}$$
$$\text{varians} = \sigma^2$$

Likningen $\mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ kalles ofte for populasjonsregresjonslikningen for Y m.h.t. x_1 *og* x_2. Denne skal vi prøve å estimere ved hjelp av et utvalg av n observasjonstripler (x_{1i}, x_{2i}, y_i), i=1,2,...,n. Vi kan da finne en såkalt estimert regresjonslikning (utvalgsregresjonslikning) som betegnes med

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

Dette er likningen for et plan som ligger i rommet. Denne vil kunne brukes til å estimere verdier av Y, dvs. å lage prognoser.
b_0, b_1 *og* b_2 er estimater for henholdsvis β_0, β_1 *og* β_2. Disse finner en ved hjelp av **minste kvadraters metode** (som i enkel regresjon) , som vi husker går ut på først å beregne avvikene

$$e_i = \textit{observert } y\textit{ verdi} - \textit{estimert } y\textit{ verdi} = y_i - \hat{y}_i \text{ for all de n punktene}$$

Geometrisk betyr det nå at vi prøver å finne det planet som passer best mulig til de n punktene.
I hvert eneste punkt i rommet så beregnes avviket mellom den observerte y-verdien og y-verdien til det ukjente planet $\hat{y} = b_0 + b_1 x_1 + b_2 x_2$ (det som er ukjent er b_0, b_1 *og* b_2). Man beregner m.a.o. avvikene

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i}) \text{ for i = 1, 2, 3,.....,n}$$

Deretter beregner en summen av de kvadrerte avvikene $\sum_i e_i^2$ som blir en funksjon *f, av* b_0, b_1 *og* b_2. Mao. man beregner :

$$f(b_0, b_1, b_2) = \sum_i e_i^2 = \sum_i (y_i - \hat{y}_i)^2 = \sum_i (y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i}))^2$$

Som i enkel regresjon så beregner en minimum av dette utrykket ved matematisk å derivere f (partielt) mhp. b_0, b_1 *og* b_2 og sette disse lik 0. M.a.o:

$$\frac{\partial f(b_0,b_1,b_2)}{\partial b_0}=0, \quad \frac{\partial f(b_0,b_1,b_2)}{\partial b_1}=0 \quad og \quad \frac{\partial f(b_0,b_1,b_2)}{\partial b_2}=0$$

Dette resulterer i følgende 3 likninger med de 3 ukjente b_0, b_1 *og* b_2:

$$\begin{aligned} b_0 n \quad &+b_1(\sum_i x_1) \quad +b_2(\sum_i x_2) \quad =\sum_i y \\ b_0(\sum_i x_1)+b_1(\sum_i x_1^2) \quad &+b_2(\sum_i x_1x_2)=\sum_i x_1y \\ b_0(\sum_i x_2)+b_1(\sum_i x_1x_2)+b_2(\sum_i x_2^2) \quad &=\sum_i x_2y \end{aligned}$$

Hvis man bruker tallene fra eksempelet over finner en følgende likningssett

$$\begin{aligned} b_0 \cdot 5 \quad &+b_1 \cdot 22 \quad +b_2 \cdot 9 \quad =16{,}9 \\ b_0 \cdot 22 &+b_1 \cdot 112 + b_2 \cdot 44 = 79{,}6 \\ b_0 \cdot 9 \quad &+b_1 \cdot 44 \quad +b_2 \cdot 19 \quad =32{,}7 \end{aligned}$$

Kontroller at disse summene stemmer slik at du får samme likningssystem (bruk enten kalkulator eller SPSS). Løs så likningssystemet ved hjelp av en kalkulator. Du vil da se at

$$b_0 = 1{,}6 \quad , \quad b_1 = 0{,}2 \quad og \quad b_2 = 0{,}5$$

M.a.o. at den estimerte regresjonslikningen til Y mhp. X_1 og X_2

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 = 1{,}6 + 0{,}2x_1 + 0{,}5x_2$$

Legger så tallene inn i SPSS, og bruker følgende kommandoer:

ANALYZE
REGRESSION
LINEAR

Velg så y som dependent variable (avhengig variabel) og x_1 *og* x_2 som uavhengige variable. SPSS gir da bl.a. følgende resultater:

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1,600	1,287		1,244	,340
	VAR00001	,200	,347	,381	,576	,623
	VAR00002	,500	,809	,409	,618	,600

a Dependent Variable: VAR00003

Herav ser en bl.a. (i 1. kolonne med tall (ustandardiserte B-koeffisienter)) at bergningene over stemmer.

k forklaringsvariable.

