Arbeidsnotater fra Høgskolen i Buskerud nr. 67

ARBEIDSNOTAT

ARBEIDSNOTAT

Anvendt statistikk

Jon Reinertsen

HØGSKOLEN i Buskerud

Arbeidsnotater fra Høgskolen i Buskerud

Nr. 67

Anvendt statistikk

Av

Jon Reinertsen

Hønefoss 2008

ISSN 0807-447X

INNHOLD

1. Innledning.

Dette heftet danner utgangspunkt for et kurs i anvendt statistikk som utgjør halvparten av kurset MAT 420 (grunnleggende og anvendt statistikk).
MAT 420 utgjør ¼-del av årsenheten i matematikk på valgfag på almenlærerutdanningen. Før man starter på dette kurset i anvendt statistikk har studentene som et minimum vært igjennom et kurs i grunnleggende matematisk analyse og didaktikk (15 studiepoeng) i høstsemesteret samt den første delen av kurset MAT 420 i vårsemesteret. I den grunnleggende statistikkdelen av kurset har man behandlet begreper som diskrete sannsynlighetsmodeller (generelt), kombinatorikk og utvalgsmodeller, betinget sannsynlighet og uavhengighet, stokastiske variable, forventning og varians, noen vanlige sannsynlighetsfordelinger (binomisk-, hyper-geometrisk-, Poisson- og normalfordeling), estimering og hypoteseprøving. Disse temaene forutsettes kjent når man starter på dette heftet. Jeg har allikevel valgt å legge inn noe av det som er behandlet i den grunnleggende statistikken slik at de som eventuelt ønsker å lese dette separat kan gjøre det uten i for stor utstrekning å måtte slå opp i en grunnbok. Dette gjelder spesielt teorien knyttet til estimering og hypoteseprøving.

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)

Α	α	*alfa*	Ν	ν	*ny*
Β	β	*beta*	Ξ	ξ	*ksi*
Γ	γ	*gamma*	Ο	o	*omikron*
Δ	δ	*delta*	Π	π	*pi*
Ε	ε	*epsilon*	Ρ	ρ	*rho*
Ζ	ς	*zeta*	Σ	σ	*sigma*
Η	η	*eta*	Τ	τ	*tau*
Θ	θ	*teta*	Υ	υ	*ypsilon*
Ι	ι	*iota*	Φ	ϕ	*fi*
Κ	κ	*kappa*	Χ	χ	*kji*
Λ	λ	*lambda*	Ψ	ψ	*psi*
Μ	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 (”Farmasia”). En, to, tre på gresk er ENA (”*ena*”), ΔΥΟ ("*dyo*"), ΤΡΙΑ ("*tria*")

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

3. Beskrivende statistikk.

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

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

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

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

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 ennå kortere $\bar{x} = \frac{\sum f \cdot x}{n}$

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.

I små tallmaterialer (n<100) så beregner en medianen ved å finne observasjon nr. $\frac{n+1}{2}$ i det ordnede tallmaterialet. Tilsvarende finner en kvartilene Q_1 og Q_3 som henholdsvis observasjon nr. $\frac{n+1}{4}$ og nr. $3 \cdot \frac{n+1}{4}$ i små tallmaterialer.

I større tallmaterialer ($n \geq 100$) så leter en tilsvarende etter observasjon nr. $\frac{n}{4}$, nr. $\frac{n}{2}$ og nr. $3 \cdot \frac{n}{4}$ i det ordnede tallmaterialet. Grunnen til dette er at det liten forskjell på (for eksempel) $\frac{n}{4}$ og $\frac{n+1}{4}$ når *n* er stor . Denne tankemåten er praktisk når man skal bruke andre mål enn kvartiler.

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

Densilene $D_1, D_2, ..., D_{10}$ 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. En beregner nå tilsvarende her observasjon nr. $\frac{n}{10}$, nr. $\frac{2n}{10}$, nr $\frac{9n}{10}$ i det ordnede tallmaterialet.

Prosentilene $P_1, P_2, ..., P_{100}$ 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. Helt analogt beregner en nå tilsvarende her observasjon nr. $\frac{n}{100}$, nr. $\frac{2n}{100}$, nr $\frac{99n}{100}$ i det ordnede tallmaterialet når man skal beregne prosentilene $P_1, P_2, ..., P_{100}$.

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 hvis 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**

Eks. Gitt tallmaterialet

$$x_i : \quad 1, 5, 4, 7, 6, 12$$

Spørsmålet er nå om $x_6 = 12$ er en outlier. Legger en inn tallene i kalkulatoren finner en Q_1 og Q_3 henholdsvis til 4 og 7. Greier du å se hvorledes kalkulatoren beregner kvartilene? En finner nå m.a.o. $IQR = 7 - 4 = 3$ og dermed $Q_3 + 1{,}5IQR = 7 + 1{,}5 \cdot 3 = 11{,}5$. Dvs at $x_6 = 12$ er en outlier siden den er > 11,5. dermed bør denne observasjonen fjernes fra tallmaterialet før en gjør noen analyser.

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 på dette er:

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

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.:

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

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: 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

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:

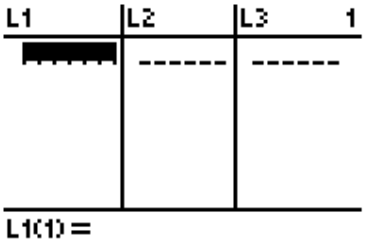
$$s^2=\frac{1}{n-1}\sum_{k=1}^{m} f_k(x_k-\bar{x})^2=\frac{1}{20-1}(3\cdot(2{,}5-8{,}3)^2+11\cdot(7{,}5-8{,}3)^2+6\cdot(12{,}5-8{,}3)^2)=11{,}3$$

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

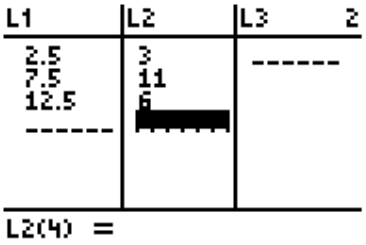
Ønsker man å legge disse tallene inn i listene i TI 84 går en fram som følger:

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

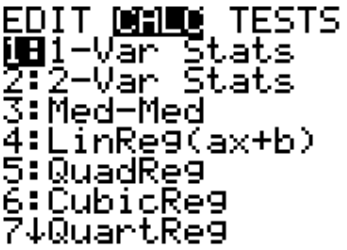
Trykk så på ENTER-tasten og du får opp følgende bilde:



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:

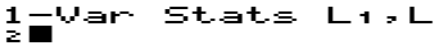


Nå trykker en så på STAT-tasten igjen, men velger nå isteden alternativet CALC (calculations). Dette gir følgende bilde:

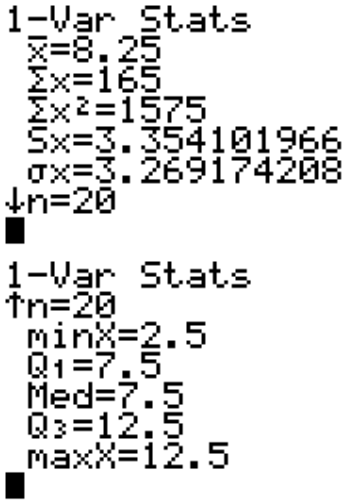


En bruker nå 1: 1-Var Stats (envariabelstatistikk) på følgende måte:

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. En ser at dette avviker en del fra beregningene i råmaterialet:

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

Hvor en fant $Q_1 = 5$ og $Q_3 = 10$, men det skyldes den forskjellen som er mellom råmaterialet og tabellmaterialet

x_i' : 2.5, 2.5, 2.5, 7.5, 7.5, 7.5, 7.5, 7.5, 7.5, 7.5, 7.5, 7.5, 7.5, 7.5, 12.5, 12.5, 12.5, 12.5, 12.5, 12.5.

En har at medianen *M* er gitt ved 7,5. Q_1 *og* Q_3 ser en blir henholdsvis 7,5 og 12,5. Legg merke til at Q_1 *og* *M* blir like. Hvordan vil du kommentere dette?

Nå kan medianen (og kvartilene) beregnes ved hjelp av tabellen på en tilnærmet måte. En har at

$$M = L + \frac{\text{Re}\,st}{f} \cdot v$$

der L = nedre klassegrense i den klassen hvor medianen ligger, f = frekvensen i den klassen hvor medianen ligger, v = klassevidden i medianklassen og $Rest$ = det en mangler for å komme fram til medianen (dvs det antall observasjoner en mangler fra L og fram til observasjon nr. observasjon nr. $\frac{n+1}{2}$ når *n* er et lite tall**)

I vårt eksempel finner en

$$M = L + \frac{\text{Re}\,st}{f} \cdot v = 5{,}0 + \frac{10{,}5 - 3}{11} \cdot 5{,}0 = 8{,}4$$

idet medianen er observasjon nr. 10,5 som ligger i klasse nr 2 som har nedre klassegrense 5, en klassevidde på 5 og en frekvens på 11. Rest blir dermed 10,5-3 der 3 er det antall observasjoner som ligger i klassene før medianklassen (her kun klasse nr 1)

Den samme teknikken kan brukes til å beregne Q_1 *og* Q_3.

En har nå tilsvarende

$$Q_1 = L + \frac{\text{Re}\,st}{f} \cdot v$$

der L = nedre klassegrense i den klassen hvor 1.kvartil ligger, f = frekvensen i den klassen hvor 1.kvartil ligger, v = klassevidden i klassen hvor medianen ligger og $Rest$ = det en mangler for å komme fram til 1.kvartil (dvs det antall observasjoner en mangler fra L og fram til observasjon nr. observasjon nr. $\frac{n+1}{4}$ når n er et lite tall)

Helt analogt finner en Q_3 ved

$$Q_3 = L + \frac{\text{Re}\,st}{f} \cdot v$$

bortsett fra at man nå leter etter observasjon nr. $3 \cdot \frac{n+1}{4}$. Beregn nå selv 3. kvartil og vurder om svaret ditt er rimelig.

** At n er et lite tall skal her bety at n<100. Hvis $n \geq 100$ så leter en etter observasjon nr $\frac{n}{4}, \frac{n}{2}, \frac{3n}{4}$,…..osv når man skal beregne kvartilene. Dette er en mye enklere og logisk tenkemåte. Når man for eksempel skal finne densilene (de verdiene som deler tallmaterialet i 10 deler) leter en etter observasjon nr. $\frac{n}{10}, \frac{2n}{10}, \frac{3n}{10}, \cdots, \frac{9n}{10}$.

Det er dessuten veldig liten forskjell på de to metodene når n er stor Anta for eksempel at n = 250, at L = nedre klassegrense i den klassen hvor 1.kvartil ligger =39,5 ; at f = frekvensen i den klassen hvor 1.kvartil ligger = 47 , at v = klassevidden medianklassen = 10, og at det ligger 25 observasjoner før klassen som inneholder første kvartil. Nå blir $\frac{n+1}{4} = 62.75$ og $\frac{n}{4} = 62{,}5$. Dermed blir $Rest$ = det en mangler for å komme fram til 1.kvartil = 62,75-25 = 37,75 eller 62,5-25 = 37,5. Det betyr at 1. kvartil beregnet ved de metodene blir henholdsvis

$$Q_1 = L + \frac{\text{Re}\,st}{f} \cdot v = 39{,}5 + \frac{37{,}75}{47} \cdot 10 = 47{,}53$$

$$Q_1 = 39{,}5 + \frac{37{,}75}{47} \cdot 10 = 47{,}48$$

Det blir mao. en forskjell på 0,05 ved de to beregningsmetodene.

Skjevhet *

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

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

hvis alle observasjonene er forskjellige, eller ved

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

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 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 $m_1 = 0$ uansett tallmateriale (se regneregler for summer i Sydsæter App.A)

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

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.

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.))

Nå skal vi prøve å kontrollregne denne verdien, og vi trenger altså både m_2 og m_3.
Vi har tidligere funnet $s^2 = 11{,}25 \Rightarrow m_2 = \frac{19}{20} \cdot 11{,}25 = 10{,}69$. 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

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

Det kan vises at standardavviket til g_1 er 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

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

Dette er et tall som kan brukes til hypotesetesting og estimering (konfidensintervaller).

Spisshet *

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 MINITAB-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

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

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 MINITAB 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

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

Dette kan da igjen brukes til å gjennomføre hypotesetesting og etimering.

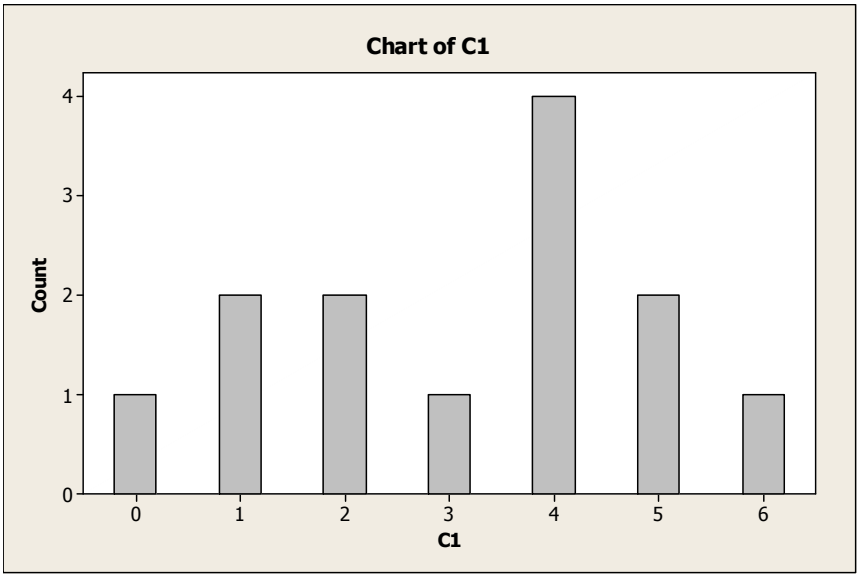
Noen grafiske framstillingsmetoder.

For å skape en oversikt og et bilde av situasjonen så bruker en ofte grafiske framstillinger av tallmaterialet. Dette kan gjøres på flere måter. Hvilke metode en bruker er delvis avhengig av tallmaterialet (dvs. om variabelen er diskret, kontinuerlig, eller en kategorivariabel) og hvilket "publikum" som skal se grafikken.

Stolpediagram (bar chart)
Anta at man har et tilfeldig utvalg på 13 karakterer i matematikk i en ungdomsskoleklasse. Variabelen "karakter" er her diskret og kan anta verdiene 0, 1, 2, 3, 4, 5 og 6. Anta at resultatet av undersøkelse ble:

Elev	1	2	3	4	5	6	7	8	9	10	11	12	13
Kar.	2	4	6	4	1	5	4	3	4	2	1	0	5

Nå bruker man et såkalt stolpediagram for å framstille disse dataene grafisk:
Legger en disse dataene inn i MINITAB vil en få følgende stolpediagram:

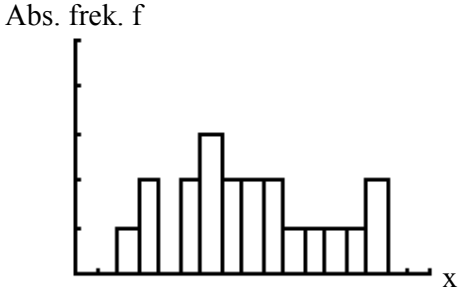


En ser at MINITAB velger å tegne rektangler. Det er også vanlig å tegne vertikale streker.

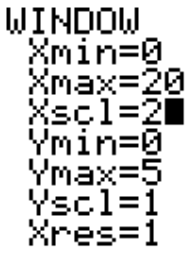
Noen av de mest brukte grafiske framstillingene for kontinuerlige tallmaterialer er **histogram** og **kurvediagram.**

Histogram.

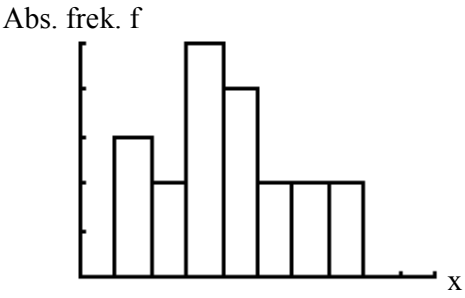
Eks. Går en nå tilbake til tallmaterialet på side 10 (aldersfordelingen på 20 barn) og legger dette inn i TI får en følgende histogram av råmaterialet:



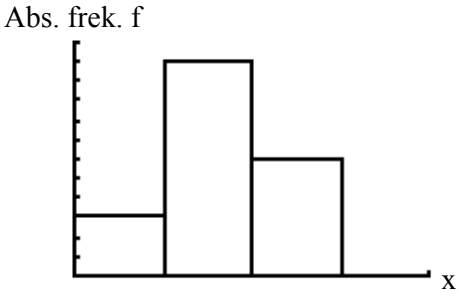
Har en mange observasjoner så blir dette fort uoversiktlig. En lager derfor et passe antall intervaller (klasser). Her må en bruke skjønn. En vanlig tommelfingerregel er å bruke 8-12 intervaller når det er en viss størrelse på tallmaterialet. Velger en for eksempel klassevidde 2 på tallmaterialet over gir TI følgende grafiske bilde: (En bør nå ha Window stilt som følger:



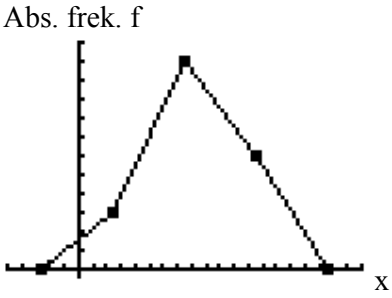
Legg spesielt merke til at Xscl nå er satt til 2)



Bruker en derimot tabellen som utgangspunkt (klassevidde = 5) får en
Velger en derimot å framstille tabellmaterialet grafisk finner en



En må egentlig i hver enkelt situasjon avgjøre hva som best for at det grafiske bildet skal representere tallene på en best mulig måte. Har en for små klassevidder vil en lage et for detaljert bilde, har en for store klassevidder vil en del detaljer bli visket bort. Prøv selv med noen andre klassevidder.
For kontinuerlige data kan en alternativt bruke **kurvediagram** istedenfor histogram. Her avsetter en punktene $(x_{k,} f_k)$, k= 1,2,….., K der K er antall klasser og tegner rette linjer mellom disse. Det er svært vanlig ta med en start-klasse og en avslutningsklasse med frekvens 0 og dermed avsette punktet $(x_{0,} 0)$ og punktet $(x_{K+1,} 0)$. Da vil arealet av histogrammet og kurvediagrammet bli like store. Tallmaterialet i tabellen vil da gi følgende kurvediagram:



Tegn histogrammet og kurvediagrammet inn i samme koordinatsystem og overbevis deg selv om at de dekker samme areal.

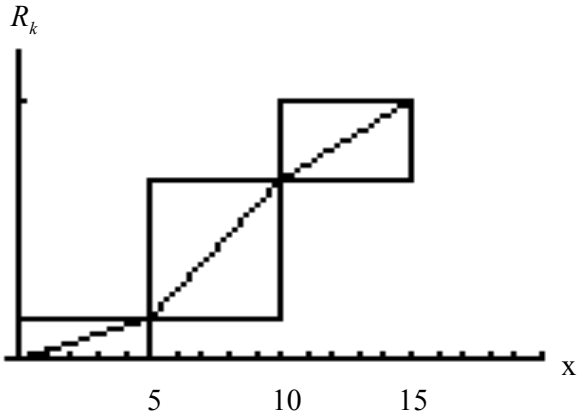
Sumfordelingskurve (sumpolygon)

En meget nyttig kurve som kan brukes til å lese av mange av de målene som vi har beregnet foran er den såkalte sumfordelingskurven. Her beregner en først de kumulative relative frekvensene $R_k = R(X \le x_k)$ (mao. summen av de relative frekvensene r_k opp til og med klasse k). Avsetter en x langs førsteaksen, R_k langs andreaksen og "tilvekstrektangelet" i hver klasse framkommer den såkalte sumfordeligskurven ved å tegne diagonalene gjennom hvert tilvekstrektangel.

Eks. Går vi nå tilbake til tabellen på side 10 har vi nå:

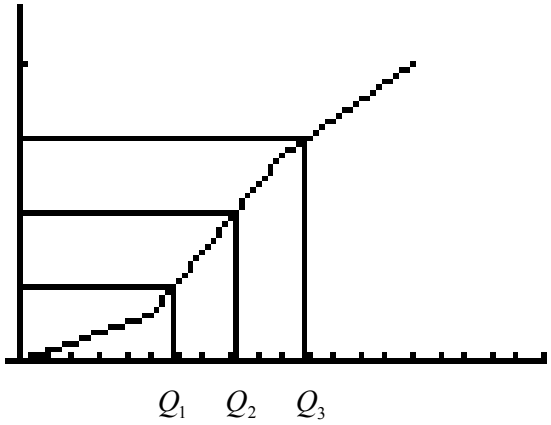
Klassegrenser	Frekvens f_k	Klassemidtpkt. x_k	Relativ frekv. r_k	Kumulativ rel. frekvens R_k
$[0,5\rangle$	3	2,5	0,15	0,15
$[5,10\rangle$	11	7,5	0,55	0,70
$[10,15\rangle$	6	12,5	0,30	1,00

Framstiller en dette grafisk får en:



Her er mao. kurven diagonalt gjennom rektanglene (som er histogrammet til dataene i tabellen bare tegnet på en litt annen måte) fra punktet (0,0) til punktet (15,1) selve sumfordelingskurven. Denne kurven skal vi nå bruke til å lese av 1., 2. og 3. kvartil. For å

finne Q_1 lar en $R_k = 0{,}25$ (det betyr at man har delt tallmaterialet i to slik at 25% ligger under og 75% ligger over det tallet vi nå søker). En går så fra 0,25 på 2.aksen og horisontalt bort til sumfordelingskurven, deretter går en vertikalt ned til 1. aksen og leser der av Q_1**:**

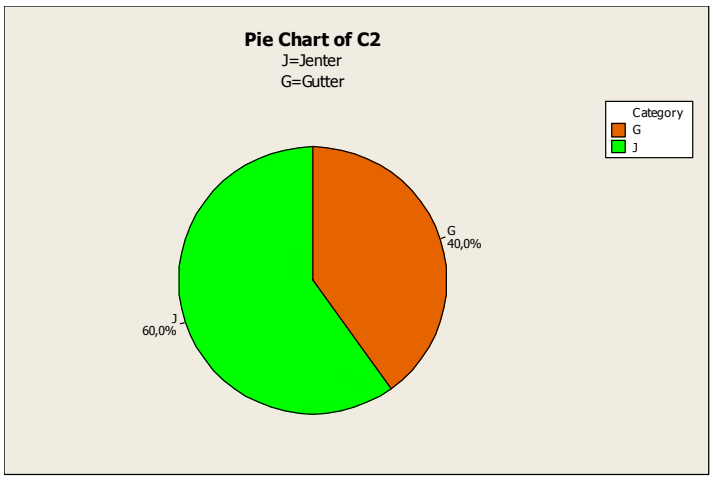


Herav finner en at $Q_1 \approx 5{,}8$; $Q_2 = medianen \approx 8{,}2$ og $Q_3 \approx 10{,}8$**.** Da har en at IQR $\approx$10,8-5,8 = 5,0 som er nøyaktig den verdien vi fant ved regning for IQR .

Noen andre grafiske framstillingsmåter er:

Kakediagram (Piechart)

Anta at det i en klasse er 15 jenter og 10 gutter. Legger en dette inn i MINITAB og ber om piechart får en følgende bilde:



Stamme- og bladdiagram (Stem and leaf)
Anta at en klasse med 30 elever har hatt en matematikkprøve og resultatene ble:

37, 45, 56, 54, 38, 23, 45, 67, 65, 43, 23, 78, 98, 75, 12, 34, 45, 59, 67, 87, 76, 51, 28, 47, 88, 77, 59, 24, 19, 90.