Anta nå at Y er avhengig av k forklaringsvariable (prediktorer) $X_1, X_2, X_3, \ldots, X_k$ der k>2. Dvs . at vi nå har n observasjons-(k+1) tupler $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$ der x_{1i} er en gitt verdi av en tilfeldig variabel X_1, x_{2i} er en gitt verdi av en tilfeldig variabel X_2 , ..., x_{ki} er en gitt verdi av en tilfeldig variabel X_k og y_i er verdien av en tilfeldig variabel Y. Modellen er nå analogt til situasjonen med 2 forklaringsvariable som følger:

Den tilfeldige variable Y (gitt de tilsvarende x-ene) har

$$\text{forventning} = \mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k \quad \text{og}$$

$$\text{varians} = \sigma^2$$

eller ekvivalent

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

der er ε et normalfordelt feilledd som har

$$\text{forventning} = 0 \text{ og varians} = \sigma^2$$

Likningen

$$\mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

kalles ofte for populasjonsregresjonslikningen for Y m.h.t. $X_1, X_2, X_3, \ldots, X_k$. Denne skal vi estimere ved hjelp av et utvalg av n observasjons-(k+1)tupler $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$, i=1,2,...,n. Vi kan da finne en såkalt estimert regresjonslikning (utvalgsregresjonslikning) som betegnes med

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

De n(k+1) observasjonene er:

x_{11}	x_{12}	x_{13}	.....	x_{1n}
x_{21}	x_{22}	x_{23}	.....	x_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_{k1}	x_{k2}	x_{k3}	.....	x_{kn}
y_1	y_2	y_3	.....	y_n

Helt analogt til tilfellet med 2 forklaringsvariable dannes avvikene

$$e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_{1i} + b_2 x_{2i} + \cdots + b_k x_{ki}) \text{ for i = 1, 2, 3,.....,n}$$

og deretter

$$f(b_0, b_1, b_2, ..., b_k) = \sum_i e_i^2$$

som minimeres (i henhold til minste kvadraters metode) og dermed gir opphav til (n+1) likninger med de (n+1) ukjente $b_0, b_1, b_2, ..., b_k$. Dette likningssettet blir på formen:

$$\begin{aligned}
&b_0 n + b_1(\sum_i x_1) + b_2(\sum_i x_2) + \cdots + b_k(\sum_i x_k) = \sum_i y \\
&b_0(\sum_i x_1) + b_1(\sum_i x_1^2) + b_2(\sum_i x_1 x_2) + \cdots + b_k(\sum_i x_1 x_k) = \sum_i x_1 y \\
&b_0(\sum_i x_2) + b_1(\sum_i x_1 x_2) + b_2(\sum_i x_2^2) + \cdots + b_k(\sum_i x_2 x_k) = \sum_i x_2 y \\
&\cdots\cdots\cdots \\
&b_0(\sum_i x_k) + b_1(\sum_i x_k x_1) + b_2(\sum_i x_k x_2) + \cdots + b_k(\sum_i x_k^2) = \sum_i x_k y
\end{aligned}$$

Som ved 2 forklaringsvariable så har en sløyfet indeksen "i" i summene slik at for eksempel $\sum_i x_1$ betyr at man skal summere de n verdiene $x_{1i}, i = 1,2,...,n$ av forklaringsvariabelen X_1, dvs at $\sum_i x_1$ egentlig burde vært skrevet $\sum_i x_{1i}$. Hvis man bare er klar over dette så blir likningssystemet over litt enklere og mer oversiktig med denne skrivemåten.

Se om du greier å se noen likhetspunkter med regresjon med en og to forklaringsvariable. Kan du finne noe mønster i likningssystemet?

Hvis man for eksempel har 5 forklaringsvariable vil dette utgjøre 6 likninger med 6 ukjente. For å bestemme likningssystemet må man først finne alle de nødvendige summene før man løser selve likningssystemet. En skjønner at dette fort blir svært arbeidskrevende selv med kalkulator, og at det derfor er nyttig å kunne bruke SPSS.