Legger en disse tallene inn i en kolonne C1 i MINITAB og gir følgende kommandoer

GRAPH
STEM AND LEAF
C1
OK

Får en følgende "grafiske " framstilling:

```
 2    1  29
 6    2  3348
 9    3  444
14    4  35557
(5)   5  14699
11    6  577
8     7  5678
4     8  78
2     9  08
```

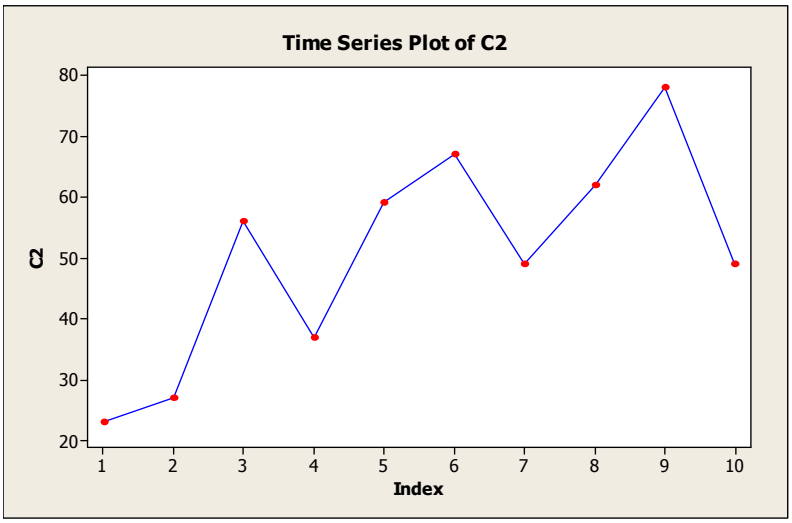
Av første rad ser en av tallet i første kolonne at det er 2 observasjoner i den første klassen og at disse er henholdsvis 12 (1-tallet fra kolonne 2 og 2 tallet fra kolonne 3) og 19 (1-tallet fra kolonne 2 og 9-tallet fra kolonne 3). Tallene i kolonne 2 kalles for stammen og tallene i kolonne 3 kalles for bladene. Tallene i kolonne 1 forteller hvor mange observasjoner det er over/under den aktuelle klassen med unntak av klassen hvor medianen ligger. Her indikerer (5) at medianen ligger. Hvis man tegner histogrammet for denne situasjonen vil det ha samme form som tallene helt til høyre (bladene) bare disse dreies 90 grader.

Tidsrekke (tidsserie)-plot

Anta at en mindre bedrift har notert salget av et produkt de siste 8 månedene og funnet

56, 37, 59, 67, 49, 62, 78, 49

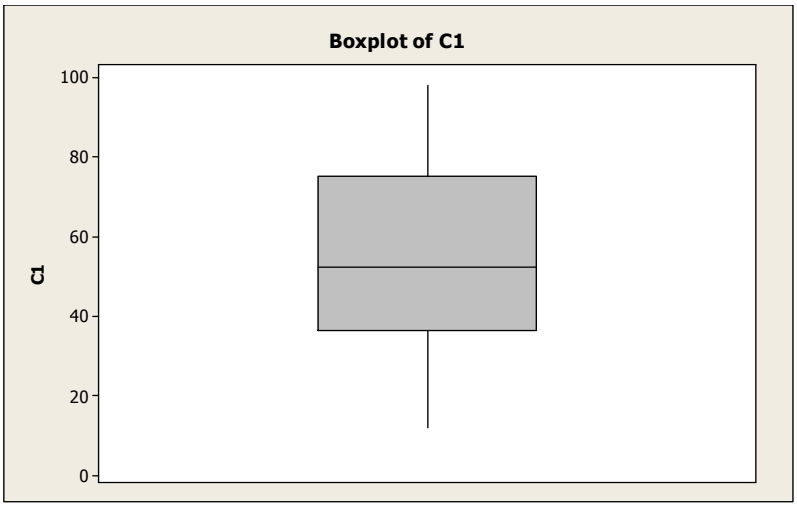
Fremstiller en dette i MINITAB via Time Series Plot får en følgende bilde:



Dette er i realiteten ikke noe annet en et kurvediagram hvor en avsetter tiden langs førsteaksen og de observerte verdiene langs andreaksen. Forskjellen er nå imidlertid at kurven starter i 1.punkt (salget i måned 1) og ikke på selve 1.-aksen som vi gjorde for kurvediagrammet.

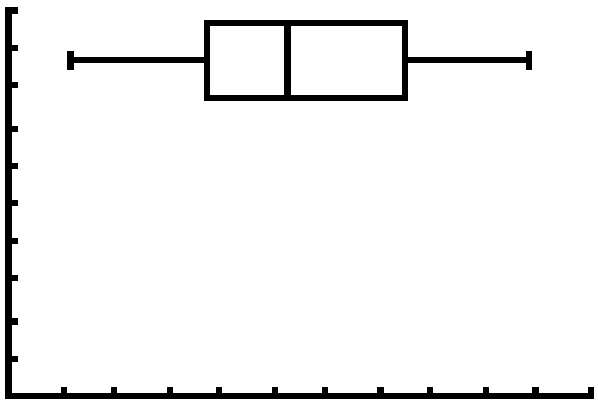
Boksplot

En mye brukt enkel figur som samtidig viser minste x-verdi, største x-verdi, og de 3 kvartilene er det såkalte boksplottet som for de 30 karakterene i eksempelet foran blir seende ut som følger:



De 5 målene på sentral tendens som angis i boksplottet kalles ofte på engelsk for **the five-number summary** i en datamengde.

Selve boksen starter ved Q_1 og slutter ved Q_3, den horisontale streken gjennom boksen viser medianen de vertikale strekene over og under boksen starter ved den minste x-verdien og slutter ved den største x-verdien. På kalkulatoren blir dette seende ut som følger:

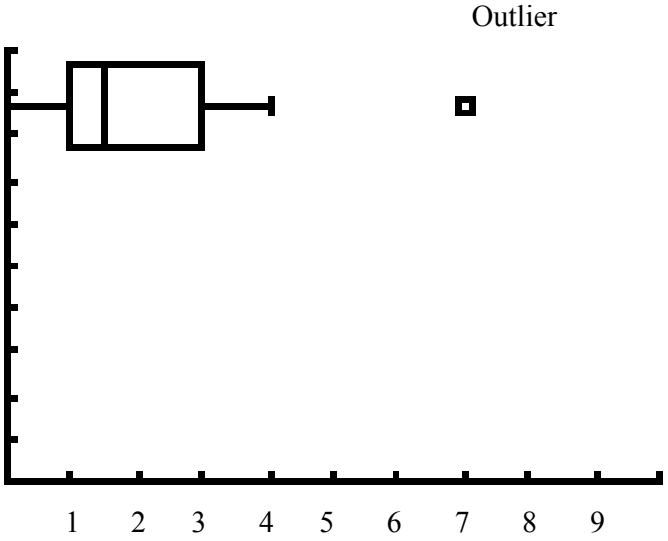


0 20 40 60 80 100

Enheten på 1.aksen er her 10, mens den på 2.aksen er 1. En ser nå at boksen er lik på de 2 figurene bortsett fra at den siste ligger vannrett.

Det såkalte **modifiserte boksplottet** angir også eventuelle "outliere" i datamaterialet.
Anta vi har spurt 15 barn om hvor mange timer de driver med dataspill på en vanlig hverdag. Resultatet av undersøkelsen ble:

$$x_i : 0, 2, 3, 2, 1, 0, 3, 1, 1, 1, 2, 4, 7, 2, 0$$



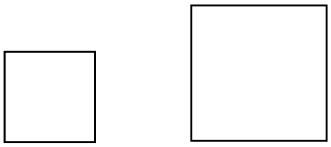
En ser her at verdien 7 er en outlier og dermed blir fjernet fra dataene før boksplottet tegnes (dette ser en blant annet av at største verdi nå er 4). Kontroller nå ved hjelp av regning 7 er en outlier og at dette boksplottet er riktig.

Geometriske figurer.

I mange sammenhenger så bruker man geometriske figurer (trekanter, firkanter, sirkler, tegning av hus, mennesker,…..) når man skal beskrive et tallmateriale. Spesielt ofte brukes det når man skal sammenlikne data for to forskjellige tidspunkt. Figurene tegnes slik at arealet er proporsjonalt med de gitte tallene

Anta for eksempel at omsetningen i år 2004 var 1 million kroner, mens den i 2005 økte til 2 millioner. Hvordan skal dette tegnes ved hjelp av to kvadrater?

Velger man for eksempel en firkant for 2004 som har side 1cm, så må firkanten for 2005 være $\sqrt{2}$ cm (hvorfor det?)



2004 2005

Hvorfor blir det galt å velge side = 2cm for kvadratet for 2005? Hvordan blir dette seende ut hvis man velger sirkler isteden og radius for 2004-sirkelen skal være 2cm?

Normal kvantilplot (Normal quantileplot)

I mange sammenhenger når vi senere skal drive med estimering og hypotesetesting så er en betingelse at tallmaterialet er normalfordelt (eller tilnærmet normalfordelt)

Det er flere måter å sjekke dette på. En måte er å tegne det såkalt **kvantilplottet**. Dette gjøres ved først å ordne tallmaterialet fra den minste til den største. Har en for eksempel de $n = 4$ observasjonene

$$6{,}9;\ 5{,}8;\ 6{,}7;\ 7{,}6$$

(som er trukket på kalk. randNorm(7,1,5) med 1 desimal) så blir dette ordnet:

$$x_{(i)}: 5{,}8;\ 6{,}7;\ 6{,}9;\ 7{,}6$$

De $n = 4$ ordnede observasjonene deler nå arealet under normalfordelingen i $(n+1) = 5$ like store deler som hver har et areal på $\frac{1}{(n+1)} = \frac{1}{5} = 0{,}2$

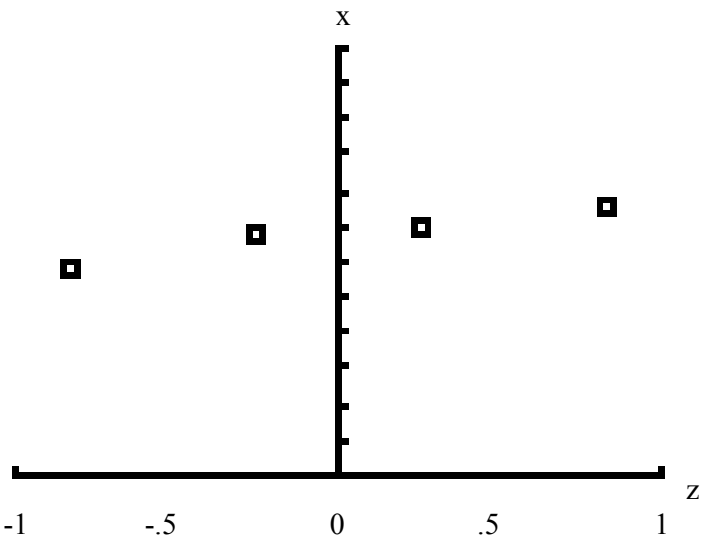
En beregner så z-scoren svarende til hver av disse 4 punktene. De blir som følger

$$z_1 = z_{0.20} = invNorm(0.20) = -0{,}84$$
$$z_2 = z_{0.40} = invNorm(0.40) = -0{,}25$$
$$z_3 = z_{0.20} = invNorm(0.60) = 0{,}25$$
$$z_4 = z_{0.20} = invNorm(0.80) = 0{,}84$$

En framstiller så punktene $(z_i, x_{(i)})$ i et zx- koordinatsystem, mao. punktene

(-0.84, 5.8), (-0.25, 6.7), (0.25, 6.9) og (0.84, 7.6)

Hvis disse punktene blir liggende tilnærmet på en rett linje så konkluderer en med at tallene kommer fra en normalfordelt populasjon. Legger en tallene inn på kalkulatoren, får en følgende grafiske bilde:

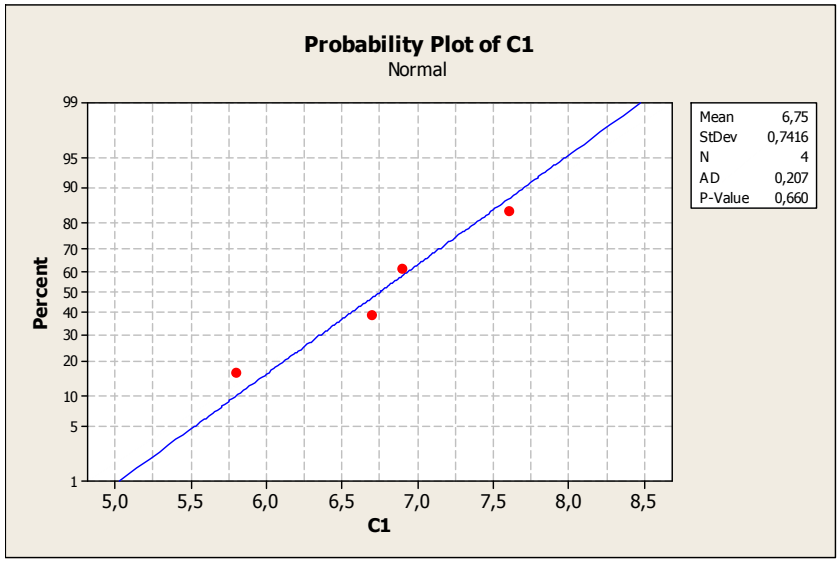


En ser at punktene tilnærmet ligger på svakt stigende rett linje, og konkluderer dermed at dataene er trukket fra en normalfordelt populasjon.

Hvis man nå isteden velger å bruke MINITAB ved først å legge de 4 observasjonene inn i kolonne 1 (=C1) og så bruke kommandoene:

STAT
BASIC STAT
Normality test
Select C1
Anderson-Darling

så får en følgende resultat:



Legg merke til at MINITAB avsetter de observerte x-verdiene langs førsteaksen og de kumulerte prosentvise z-scorene langs andreaksen. En ser også her etter om punktene blir liggende (ev tilnærmet liggende) på en rett linje. Det testes (se hypoteseprøving) også om dataene er normalfordelte gjennom følgende nullhypotese og alternativ hypotese:

H_0 : Dataene er trukket fra en normalfordelt populasjon

mot

H_A : Dataene er ikke trukket fra en normalfordelt populasjon

En ser at MINITAB angir en P-verdien på 0,66. Dette betyr mao. at H_0 ikke kan forkastes. Det er ganske sterke signaler på at H_0 er rett.
De to andre testene i MINITAB, Ryan-Joiner og Kolmogorov-Smirnov, gir helt tilsvarende resultater dog med litt lavere P-verdi.

Hvis dataene ikke er normalfordelte så kan man prøve om det hjelper med en transformasjon. Noen vanlige transformasjonene hvis dataene inneholder for mange store verdier er:

Logaritmisk transformasjon, dvs. beregn $y = ln(x)$

Kvadratrottransformasjon, dvs. beregn $y = \sqrt{x}$

Invers transformasjon, dvs. beregn $y = \frac{1}{x}$

Test så om de nye dataene (y-verdiene) er normalfordelte ved en av testmetodene over.

Noen andre vanlige transformasjoner hvis tallmaterialet inneholder for mange små verdier er:

Potenstransformasjon, dvs. beregn $y = x^a$ der $a > 1$

Eksponentiell transformasjon, dvs. beregn $y = a^x$ der $a > 1$

Test så om de nye dataene (y-verdiene) er normalfordelte ved en av testmetodene over.

Hvis ikke noe av dette fører fram så finnes det såkalte ikkeparametriske tester som kan brukes.

4. 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.

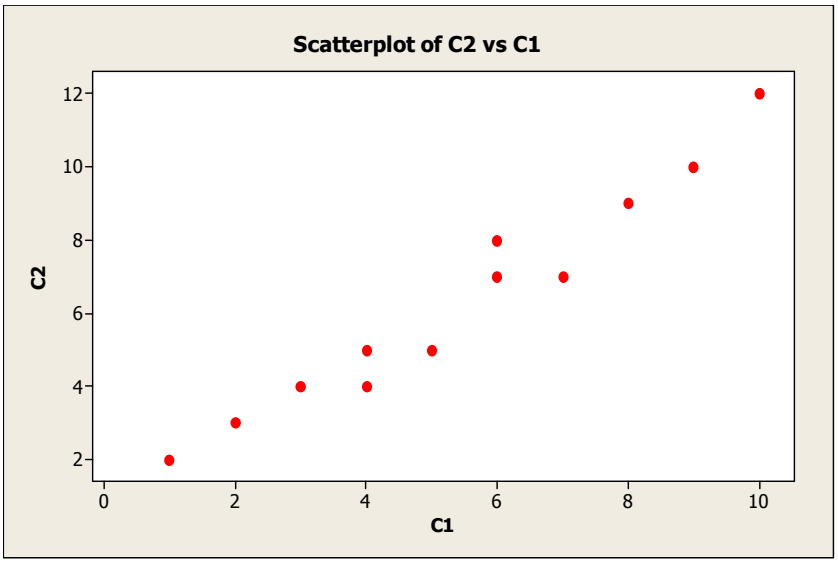
x_1	x_2	x_3	……..	x_n
y_1	y_2	y_3	……..	y_n

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

Spredningsdiagrammet blir i dette tilfellet



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

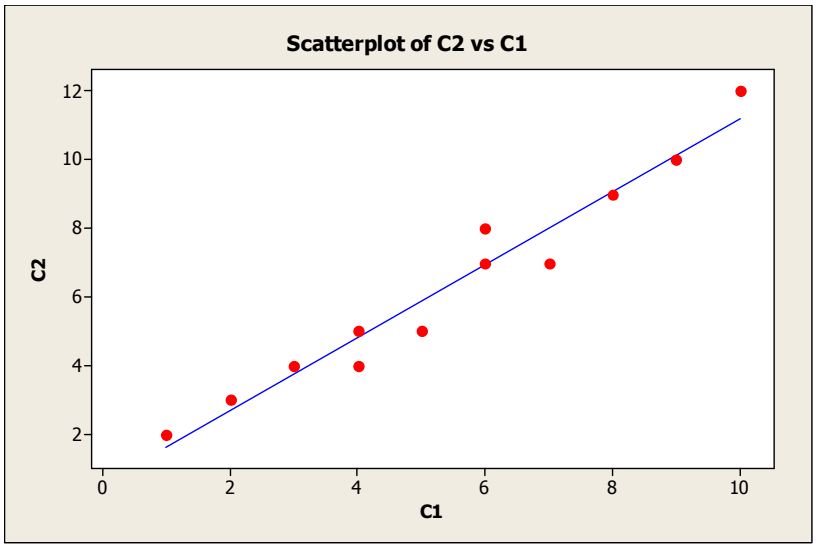
$$\text{forventning} = 0 \text{ og varians} = \sigma^2$$

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

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



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?.
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:

$$\begin{aligned} a\,n \qquad & + b(\sum_i x_i) = \sum_i y_i \\ a(\sum x_i) & + b(\sum_i x_i^2) = \sum_i x_i y_i \end{aligned}$$

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

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

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)

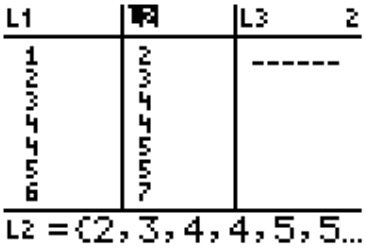
Nå skal vi kontrollere disse beregningene ved hjelp av kalkulatorens statistikkprogram og MINITAB.

Regresjonsanalyse ved hjelp av kalkulator:

Med TI-83 gjør man følgende :

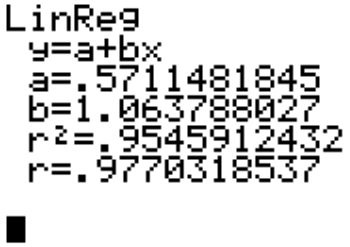
i) Trykk på STAT-tasten.
ii) Trykk på ENTER.

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 MINITAB:

STAT
REGRESSION
REGRESSION

Responce C2 (y-verdiene), Predictors C1 (x-verdiene)
OK

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)

Regression Analysis: C2 versus C1

The regression equation is
C2 = 0,571 + 1,06 C1

Predictor	Coef	SE Coef	T	P
Constant	0,5711	0,4428	1,29	0,226
C1	1,06379	0,07337	14,50	0,000

S = 0,676103 R-Sq = 95,5% R-Sq(adj) = 95,0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	96,096	96,096	210,22	0,000
Residual Error	10	4,571	0,457		
Total	11	100,667			

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

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

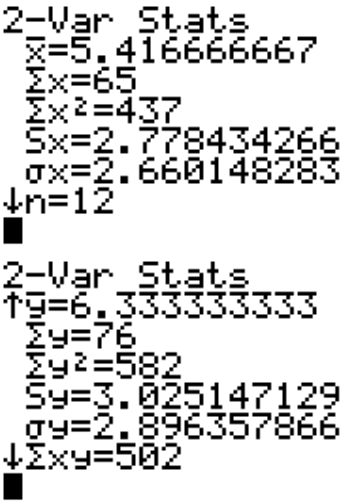
og at

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

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:

$$a = \bar{y} - b\bar{x} = \frac{76}{12} - 1{,}0638\cdot\frac{65}{12} = 0{,}571$$

5. 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

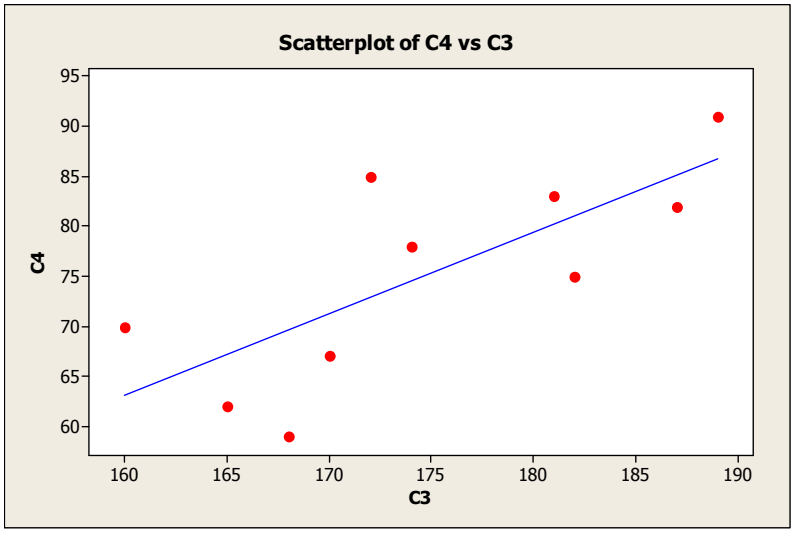
Merk nå at x_i ikke er gitte verdier av X som under regresjon, men verdier av en tilfeldig variabel, og det betyr at de ikke kan bestemmes på forhånd. M.a.o. er nå både x_i *og* y_i verdier av tilfeldige variable.

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 med regresjonslinje vil da se ut som følger:



Hvis man nå ber MINITAB å regne ut korrelasjonskoeffisienten ved kommandoene

STAT
BASIC STATISTICS
CORRELATION

får man følgende utskrift:

Correlations: Høyde; Vekt

```
Pearson correlation of Høyde and Vekt = 0,748
P-Value = 0,013
```

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 :

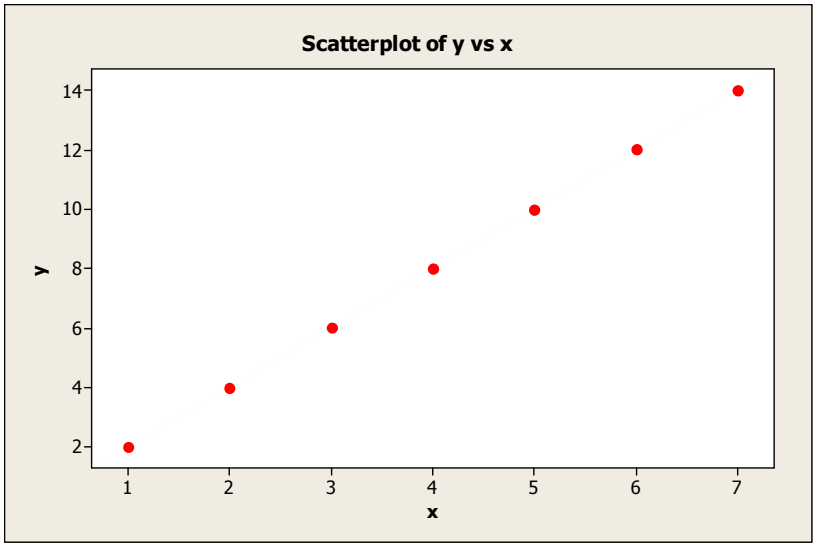
$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{75{,}6}{\sqrt{92{,}622}\sqrt{110{,}177}} = 0{,}748$$

som stemmer overens med MINITAB-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 korrelasjonskoeffisienten =1. Dette bekrefter også MINITAB:

Correlations: x; y

```
Pearson correlation of x and y = 1,000
P-Value = *
```

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 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

MINITAB gir nå:

Correlations: x; y

```
Pearson correlation of x and y = -0,954
P-Value = 0,003
```

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.

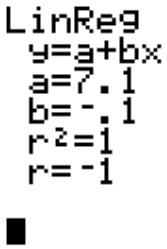
x	1	2	3	4	5
y	7,0	6,9	6,8	6,7	6,6

MINITAB viser nå:

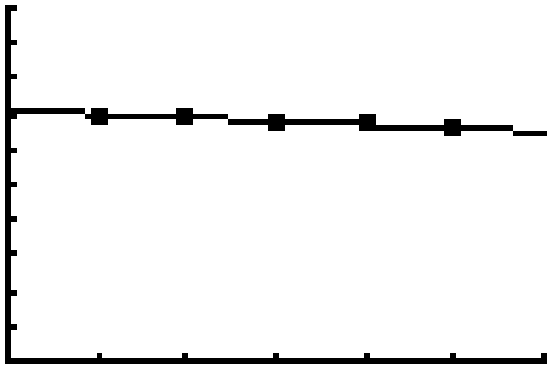
Correlations: x; y

```
Pearson correlation of x and y = -1,000
P-Value = *
```

Bruker en kalkulatoren finner en:



Tegner en spredningsdiagrammet sammen med regresjonslikningen finner en:



Her har man en perfekt lineær negativ sammenheng (hvis man kan se bort ifra den dårlige grafikken på kalkulatoren) , og finner dermed en korrelasjonskoeffisienten lik -1 (Kontroller selv at på etter eller annet vis at dette stemmer)

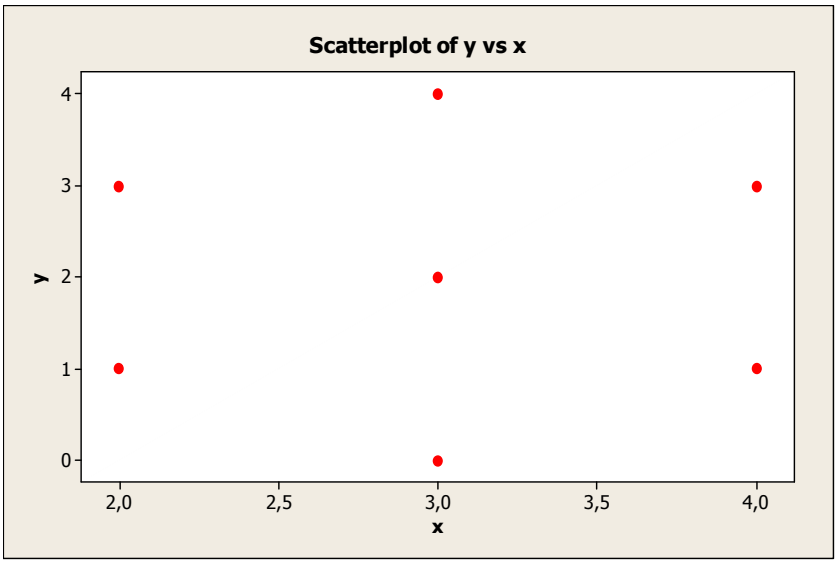
I de situasjonene vi har sett på her har r vært i nærheten av 1 eller -1, men i mange situasjoner er r i nærheten av 0. Det betyr at utvalget viser at det tyder på at det ikke er noen grad av lineær sammenheng mellom de to variablene som er involvert, og det er en viktig konklusjon og eventuelt komme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

Spredningsdiagrammet blir i denne situasjonen:



MINITAB gir nå:

Correlations: x; y

```
Pearson correlation of x and y = 0,000
P-Value = 1,000
```

En ser av tabellen og spredningsdiagrammet 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 andre 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)

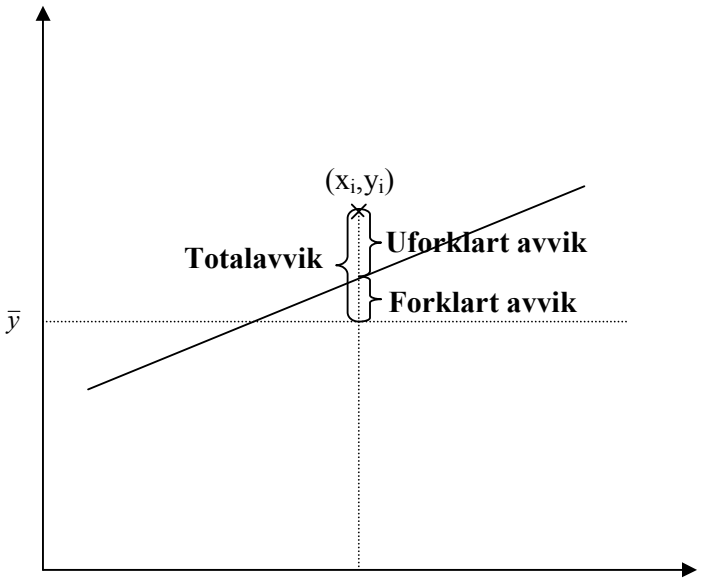
det såkalte **forklarte avviket** ved $\hat{y}_i - \bar{y}$ (eng.: explained deviation) og

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

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

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 MINITAB så får man bl.a. følgende utskrifter ved å gjøre en regresjonsanalyse

Regression Analysis: y versus x

```
The regression equation is
   y = - 67,5 + 0,816 x
```

```
Predictor     Coef  SE Coef      T      P
Constant   -67,48    44,77  -1,51  0,170
x          0,8162   0,2558   3,19  0,013


S = 7,38448   R-Sq = 56,0%   R-Sq(adj) = 50,5%


                Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  555,36  555,36  10,18  0,013
Residual Error   8  436,24   54,53
Total            9  991,60
```

Herav ser en bl.a. angivelsen av den kvadrerte korrelasjonskoeffisienten $R^2 = R\ Square = 0{,}56$ som stemmer med $R = \sqrt{0{,}56} = 0{,}748$ som vi beregnet tidligere.

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.

6. Ikkelineær regresjon

Vi skal nå se på hvorledes vi kan bruke kalkulatoren og MINITAB til å bestemme andre trender enn lineære. Dvs. vi skal se på en del enkle ikkelineære funksjoner f som passer til de gitte dataene. Utgangspunktet er igjen minste kvadraters metode, dvs vi bestemmer funksjonen f slik at

$$\sum_i (\hat{y} - f(x))^2$$

minimeres. Dette gjøres igjen ved å partiellderivere dette utrykket med hensyn på de forskjellige parametrene som inngår i utrykket og sette alle disse utrykkene lik 0. Dette gir p likninger med p ukjente hvis det er p parametre som inngår i modellen.

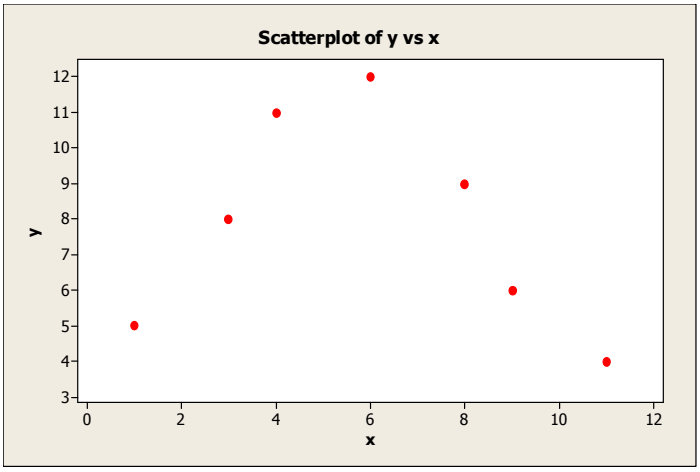
Kvadratisk regresjon.

Anta at vi nå har et tallmateriale som har en trend hvor det er tydelig å se at det passer dårlig med en lineær tilnærming.

Eks. Anta at vi har observert følgende sammenheng mellom x og y:

x	1	3	4	6	8	9	11
y	5	8	11	12	9	6	4

Tegner en spredningsdiagrammet ser at det passer dårlig med en rettlinjet modell, men derimot bedre med en kvadratisk modell.

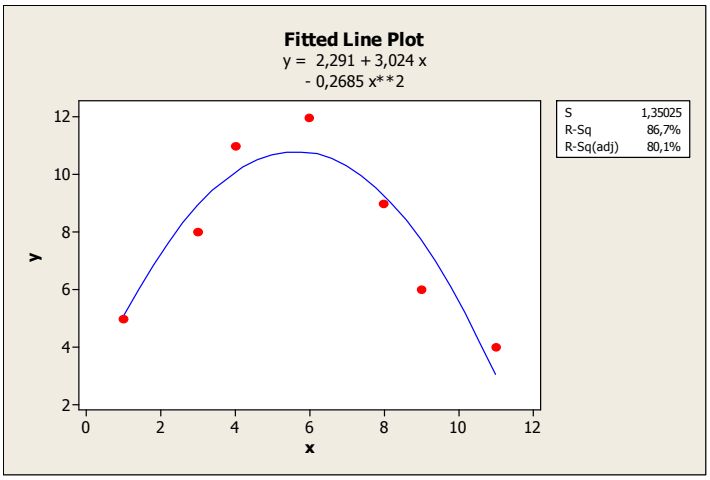


Vi antar nå at vi har modellen

$$\hat{y} = f(x) = ax^2 + bx + c$$

Lar nå først MINITAB prøve å finne denne annengradsfunksjonen . En bruker da følgende kommandoer:

STAT
REGRESSION
FITTED LINE PLOT
QUADRATIC og finner.



En ser mao. at MINITAB foreslår at y kan estimeres ved hjelp av

$$\hat{y} = f(x) = ax^2 + bx + c = -0{,}2685x^2 + 3{,}024x + 2{,}291$$

Er det mulig å kontrollere denne modellen? Ifølge minste kvadraters metode skal altså modellen bestemmes slik at

$$\sum_i (y - \hat{y})^2 = \sum_i (y - (ax^2 + bx + c))^2$$

minimeres. Dette utrykket er en funksjon av de tre parametrene *a, b* og *c*. Minimum av denne funksjonen finner en av de tre likningene :

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

Nå kan det vises at dette fører til følgende normallikninger:

$$\begin{aligned} an + b\sum_i x_i + c\sum_i x_i^2 &= \sum_i y_i \\ a\sum_i x_i + b\sum_i x_i^2 + c\sum_i x_i^3 &= \sum_i x_i y_i \\ a\sum_i x_i^2 + b\sum_i x_i^3 + c\sum_i x_i^4 &= \sum_i x_i^2 y_i \end{aligned}$$

Dette er mao. tre likninger med de tre ukjente *a, b* og *c*.

Å finne disse summene og løse dette likningssystemet er en teknologisk men overkommelig utfordring på kalkulatoren:
Start med å legge inn x- og y-verdiene i liste 1 og 2. 2-variabelstatistikk på liste 1 og liste 2 gir da følgende summer:

$$\sum_i x_i = 42,\ \sum_i x_i^2 = 328,\ \sum_i y_i = 57 \text{ og } \sum_i x_i y_i = 323$$

For å finne de resterende 3 summene $\sum_i x_i^3$, $\sum_i x_i^4$ og $\sum_i x_i^2 y_i$ må en lage nye lister ved hjelp av L_1 og L_2. Lag så $L_3 = L_1^3$ og utfør 1-variabelstatistikk på L_3. Dette gir da

$$\sum_i x_i^3 = 2880$$

Lag deretter $L_4 = L_1^3$ og utfør 1-variabelstatistikk på L_4. Dette gir:

$$\sum_i x_i^4 = 26932$$

Tilslutt finner en (prøv selv)

$$\sum_i x_i^2 y_i = 154327$$

Likningssystemet blir da på formen:

$$\begin{aligned} 7a + \ 42b + \ 328c &= 57 \\ 42a + \ 328b + \ 2880c &= 323 \\ 328a + 2880b + 26932c &= 154327 \end{aligned}$$

Dette kan løses på mange måter ved hjelp av kalkulatoren. Bruker en matrisenotasjon (forutsetter MAT 415) har en:

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 7 & 42 & 328 \\ 42 & 328 & 2880 \\ 328 & 2880 & 26932 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 57 \\ 323 \\ 154327 \end{bmatrix}$$

Ved hjelp av kalkulatoren finner en nå

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} -0{,}2685 \\ 3{,}024 \\ 2{,}291 \end{bmatrix}$$

som stemmer med beregningene i MINITAB.

Kubisk regresjon.

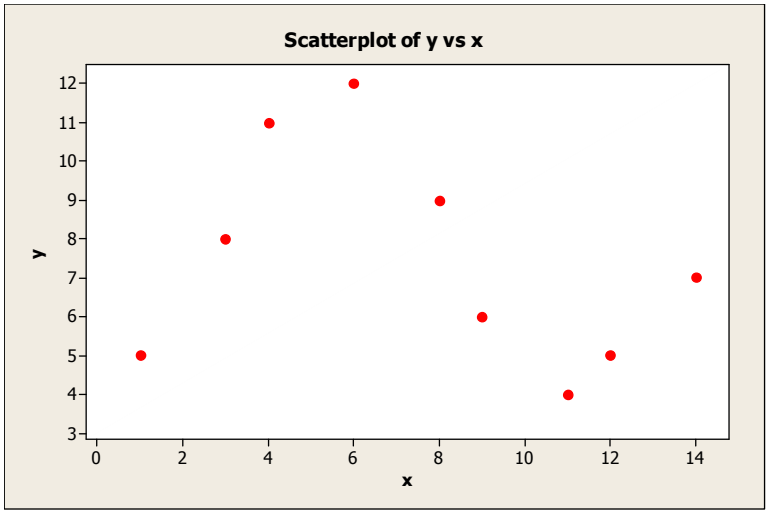
Anta nå at spredningsdiagrammet har en form som gjør at det er mer naturlig å bruke en tredjegradskurve som modell, dvs anta at

$$\hat{y} = f(x) = ax^3 + bx^2 + cx + d$$

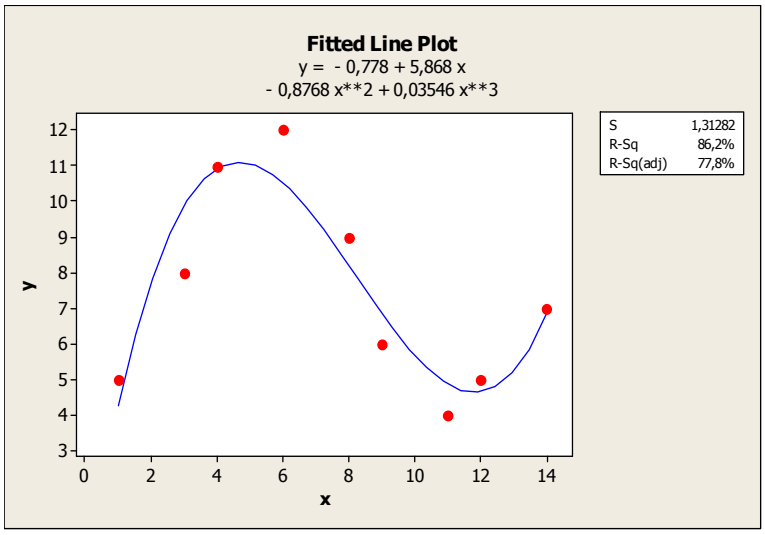
Eks. Anta at man nå har observert følgende sammenhørende verdier mellom X og Y:

X	1	3	4	6	8	9	11	12	14
Y	5	8	13	12	9	6	4	5	7

Spredningsdiagrammet blir da:



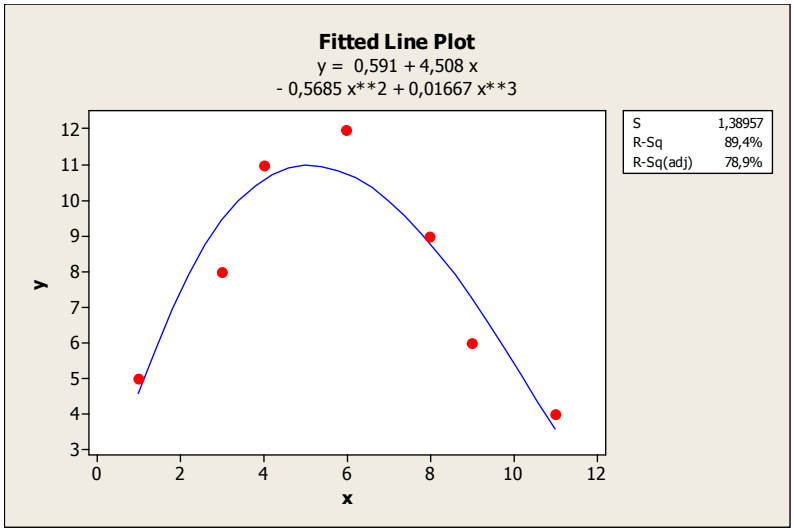
MINITAB gir nå følgende modell med Fitted Line Plot:



En finner mao.

$$\hat{y} = f(x) = ax^3 + bx^2 + cx + d = 0{,}03546x^3 - 0{,}8768x^2 + 5{,}868x + 0{,}778$$

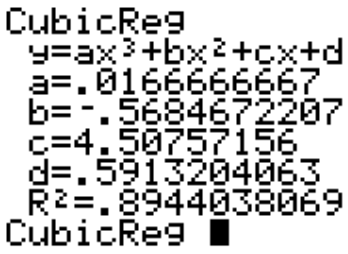
I noen sammenhenger vil også en kubisk modell kunne være bedre enn en kvadratisk modell. Prøver vi nå å tilpasse en tredjegradskurve til dataene i eksempelet under kvadratisk regresjon finner en:



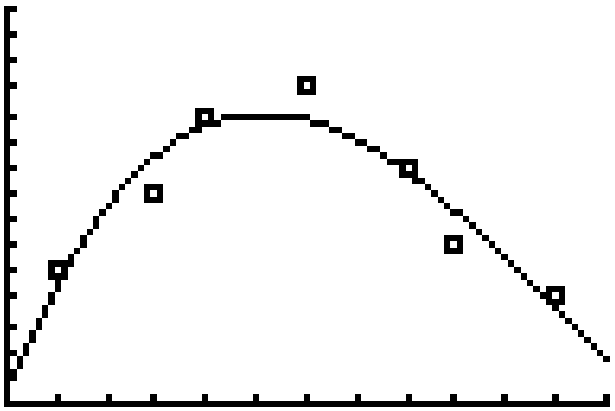
Nå ser en at forklaringsgraden har økt fra 86,7% (kvadratisk modell) til 89,4% (kubisk modell)

Hvis man prøver det tilsvarende på kalkulatoren finner en etter å ha utført kommandoene

STAT
CALC
6:CubicReg
ENTER
L1,L2
ENTER



Tegner en spredningsdiagrammet sammen med den funnene tredjegradskurven viser TI:

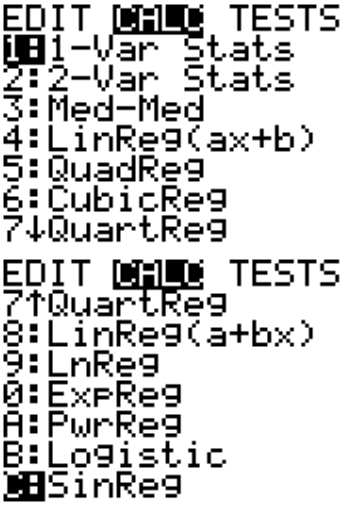


En ser at dette stemmer perfekt med det som en fant i MINITAB.

Ser en nærmere på TI ser en at kalkulatoren har et mye bedre utbygd apparat (flere muligheter) enn det en finner på LineFit i MINITAB.

STAT
CALC

viser følgende muligheter:



En har i tillegg til lineær regresjon (både y=ax+b og y=a+bx), kvadratisk regresjon (modell y=ax^2+bx+c) og kubisk regresjon (modell y=ax^3+bx^2+cx+d) som MINITAB har også følgende:

Med-Med regresjon som finner en lineær regresjonslinje som baser seg på å beregne medianen for x-ene og for y-ene og la disse legge grunnlaget for avviksmålene i regresjonsanalysen i motsetning til vanlig lineæer regresjon hvor en bruker gjennomsnittlig x-verdi og gjennomsnittlig y-verdi. Dette er fordelaktig når en har enkelte ekstremverdier idet medianene ikke lar seg så lett påvirke av slike som gjennomsnittene.
Ser en nå på eksempelet på side 43

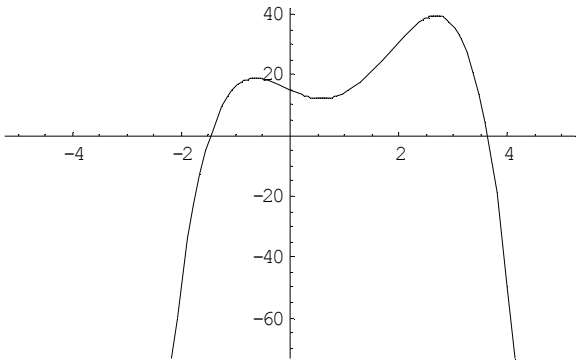
x	160	165	189	187	182	168	181	170	174	172
y	70	62	91	82	75	59	83	67	78	85

hvor vi brukte lineær regresjon finner vi nå ved hjelp av Med-Med –analyse

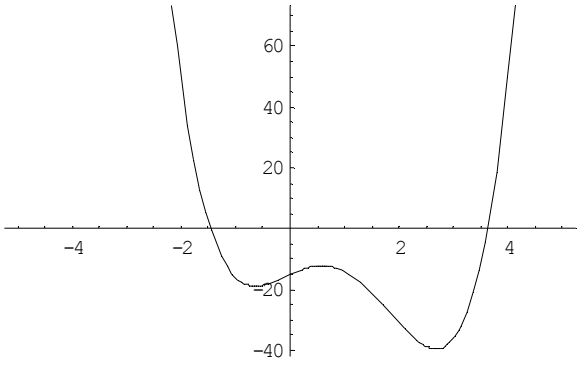
```
Med-Med
 y=ax+b
 a=.9090909091
 b=-84.25757576
```

Mao. En finner nå y = -84,3 + 0,909x mot y = -67,5 + 0,816x. Prøv selv å finne ut hvilken av de to analysene som har størst forklaringsgrad.