Hvis du kan litt om matriseregning (hvis ikke kan du bare hoppe over dette avsnittet) så vet du at likningssystemet over kan skrives på formen:

$$A \cdot B = Y$$

der matrisene *A, B og Y* er gitt ved

$$A = \begin{bmatrix} n & \sum_i x_1 & \sum_i x_2 & \cdots & \sum_i x_k \\ \sum_i x_1 & \sum_i x_1^2 & \sum_i x_1 x_2 & \cdots & \sum_i x_1 x_k \\ \sum_i x_2 & \sum_i x_2 x_1 & \sum_i x_2^2 & \cdots & \sum_i x_2 x_k \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \sum_i x_k & \sum_i x_k x_1 & \sum_i x_k x_2 & \cdots & \sum_i x_k^2 \end{bmatrix}$$

$$B = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} \quad \text{og} \quad Y = \begin{bmatrix} \sum_i y \\ \sum_i x_1 y \\ \vdots \\ \sum_i x_k y \end{bmatrix}$$

Dermed har en følgende løsning på likningssystemet (mhp. de (n+1) ukjente $b_0, b_1, b_2, ..., b_k$.)

$$B = A^{-1} \cdot Y$$

Bergning av invers matrise og matrise-produkter gjøres greitt på en kalkulator (som for eksempel TI-84), men fortsatt står en del regning igjen i forhold til å finne alle summene i A og Y.

Bruker en nå tallene på side 71 så kan normallikningene $A \cdot B = Y$
skrives på følgende måte:

$$\begin{bmatrix} 5 & 22 & 9 \\ 22 & 112 & 44 \\ 9 & 44 & 19 \end{bmatrix} \cdot \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 16{,}9 \\ 79{,}6 \\ 32{,}7 \end{bmatrix}$$

Herav har en nå

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 5 & 22 & 9 \\ 22 & 112 & 44 \\ 9 & 44 & 19 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 16{,}9 \\ 79{,}6 \\ 32{,}7 \end{bmatrix} = \begin{bmatrix} 1{,}655 & -0{,}190 & -0{,}345 \\ -0{,}190 & 0{,}121 & -0{,}190 \\ -0{,}345 & -0{,}190 & 0{,}655 \end{bmatrix} \cdot \begin{bmatrix} 16{,}9 \\ 79{,}6 \\ 32{,}7 \end{bmatrix} = \begin{bmatrix} 1{,}6 \\ 0{,}2 \\ 0{,}5 \end{bmatrix}$$

som stemmer med resultatene foran.

Tolkning av regresjonskoeffisientene.

Anta at vi har 2 forklaringsvariable X_1 *og* X_2 slik at

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

Anta nå at variabelen X_2 holdes konstant og at variabelen X_1 endres med ΔX_1 . Hva blir da $\Delta\hat{y}$ = endringen i $\hat{y}$?
En har da:

$$\Delta\hat{y} = \text{(Ny y-verdi) - (Opprinnelig y-verdi)}$$

Dvs.

$$\Delta\hat{y} = (b_0 + b_1(x_1 + \Delta x_1) + b_2 x_2) - (b_0 + b_1 x_1 + b_2 x_2) = b_1 \Delta x_1$$

ved elementær algebra.

Setter en spesielt $\Delta x_1 = 1$ ser en at

$$\Delta\hat{y} = b_1$$

Mao. b_1 er altså den endringen som blir i $\hat{y}$ når x_2 er konstant og x_1 endres med 1.

Tilsvarende ser en hvis X_1 holdes konstant og X_2 endres med ΔX_2 så blir

$$\Delta\hat{y} = b_2 \Delta x_2$$

Setter en spesielt $\Delta x_2 = 1$ ser en at

$$\Delta\hat{y} = b_2$$

Mao. b_2 er altså den endringen som blir i $\hat{y}$ når x_1 er konstant og x_2 endres med 1.

Anta nå at både x_1 og x_2 endres med hhv. Δx_1 og Δx_2 . Da blir

$$\Delta\hat{y} = (b_0 + b_1(x_1 + \Delta x_1) + b_2(x_2 + \Delta x_2) - (b_0 + b_1 x_1 + b_2 x_2) = b_1 \Delta x_1 + b_2 \Delta x_2$$

Er spesielt både Δx_1 og Δx_2 lik 1 så blir

$$\Delta\hat{y} = b_1 + b_2$$

Anta nå at vi har k forklaringsvariable slik at

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

Anta så at $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots . x_k$ er konstante mens x_i endres med Δx_i . Da har en analogt at

$$\Delta\hat{y} = b_i \Delta x_i$$

Er spesielt $\Delta x_i = 1$ så har en at

$$\Delta\hat{y} = b_i$$

Mao. b_i er altså den endringen som blir i $\hat{y}$ når $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots . x_k$ er konstante

mens x_i endres med 1.

Anta nå at x_1, x_2,..., x_k endres med hhv. Δx_1, Δx_2,....,. Δx_k Da blir analogt

$$\Delta\hat{y} = b_1\Delta x_1 + b_2\Delta x_2 + \cdots + b_k\Delta x_k$$

Eksempel:

Anta at sammenhengen mellom inntekt ($=X_1$), antall barn ($=X_2$) og sparebeløp pr. år *(=Y)* er gitt ved :

Totalinntekt i 1000000 kr. X_1	Antall barn X_2	Sparebeløp pr.år i 10000kr Y
6,6	2	2,1
5,1	1	3,0
4,9	3	1,6
4,5	1	2,1
4,2	4	1,2

Legger en disse dataene inn i SPSS og ber om en multippel regresjonsanalyse av dette med X_1 *og* X_2 som forklaringsvariable og Y som avhengig variabel får en bl.a.:

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	2,536	1,414		1,793	,215
	VAR00001	,085	,246	,117	,346	,763
	VAR00002	-,439	,175	-,848	-2,507	,129

a Dependent Variable: VAR00003

Det betyr bl.a. at en har følgende sammenheng

$$\hat{y} = 2{,}536 + 0{,}085\,x_1 - 0{,}439x_2$$

Hvis nå for eksempel

$\Delta X_1 = 0{,}2$ (dvs. totalinntekten øker med 20000 kr)

og at X_2 = antall barn holdes konstant

Da er endringen i y

$$\Delta\hat{y} = b_1\Delta x_1 = 0{,}085 \cdot 0{,}2 = 0{,}017$$

som altså betyr at sparingen pr. år øker med 170 kr.

Hvis nå isteden

$\Delta X_2 = 1$ (dvs. man får 1 barn til)

$$\text{og } X_1 = \text{ totalinntekten er konstant}$$

så er endringen i y

$$\Delta\hat{y} = b_2\Delta x_2 = -0{,}439 \cdot 10 = -0{,}439$$

som altså betyr at sparingen pr. år avtar med 4390 kr.

Hvis en nå tilslutt lar

$$\Delta X_1 = 0{,}2 \text{ (dvs.totalinntekten øker med 20000 kr)}$$
$$\text{og } \Delta X_2 = 1 \text{ (dvs. man får 1 barn til)}$$

så er endringen i y

$$\Delta\hat{y} = b_1\Delta x_1 + b_2\Delta x_2 = 0{,}085 \cdot 0{,}2 + (-0{,}439) \cdot 1 = -0{,}422$$

som altså betyr at sparingen pr. år avtar med 4220 kr

Estimering og hypoteseprøving i multippel regresjon.

Anta at man nå har modellen

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

der ε er et normalfordelt feilledd som har

$$\text{forventning} = 0 \text{ og varians} = \sigma^2$$

La som foran

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

være den estimerte regresjonslikningen.

Helt analogt til situasjonen i enkel regresjon (se s. 53) kan en nå utlede at 95% konfidensintervall for β_i, i = 1,2,3,….,k blir

$$b_i - t_{\alpha/2,n-(k+1)}SE(b_i) < \beta < b_i + t_{\alpha/2,n-(k+1)}SE(b_i)$$

Legg imidlertid merke til at antall frihetsgrader har endret seg fra $(n\text{-}2)$ til $(n\text{-}(k+1))$ idet det i enkel regresjon er 2 parametre som skal estimeres, mens det i den multiple situasjonen er $(k+1)$ parametre som skal estimeres.
$SE(b_i)$ er imidlertid noe mer komplisert å beregne enn i enkel regresjon. Formelen for denne baserer seg en del matriseregning som vi ikke skal komme inn på her.

95% konfidensintervaller i eksempelet på side 70 blir nå (ved SPSS)

Coefficients(a)

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2,536	1,414		1,793	,215	-3,550	8,622
	VAR00001	,085	,246	,117	,346	,763	-,972	1,142
	VAR00002	-,439	,175	-,848	-2,507	,129	-1,192	,314

a Dependent Variable: VAR00003

Herav ser en bl.a. at 95% konfidensintervall for β_1 *og* β_2 er gitt ved henholdsvis

$$-0{,}972 < \beta_1 < 1{,}142$$

og

$$-1{,}192 < \beta_2 < 0{,}314$$

Noen vil kanskje reagere på den store vidden på konfidensintervallene. Dette skyldes bl.a. at det er veldig få observasjonstripler (n=5). Hadde det vært flere observasjoner ville konfidensintervallene blitt smalere.
Vær oppmerksom på at konfidensintervallene gjelder kun hver for seg med en pålitelighet på 95% og ikke samtidig. Dette er en veldig vanlig misforståelse. Det man kan si er at begge intervallene gjelder med en konfidenskoeffisient som er mindre enn 95%. Det finnes imidlertid metoder som også løser dette problemet (bl.a. Bonferroni)