7: QuartReg hvor modellen er den generelle 4.gradskurven y=ax^4+bx^3+cx^2+dx+e finner en ikke i MINITAB. En slik kurve kan være (generelt) enten på formen



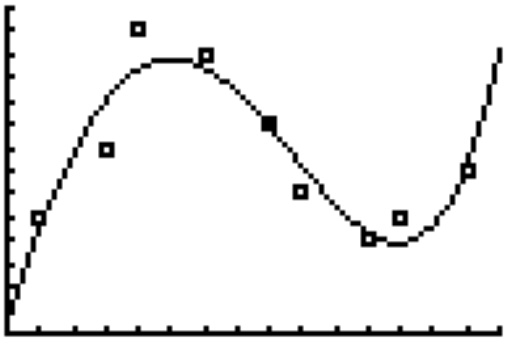
eller



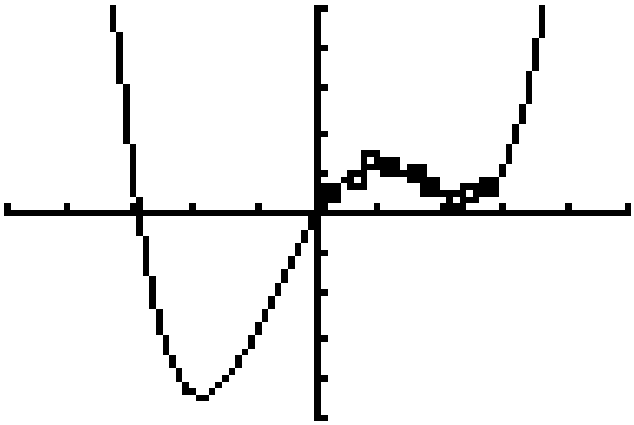
En behøver imidlertid ikke ha et forløp som følger hele kurven. Prøver en nå med en fjerdegradsmodell på eksempelet på side 47 finner en ifølge TI:

$$y = 0{,}00202x^4 - 0{,}02016x^3 - 0{,}40178x^2 + 4{,}51904x + 0{,}488707$$

med en $R^2 = 0{,}8729$ mot $R^2 = 0{,}8944$ med en tredjegradsmodell. Mao. en litt dårligere forklarigsgrad.
Spredningsdiagrammet og den tilpassede fjerdegradskurven i 1. kvadrant viser da følgende.



Tegner en "hele" kurven ser en at kurvens forløp blir som følger:

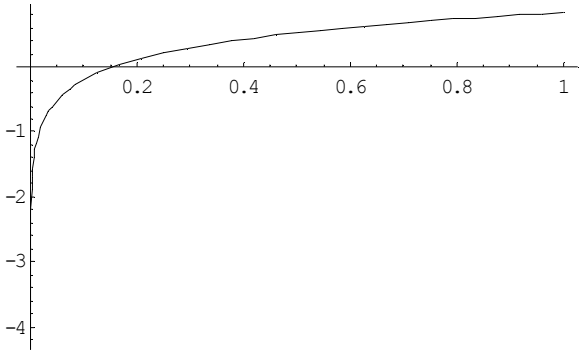


Nå skal man imidlertid være svært forsiktig med å bruke modellen for langt ut over det området hvor man har gjort sine observasjoner.

9: LnReg (logaritmisk regresjon) bruker modellen $y = a+b\ln x$. En antar nå med andre ord at det er en lineær sammenheng mellom Y og den naturlige logaritmen til X. Det man nå gjør er først å beregne de naturlige logaritmene til de observerte x-verdiene og deretter utfører en vanlig lineær regresjon på det nye tallmaterialet (y og lnx). Nå har vi vist i matematikkurset i høst at

$$y = a + b\ln x \Leftrightarrow y = \ln A + \ln x^b \Leftrightarrow y = \ln(A \cdot x^b)$$

der en har satt $a = \ln A$
Har en $A = 2{,}3$ og $b = 0{,}45$ vil kurven se ut som følger:



0: ExpReg (eksponentiell regresjon) tar utgangspunkt i modellen

$$y = ab^x$$

Herav har en

$$\ln y = \ln a + x \ln b \Leftrightarrow \ln y = A + Bx$$

Der en har satt $\ln a = A \quad og \quad \ln b = B$. Det betyr at nå kan en utføre lineær regresjon på tallmaterialet bestående av de opprinnelige x-verdiene og den naturlige logaritmen til de observerte y-ene, og deretter regne seg tilbake til den opprinnelige modellen Dette er ikke nødvendig med TI idet en får resultatet direkte.

Anta for eksempel at man har observert følgende sammenhang mellom x og y:

x	1	2	3	4	6	7
y	6	17	57	160	480	1463

Bruker en nå kommandoene

STAT
CALC
0:ExpReg
L_1, L_2
ENTER

gir TI

$$y = 1{,}975x^{3{,}007}$$

A:PwrReg (potensregresjon) bruker modellen

$$y = ax^b$$

Herav har en

$$\ln y = \ln a + b \ln x \Leftrightarrow z = A + Bw$$

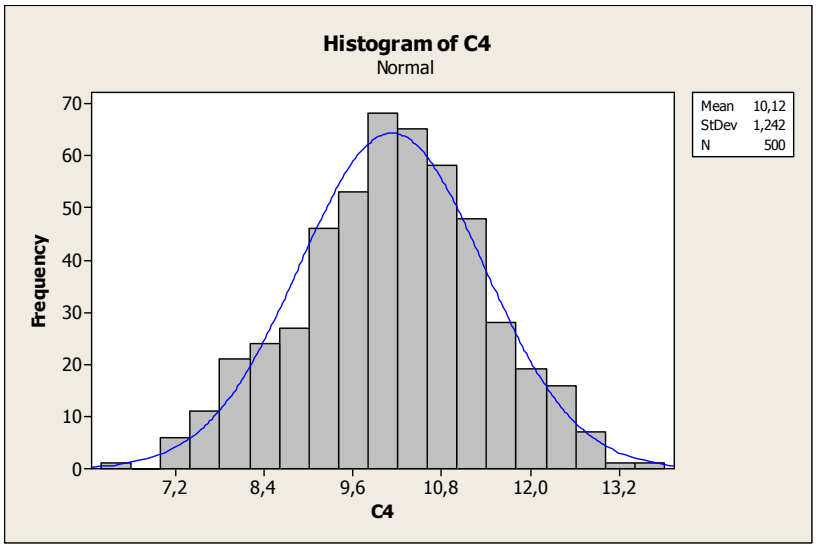
dvs. at man nå kan utføre lineær regresjon dataene $(w, z) = (\ln x, \ln y)$. Igjen så kan modellen finnes direkte av TI uten å gå veien om lineær regrsjon.

Det finnes også logistisk regresjon og sinustilpasset regresjon, men disse skal vi ikke komme inn på her.

7. Noen viktige kontinuerlige fordelinger. Sentralgrense-teoremet

Vi skal nå kort repetere noen viktige temaer fra den grunnleggende delen av statistikkurset (normalfordelingen og sentralgrensesetningen), men også ta opp en del andre viktige kontinuerlige fordelinger (t-fordelingen, kjikvadratfordelingen og Fisherfordelingen) som vi ikke har drøftet.

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.



Dataene over er fremkommet ved å la MINITAB simulere 500 normalfordelte data med μ = 10,0 og σ =1,2. dette gjøres ved å bruke kommandoene:

CALC
RANDOM DATA
Normal
Generate 500 rows of data
Sett Mean=10,0 og standarddeviation=1,2
Store in (for eksempel) C4
OK

En ser at de 500 simulerte dataene gir et gjennomsnitt på 10,12 og et standardavvik på 1,241. En normalfordeling er tegnet inn sammen med histogrammet.

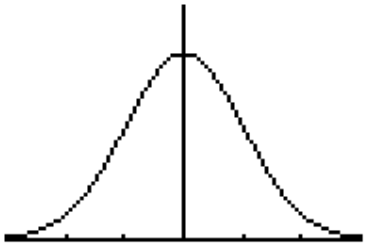
.
Vi skal nå bare innledningsvis se på noen egenskaper knyttet til normalfordelingen.

- 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 μ = 0 og σ =1

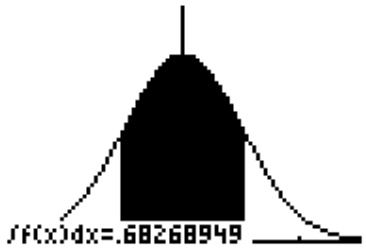
- Sannsynligheten for at utvalg av normalfordelte data skal ligge mellom det teoretiske gjennomsnittet minus et standardavvik og det teoretiske gjennomsnittet pluss et standardavvik er tilnærmet 0,68. Det vil m.a.o. si at når man har normalfordelte data så

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

Da er $P(X \leq 45) = 0{,}8689$. 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):

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

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

I mange av de anvendelsene vi skal se på bryr man seg ikke om denne 0,5-korreksjonen.

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. Se for øvrig Lillestøls bok.

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

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

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

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

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

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

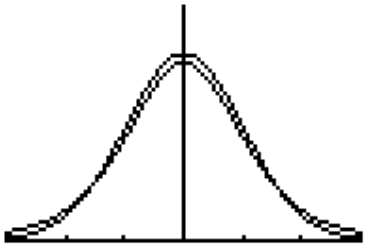
Dette gjør det lettere å finne utrykket for t-fordelingen når v er gitt. Anta f..eks. at antall frihetsgrader v = 5. Da blir sannsynlighetstettheten til t

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

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 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, …, 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 $v = 6$ (= antall frihetsgrader) så er

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

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

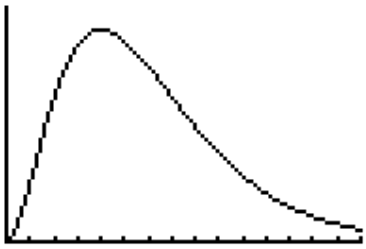
$Y =$

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 v frihetsgrader så vil en ha at for store v så er tilnærmet

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

Prøv for eksempel med $v = 50$ 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:

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

Fisherfordelingen.

Den siste fordelingen som vi skal se på er den såkalte Fisherfordelingen som er oppkalt etter en av verdens mest berømte statistikere gjennom tidene (Sir Ronald Fisher). Den blir brukt i forbindelse med såkalt variansanalyse, som skal brukes i forbindelse med testing av flere (enn tre) gjennomsnitt opp mot hverandre og i forbindelse med regresjonsanalyseanalyse.

Hvis $X \sim \chi^2_{v_1} \quad og \quad Y \sim \chi^2_{v_2}$ så kan det vises at $Z = \frac{X}{Y} \sim F_{v_1, v_2}$ (leses: er Fisherfordelt med v_1 *og* v_2 frihetsgrader. Mao. En Fisherfordelt variabel fremkommer ved å dele to kjikvadratfordelte variable på hverandre. Sannsynlighetstettheten til Z er gitt ved :

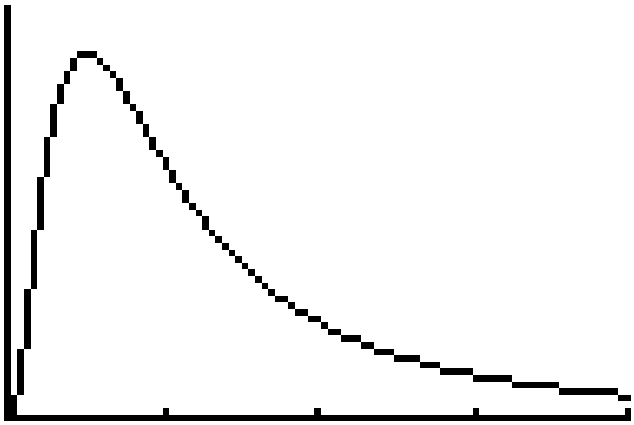
$$f(z) = \frac{\Gamma(\frac{v_1 + v_2}{2})}{\Gamma(\frac{v_1}{2}) \cdot \Gamma(\frac{v_2}{2})} \cdot (\frac{v_1}{v_2})^{\frac{v_1}{2}} \cdot z^{\frac{v_1}{2}-1} \cdot (1 + \frac{v_1}{v_2} \cdot z)^{-\frac{1}{2}(v_1+v_2)} \text{ for } z > 0.$$

For øvrig så er $f = 0$.

Har en for eksempel $v_1 = 8 \; og \; v_2 = 4$ får en

$$f(z) = \frac{\Gamma(6)}{\Gamma(4) \cdot \Gamma(2)} \cdot (2)^4 \cdot z^3 \cdot (1 + 2z)^{-6} = 320 z^3 (1 + 2z)^{-6} \text{ for z} > 0$$

Legger en dette utrykket inn på kalkulatoren fremkommer følgende graf :



En ser at grafen har en form som likner litt på kjikvadratfordelingen. Den er entoppet og nærmer seg asymptotisk mot 0.

Nå kan også Fisherfordelingen hentes direkte på kalkulatoren. Ved hjelp av kommandoene

Y=
2ND VARS
DISTR

8: FPdf(
ENTER
$Y_1 = F_P(X,8,4)$
GRAPH

får en nøyaktig samme graf som over.

8. Statistisk inferens.

Vi skal nå innledningsvis repetere noen av de viktigste begrepene i inferens fra det grunnleggende statistikkpensumet. Statistisk inferens består av to hoveddeler som ble betegnet med estimering og hypoteseprøving. Estimering går ut på å anslå verdier knyttet til ukjente størrelser i populasjonen. Hypoteseprøving går ut på teste påstander knyttet til de samme ukjente størrelsene.

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 ofte (ved symmetriske fordelinger) 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

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

der $S\hat{E}(\hat{\theta})$ 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

$$\hat{\theta} \pm f_{\alpha/2} S\hat{E}(\hat{\theta})$$

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 konfidensintervallet 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.

9. 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: Soppen er giftig

mot

H_A: 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 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 $X \leq k$. Kritisk verdi k bestemmes slik at

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

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

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

Herav får en da følgende likning

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

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

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

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.

Vi skal nå se på en del viktige klassiske situasjoner som er mye i bruk i anvendelser. De bygger alle på denne innledende teorien i estimering og hypoteseprøving, men bærer mye preg av en "kokebokoppskrift".

10. Inferens for ett populasjonsgjennomsnitt.

Hypotesetesting.

Anta at vi har en populasjon hvor det er en eller annen interessant størrelse med et gjennomsnitt på μ (ukjent eller eventuelt kjent fra tidligere målinger). Anta at man nå mener at det har skjedd endringer i populasjonen slik at μ har endret seg på den ene eller andre måten. Dette er en situasjon hvor en kan bruke inferens. Anta at populasjonsvariansen σ^2 er kjent (fra tidligere eller liknende målinger).

Anta at vi nå ønsker å teste

$$H_0 : \mu = \mu_0$$

mot et eller annet alternativ (ensidig eller tosidig)

Vår testobservator er nå

$$Z = \frac{Estimat - hypoteseprøvingsverdien}{s\tan dardavviket\ til\ estimatet} = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

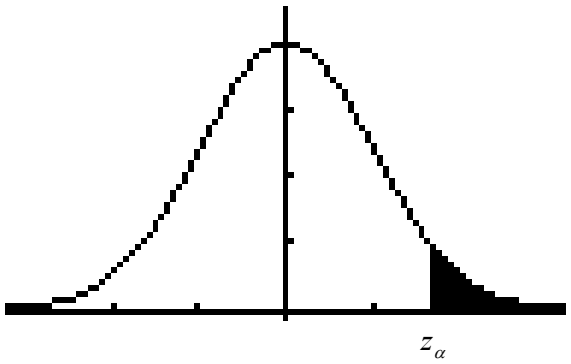
Legg merke til at den måler differansen mellom det gjennomsnittet som man nå finner i utvalget og det gjennomsnittet man har hatt i populasjonen. I henhold til sentralgrensesetningen er *Z* normalfordelt. Dette er viktig i forhold til de beregningene som vi etter hvert skal gjøre.

I: Vi ser nå først på testing av

$$H_0 : \mu = \mu_0 \text{ mot } H_A : \mu > \mu_0$$

i) Klassisk testing: H_0 forkastes (idet store verdier av Z er signifkante) hvis *z* (=verdien av Z) $\geq z_\alpha$ = kritisk verdi på nivået α (z_α =1,645 hvis $\alpha = 0{,}05$)

ii) Beregning av signifikanssannsynlighet (=P-verdi). P-verdien = $P_{H_0}(Z \geq z)$



Eks. Anta at vi ønsker å teste

$$H_0 : \mu = 4{,}5 \text{ mot } H_A : \mu > 4{,}5$$

med nivået $\alpha = 0{,}05$. Et tilfeldig utvalg på n = 25 viste et gjennomsnitt på 4,8. Kristisk verdi $z_{0{,}05}$ er nå 1,645. Den finner en som vist i grunnleggende statistikk ved hjelp av TI og følgende kommandoer:

2ND VARS
3: invNorm
ENTER
0.95) ENTER

Kalkulatoren viser da

```
invNorm(.95)
_     1.644853626
```

Nå blir verdien av testobservatoren

$$z = \frac{4{,}8 - 4{,}5}{\frac{0{,}6}{\sqrt{25}}} = 2{,}5$$

En ser da at $z = 2{,}5 > 1{,}645$ = kritisk verdi. Konklusjonen blir dermed at $H_0 : \mu = 4{,}5$ forkastes på 5%-nivået og vi påstår dermed $H_A : \mu > 4{,}5$.

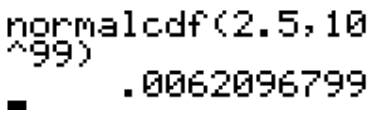
Beregner en nå alternativt P-verdien finner en ved hjelp av kalkulatoren følgende :

$$P_{H_0}(Z \geq z) = P_{H_0}(Z \geq 2{,}5) = 0{,}0062$$

Siden denne er mindre enn nivået på 0,05 så ser en at nullhypotesen forkastes på 5%-nivået. Det en imidlertid også ser er at H_0 forkastes helt ned til 0,62%-nivået. Forklar hvorfor.
Beregningen av P-verdien gjøres med følgende kommandoer:

2ND VARS
2:normalcdf(
ENTER
$2.5,10^{99}$)
ENTER

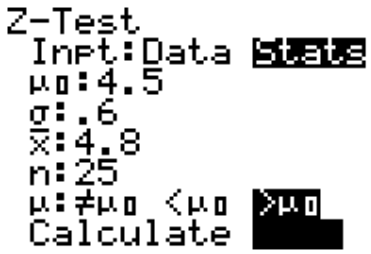
Kalkulatoren viser nå:



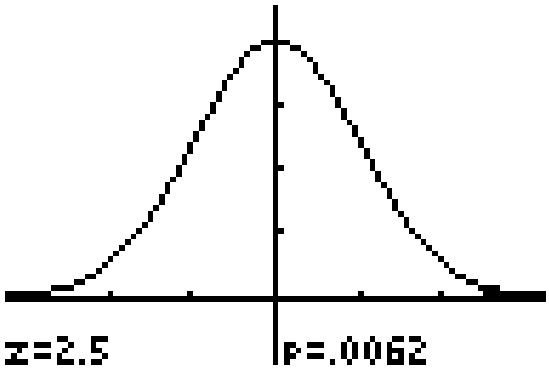
Velger en nå å bruke kalkulatorens "testpakke" finner en ved følgende kommandoer:

STAT
TESTS
1:Z-Test
ENTER

Her må en nå legge inn de gitte verdiene for $\mu_0, \sigma, \bar{x}$ *og* n. Dessuten må kalkulatoren få vite hva alternativet skal være. En får da følgende bilde på kalkulatoren:



Her fortelles kalkulatoren at det skal tegnes en normalfordeling og angis en p-verdi.(kommandoen Draw er skjult bak det svarte feltet helt nederst til høyere, kurser står her og blinker, og dette resulterer i det svarte feltet) Trykker en ENTER får en følgende resultat:



Velger en isteden kommandoen CALCULATE får en følgende resultat:

```
Z-Test
 μ>4.5
 z=2.5
 p=.0062096799
 x̄=4.8
 n=25
```

Prøver nå å bruke MINITAB for å løse den samme oppgaven. En gjør da bruk av følgende kommandoer:

Stat
Basic Statistics
1 Z 1–Sample Z…..

Her må en så velge alternativet Summarized data, og deretter legge inn

Sample size på 25, Mean på 4,8, Standarddeviation på 0,6, test mean på 4,5 og tilslutt under options velge alternativet greater than. Velger en så OK får en følgende utskrift:

One-Sample Z

```
Test of mu = 4,5 vs > 4,5
The assumed standard deviation = 0,6

                          95%
                          Lower
N     Mean  SE Mean    Bound     Z      P
25  4,80000  0,12000  4,60262  2,50  0,006
```

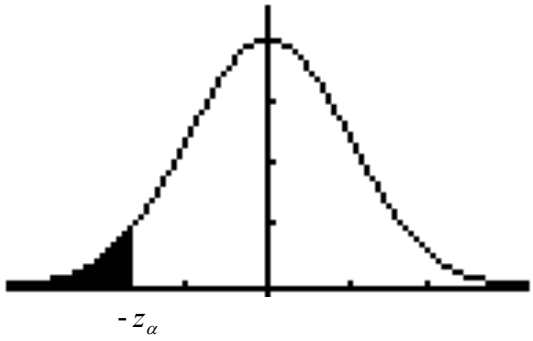
MINITAB gir ingen grafikk ved Summarized data.

II: Anta at vi nå ønsker å teste

$$H_0 : \mu = \mu_0 \text{ mot } H_A : \mu < \mu_0$$

Ved klassisk testing så forkastes H_0 hvis z (=verdien av Z) $\leq -z_\alpha$ = kritisk verdi på nivået α

i) .(Nå er små verdier av testobservatoren signifikante)



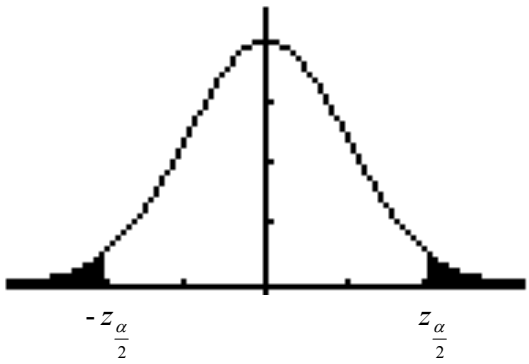
Kritisk verdi er nå det samme som ved testingen under I, men med motsatt fortegn.

ii) P-verdien blir nå = $P_{H_0}(Z \le z)$

III: Anta at vi nå ønsker å teste

$$H_0 : \mu = \mu_0 \text{ mot } H_A : \mu \neq \mu_0$$

i) Ved klassisk testing så forkastes nå H_0 hvis $z \le -z_{\frac{\alpha}{2}}$ eller hvis $z \ge z_{\frac{\alpha}{2}}$.(Nå er både små og store verdier av testobservatoren signifikante) Prøv å forklare hvorfor!



ii) P-verdien blir nå = $2\,P_{H_0}(Z \le z)$ der z som foran er den observerte verdien av *Z*. Legg merke til at en multipliserer med 2 fordi man nå driver med tosidig testing.

Eks. Anta at vi ønsker å teste

$$H_0 : \mu = 7{,}3 \text{ mot } H_A : \mu \neq 7{,}3$$

med nivået $\alpha = 0{,}05$. Et tilfeldig utvalg på n = 30 viste et gjennomsnitt på 6,9. Kritisk verdier $-z_{0,025}$ og $z_{0,025}$ er nå -1,96 og 1,96. $z_{0,025}$ finner en som vist i grunnleggende statistikk ved hjelp av TI og følgende kommandoer:

2ND VARS
3: invNorm
ENTER
0.975) ENTER

Hvorfor brukes verdien 0,975? Den tilsvarende negative kritiske verdien $-z_{0,025}$ er den som (p.g.a. symmetri) har motsatt fortegn av den kritiske verdien vi fant.) Ved klassisk testing så forkastes mao. nå H_0 hvis $z \leq -1{,}96$ eller hvis $z \geq 1{,}96$.
Nå blir verdien av testobservatoren

$$z = \frac{6{,}9 - 7{,}3}{\frac{1{,}2}{\sqrt{30}}} = -1{,}83$$

En ser da at $z = -1{,}83 > -1{,}96 =$ kritisk verdi. Konklusjonen blir dermed at $H_0 : \mu = 7{,}3$ ikke kan forkastes på 5%-nivået.

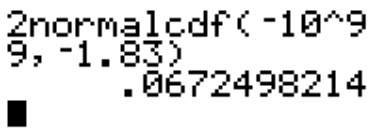
Beregner en nå alternativt P-verdien finner en ved hjelp av kalkulatoren følgende :

$$2\,P_{H_0}(Z \geq z) = 2\,P_{H_0}(Z \leq -1{,}83) = 0{,}0679$$

Dette finner en ved hjelp av følgende kommandoer på kalkulatoren

2
2ND VARS
2:normalcdf(
ENTER
-10^{99},-1.83)
ENTER

Kalkulatoren viser nå:



Konfidensintervall.

Anta nå som foran at vi har en normalfordelt populasjon hvor det er en eller annen interessant størrelse med et gjennomsnitt på μ (ukjent eller eventuelt kjent fra tidligere målinger). Anta at man nå mener at det har skjedd endringer i populasjonen slik at μ har endret seg på den ene eller andre måten. Dette er en situasjon hvor en kan bruke inferens. Anta at populasjonsvariansen σ^2 er kjent (fra tidligere eller liknende målinger) og at vi nå ønsker å finne et konfidensintervall for μ. Vi tar også nå utgangspunkt i variabelen

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

som vi gjorde ved hypoteseprøvingen. Hvis vi nå ønsker et $(1-\alpha)100\%$ konfidensintervall for μ og Z er normalfordelt må følgende gjelde:

$$P(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}) = 1-\alpha$$

Herav har en da

$$P(-z_{\frac{\alpha}{2}} \le \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le z_{\frac{\alpha}{2}}) = 1-\alpha$$

Vi skal nå behandle den dobbelt ulikheten inne i parentesen som vi løser en hvilken som helst vanlig dobbel ulikhet an typen

$$-2 \le \frac{5-x}{4} \le 2 \quad \text{m.h.p. } x$$

slik at vi nå blir stående med en ulikhet knyttet til μ tilsvarende at ulikheten med x gir

$$-8 \le 5-x \le 8 \qquad \text{dvs}$$

$$-8-5 \le -x \le 8-5 \qquad \text{dvs}$$

$$-8+5 \le x \le 8+5 \qquad \text{dvs}$$

$$-3 \le x \le 13$$

Løs ulikheten selv både algebraisk og geometrisk. Legg merke til at ulikheten $-8+5 \le x \le 8+5$ kunne vært skrevet $5-2\cdot 4 \le x \le 5+2\cdot 4$

Nå skal vi bruke den samme framgangsmåten på ulikheten

$$-z_{\frac{\alpha}{2}} \le \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \le z_{\frac{\alpha}{2}}$$

hvor nå μ svarer til x i ulikheten over. En får nå:

$$-z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \overline{X}-\mu \le z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

ved å multiplisere med $\frac{\sigma}{\sqrt{n}}$ på begge sider. Deretter trekker vi fra $\overline{X}$ over hele ulikheten og får

$$-\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le -\mu \le -\overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Ved så til slutt å multiplisere med (-1) over hele ulikheten fremkommer

$$\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Det betyr mao. ifølge $P(-z_{\frac{\alpha}{2}} \le \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \le z_{\frac{\alpha}{2}}) = 1 - \alpha$ at en dermed må ha at

$$P(\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

Man sier da at intervallet

$$[A, B] = \left[\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

danner et (1-α)100% konfidensintervall for μ. Legg merke til at i utrykket $P(\overline{X} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$ er det $\overline{X}$ som er variabelen, mens μ er en konstant (den ukjente parameteren). Når utvalget er kjent er verdien av $\overline{X}$, $\bar{x}$, også kjent. Dermed blir det beregnede $(1-\alpha)$100% konfidensintervallet for μ:

$$[a, b] = \left[\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

der $a = \bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ og $b = \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$ kalles henholdsvis nedre og øvre konfidensgrense.

Forutsetningen for at dette gjelder er at populasjonen er normalfordelt (se side..) eller at $n \ge 30$. Det siste er ikke nevnt tidligere men bygger på det faktum at (se grunnleggende statistkk, Lillestøl side….)

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

er normalfordelt (tilnærmet) når $n \ge 30$ (jo større *n* jo bedre tilnærmelse) selv om populasjonen ikke er normalfordelt. Det betyr at

$$[a,b]=\left[\bar{x}-z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}},\bar{x}+z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right]$$

er et (1-α)100% konfidensintervall for μ selv om populasjonen ikke er normalfordelt forutsatt at σ er kjent og $n \geq 30$.
Dessuten vil da

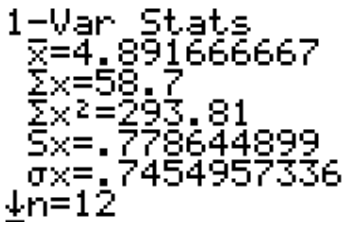
$$[a,b]=\left[\bar{x}-z_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}},\bar{x}+z_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}}\right]$$

være et (1-α)100% konfidensintervall for μ når populasjonen ikke er normalfordelt, σ er **u**kjent og $n \geq 30$.

Eks. Anta at det er tatt et utvalg på n = 12 observasjoner fra en normalfordelt populasjon hvor σ = 0,8 og at man fant følgende verdier av variabelen X:

x_i : 4,5; 3,7; 4,9; 5,3; 6,1; 3,9; 5,0; 5,5; 5,9; 3,9; 4,8; 5,2

Legger en dette inn i en liste på kalkulatoren gir 1-Vars Stats blant annet følgende:



Det betyr at et 95% konfidensintervall for μ dermed blir:

$$[a,b]=\left[\bar{x}-z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}},\bar{x}+z_{\frac{\alpha}{2}}\cdot\frac{\sigma}{\sqrt{n}}\right]=\left[4{,}89-1{,}96\cdot\frac{0{,}8}{\sqrt{12}},4{,}89+1{,}96\cdot\frac{0{,}8}{\sqrt{12}}\right]=[4{,}44;\ 5{,}34]$$

Vær oppmerksom på at det ikke er μ som med en sannsynlighet på 0,95 ligger mellom 4,44 og 5,34. Dette er en vanlig misforståelse. Grunnen til det er at μ er en konstant og at det dermed ikke er mulig å knytte en sannsynlighet til denne. Dette er forbeholdt variabler. Hva betyr så resultatet over? Resultatet skal tolkes som følger: Det er metodens pålitelighet som er angitt ved 0,95. Alternativt kan man si at

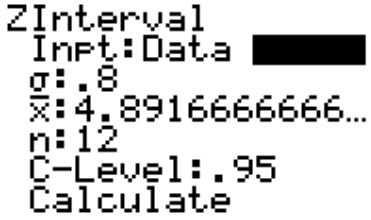
$$P(\bar{X}-1{,}96\cdot\frac{0{,}8}{\sqrt{12}}\leq\mu\leq\bar{X}+1{,}96\cdot\frac{0{,}8}{\sqrt{12}})=0{,}95$$

idet det her inngår en variabel nemlig $\bar{X}$.

Hvis man nå ønsker å bruke kalkulatorens statistikkfunksjoner optimalt bruker en nå følgende kommandoer:

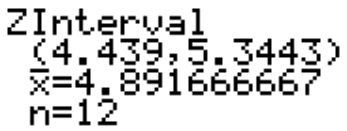
STAT
TESTS
7: ZInterval
Data

Legger en her inn sigma = 0,8 ; $\bar{x}$ =4,89(16666666) vil en ha følgende bilde:



Bak det markerte feltet har kommandoen Stats. Ved kommandoen Stats må gjennomsnittet og standardavviket regnes ut først og så legges inn her.

Går en nå ned til Calculate og trykker ENTER får en følgende bilde:



En ser at dette stemmer med de resultatene over. hvis man har lagt inn dataene i en liste (f.eks.liste 1) kan en like gjerne bruke kommandoen Data. Prøv dette selv. Legg da merke til at kalkulatoren også beregner utvalgsstandardavviket s_x =0,78.

I MINITAB er de tilsvarende kommandoene:

Stats
Basic Statistics
1Z: 1-Sample Z…
Summarized data

En legger så inn antall observasjoner, deres gjennomsnitt og verdien av sigma. Resultatet av dette blir da:

One-Sample Z

```
The assumed standard deviation = 0,8

 N     Mean  SE Mean        95% CI
12  4,89167  0,23094  (4,43903; 5,34430)
```

som igjen stemmer med resultatet foran.

Også i MINITAB kan man legge inn resultatene i en liste og referere til denne når man skal lage konfidensintervall.

Forutsetningen for at formelen for konfidensintervallet gjelder er at populasjonen er normalfordelt (se side..) eller at $n \geq 30$. Det siste er ikke nevnt tidligere men bygger på det faktum at (se grunnleggende statistkk, Lillestøl side….)

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

er normalfordelt (tilnærmet) når $n \geq 30$ (jo større *n* jo bedre tilnærmelse) selv om populasjonen ikke er normalfordelt. Det betyr at

$$[a,b] = \left[\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \right]$$

er et (1-α)100% konfidensintervall for μ selv om populasjonen ikke er normalfordelt forutsatt at σ er kjent og $n \geq 30$.
Dessuten vil da

$$[a,b] = \left[\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \right]$$

være et (1-α)100% konfidensintervall for μ når populasjonen ikke er normalfordelt, σ er **u**kjent og $n \geq 30$.

Eks. Anta at man har tatt et utvalg på $n = 40$ fra en populasjon (som ikke nødvendigvis ikke er normalfordelt) og hvor populasjonsstandardavviket er ukjent. Man fant et gjennomsnitt $\bar{x} = 23{,}6$ og et standardavvik $s_x = 4{,}7$ i utvalget. Da får en følgende 95% konfidensintervall for populasjonsgjennomsnittet μ:

$$[a,b] = \left[\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \right] = \left[23{,}6 - 1{,}96 \cdot \frac{4{,}7}{\sqrt{40}}; 23{,}6 + 1{,}96 \cdot \frac{4{,}7}{\sqrt{40}} \right] = [22{,}14; 25{,}06]$$

Anta nå at populasjonen er normalfordelt , $n < 30$ og at σ er **u**kjent. Da betrakter man

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \quad \text{istedenfor} \quad z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

der t er verdien av variabelen

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

som ifølge teorien side… følger t-fordelingen med (n -1) frihetsgrader (ofte skrevet som d.f. = (n -1); der d.f. er forkortelsen for degrees of freedom)

I: Vi ser nå igjen først på testing av

$$H_0 : \mu = \mu_0 \text{ mot } H_A : \mu > \mu_0$$

i) Klassisk testing: H_0 forkastes (idet store verdier av T er signifkante) hvis t (=verdien av T) $\geq t_\alpha$ = kritisk verdi på nivået α (t_α =1,729 hvis $\alpha = 0{,}05$ og $n = 20$ (dvs. at d.f.=19)

ii) Beregning av signifikanssannsynlighet (=P-verdi). P-verdien = $P_{H_0}(T \geq t)$

Eks. Anta at man har tatt et utvalg på 20 fra en normalfordelt populasjon hvor det til nå har vært et gjennomsnitt på 7,5. Det er signaler som nå tyder på at dette har økt. Man ønsker nå gjennom hypotesetesting å avgjøre om dette er tilfellet. Utvalget gir et gjennomsnitt på 7,9 og et standardavvik på 2,1. Vi tester nå

$$H_0 : \mu = 7{,}5 \text{ mot } H_A : \mu > 7{,}5$$

Det betyr nå at H_0 forkastes hvis verdien av T blir minst 1,729, dvs hvis $t \geq 1{,}729$. En finner nå

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{7{,}9 - 7{,}5}{\frac{2{,}1}{\sqrt{20}}} = 0{,}85 < 1{,}729$$

Det er mao. ikke sterke nok signaler om at det er grunn til å påstå at $\mu > 7{,}5$.

Regner en alternativt ut P-verdien finner en

$$P_{H_0}(T \geq t) = P_{H_0}(T \geq 0{,}85) = 0{,}203$$

Det siste finner en ved å bruke kalkulatoren og følgende kommandoer:

2ND VARS
5:tcdf(
ENTER
.85, 10^99, 19
ENTER

Dette gir følgende bilde:

```
tcdf(.85,10^99,1
9)
_     .2029545711
```

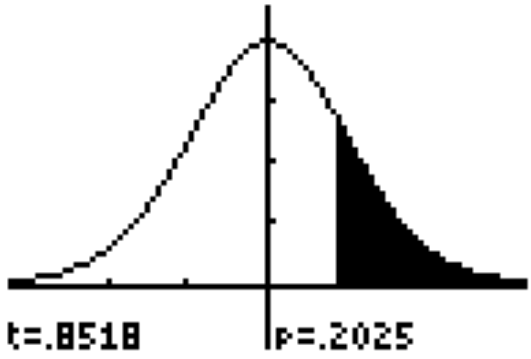
Velger en alternativt å bruke kalkulatorens statistikkfunksjoner optimalt finner en ved følgende kommandoer:

STAT
TESTS
2:T-Test
ENTER
Stats

Legger en så inn $\mu_0 = 7.5$, $\bar{x} = 7{,}9$, $s_x = 2.1$, $\mu :> \mu_0$ *og Calculate* får en følgende bilde:

```
T-Test
 μ>7.5
 t=.85183542
 p=.2024573994
 x̄=7.9
 Sx=2.1
_n=20
```

Velger en isteden kommandoen Draw istedenfor Calculate får en følgende bilde:



Her er t-fordelingen med 19 frihetsgrader tegnet. Det skraverte arealet er lik P-verdien som en ser er 0,2025.

Ønsker en å bruke MINITAB må en gi følgende kommandoer:

Stat
Basic Statistics
1-t sample
Summarized data

og legge inn Sample size = 20, Mean = 7,9, Standard deviation 2,1, Test Mean =7,5 og velg så tilslutt under Options alternativet greater than. Dette resulterer i følgende utskrift:

One-Sample T

```
Test of mu = 7,5 vs > 7,5

                                          95%
                                        Lower
 N     Mean    StDev  SE Mean    Bound     T      P
20  7,90000  2,10000  0,46957  7,08804  0,85  0,202
```

Dette ser en stemmer med resultatene foran. I tillegg er det angitt et 95% ensidig konfidensintervall. Prøv selv å beregne dette. Vink: Ta utgangspunkt i

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}$$

og bruk

$$P(-z_\alpha \le \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}}) = 1 - \alpha$$

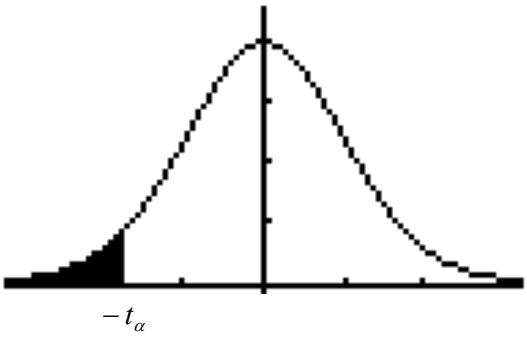
til å vise at et $(1-\alpha)100\%$ ensidig konfidensintervall for en nedre grense for μ blir

$$[a, \infty) = \left[\bar{x} - t_\alpha \cdot \frac{s}{\sqrt{n}}, \infty \right)$$

II: Vi ser så på testing av

$$H_0 : \mu = \mu_0 \text{ mot } H_A : \mu < \mu_0$$

i) Klassisk testing: H_0 forkastes (idet små verdier av T er signifkante) hvis t (=verdien av T) $\le -t_\alpha$ = kritisk verdi på nivået α



ii) Beregning av signifikanssannsynlighet (=P-verdi). P-verdien = $P_{H_0}(T \le t)$. Her vil vanligvis *t* være negativ (hvorfor det?)

III. Tilslutt ser vi så på testing av

$$H_0 : \mu = \mu_0 \text{ mot } H_A : \mu \neq \mu_0$$

som vi tidligere har kalt 2-sidig testing. Her er både store og små verdier signifkante. Det betyr ved testing på nivået α :

i) Ved klassisk testing at H_0 forkastes hvis *t* (=verdien av T) $\le -t_{\frac{\alpha}{2}}$ eller hvis $t \ge t_{\frac{\alpha}{2}}$ = de kritiske verdiene på nivået α

Eks.Anta at man har tatt et utvalg på 15 fra en normalfordelt populasjon hvor det til nå har vært et gjennomsnitt på 6. Det er signaler som nå tyder på at dette har endret seg. Man ønsker nå gjennom hypotesetesting å avgjøre om dette er tilfellet. Utvalget gir et gjennomsnitt på 5,1 og et standardavvik på 1,2. Vi tester nå

$$H_0 : \mu = 6{,}0 \text{ mot } H_A : \mu \neq 6{,}0$$

Hvis vi nå ønsker et nivå 5% betyr det at H_0 forkastes hvis verdien av T blir minst 2,145 eller høyst -2,145; dvs hvis $t \ge 2{,}145$ eller $t \le -2{,}145$. En finner nå

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{5{,}1 - 6{,}0}{\frac{1{,}2}{\sqrt{15}}} = -2{,}905 < -2{,}145$$

En ser mao. at resultatet er signifikant på 5%-nivået, dvs. at $H_0 : \mu = 6{,}0$ forkastes til fordel for $H_A : \mu \neq 6{,}0$ med en sannsynlighet på høyst 0,05 for at vi tar feil.

Hvis vi isteden beregner P-verdien finner vi denne ved:

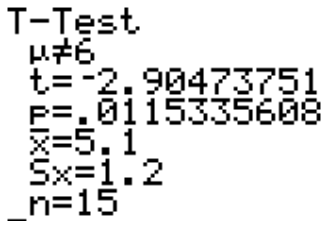
$$2P_{H_0}(T \le -2{,}905) = 2tcdf(-10\text{^}99, -2.905{,}14) = 0{,}0115$$

som gir samme konklusjon som over men med den forskjell at sannsynligheten for at vi nå tar feil nå kun er 0,0115. Hvorfor multipliseres det med 2?

Hvis man nå bruker kalkulatoren optimalt med kommandoene:

STAT
TESTS
2:T-Test
Stats

,legger inn de gitt dataene, velger alternativet $\mu \neq \mu_0$ og tilslutt Calculate får en følgende resultat:



Gå nå selv inn og bruk MINITAB for å se at du får samme resultat.

Konfidensintervall.

Vi skal nå se på hvorledes konfidensintervallene ser ut i de tilfellene vi må bruke t-fordelingen.
Anta nå som foran at vi har en normalfordelt populasjon hvor det er en eller annen interessant størrelse med et gjennomsnitt på μ (ukjent eller eventuelt kjent fra tidligere målinger). Anta at man nå mener at det har skjedd endringer i populasjonen slik at μ har endret seg på den ene eller andre måten. Dette er en situasjon hvor en kan bruke inferens. Anta at populasjonsvariansen σ^2 er **u**kjent og hvor n<30. Vi ønsker nå igjen å finne et konfidensintervall for μ. Vi tar utgangspunkt i variabelen

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

som vi gjorde ved hypoteseprøvingen. Hvis vi nå ønsker et $(1-\alpha)100\%$ konfidensintervall for μ og T er t-fordelt med (n-1) frihetsgrader må følgende gjelde:

$$P(-t_{\frac{\alpha}{2}} \leq T \leq t_{\frac{\alpha}{2}}) = 1 - \alpha$$

Herav har en da

$$P(-t_{\frac{\alpha}{2}} \leq \frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \leq t_{\frac{\alpha}{2}}) = 1 - \alpha$$

Helt analogt til hva vi gjorde på side 77/78 kan vi nå utlede følgende $(1-\alpha)100\%$ konfidensintervall for μ:

$$[a,b] = \left[\bar{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}, \bar{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}} \right]$$

Prøv selv og se om du får det til (uten å se etter på side 78)

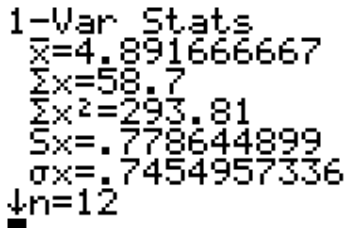
Eks. Anta at det er tatt et utvalg på $n = 12$ observasjoner fra en normalfordelt populasjon hvor σ er ukjent og at man fant følgende verdier av variabelen X : (se eks. side 80)

$$x_i : 4{,}5;\ 3{,}7;\ 4{,}9;\ 5{,}3;\ 6{,}1;\ 3{,}9;\ 5{,}0;\ 5{,}5;\ 5{,}9;\ 3{,}9;\ 4{,}8;\ 5{,}2$$

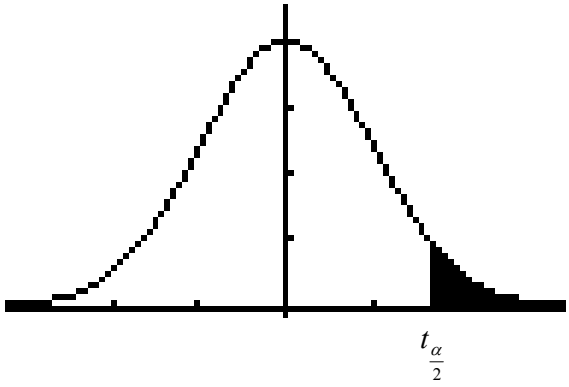
Legger vi disse tallene inn i en liste på kalkulatoren (for eksempel) liste 1gir (som før) gir

STAT
CALC
1:1-Var Stats
ENTER
2ND1
ENTER

(bl.a.) følgende :



Herav finner en de nødvendige målene en trenger for å beregne konfidensintervallet med unntak av $t_{\frac{\alpha}{2}}$ (= $\frac{\alpha}{2}$ fraktilen i t-fordelingen).



Ønsker en nå et 95% konfidensintervall finner en

$$t_{\frac{\alpha}{2}} = t_{0{,}025} = 2{,}201$$

når en som her har d.f.=11 (hvorfor)

En finner nå følgende 95% konfidensintervall for μ.

$$[a,b]=\left[\bar{x}-t_{0,025}\cdot\frac{s}{\sqrt{n}},\bar{x}+t_{0,025}\cdot\frac{s}{\sqrt{n}}\right]=\left[4{,}89167-2{,}201\cdot\frac{0{,}77864}{\sqrt{12}},4{,}89167+2{,}201\cdot\frac{0{,}77864}{\sqrt{12}}\right]$$
$$=[0{,}397;1{,}386]$$

Hypotesetesting og konfidensintervall.

Det er en nær sammenheng mellom tosidig hypoteseprøving og konfidensintervaller. (Det er også en sammenheng mellom ensidig hypoteseprøving og konfidensintervaller, men da må vi først utlede såkalt ensidige konfidensintervall og det skal vi ikke komme inn på her i dette heftet.

Anta at vi ønsker å teste

$$H_0:\mu=\mu_0 \text{ mot } H_A:\mu\neq\mu_0$$

På nivået α. Anta at vi har funnet et $(1-\alpha)100\%$ for μ. Anta at dette intervallet basert på dataene man har blir $[a,b]$ (mao. fra og med a tom. b). Da gjelder følgende:

Hvis $\mu_0\notin[a,b]$ (mao. hvis μ_0 ikke faller innenfor konfidensintervallet) så forkastes $H_0:\mu=\mu_0$ på nivået α. Siden vi ved hypotesetesting kun har to alternativer så vil vi ikke forkaste $H_0:\mu=\mu_0$ på nivået α hvis $\mu_0\in[a,b]$.