Bruker en nå

$$b_1 - t_{\alpha/2,n-(k+1)}SE(b_1) < \beta_1 < b_1 + t_{\alpha/2,n-(k+1)}SE(b_1),$$

SPSS-utskriften hvor en ser at $SE(b_1) = 0{,}246$ og $b_1 = 0{,}085$, og en tabell over t-fordelingen hvor en finner 0,025 fraktilen med d.f.=5-(2+1)=3 til 4,303. Dermed ser en at 95% konfidensintervall for β_1 blir

$$0{,}085 - 4{,}303 \cdot 0{,}246 < \beta_1 < 0{,}085 + 4{,}303 \cdot 0{,}246$$

som gir

$$-0{,}974 < \beta_1 < 1{,}144$$

som stemmer godt overens med de direkte beregningene gjort i SPSS.

Hvis man ønsker å teste

$$H_0 : \beta_i = \beta_0$$

mot

$$H_A : \begin{cases} \beta_i < \beta_0 \ eller \\ \beta_i > \beta_0 \ eller \\ \beta_i \neq \beta_0 \end{cases}$$

for i=0,1,2,…,k så bruker en testobservatoren

$$t = \frac{\hat{\beta}_i - \beta_0}{SE(\hat{\beta}_i)}$$

Svært ofte så lar en β_0=0 idet en ønsker å teste om det er noen sammenheng mellom indikator nr. i, X_i, og Y. Testobservatoren blir da

$$t = \frac{\hat{\beta}_i}{SE(\hat{\beta}_i)}$$

Ser en nå på talleksempelet over og for eksempel tester $H_0 : \beta_2 = 0$ mot $H_A : \beta_2 \neq 0$ finner en (se SPSS-utskriften)

$$t = \frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{-0{,}439}{0{,}175} = -2{,}509$$

som stemmer bra overens med SPSS sin verdi på -2,507. P-verdien for dette blir nå (pga. 2-sidig testing)

$$\mathrm{P}(|t| \geq 2{,}509) = 2 \cdot P(t \geq 2{,}509) = 2 \cdot tcdf(2{,}509, 10^{99}, 2) = 0{,}129$$

som stemmer presist med SPSS utskriften. Konklusjonen blir altså at $H_0 : \beta_2 = 0$ ikke kan forkastes på 5%-nivået.
I SPSS-utskriften knyttet til regresjonsanalysen av dataene over har en også følgende tabell:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,888(a)	,789	,578	,43809

a Predictors: (Constant), VAR00002, VAR00001

Her forteller den kvadrerte R = (R Square (som også kalles for "The coefficient of Determination")) hvor god forklaringsgrad analysen har. Det betyr mer presist i denne sammenhengen at R^2 forteller hvor stor andel av totalvariasjonen i Y som blir forklart ved X-ene. Her ser en at $R^2 = 0{,}789$, hvilket betyr at 78,9% av totalvariasjonen i Y-ene blir forklart ved hjelp av X-ene. Jo nærmere dette tallet kommer 1 (eller 100%) jo bedre er mao. analysen. Resten, dvs 21,1% forblir uforklart.
Vær imidlertid oppmerksom på at regresjonsanalysen ikke viser noe om årsakssammenheng.

R er her den multiple korrelasjonskoeffisienten som vi kommer tilbake til i neste avsnitt.

Det som her er nevnt om multippel regresjon utgjør bare en bitte liten del av stoffet. Hvis man ønsker å få en grundig innføring kan ta for seg en bok som "Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences" av J. Cohen, P. Cohen, S.G. West og L.S. Aiken på kun(!) 642 sider. En annen litt mer avansert bok (bruker matrise regning) er "Applied Regression Analysis" av N.R. Draper og H. Smith på kun (!) 591 sider. Litt senere i hovedfagstudiet skal man ta opp mer om regresjon og korrelasjon i en kortfattet bok med tittelen "Understanding Regression Assumptions" av William D. Berry. Denne behandler en del viktige sider ved regresjon som ikke er behandlet her. Det er bl.a. om multikollineæritet, variabel utvelgelse (hvilke av flere forklaringsvariable skal være med, og hvilke skal ikke være med?), er modell forutsetningene oppfylt? (lineæritet, normalitet, konstant varians).

15. Multippel og partiell korrelasjon.

Partiell korrelasjon.