Eks. Hvis vi nå gjennomfører testingen av

$$H_0:\mu=6{,}0 \text{ mot } H_A:\mu\neq 6{,}0$$

(se eks. side 86) ved hjelp av et 95% konfidensintervall finner en først konfidensintervallet Eks.Anta at man har tatt et utvalg på 15 fra en normalfordelt populasjon hvor det til nå har vært et gjennomsnitt på 6. Det er signaler som nå tyder på at dette har endret seg. Man ønsker nå gjennom hypotesetesting å avgjøre om dette er tilfellet. Utvalget gir et gjennomsnitt på 5,1 og et standardavvik på 1,2. Vi tester nå

$$H_0:\mu=6{,}0 \text{ mot } H_A:\mu\neq 6{,}0$$

Hvis vi nå ønsker et nivå 5% betyr det at H_0 forkastes hvis verdien av T blir minst 2,145 eller høyst -2,145; dvs hvis $t\geq 2{,}145$ eller $t\leq -2{,}145$. En finner nå

$$[a,b]=\left[\bar{x}-t_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}},\bar{x}+t_{\frac{\alpha}{2}}\cdot\frac{s}{\sqrt{n}}\right]=\left[5.1-2.145\cdot\frac{1{,}2}{\sqrt{15}},\ 5.1+2.145\cdot\frac{1{,}2}{\sqrt{15}}\right]=[4.44,\ 5.76]$$

Det betyr mao. at $H_0 : \mu = 6{,}0$ forkastes på 5%-nivået idet $\mu_0 = 6{,}0 \notin [a,b] = [4.44, 5.76]$.

11. Inferens for to populasjonsgjennomsnitt.

Anta at vi har to normalfordelte populasjoner med hvert sitt gjennomsnitt (ukjent eller eventuelt kjent fra tidligere målinger). Anta at disse er henholdsvis μ_1 og μ_2. Til nå er det all grunn til å tro at $\mu_1 = \mu_2$, men nå mener man at det har skjedd endringer i populasjonene slik at μ_1 og μ_2 har endret seg på den ene eller andre måten. Dette er en situasjon hvor en kan bruke inferens. Anta at populasjonsvariansene σ_1^2 og σ_2^2 er kjente (fra tidligere eller liknende målinger), og at det tas uavhengige tilfeldig utvalg fra de to populasjonene på henholdsvis n_1 og n_2.

Anta at vi nå ønsker å teste

$$H_0 : \mu_1 = \mu_2$$

(det er ingen forskjell på de to populasjonsgjennomsnittene)

mot et eller annet alternativ (ensidig eller tosidig)

Vår testobservator er nå

$$Z = \frac{Estimat - hypoteseprøvingsverdien}{s\tan dardavviket\ til\ estimatet} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

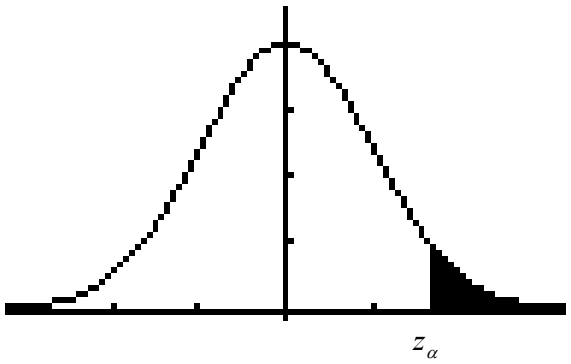
når H_0 er riktig (hvorfor det?). Legg merke til at testobservatoren måler differansen mellom differansen av utvalgsgjennomsnittene og differansen av populasjonsgjennomsnittene når H_0 er riktig. Dette er en størrelse som i henhold til sentralgrenseteoremet er normalfordelt.

I: Vi ser nå først på testing av

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 > \mu_2$$

i) Klassisk testing: H_0 forkastes (idet store verdier av Z er signifkante) hvis z (=verdien av Z) $\geq z_\alpha$ = kritisk verdi på nivået α (z_α =1,645 hvis $\alpha = 0{,}05$)

ii) Beregning av signifikanssannsynlighet (=P-verdi). P-verdien = $P_{H_0}(Z \geq z)$



Eks. Anta at man ønsker å teste

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 > \mu_2$$

med nivået $\alpha = 0{,}05$. Det blir tatt to uavhengige tilfeldige utvalg på henholdsvis 13 og 12 fra de to populasjonene og man finner $\overline{X}_1 = 12{,}6$ og $\overline{X}_2 = 11{,}8$. Anta at de to populasjonsvariansene σ_1^2 og σ_2^2 er henholdsvis 2,4 og 3,1.

Ved klassisk testing finner en nå at $H_0 : \mu_1 = \mu_2$ ikke forkastes på 5%-nivået idet

$$z = \frac{12{,}6 - 11{,}8}{\sqrt{\frac{2{,}4}{13} + \frac{3{,}1}{12}}} = 1{,}20 < 1{,}645 \text{ (=kritisk verdi på 5\%-nivået)}$$

Beregner en nå alternativt P-verdien finner en ved hjelp av kalkulatoren følgende :

$$P_{H_0}(Z \geq z) = P_{H_0}(Z \geq 1{,}20) = 0{,}115$$

Siden denne er større enn nivået på 0,05 så ser en at nullhypotesen ikke kan forkastes på 5%-nivået.

Beregningen av P-verdien gjøres med følgende kommandoer:

2ND VARS
2:normalcdf(
ENTER
$1{,}20, 10^{99}$)
ENTER

Kalkulatoren viser nå:

```
normalcdf(1.20,1
0^99)
     .1150697316
■
```

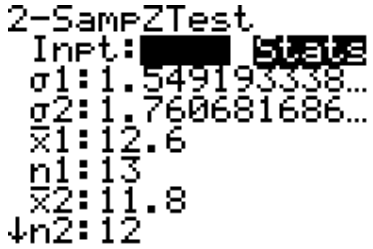
Velger en nå å bruke kalkulatorens "testpakke" finner en ved følgende kommandoer:

STAT
TESTS
3: 2-SampZtest
ENTER

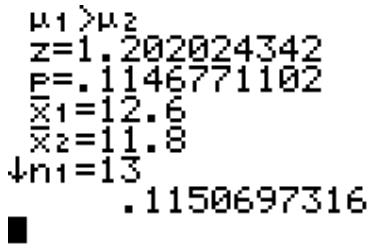
Her må en nå legge inn de gitte verdiene for $\overline{X}_1$ og $\overline{X}_2$, de to populasjonsvariansene σ_1^2 og σ_2^2 og tilslutt verdiene av n_1 og n_2
Dessuten må kalkulatoren få vite hva alternativet skal være. En velger her

$$H_A : \mu_1 > \mu_2$$

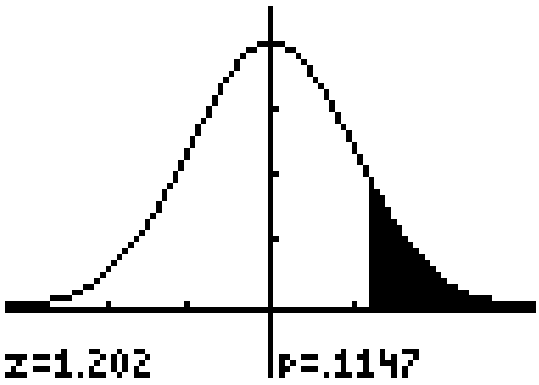
Da får en følgende bilde på kalkulatoren:



Velger en nå kommandoen CALCULATE får en følgende resultat.



Velger en isteden kommandoen Draw får en følgende resultat:



Prøver nå å bruke MINITAB for å løse den samme oppgaven. En gjør da bruk av følgende kommandoer:

Stat
Basic Statistics
2t 2–Sample t…..

Her må en så velge alternativet Summarized data, og deretter legge inn sample size på 13 og 12, mean på 12,6 og 11,8, standarddeviation på 1,55 og 1,76 og tilslutt under options velge alternativet greater than. Velger en så OK får en følgende utskrift:

Two-Sample T-Test and CI

```
Sample   N   Mean  StDev  SE Mean
1       13  12,60   1,55     0,43
2       12  11,80   1,76     0,51


Difference = mu (1) - mu (2)
Estimate for difference:  0,800000
95% lower bound for difference:  -0,342825
T-Test of difference = 0 (vs >): T-Value = 1,20  P-Value =
0,121  DF = 22
```

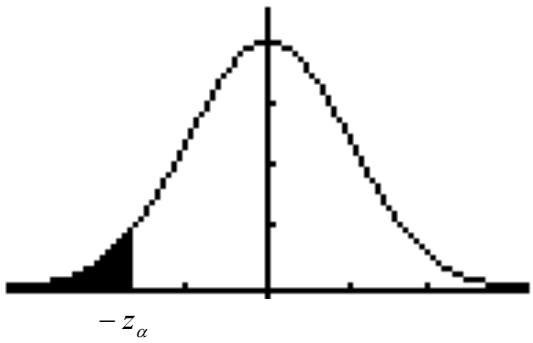
Legg merke til at vi her må bruke en t-test idet MINITAB ikke dekker denne situasjonen. Det betyr at man ifølge utskriften brukes df.=22 istedenfor df.=∞ som er situasjonen hvis man skal bruke normalfordelingen som gjort med kalkulatoren over. Dette blir dermed en tilnærming. Dette ser en bl.a. av forskjellen i P-verdiene som er 0,1147 ved normalfordelingen (kalkulator) og 0,121 ved t-fordelingen(MINITAB). Grunnen til at MINITAB ikke dekker denne situasjonen er den sjelden brukes i praksis. Dette skyldes at det er svært sjelden at man kjenner standardavviket i en populasjon mens gjennomsnittet er utkjent.

II: Vi ser så på testing av

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 < \mu_2$$

i) Klassisk testing: H_0 forkastes (idet små verdier av Z er signifikante) hvis z (=verdien av Z) $\leq -z_\alpha$ = kritisk verdi på nivået α (z_α = -1,645 hvis $\alpha = 0{,}05$)

ii) Beregning av signifikanssannsynlighet (=P-verdi). P-verdien = $P_{H_0}(Z \leq z)$



Eks. Anta at man ønsker å teste

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 < \mu_2$$

med nivået $\alpha = 0{,}05$. Det blir tatt to uavhengige tilfeldige utvalg på henholdsvis n_1=25 og n_2=27 fra de to populasjonene og man finner $\bar{x}_1 = 8{,}2$ og $\bar{x}_2 = 8{,}8$. Anta at de to populasjonsvariansene σ_1^2 og σ_2^2 er henholdsvis 1,5 og 1,7.

Ved klassisk testing finner en nå at $H_0 : \mu_1 = \mu_2$ forkastes på 5%-nivået idet

$$z = \frac{8{,}2 - 8{,}8}{\sqrt{\frac{1{,}5}{25} + \frac{1{,}7}{27}}} = -1{,}711 < -1{,}645 \text{ (=kritisk verdi på 5\%-nivået)}$$

Beregner en nå alternativt P-verdien finner en ved hjelp av kalkulatoren følgende :

$$P_{H_0}(Z \le z) = P_{H_0}(Z \le -1{,}711) = 0{,}0435$$

Siden denne er mindre enn nivået på 0,05 så ser en at nullhypotesen forkastes på 5%-nivået.

Beregningen av P-verdien gjøres med følgende kommandoer:

2ND VARS
2:normalcdf(
ENTER
-10^{99},-1,711)
ENTER

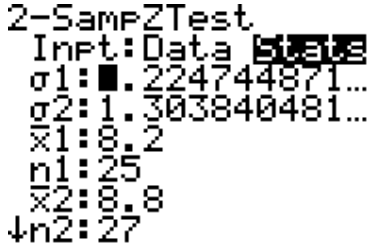
Velger en nå å bruke kalkulatorens "testpakke" finner en ved følgende kommandoer:

STAT
TESTS
3: 2-SampZtest
ENTER

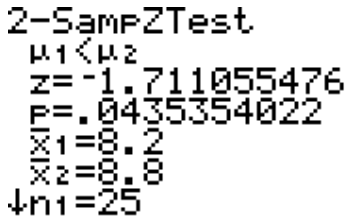
Her må en nå legge inn de gitte verdiene for $\overline{X}_1$ og $\overline{X}_2$, de to populasjonsvariansene σ_1^2 og σ_2^2 og tilslutt verdiene av n_1 og n_2
Dessuten må kalkulatoren få vite hva alternativet skal være. En velger her

$$H_A : \mu_1 < \mu_2$$

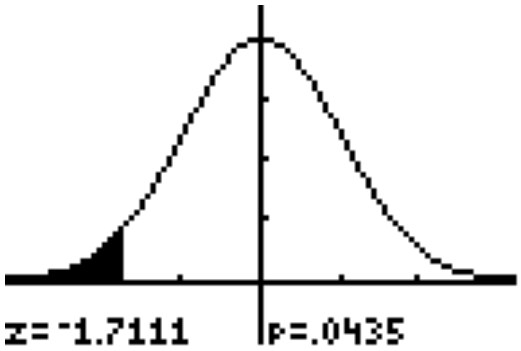
En får da følgende bilde på kalkulatoren:



Velger en så kommandoen CALCULATE viser kalkulatoren:



Velger en alternativt kommandoen får en følgende bilde:



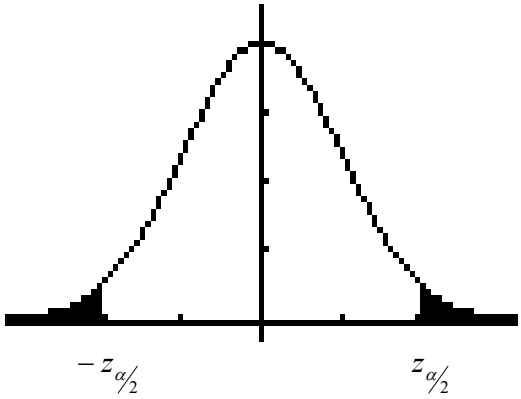
III: Vi ser så tilslutt på testing av

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 \neq \mu_2$$

Dette er en test som brukes når en har mistanke om at gjennomsnittene er forskjellige, men en ikke vet noe apriori om hvilken av de som er størst. Dette kan som foran gjennomføres enten ved klassisk testing eller ved beregning av p-verdien eller eventuelt bruk av konfidensintervall.

i) Klassisk testing: H_0 forkastes (idet både store og små verdier av Z er signifikante) hvis z (=verdien av Z) $\leq -z_{\alpha/2}$ eller $z \geq z_{\alpha/2}$ ($z_{\alpha/2}$ = 1,96 hvis $\alpha = 0{,}05$) (se figur neste side)

ii) Beregning av signifikanssannsynlighet (=P-verdi). P-verdien = $2\,P_{H_0}(Z \geq |z|)$



iii) Beregn $(1-\alpha)100\%$ konfidensintervall.

Eks. . Anta at man ønsker å teste

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 \neq \mu_2$$

med nivået $\alpha = 0{,}05$. Det blir tatt to uavhengige tilfeldige utvalg på henholdsvis n_1=35 og n_2=38 fra de to populasjonene og man finner $\bar{x}_1 = 18{,}5$ og $\bar{x}_2 = 19{,}0$. Anta at de to populasjonsvariansene σ_1^2 og σ_2^2 er henholdsvis 2,5 og 3,7.

Ved klassisk testing finner en nå at $H_0 : \mu_1 = \mu_2$ ikke kan forkastes på 5%-nivået idet

$$\frac{18{,}5 - 19{,}0}{\sqrt{\frac{2{,}5}{35} + \frac{3{,}7}{38}}} = -1{,}217 > -1{,}96 = z_{0.025}$$

Hvis en isteden ønsker å beregne 95% konfidensintervall må dette først utledes:

Konfidensintervall.

Anta nå som foran at vi har to normalfordelte populasjoner med gjennomsnitt henholdsvis μ_1 og μ_2. Anta at populasjonsvariansene σ_1^2 og σ_2^2 er kjente (fra tidligere eller liknende målinger) og at vi nå ønsker å finne et konfidensintervall for differansen mellom μ_1 og μ_2. Vi tar nå utgangspunkt i variabelen

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

som vi gjorde ved hypoteseprøvingen. A vi nå ønsker et $(1-\alpha)100\%$ konfidensintervall for μ_1 - μ_2. Fordi Z er normalfordelt må en ha :

$$P(-z_{\frac{\alpha}{2}} \le Z \le z_{\frac{\alpha}{2}}) = 1-\alpha$$

Herav har en da

$$P(-z_{\frac{\alpha}{2}} \le \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \le z_{\frac{\alpha}{2}}) = 1-\alpha$$

Dvs.

$$P(-z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le \overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2) \le z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) = 1-\alpha$$

Dvs.

$$P(-(\overline{X}_1 - \overline{X}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le -(\mu_1 - \mu_2) \le -(\overline{X}_1 - \overline{X}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) = 1-\alpha$$

Dvs.

$$P((\overline{X}_1 - \overline{X}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \le (\mu_1 - \mu_2) \le (\overline{X}_1 - \overline{X}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) = 1-\alpha$$

Herav har en mao. at $(1-\alpha)100\%$ konfidensintervall for μ_1 - μ_2 (når dataene er gitte) er

$$\left[(\overline{x}_1 - \overline{x}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\overline{x}_1 - \overline{x}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

I vårt eksempel finner en dermed følgende 95% konfidensintervall:

$$\left[(18{,}5 - 19{,}0) - 1{,}96\sqrt{\frac{2{,}5}{35} + \frac{3{,}7}{38}}, 18{,}6 - 19{,}0 + 1{,}96\sqrt{\frac{2{,}5}{35} + \frac{3{,}7}{38}}\right]$$

$$\left[-1{,}31;\ 0{,}31\right]$$

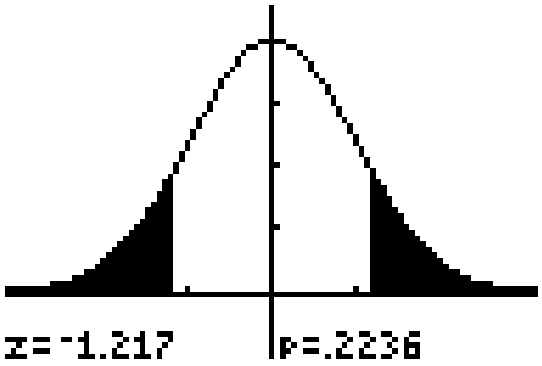
Her av ser en at verdien 0 av μ_1 - μ_2 under H_0 er inneholdt i konfidensintervallet og dermed har en at H_0 ikke kan forkastes på 5%-nivået.

Prøv selv å beregne P-verdien og se at du får samme konklusjon som over.

Hvis man ønsker å kontrollere disse beregningene kan det gjøres ved hjelp av kalkulatorens statistikkfunksjoner:

STAT
TESTS
3: 2-SampZtest
ENTER

Her må en nå huske å be om tosidig testing, dvs velge $H_A : \mu_1 \neq \mu_2$. Dette gir da følgende resultat:



Legg merke til at det nå er skravert to arealer med en samlet verdi på 0,2236. dette kommer av man tester tosidig og da er både store og små verdier av testobservatoren signifikante. Sjekk at den P-verdien du fant ved direkte regning stemmer med dette resultatet. (Vink: Husk at P-verdien = $2\,P(Z \geq |-1{,}217|)$

Hvis nå σ_1 og σ_2 er ukjente har man et problem som løses ved at man i testobservatoren

$$Z = \frac{Estimat - hypoteseprøvingsverdien}{s\tan dardavviket\ til\ estimatet} = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

erstatter σ_1 og σ_2 med de respektive utvalgsstandardavvikene s_1 og s_2. Dette kan en gjøre forutsatt at utvalgene er "store" og det vil som vanlig si at utvalgene n_1 og n_2 begge er større enn eller lik 30. Det betyr mao. at testobservatoren nå under forutsetning at H_0 er riktig antar formen

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

der Z som vanlig står for en variabel som er $N(0,1)$. Testingen gjennomføres nå nøyaktig på samme måte som over. Jeg velger derfor ikke å gi noen eksempler i denne situasjonen.

Anta så at σ_1 og σ_2 er ukjente og n_1 og n_2 ikke nødvendigvis er større enn eller lik 30. Da må en som i ett-utvalgssituasjonen bruke t-fordelingen istedenfor normalfordelingen. I denne situasjonen skal vi betrakte to muligheter.

i) Vi skal anta at $\sigma_1 = \sigma_2$ ($= \sigma$), men fortsatt at de er ukjente.
ii) Vi skal ikke gjøre antagelsen om at standardavvikene er like, men fortsatt at de er ukjente.

Vi ser nå først på situasjon i). Vi tar nå utgangspunkt i følgende testobservator:

$$\frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{\overline{X}_1 - \overline{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ idet } \sigma_1 = \sigma_2 \ (= \sigma)$$

Fordi σ er ukjent så må denne estimeres. Dette gjøres med den såkalte sammenslåtte variansen s_p^2 som er et veiet gjennomsnitt av utvalgsvariansene s_1^2 og s_2^2. En har at

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

og dermed følgende testobservator:

$$T = \frac{\overline{X}_1 - \overline{X}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

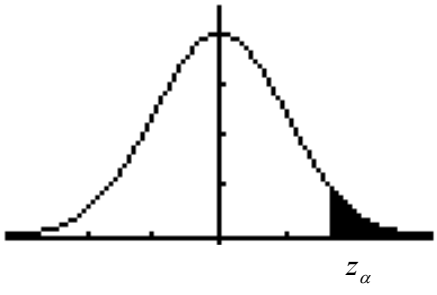
som er t-fordelt med $(n_1 + n_2 - 2)$ frihetsgrader, mao. d.f.= $n_1 + n_2 - 2$.

I: Vi ser nå først på testing av

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 > \mu_2$$

i) Klassisk testing: H_0 forkastes (idet store verdier av T er signifkante) hvis t (=verdien av T) $\geq t_\alpha$ = kritisk verdi på nivået α.

ii) Beregning av signifikanssannsynlighet (=P-verdi). P-verdien = $P_{H_0}(T \geq t)$



Eks. Anta at man ønsker å teste

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 > \mu_2$$

med nivået $\alpha = 0{,}05$. Det blir tatt to uavhengige tilfeldige utvalg på henholdsvis 13 og 10 fra de to populasjonene og man finner $\overline{X}_1 = 9{,}5$ og $\overline{X}_2 = 9{,}1$. Anta at de to populasjonsvariansene s_1^2 og s_2^2 er henholdsvis 1,4 og 1,6. Da blir

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(13-1)1{,}4 + (10-1)1{,}6}{(13+10-2)} = 1{,}4857...$$

Legg merke til at s_p^2 kan skrives som

$$s_p^2 = \frac{12}{21} \cdot 1{,}4 + \frac{9}{21} \cdot 1{,}6$$

Herav ser en tydelig at s_p^2 er det veiede gjennomsnittet av 1,4 ($=s_1^2$) og 1,6 ($=s_2^2$). Vektene er henholdsvis $\frac{12}{21}$ og $\frac{9}{21}$. Hvorfor legges det mer vekt på 1,4 ($=s_1^2$) enn 1,6 ($=s_2^2$)?

Bruker en nå klassisk testing finner en først

$$t = \frac{9{,}5 - 9{,}1}{\sqrt{1{,}4857}\sqrt{\frac{1}{13} + \frac{1}{10}}} = 0{,}780$$

Med df.= $n_1 + n_2 - 2 = 21$ finner en kritisk verdi $t_{0.05}$ på 5%-nivået av tabell over *t*-fordelingen (eller eventuelt av invers *t* på kalkulatoren) til $t_{0.05}$ =1,721. Dermed ser en at $t = 0{,}780 < 1{,}721 = t_{0.05}$ og konklusjonen blir dermed at $H_0 : \mu_1 = \mu_2$ ikke kan forkastes på 5%-nivået.