På side 28-29 innførte vi korrelasjonsbegrepet og definerte den såkalte korrelasjonskoeffisienten r_{xy} (Pearsons korrelasjonskoeffisient) ved

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

r_{xy} måler graden av lineær sammenheng mellom variablene X og Y. Anta at vi nå har en situasjon med 3 variable : X_1, X_2 og Y der en antar at det er en lineær sammenheng mellom disse. Fortsatt kan vi finne de vanlige korrelasjonskoffisientene(Pearson) mellom X_1 og $X_2 (=r_{x_1 x_2})$, mellom X_1 og $Y (=r_{x_1 y})$, og mellom X_2 og $Y (=r_{x_2 y})$.
Anta at vi har følgende observasjoner av de tre variablene X_1, X_2 og Y

X_1	X_2	Y
4,00	3,00	9,00
3,00	5,00	10,00
5,00	6,00	9,00
7,00	5,00	5,00
6,00	8,00	7,00
5,00	9,00	3,00
7,00	10,00	2,00

Kommandoene

ANALYZE
CORRELATE
BIVARIATE
PEARSON
SELECT variabel 1,2 og 3 (h.h.v. X_1, X_2 og Y)

i SPSS gir nå:

Correlations

		VAR00001	VAR00002	VAR00003
VAR00001	Pearson Correlation	1	,527	-,737
	Sig. (2-tailed)		,224	,059
	N	7	7	7
VAR00002	Pearson Correlation	,527	1	-,774(*)
	Sig. (2-tailed)	,224		,041
	N	7	7	7
VAR00003	Pearson Correlation	-,737	-,774(*)	1
	Sig. (2-tailed)	,059	,041	
	N	7	7	7

* Correlation is significant at the 0.05 level (2-tailed).

Herav ser en at $r_{x_1 x_2} = 0{,}527$, at $r_{x_1 y} = -0{,}737$, og at $r_{x_2 y} = -0{,}774$.
Disse er viktige nok (spesielt som et verktøy i regresjonsanalyse til å kunne vurdere om en variabel skal være med i analysen eller ikke) men ved for eksempel kun å regne ut $r_{x_1 y}$ så kaster en bort den informasjonen som verdiene av X_2 gir. Det kan nemlig være slik at det hovedsakelig er X_2 som påvirker Y ved at den påvirker X_1 som så påvirker Y. Man beregner derfor ofte de såkalte **partielle korrelasjonskoeffisienter** (jfr. partielle regresjonskoeffisienter):

$r_{x_1 y \bullet x_2}$ = korrelasjonskoeffisienten mellom X_1 og Y når X_2 holdes konstant

$r_{x_2 y \bullet x_1}$ = korrelasjonskoeffisienten mellom X_2 og Y når X_1 holdes konstant

man kan selvfølgelig også beregne korrelasjonskoeffisienten mellom X_1 og X_2 når Y holdes konstant, men i en multippel regresjonsanalysesituasjon er ikke det så interessant idet Y er den avhengige variabelen som skal forklares ved hjelp av uavhengige variable X_1 og X_2.
De partielle korrelasjonskoeffisientene er definert ved :

$$r^2_{X_1 Y \bullet X_2} = \frac{SSE_{X_2} - SSE_{X_1 X_2}}{SSE_{X_2}} = 1 - \frac{SSE_{X_1 X_2}}{SSE_{X_2}}$$

der SSE_{X_2} er restvariasjonen i modellen når bare X_2 er med, og $SSE_{X_1 X_2}$ er restvariasjonen i modellen når bare både X_1 og X_2 er med. SSE – leddene er dermed gitt ved:

$$SSE_{X_2} = \sum_i (y - \hat{y}_{X_2})^2$$

og

$$SSE_{X_2 X_2} = \sum_i (y - \hat{y}_{X_2 X_2})^2$$

Hvis man nå beregner regresjonslikningene i dataene over finner en:

$$\hat{y}_{X_2} = 12{,}826 - 0{,}973x_2$$

og

$$\hat{y}_{X_2X_2} = 15{,}917 - 0{,}960x_1 - 0{,}973x_2$$

Dermed finner en følgende tabell for dataene over:

X_1	X_2	Y	$\hat{Y}_{X_2}$	$\hat{Y}_{X_1X_2}$
4	3	9	9,907	10,061
3	5	10	7,961	9,677
5	6	9	6,988	7,085
7	5	5	7,961	5,837
6	8	7	5,042	4,781
5	9	3	4,069	5,069
7	10	2	3,096	2,477

Herav finner en nå

$$SSE_{X_2} = \sum_i (y - \hat{y}_{X_2})^2 = (9 - 9{,}907)^2 + \cdots + (2 - 3{,}096)^2 = 23{,}974$$

og

$$SSE_{X_2X_2} = \sum_i (y - \hat{y}_{X_2X_2})^2 = (9 - 10{,}061)^2 + \cdots + (2 - 2{,}477)^2 = 15{,}030$$

Dermed blir

$$r^2_{X_1Y \bullet X_2} = \frac{SSE_{X_2} - SSE_{X_1X_2}}{SSE_{X_2}} = \frac{23{,}974 - 15{,}030}{23{,}974} = 0{,}373$$

Det betyr da at

$$\left| r_{X_1Y \bullet X_2} \right| = \sqrt{0{,}373} = 0{,}611$$

Nå skal $r_{X_1Y \bullet X_2}$ ha samme fortegn som den tilsvarende partielle regresjonskoeffisienten som er -0,960. M.a.o. $r_{X_1Y \bullet X_2}$ = -0,611.

Kommandoene

ANALYZE
CORRELATE
PARTIEL
PEARSON

SELECT variabel 1,2 og 3 (h.h.v. X_1, X_2 og Y)

i SPSS gir nå:

Correlations

Control Variables			VAR00001	VAR00003
VAR00002	VAR00001	Correlation	1,000	-,611
		Significance (2-tailed)	.	,198
		df	0	4
	VAR00003	Correlation	-,611	1,000
		Significance (2-tailed)	,198	.
		df	4	0

Correlations

Control Variables			VAR00003	VAR00002
VAR00001	VAR00003	Correlation	1,000	-,671
		Significance (2-tailed)	.	,145
		df	0	4
	VAR00002	Correlation	-,671	1,000
		Significance (2-tailed)	,145	.
		df	4	0

Herav ser en at $r_{x_1 y \bullet x_2}$ = -0,611 og $r_{x_2 y \bullet x_1}$ = -0,671. En ser at det første stemmer helt med beregningene over. Prøv selv å kontrollere den siste partielle korrelasjonskoffesienten $r_{x_2 y \bullet x_1}$ ved det tilsvarende uttrykket

$$r^2_{X_2 Y \bullet X_2} = \frac{SSE_{X_1} - SSE_{X_1 X_2}}{SSE_{X_1}}$$

Nå kan det også vises at

$$r_{x_1 y \bullet x_2} = \frac{r_{x_1 y} - r_{x_2 y} r_{x_1 x_2}}{\sqrt{(1 - r^2_{x_2 y})(1 - r^2_{x_1 x_2})}}$$

og at

$$r_{x_2 y \bullet x_1} = \frac{r_{x_2 y} - r_{x_1 y} r_{x_1 x_2}}{\sqrt{(1 - r^2_{x_1 y})(1 - r^2_{x_1 x_2})}}$$

Av det første av utrykkene og SPSS-utskriften side 75 finner en nå

$$r_{x_1 y \bullet x_2} = \frac{r_{x_1 y} - r_{x_2 y} r_{x_1 x_2}}{\sqrt{(1 - r^2_{x_2 y})(1 - r^2_{x_1 x_2})}} = \frac{-0,737 - (-0,774) \cdot 0,527}{\sqrt{(1 - (-0,737^2))(1 - 0,527^2)}} = -0,612$$

som stemmer med resultatene over. Prøv selv å kontrollere om du får det til å stemme for $r_{x_2 y \bullet x_1}$

Multippel korrelasjon.

Anta at vi nå har en multippel regresjonsmodell med 2 forklaringsvariable, dvs. at

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

der ε er et normalfordelte feilledd som har

$$\text{forventning} = 0 \text{ og varians} = \sigma^2$$

Den estimerte regresjonslikning er da på formen

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2$$

I denne multiple situasjonen har vi tidligere beregnet flere typer korrelasjonskoeffisienter:

i) De vanlige korrelasjonskoeffisientene(Pearson) mellom X_1 og X_2 $(=r_{x_1x_2})$, mellom X_1 og $Y(=r_{x_1y})$, og mellom X_2 og Y $(=r_{x_2y})$ og

ii) De partielle korrelasjonskoeffisientene $r_{x_1y\bullet x_2}$ og $r_{x_2y\bullet x_1}$