Beregner en nå alternativt P-verdien finner en ved hjelp av kalkulatoren følgende :

$$P_{H_0}(T \geq t) = P_{H_0}(T \geq 0{,}780) = 0{,}222$$

Siden denne er større enn nivået på 0,05 så ser en at nullhypotesen ikke kan forkastes på 5%-nivået. En annen måte å si dette resultatet på er man måtte hatt et nivå på 22,2% for at nullhypotesen skal forkastes, eller mao. hvis man hadde valgt å forkaste $H_0 : \mu_1 = \mu_2$ så hadde en hatt en risiko på 22,2% for at man trekker gal konklusjon.

Beregningen av P-verdien gjøres med følgende kommandoer:

2ND VARS

5:tcdf(

ENTER

$0,780,10^{99},21)$

ENTER

Kalkulatoren viser nå:

```
tcdf(0.780,10^99
,21)
      .2220469637
```

Velger en nå å bruke kalkulatorens "testpakke" finner en ved følgende kommandoer:

STAT

TESTS

4: 2-SampTTest

ENTER

Her må en nå legge inn de gitte verdiene for $\overline{X}_1$ og $\overline{X}_2$, de to populasjonsvariansene σ_1^2 og σ_2^2 og tilslutt verdiene av n_1 og n_2
Dessuten må kalkulatoren få vite hva alternativet skal være. En velger her

$$H_A : \mu_1 > \mu_2$$

Da får en følgende bilde på kalkulatoren:

```
2-SampTTest
 Inpt:Data Stats
 x̄1:■.5
 Sx1:1.18321595…
 n1:13
 x̄2:9.1
 Sx2:1.26491106…
↓n2:10
↑n1:13
 x̄2:9.1
 Sx2:1.26491106…
 n2:10
 μ1:≠μ2 <μ2 >μ2
 Pooled:No Yes
 Calculate Draw
```

Legg merke til at en her blir spurt om det er rimelig å tro at $\sigma_1 = \sigma_2$ gjennom nest siste linje: Pooled: No Yes, mao kan man slå sammen variansene eller ikke ? Vi velger her alternativet Yes.

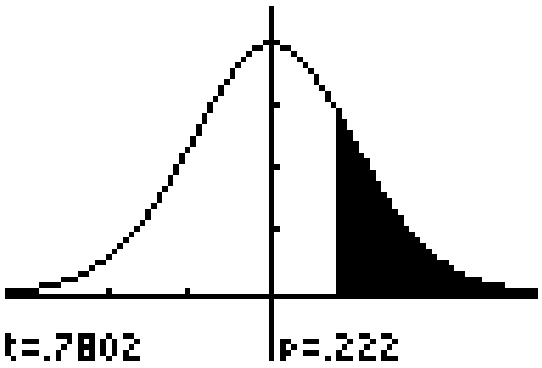
Velger en nå kommandoen CALCULATE får en følgende resultat.

```
2-SampTTest
 μ1>μ2
 t=.7801894976
 p=.2219924035
 df=21
 x̄1=9.5
↓x̄2=9.1

↑Sx1=1.18321596
 Sx2=1.26491106
 Sxp=1.2188988
 n1=13
 n2=10
```

En ser at beregningene foran stemmer med kalkulatorutregningene.

Velger en isteden kommandoen Draw får en følgende resultat:



Prøver nå å bruke MINITAB for å løse den samme oppgaven. En gjør da bruk av følgende kommandoer:

Stat
Basic Statistics
2t 2–Sample t…..

Her må en så velge alternativet Summarized data, og deretter legge inn sample size på 13 og 10, mean på 9,5 og 9,1, standarddeviation på $\sqrt{1.4}$ og $\sqrt{1.6}$ og tilslutt under options velge alternativet greater than. Velger en så OK får en følgende utskrift:

Two-Sample T-Test and CI

```
Sample   N  Mean  StDev  SE Mean
1       13  9,50   1,18     0,33
2       10  9,10   1,26     0,40


Difference = mu (1) - mu (2)
Estimate for difference:  0,400000
95% lower bound for difference:  -0,482218
```

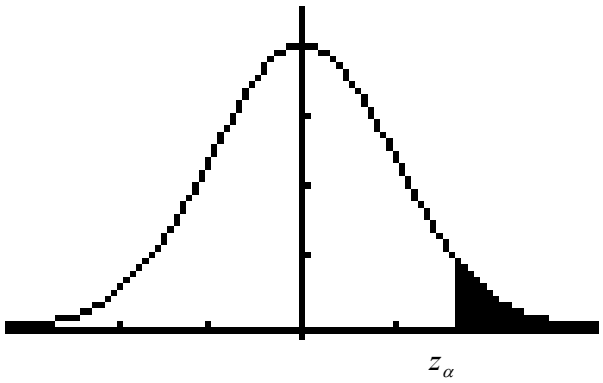
```
T-Test of difference = 0 (vs >): T-Value = 0,78
P-Value = 0,222  DF = 21
Both use Pooled StDev = 1,2189
```

II: Vi ser så på testing av

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 < \mu_2$$

i) Klassisk testing: H_0 forkastes (idet små verdier av T er signifkante) hvis t (=verdien av T) $\leq -t_\alpha$ = kritisk verdi på nivået α .

ii)Beregning av signifikanssannsynlighet (=P-verdi). P-verdien = $P_{H_0}(T \leq t)$



Eks. Anta at man ønsker å teste

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 < \mu_2$$

med nivået $\alpha = 0{,}05$. Det blir tatt to uavhengige tilfeldige utvalg på henholdsvis 16 og 15 fra de to populasjonene og man finner $\overline{X}_1 = 24{,}5$ og $\overline{X}_2 = 26{,}7$. Anta at de to populasjonsvariansene s_1^2 og s_2^2 er henholdsvis 2,1 og 2,2. Da blir

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(16-1)2{,}1 + (15-1)2{,}2}{(16+15-2)} = 2{,}148...$$

Bruker en nå klassisk testing finner en først

$$t = \frac{24{,}5 - 26{,}7}{\sqrt{2{,}418}\sqrt{\frac{1}{16} + \frac{1}{15}}} = \text{-}4{,}176$$

Med df.= $n_1 + n_2 - 2 = 29$ finner en kritisk verdi -$t_{0.05}$ på 5%-nivået av tabell over t-fordelingen (eller eventuelt av invers t på kalkulatoren) til -$t_{0.05}$ =-1,699. Dermed ser en at

t = -4,176 < -1.699 = $-t_{0.05}$ og konklusjonen blir dermed at $H_0 : \mu_1 = \mu_2$ forkastes på 5%-nivået.

Beregner en nå alternativt P-verdien finner en ved hjelp av kalkulatoren følgende :

$$P_{H_0}(T \le t) = P_{H_0}(T \le -4{,}176) = 0{,}00012$$

Mao. resultatene er signifikante dvs $H_0 : \mu_1 = \mu_2$ forkastes til fordel for $H_A : \mu_1 < \mu_2$

Beregningen av P-verdien gjøres med følgende kommandoer:

2ND VARS
5:tcdf(
ENTER
-10^{99},-4.176, 29)
ENTER

Kalkulatoren viser nå:

```
tcdf(-10^99,-4.1
76,29)
  1.237960246E-4
■
```

Velger en nå å bruke kalkulatorens "testpakke" finner en ved følgende kommandoer:

STAT
TESTS
4: 2-SampTTest
ENTER

Her må en nå legge inn de gitte verdiene for $\overline{X}_1$ og $\overline{X}_2$, de to populasjonsvariansene σ_1^2 og σ_2^2 og tilslutt verdiene av n_1 og n_2
Dessuten må kalkulatoren få vite hva alternativet skal være. En velger her en

$$H_A : \mu_1 < \mu_2$$

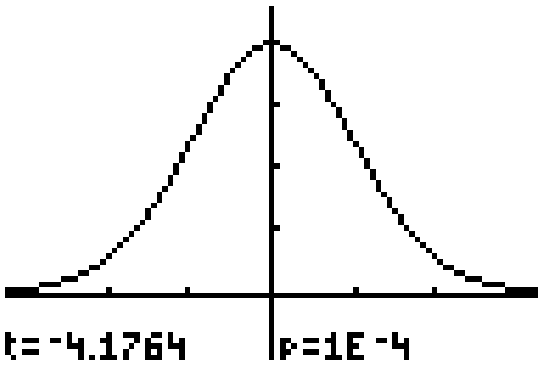
Velger en nå kommandoen CALCULATE får en følgende resultat.

```
2-SampTTest
 μ1<μ2
 t=-4.176403301
 p=1.2365947E-4
 df=29
 x̄1=24.5
↓x̄2=26.7
```

```
↑Sx1=1.44913767
 Sx2=1.4832397
 SxP=1.46569979
 n1=16
 n2=15
```

En ser at beregningene foran stemmer med kalkulatorutregningene.

Velger en kommandoen Draw får en følgende resultat:



Prøver nå å bruke MINITAB for å løse den samme oppgaven. En gjør da bruk av følgende kommandoer:

Stat
Basic Statistics
2t 2–Sample t…..

Legger en nå inn de gitte dataene får en

Two-Sample T-Test and CI

Sample	N	Mean	StDev	SE Mean
1	16	24,50	1,45	0,36
2	15	26,70	1,48	0,38

```
Difference = mu (1) - mu (2)
Estimate for difference:  -2,20000
95% upper bound for difference:  -1,30495
T-Test of difference = 0 (vs <): T-Value = -4,18
P-Value = 0,000  DF = 29
Both use Pooled StDev = 1,4657
```

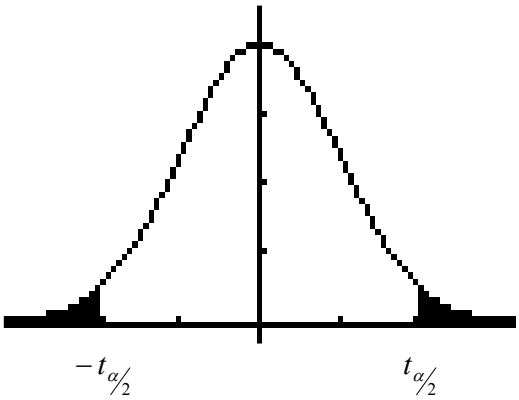
III: Vi ser så tilslutt på testing av

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 \neq \mu_2$$

Igjen så har vi 3 muligheter for å gjennomføre hypoteseprøvingen:

i) Klassisk testing: H_0 forkastes (idet både store og små verdier av T er signifikante) hvis t (=verdien av T) $\le -t_{\alpha/2}$ eller $t \ge t_{\alpha/2}$

ii) Beregning av signifikanssannsynlighet (=P-verdi). P-verdien = $2\,P_{H_0}(T \ge |t|)$



iii)Beregn $(1-\alpha)100\%$ konfidensintervall.

Eks. . Anta at man ønsker å teste

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 \neq \mu_2$$

med nivået $\alpha = 0{,}05$. Det blir tatt to uavhengige tilfeldige utvalg på henholdsvis n_1=14 og n_2=13 fra de to populasjonene og man finner $\bar{x}_1 = 11{,}9$ og $\bar{x}_2 = 11{,}2$. Anta at de to utvalgssvariansene s_1^2 og s_2^2 er henholdsvis 2,5 og 2,3.

Ved klassisk testing finner en nå at $H_0 : \mu_1 = \mu_2$ forkastes på 5%-nivået hvis $t \le -t_{0.05}$ = -2,060 eller $t \ge t_{0.05} = 2{,}060$. Nå blir

$$t = \frac{11{,}9 - 11{,}2}{1{,}845\sqrt{\frac{1}{14} + \frac{1}{13}}} = 0{,}985 < 2{,}060$$

Mao. $H_0 : \mu_1 = \mu_2$ kan ikke forkastes på 5%-nivået.

Hvis en isteden ønsker å beregne 95% konfidensintervall må dette først utledes:

Konfidensintervall.

Helt analogt til side 97 hvor det ble utledet $(1-\alpha)100\%$ konfidens intervall for $\mu_1 - \mu_2$ til

$$\left[(\bar{x}_1-\bar{x}_2)-z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},(\bar{x}_1-\bar{x}_2)+z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$$

forutsatt at σ_1 og σ_2 var kjente, kan en nå utlede $(1-\alpha)100\%$ konfidens intervall for $\mu_1-\mu_2$ når σ_1 og σ_2 er ukjente (men like).

En tar nå utgangspunkt i

$$T=\frac{\bar{X}_1-\bar{X}_2-(\mu_1-\mu_2)}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$$

og at

$$P(-t_{\frac{\alpha}{2}}\le T\le t_{\frac{\alpha}{2}})=1-\alpha$$

Da finner en følgende $(1-\alpha)100\%$ konfidensintervall for $\mu_1-\mu_2$ (prøv å utlede selv):

$$\left[(\bar{x}_1-\bar{x}_2)-t_{\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}},(\bar{x}_1-\bar{x}_2)+t_{\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$$

Hvis en bruker tallene fra eksempelet over finner følgende 95% konfidensintervall:

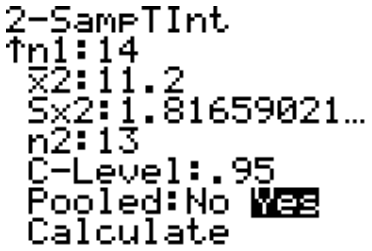
$$\left[(11{,}9-11{,}2)-2{,}060\cdot 1{,}845\sqrt{\frac{1}{14}+\frac{1}{13}},(11{,}9-11{,}2)+2{,}060\cdot 1{,}845\sqrt{\frac{1}{14}+\frac{1}{13}}\right]$$

$$=[-0{,}76;\ 2{,}16]$$

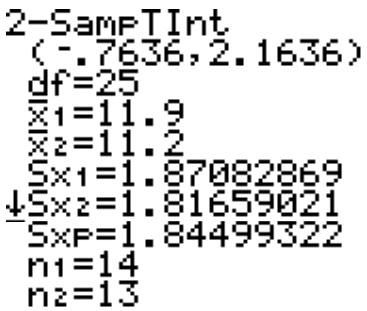
Bruker en nå kalkulatoren direkte med kommandoene

STAT
TESTS
0:2-SampTInt…
ENTER

får en følgende bilde



En velger også nå at variansene skal slås sammen ved å velge Yes på alternativet Pooled. Siden vi ønsker et 95% konfidensintervall velger vi alternativet .95 på C-Level. Calculate gir nå:



En ser mao. at man får samme resultat som over.

Velger en nå til sist å bruke MINITAB må en bruke følgende kommandoer:

Stat
Basic Statistics
2t 2-Sample t…
Summarized data (legg inn de gitte verdiene)
Options (velg her Confidence level 95,0, testdifferance 0,0 og Alternative: not equal)

Da får en opp følgende bilde:

Two-Sample T-Test and CI

Sample	N	Mean	StDev	SE Mean
1	14	11,90	1,87	0,50
2	13	11,20	1,82	0,50

```
Difference = mu (1) - mu (2)
Estimate for difference:  0,700000
95% CI for difference:  (-0,764999; 2,164999)
T-Test of difference = 0 (vs not =):
T-Value = 0,99  P-Value = 0,334  DF = 24
```

Herav ser en i tredje nederste linje at konfidensintervallet blir det samme som foran.

Nå skal vi se på den siste muligheten. Hva gjør man hvis det **ikke** er rimelig å anta at $\sigma_1 = \sigma_2 (= \sigma)$. Anta som foran at σ_1 og σ_2 er ukjente og n_1 og n_2 ikke nødvendigvis er >30.

En må nå bruke testobservatoren

$$T = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

som er *t*- fordelt med *a* frihetsgrader. En har som før at utvalgsvariansene, s_1^2 og s_2^2, er estimater for de ukjente populasjonsvariansene σ_1^2 og σ_2^2.

En har nå to alternativer med hensyn til antall frihetsgrader:

i) Det forsiktige (konservative) alternativet er å velge $a = \min(n_1 - 1, n_2 - 1)$ (a velges mao. som det minste av tallene $n_1 - 1$ og $n_2 - 1$. Det betyr imidlertid at man kaster bort en del informasjon (data). På den andre siden så vil det resultatet en kommer fram til være minst være så godt som det man finner (P-verdien vil mao. være mindre enn den vi kommer fram til)

ii) Det kan vises at antall frihetsgrader *a* tilnærmet er gitt ved

$$a = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Eks. La oss nå se på et eksempel fra side … uten å anta at $\sigma_1 = \sigma_2$. Vi testet her

$$H_0 : \mu_1 = \mu_2 \text{ mot } H_A : \mu_1 < \mu_2$$

med nivået $\alpha = 0{,}05$. Det blir tatt to uavhengige tilfeldige utvalg på henholdsvis 16 og 15 fra de to populasjonene og man finner $\overline{X}_1 = 24{,}5$ og $\overline{X}_2 = 26{,}7$. Anta at de to populasjonsvariansene s_1^2 og s_2^2 er henholdsvis 2,1 og 2,2.

En finner da først

$$t = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{24{,}5 - 26{,}7}{\sqrt{\frac{2{,}1}{16} + \frac{2{,}2}{15}}} = -4{,}173$$

som en ser avviker svært lite fra -4,176 som vi fant når vi slo sammen variansene.
Nå må en imidlertid bestemme antall frihetsgrader før en kan trekke noen konklusjon. Dette kan en som nevnt gjøre på to måter:

i) Bruk antall frihetsgrader $a = \min(16-1, 15-1) = \min(15,14) = 14$ (a velges mao. som det minste av tallene $n_1 - 1$ og $n_2 - 1$.
Med 14 frihetsgrader finner en kritisk verdi $-t_{0.05}$ på 5%-nivået av tabell over t-fordelingen (eller eventuelt av invers t på kalkulatoren) til $-t_{0.05} = -1{,}761$. Dermed ser en at $t = -4{,}173 < -1.761 = -t_{0.05}$ og konklusjonen blir dermed som foran at $H_0 : \mu_1 = \mu_2$ forkastes på 5%-nivået.

ii) Alternativt regner en ut antall frihetsgrader ved hjelp av formelen på side 109.

$$a = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2} = \frac{\left(\frac{2{,}1^2}{16} + \frac{2{,}2^2}{15}\right)^2}{\frac{1}{16-1}\left(\frac{2{,}1^2}{16}\right)^2 + \frac{1}{15-1}\left(\frac{2{,}2^2}{15}\right)^2} = 28{,}63$$

Dette runder en ned til 28. Da blir kritisk verdi -1,701 og en får samme konklusjon som over. En ser at det blir en viss forskjell på den konservative (og enkle) metoden som gir 14 frihetsgrader og denne metoden som gir 28 frihetsgrader.

Beregner en nå alternativt P-verdien finner en ved hjelp av kalkulatoren følgende :

$$P_{H_0}(T \le -4{,}173) = tcdf(-10^{99}, -4{,}173, 14) = 0{,}00047$$

hvis en bruker metode i) og

$$P_{H_0}(T \le -4{,}173) = tcdf(-10^{99}, -4{,}173, 28) = 0{,}000132$$

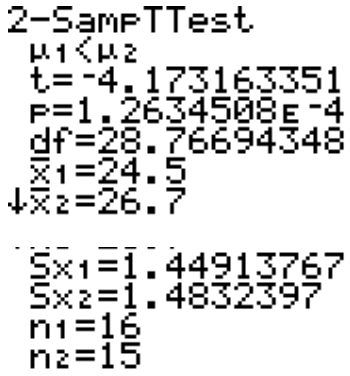
hvis en bruker metode ii). Bruker 28,63 istedenfor å runde av til 28 (det er ikke noe problem verken for kalkulatoren eller MINITAB å regne med df =16,48) finner en

$$P_{H_0}(T \le -4{,}173) = tcdf(-10^{99}; -4{,}173;\ 28{,}63) = 0{,}000127$$

Velger en alternativt å bruke kalkulatorens statistikkpakke med kommandoene

STAT
TESTS
4: 2- SampTTest
ENTER

og så legge inn de gitte verdiene av gjennomsnittene og standardavvikene, velge riktig alternativ , be om at variansene ikke skal slås sammen (Pooled: No) og til slutt CALCULATE finner en:



Velger en tilslutt å bruke MINITAB finner en følgende:

Two-Sample T-Test and CI

```
Sample   N   Mean  StDev  SE Mean
1       16  24,50   1,45     0,36
2       15  26,70   1,48     0,38


Difference = mu (1) - mu (2)
Estimate for difference:  -2,20000
95% upper bound for difference:-1,30320
T-Test of difference = 0 (vs <):
T-Value = -4,17  P-Value = 0,000  DF = 28
```

som stemmer meget bra med kalkulatorens beregninger.
Mao. resultatene er signifikante dvs $H_0 : \mu_1 = \mu_2$ forkastes til fordel for $H_A : \mu_1 < \mu_2$

En ser at det blir en viss forskjell på resultatene selv om en får samme konklusjon. Det kan imidlertid lett bli slik at en får forskjellig konklusjon med de forskjellige metodene.

12. Kjikvadrattester.

Vi skal nå se på såkalte kjikvadrattester. Dette er tester hvor en sammenlikner de observerte verdiene med de forventede verdiene som man kan regne ut når H_0 er riktig. Det finnes to hovedtyper av kjikvadrattester:

A. Kjikvadrattester for modellkontroll.
   - A.1 For helspesifiserte modeller/hypoteser
   - A.2 For delvis spesifiserte modeller/hypoteser

B. Kjikvadrattester for uavhengighet.

Vi skal nå se på A.1 som brukes til avgjøre om en statistisk modell er god/brukbar eller ikke.

Vi ser først på et ganske banalt, men allikevel instruktivt eksempel.
Anta at en person skal kaste en mynt 100 ganger.

Vi tester da

$$H_0: \text{ Modellen for myntkast (P(M) = P(K) = 0,5) er holdbar}$$

mot

$$H_A: \text{ Modellen for myntkast (P(M) = P(K) = 0,5) er ikke holdbar.}$$

Anta at man nå observerer

Res.	Mynt	Kron	Sum
Hyp. 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 en følgende forventede verdier:

Res.	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 α = 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.

A.2 Anta nå at vi ønsker å teste om vi har en binomisk modell hvor p er ukjent, men hvor $n = 3$. Det betyr at vi har en delvis speifisert modell.

Vi tester derfor:

$$H_0 : X \sim bin(3, p)$$

mot

$$H_A : \ X \ ikke \sim bin(3, p)$$

Der X som vanlig representerer antall suksesser på 3 forsøk.

Anta at det nå gjøres 200 forsøksserier med $n = 3$ forsøk i hver serie og at man observerer

x	0	1	2	3	SUM
O_i	40	95	45	20	200

Når man nå skal begynne å regne ut de forventede verdiene får en et problem. p er ukjent og dermed kan ikke finne de forventede verdiene. Dette løser en som alltid i statistikken med å finne et estimat, $\hat{p}$, for p, og så bruke dette isteden.

p estimeres nå ved

$$\hat{p} = \frac{0 \cdot 40 + 1 \cdot 95 + 2 \cdot 45 + 3 \cdot 20}{600} = \frac{245}{600} = 0{,}40833... = 0{,}41$$

Vår delvis spesifiserte modell og nullhypotese blir da gitt ved:

$$H_0 : X \sim bin(3; 0{,}41)$$

Dvs. at sannsynlighetsfordelingen p til X er gitt ved

$$p(x) = \binom{3}{x} 0{,}41^x 0{,}59^{3-x} \quad , x = 0,1,2,3$$

Dermed har en

x	0	1	2	3	SUM
$p(x)$	0,2054	0,4282	0,2975	0,0689	1
E_i	41,08	85,64	59,50	13,78	200
O_i	40	95	45	20	200

Nå forkastes $H_0 : X \sim bin(3; 0{,}41)$ hvis

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} \geq \chi^2_{\alpha, m-t-1}$$

Legg nå merke til at antall frihetsgrader ikke lenger er m – 1, men nå er gitt ved

$$\text{df.} = m - t - 1$$

der m som før er antall mulige utfall i modellen og t er antall parametre som må estimeres først for å kunne regne ut forventningsverdiene. Dette er kun aktuelt der man har en delvis spesifisert modell. Da må man bruke litt av "energien" i tallmaterialet først og antall frihetsgrader blir dermed redusert. I eksempelet over er t =1 og man har dermed

$$\text{df.} = m - t - 1 = 4 - 1 - 1 = 2$$

Av tabellen over χ^2 – fordelingen finner en nå følgende kritisk verdi på 5%-nivået :

$$\chi^2_{0.05,2} = 5{,}991$$

Vi må nå regne ut χ^2 – tallet og sammenlikne det med 5,991. En finner nå :

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{(40 - 41{,}08)^2}{41{,}08} + \cdots + \frac{(20 - 13{,}78)^2}{13{,}78} = 7{,}39$$

Konklusjon: $H_0 : X \sim bin(3;\, 0{,}41)$ forkastes på 5%-nivået siden $\chi^2 = 7{,}39 > 5{,}991$. Det tyder på at modellen $X \sim bin(3;\, 0{,}41)$ er ikke holdbar (kan ligge til grunn for de observerte dataene.)