Nå skal vi se på en tredje type korrelasjonskoeffisient, nemlig den såkalte multiple korrelasjonskoeffisienten R $(=R_{y\bullet x_2,x_2})$ som måler graden av lineær sammenheng mellom Y og X_1 og X_2.
Den defineres ved :

$$R^2_{y\bullet x_1,x_2} = \frac{\textit{Variasjonen i Y forklart ved regressorene } X_1 \textit{ og } X_2}{\textit{Total variasjonen i } Y} =$$

$$= \frac{\textit{Total variasjonen i } Y - \text{Rest}\textit{variasjonen}}{\textit{Total variasjonen i } Y} = \frac{SSY - SSE_{X_1X_2}}{SSY}$$

der

$$SSE_{X_2X_2} = \sum_i (y - \hat{y}_{X_2X_2})^2$$

som over, og

$$\textit{totalvariasjonen } SSY = \sum_i (y - \bar{y})^2$$

R kan også defineres ved $R_{y\hat{Y}}$, mao. som korrelasjonskoeffesienten mellom Y og $\hat{Y}$.

Bruker en nå tallene fra side 76 har en

$$SSE_{X_2X_2} = \sum_i (y - \hat{y}_{X_2X_2})^2 = 15{,}030$$

og

$$SSY = \sum_i (y - \bar{y})^2 = (9 - \frac{45}{7})^2 + \cdots + (2 - \frac{45}{7})^2 = 59{,}714$$

Dermed blir den kvadrerte multiple korrelasjonskoeffesienten

$$R^2_{y\bullet x_1,x_2} = \frac{SSY - SSE_{X_1X_2}}{SSY} = \frac{59{,}714 - 15{,}030}{59{,}714} = 0{,}748$$

Herav finner en da $R_{y\bullet x_1,x_2} = \sqrt{0{,}748} = 0{,}865$

Går man nå inn i SPSS og legger inn dataene og bruker

ANALYZE
REGRESSION
LINEAR

finner man bl.a.

ANOVA(b)

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	44,684	2	22,342	5,946	,063(a)
	Residual	15,030	4	3,758		
	Total	59,714	6			

a Predictors: (Constant), VAR00002, VAR00001
b Dependent Variable: VAR00003

og

Modell Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,865(a)	,748	,622	1,93843

a Predictors: (Constant), VAR00002, VAR00001

Herav ser en at beregningene over stemmer perfekt.
I den siste tabellen ser en også at det er med en korrigert R^2 (Adjusted R Square) som er beregnet til 0,622. Hvis man tar med en forklaringsvariabel til i analysen vil dette øke R^2 (litt) selv om den er irrelevant (i den betydning at den har en β som er 0). Dette kan man korrigere for ved å redusere R^2. Hvis det er k forklaringsvariable så er den korrigerte R^2 i SPSS gitt ved

$$\overline{R}^2 = \frac{(n-1)R^2 - k}{n-k-1}$$

Det betyr at den i eksempelet over blir

$$\overline{R}^2 = \frac{(7-1)\cdot 0{,}748 - 2}{7-2-1} = 0{,}622$$

som stemmer med tabellen over.

$\overline{R}^2$ kan brukes i såkalt trinnvis regresjon (stepwise regression) hvor en i utgangspunktet har mange forklaringsvariable (= k) og ønsker å finne de viktigste. Dette gjøres ved at det regnes ut hvilken av de k forklaringsvariablene som har størst innvirkning på $\overline{R}^2$, deretter den som har nest størst innvirkning på $\overline{R}^2$ osv..., Dette gjentas inntil en ser at det ikke blir noen nevneverdig økning i $\overline{R}^2$.

Litteraturliste

(1) David S. Moore, George P. McCabe : Introduction to the Practice of Statistics (4.ed.2003)

(2) John E. Freund: Mathematical Statistics with Applications (7.ed.2004)

(3) Jostein Lillestøl: Sannsynlighetsregning og statistikk med anvendelser (5.utg.1997)

(4) Ajit C. Tamhane, Dorothy D. Dunlop: Statistics and Data Analysis from Elementary to Intermediate (2000)

(5) Kleinbaum, Kupper, Muller: Applied Regression Analysis and other Multivariate Methods (2.ed. 1988)

(6) Alan Agresti, Barbara Finlay : Statistical Methods for the Social Sciences (2.ed 1986)

(7) Ronald J. Wonnacott, Thomas H. Wonnacott : Introductory Statistics (4.ed. 1985)

(8) Gunnar G. Løvås : Statistikk for unniversiteter og høgskoler (1.utg. 1999)

(9) Jøreskog : Formulas for Skewness and Kurtosis (1999)

(10)SPSS support: http://support.spss.com/tech/default.asp

Høgskolen i Buskerud
Postboks 235
3603 Kongsberg
Telefon: 32 86 95 00
Telefaks: 32 86 98 83

www.hibu.no

HØGSKOLEN
i Buskerud