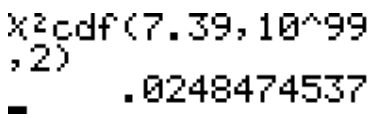
Beregner en nå isteden P- verdien finner en ved hjelp av kalkulatoren

$$P = P(\chi^2 \geq 7{,}39) = 0{,}0248$$

Dette finner en ved hjelp av følgende kommandoer:

2ND VARS
7: χ^2 cdf(
ENTER
7.39, 10^99, 2
ENTER

En får da følgende resultat:



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(18,2)	25(21,1)	30(25,7)	65
30-40 år	18(20,2)	20(23,4)	34(28,8)	72
40-50 år	28(17,6)	20(20,5)	15(24,9)	63
SUM	56	65	79	200

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 $E_1 = 56 \cdot \frac{65}{200} = 18{,}2$ personer forutsatt at nullhypotesen er riktig……osv. I aldersgruppen 40 til 50 år så er det $O_9 = 15$ personer som mener at produktet er bra. I denne gruppen kan vi forvente $E_9 = 79 \cdot \frac{63}{200} = 24{,}9$ personer forutsatt at nullhypotesen er riktig. Beregner en så kjikvadrattallet for hele matrisen finner en

$$\chi^2 = \sum_{i=1}^{m} \frac{(O_i - E_i)^2}{E_i} = \frac{(10-18{,}2)^2}{18{,}2} + \cdots + \frac{(15-24{,}9)^2}{24{,}9} = 16{,}97$$

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

$$P_{H_0}(\chi^2 \geq 16{,}97) = \chi^2 cdf(16.97,\ 10^{99},\ 4) = 0{,}001959.. = 0{,}002$$

idet en her har

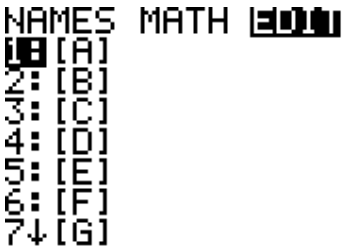
d.f.=(ant.rader-1)(ant. kolonner -1) = (3-1)(3-1) = 4

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.

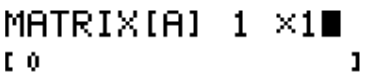
Kalkulatorens statistikkpakke kan også brukes direkte til å gjennomføre en uavhengighetstest. Det gjøres på følgende måte: Først må en legge de observerte verdiene i kalkulatorens matrisefunksjon. Til det bruker en følgende kommandoer:

2ND x^{-1}
EDIT

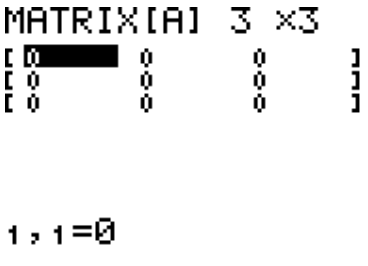
Første gang man bruker denne kommandoen vil en nå få opp følgende bilde:



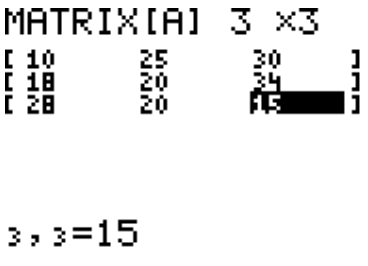
En må nå velge en av de 10 mulige matrisene (A tom. J) og legge tallene inn i. Jeg velger her å bruke den første matrisen A. Jeg trykker derfor ENTER en gang til og får opp følgende bilde:



Hvis man tidligere har brukt kalkulatoren til matriseregning vil den gamle matrisen A ligge her. Den matrisen vi nå ønsker å legge inn er en 3 x 3- matrise og dette må vi forberede kalkulatoren på. Det gjøres ved å skrive 3-tall oppe på 1-tallene og trykke ENTER. Da får en følgende bilde:



På de 9 nullene legger en nå inn et og et av de observerte tallene og får:



Nå må en så gå ut av dette ved hjelp av kommandoene 2ND MODE også deretter gå til statistikkfunksjonene:

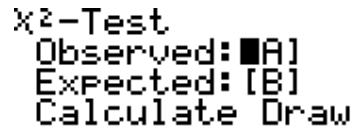
STAT

TESTS

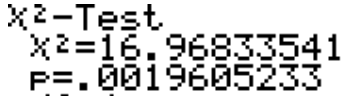
C: $\chi^2 - Test$

ENTER

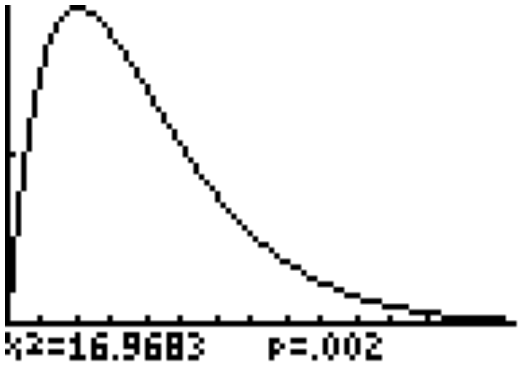
Kalkulatoren viser da følgende bilde



Velger en nå kommandoen Calculate får en følgende bilde:



Velger isteden kommandoen DRAW får en følgende bilde:



Velger en isteden å bruke MINITAB må en først legge inn de observerte verdiene i regnearket dvs. mening ”dårlig” i kolonne 1 (=C1), mening ”middels” i kolonne 2(=C2) og mening ”bra” i kolonne 3(=C3). I rad 1, 2 og 3 legger en henholdsvis aldergruppe $[15,25>$, $[25,35>$ og $[35,45>$. Deretter må en gi følgende kommandoer:

Stat

Tabels

χ^2 Chi-Square Test

og deretter legge inn den gitte observasjonsmatrisen. Valget OK gir nå

Chi-Square Test:Mening(Dårlig; Middels; Bra) mot Alder (1=(15,25), 2=(25,35), 3=(35,45))

```
Expected counts are printed below observed counts
Chi-Square contributions are printed below expected counts

       Dårlig  Middels    Bra  Total
    1      10       25     30     65
        18,20    21,13  25,68
        3,695    0,711  0,729

    2      18       20     34     72
        20,16    23,40  28,44
        0,231    0,494  1,087

    3      28       20     15     63
        17,64    20,48  24,89
        6,084    0,011  3,927

Total      56       65     79    200

Chi-Sq = 16,968; DF = 4; P-Value = 0,002
```

13. Inferens for andeler.

En andel.

Anta at vi har en populasjon hvor vi er interessert i å finne ut noe om andelen individer med en bestemt egenskap (kjennetegn) *A.* Det kan for eksempel være dysleksi, fargeblind, stemmer Arbeiderpartiet, er mot EU, har diabetes 2,……. La oss betegne denne andelen med p . Vi skal nå anta at vi har en binomisk populasjon, dvs.

i) Hvorvidt et individ har egenskapen er uavhengig av om andre har egenskapen.
ii) Enten så har man egenskapen (=”Suksess”) eller så har man den ikke (=”Fiasko”) (det er altså bare to mulige utfall (derav forstavelsen ”bi”) Suksess eller Fiasko)
iii) Sannsynligheten for at et vilkårlig valgt individ i populasjonen har egenskapen (= p) er konstant i hele populasjonen.

Det er denne parameteren p vi nå ønsker å uttale oss om enten gjennom estimering eller gjennom hypoteseprøving.

Anta at vi har tatt et utvalg på n fra denne populasjonen. La X = antall individer i utvalget med egenskapen A . Da har en at

$$X \sim bin(n, p)$$

Da har en fra kurset i grunnleggende statistikk at

$$E(X) = np \text{ og } Var(X) = np(1-p)$$

og dermed at

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

Dermed ser en at $\hat{P} = \frac{X}{n}$ er en forventningsrett estimator for p. Nå er generelt et $(1-\alpha)100\%$ konfidensintervall for θ på formen

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

der $f_{\alpha/2}$ er $\frac{\alpha}{2}$- fraktilen i fordelingen til $\hat{\Theta}$. Nå er

$$Z = \frac{\frac{X}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1) \quad \text{(tilnærmet for stor } n \text{)}$$

Dermed blir $(1-\alpha)100\%$ konfidensintervall på formen

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{(for stor } n \text{)}$$

Eks. La nå p være andelen stemmeberettigede i Norge som er for norsk medlemskap i EU. Man ønsker nå et 95% konfidensintervall for denne andelen. Anta det er tatt et tilfeldig utvalg på 1200 personer og at man her fant 495 personer som var for norsk medlemskap i EU. Et 95% konfidensintervall for p finner en da av

$$\hat{p} \pm 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{495}{1200} \pm 1{,}96\sqrt{\frac{\frac{495}{1200}(1-\frac{405}{1200})}{1200}} = 0{,}413 \pm 0{,}028$$

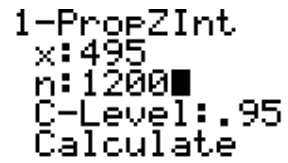
Mao. Selve intervallet for p blir dermed

$$[0{,}385;\ 0{,}441]$$

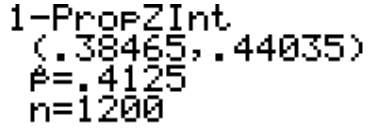
Bruker en isteden kalkulatoren direkte med følgende kommandoer:

STAT
TESTS
A: 1-PropZInt…
ENTER

Legger en så inn $x = 495$ og $n = 1200$ får en opp følgende bilde:



Velger en nå kommandoen CALCULATE får en :



som stemmer med beregningene foran.

Går en nå til MINITAB må en bruke følgende kommandoer:

Stat
Basic Statistics
1 P 1 Proportion

Her velger en Summarized data og legger inn

Number of trials =1200

Number of events = 495

Deretter velger en options og ber om Confidece level =95,0 og trykker OK får en blant annet følgende bilde (en må også her gjennomføre hypotesetesting, men jeg bryr meg ikke om dette nå) Fjerner en nå alt som har med hypoteseprøvingen å gjøre får en

CI for One Proportion

X	N	Sample p	95% CI
495	1200	0,412500	(0,384647; 0,440353)

Konfidensintervallet foran er beregnet under forutsetning av at den binomiske fordelingen kan tilnærmes med normalfordelingen. Hvis en under options velger å ikke bruke normaltilnærmelsen gir MINITAB:

CI for One Proportion

X	N	Sample p	Exact 95% CI
495	1200	0,412500	(0,384474; 0,440956)

En ser at det er marginale forskjeller mellom det eksakte konfidensintervallet over og det hvor man har gjort bruk av normalfordelingen. Det kommer av at normalfordelingen er meget god tilnærming til den binomiske fordelingen når *n* er så stor som 1200. En vanlig tommelfingerregel er at tilnærmingen er svært god når

$$Var(X) = np(1-p) \geq 10$$

Med vår $n = 1200$ og $p \approx 0{,}41$ har en

$$Var(X) = np(1-p) \approx 1200 \cdot 0.41 \cdot (1-0{,}41) = 290{,}28$$

Av dette ser en for eksempel at

$$np(1-p) \geq 10 \Leftrightarrow n \geq \frac{10}{p(1-p)}$$

Dette gir følgende sammenheng mellom *n* og p :

p	*n* bør minst være
0,1	112
0,2	63
0,3	48
0,4	42
0,5	40

Hypotesetesting med en andel.

Anta at vi nå ønsker å teste

$$H_0 : p = p_0$$

der p_0 er en spesifikk verdi av populasjonsandelen p . Vi skal som vanlig se på tre forskjellige alternative hypoteser: $p \geq p_0$, $p \leq p_0$ og $p \neq p_0$. Som testobservator skal vi nå bruke

$$Z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

der $\hat{P}$ som foran er gitt ved $\hat{P} = \frac{X}{n}$. Z er tilnærmet normalfordelt (0,1)forutsatt at *n* er stor nok.

I. Vi ser som vanlig først på testing av

$$H_0 : p = p_0 \text{ mot } H_A : p > p_0$$

Ved klassisk testing så forkastes H_0 på nivået α hvis

$$Z \geq z_\alpha \text{ (=kritisk verdi på nivået } \alpha \text{)}$$

eller hvis P-verdien

$$P_{H_0}(Z \geq z)$$

blir tilstrekkelig liten. z er som vanlig den observerte verdien av Z

Eks. Anta vi har en populasjon hvor kjennetegnet *A* (vil stemme arbeiderpartiet ved neste valg) til nå har ligget på 30%. Man mener nå at denne andelen har økt noe (mindre undersøkelser tyder på det). Vi ønsker derfor å teste

$$H_0 : p = 0{,}3 \text{ mot } H_A : p > 0{,}3$$

der *p* er andelen i populasjonen som vil stemme Arbeiderpartiet ved neste valg. Anta at et utvalg på *n* = 873 viser 303 personer med egenskapen *A*. Kritisk verdi på 5%-nivået er nå

$$z_{0,05} = 1{,}645$$

Nå blir verdien av Z :

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{303}{873} - 0{,}3}{\sqrt{\frac{0{,}3(1-0{,}3)}{873}}} = 3{,}035$$

Konklusjonen blir dermed at $H_0 : p = 0{,}3$ forkastes til fordel for $H_A : p > 0{,}3$ på 5%-nivået idet $z = 3{,}035 > z_{0,05} = 1{,}645$.

Beregner en isteden P-verdien får en

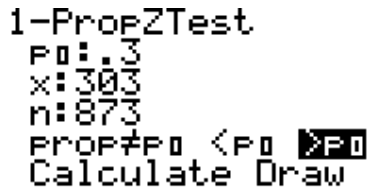
$$P_{H_0}(Z \geq z) = P_{H_0}(Z \geq 3{,}035) = \text{Normalcdf(3.035,10^99)} = 0{,}0012$$

som gir samme konklusjon som over.

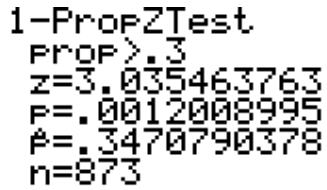
Velger en nå å bruke kalkulatoren direkte må en gjøre bruk av følgende kommandoer:

STAT
TESTS
5: 1-PropZTest
ENTER

Legger en så inn de aktuelle tallene over vil en få følgende bilde:



Velger en så til slutt kommandoen Calculate får en følgende bilde:



som bekrefter resultatene over.

Hvis man ønsker å bruke MINITAB må en bruke følgende kommandoer:

Stat
Basic Statistics
1 P 1 Proportion

Her velger en Summarized data og legger inn

Number of trials = 873

Number of events = 303

Deretter velger en Options og ber om

Test proportion	0,3
Alternative	greater than

og haker av i feltet hvor en skal avgjøre om en skal regne tilnærmet eller ikke (Use test and interval based on normal distribution or not). Trykker en til slutt på OK får en følgende resultat:

Test and CI for One Proportion

```
Test of p = 0,3 vs p > 0,3

                                          95%
                                        Lower
Sample    X    N  Sample p     Bound  Z-Value  P-Value
1       303  873  0,347079  0,320578     3,04    0,001
```

Legg merke til at man her også får et ensidig 95% konfidensintervall (selv om en ikke ber om dette). Konfidensintervallet er ensidig fordi en ønsker et ensidig alternativ under hypotesetestingen. Dette intervallet finner en av:

$$p > \hat{p} - z_{0,05}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0{,}347\text{-}1{,}645\sqrt{\frac{0{,}347(1-0{,}347)}{873}} = 0{,}3205$$

Vi påstår mao. at $p > 0{,}3205$ hvor metodens pålitelighet er 95%.

II. anta vi nå ønsker å teste

$$H_0 : p = p_0 \text{ mot } H_A : p < p_0$$

La nå som foran

$$Z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Ved klassisk testing så forkastes H_0 på nivået α hvis

$$Z \le -z_\alpha \text{ (=kritisk verdi på nivået } \alpha)$$

eller hvis P-verdien

$$P_{H_0}(Z \le z)$$

blir tilstrekkelig liten. z er som vanlig den observerte verdien av Z. I denne situasjonen vil z vanligvis være negativ. Det betyr at P-verdien alternativt kan beregnes ved

$$P_{H_0}(Z \ge |z|)$$

der $|z|$ betegner absoluttverdien til z.

Eks. Anta at andelen røykere blant kvinner over flere år har ligget stabilt på 35%. En stor kampanje mot røyking blir gjennomført både på skoler, i aviser og på TV over en 3 måneders periode og man har stor tro på at denne kampanjen skal virke. Man ønsker derfor nå å teste

$$H_0 : p = 0{,}35 \text{ mot } H_A : p < 0{,}35$$

En har nå kritsk verdi $= -z_{0,05} = -1{,}645$ hvis vi som vanlig tester på 5%-nivået. Nå finner en i et tilfeldig på $n = 750$ kvinner tatt etter at kampanjen var ferdig $x = 251$ røykere.

Nå blir

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\frac{251}{750} - 0{,}35}{\sqrt{\frac{0{,}35(1-0{,}35)}{750}}} = -0{,}880$$

En ser nå at $z = -0{,}880 > -1{,}645$ hvilket medfører at $H_0 : p = 0{,}35$ ikke kan forkastes på 5%-nivået. Det tyder altså ikke på at det noen signifikant nedgang i andelen kvinnelige røykere.

Regner en ut P-verdien ser bedre hvilken risiko det er forbundet med eventuelt feilaktig å forkaste $H_0 : p = 0{,}35$. P-verdien blir

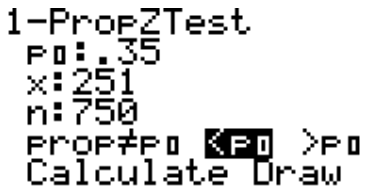
$$P_{H_0}(Z \le -0{,}880) = Normalcdf(-10^{99}, -0{,}880) = 0{,}1894$$

En ser da altså at vi måtte hatt et nivå på 18,94% før vi kunne forkaste H_0.

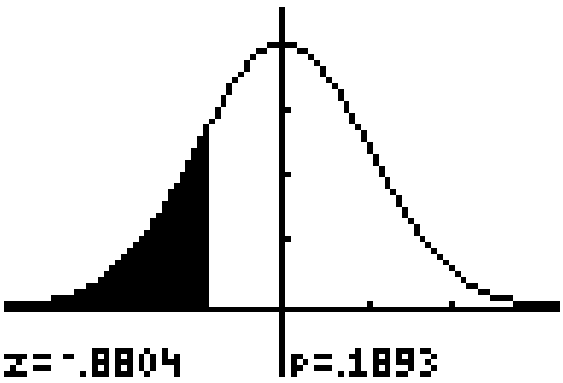
Brukes en nå kalkulatoren direkte finner en ved hjelp av kommandoene:

STAT
TESTS
5: 1-PropZTest
ENTER

Legger en så inn de aktuelle tallene over vil en få følgende bilde:



Hvis man nå velger kommandoen Draw får en følgende bilde:



Hvis man ønsker å bruke MINITAB må en igjen gjøre bruk av kommandoene over. Velger så Summarized data og legger inn

Number of trials = 873

Number of events = 303

Deretter velger en Options og ber om

Test proportion	0,3
Alternative	less than

og haker av i feltet hvor en skal avgjøre om en skal regne tilnærmet eller ikke (Use test and interval based on normal distribution or not). Jeg velger nå å prøve både eksakt og tilnærmet utregning for å se eventuelle forskjeller. Eksakt utregning blir

Test and CI for One Proportion

```
Test of p = 0,35 vs p < 0,35

                                        95%
                                      Upper    Exact
Sample    X    N  Sample p          Bound  P-Value
1       251  750  0,334667  0,364125    0,200
```

Regner en tilnærmet får en:

Test and CI for One Proportion

```
Test of p = 0,35 vs p < 0,35

                                      95%
                                    Upper
Sample    X    N  Sample p          Bound  Z-Value  P-Value
1       251  750  0,334667  0,363008    -0,88    0,189
```

En ser at det ikke er store forskjeller på (for eksempel) de to P-verdiene som begge gir samme konklusjon. I andre situasjoner vil en imidlertid kunne få forskjellig konklusjon.

III. Anta vi nå ønsker å teste

$$H_0 : p = p_0 \text{ mot } H_A : p \neq p_0$$

La nå som foran betrakte testobservatoren

$$Z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Ved klassisk testing så forkastes H_0 på nivået α hvis

$$Z \le -z_{\alpha/2} \quad \text{eller} \quad Z \ge z_{\alpha/2} \text{ (= de kritiske verdiene på nivået } \alpha)$$

eller hvis P-verdien

$$2\,P_{H_0}(Z \ge |z|)$$

blir tilstrekkelig liten. z er som vanlig den observerte verdien av Z. I denne situasjonen vil z enten være negativ eller positiv.

Eks. Anta at andelen defekte av en masseprodusert artikkel fram til nå har ligget på 7%. Etter man har begynt å importere noe billigere råstoff fra et annet land er man usikker på om dette endrer kvaliteten på artikkelen eller ikke. La p = P(En vilkårlig uttrukket artikkel er defekt med det nye råstoffet). Vi ønsker derfor å teste

$$H_0 : p = 0{,}07 \text{ mot } H_A : p \ne 0{,}07$$

Bestem nå de kritiske verdiene med hensyn på andelen (og antall) defekte i et tilfeldig utvalg på n = 500 produserte artikler med det nye råstoffet. Hvis vi velger et nivå på 5% så forkastes mao. $H_0 : p = 0{,}07$ hvis

$$Z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \le -1{,}96 \text{ eller } Z \ge 1{,}96$$

der 1,96 er funnet av tabell eller kalkulator som $z_{0{,}025}$ = invNorm(0,975). Nå er

$\hat{P} = \frac{X}{n} = \frac{X}{500}$ og $p_0 = 0{,}07$ slik at $Z = \frac{\hat{P} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{\hat{P} - 0{,}07}{\sqrt{\frac{0{,}07(1-0{,}07)}{500}}} \le -1{,}96$ betyr at

$$\hat{P} - 0{,}07 \le -1{,}96\sqrt{\frac{0{,}07(1-0{,}07)}{500}} \quad \text{dvs}$$

$$\hat{P} \le 0{,}07 - 1{,}96\sqrt{\frac{0{,}07(1-0{,}07)}{500}} = 0{,}07 - 0{,}022 = 0{,}048$$

Tilsvarende finner en at $Z \ge 1{,}96$ gir $\hat{P} \ge 0{,}07 + 0{,}022 = 0{,}092$.

Mao. $H_0 : p = 0{,}07$ forkastes til fordel for $H_A : p \neq 0{,}07$ på 5%-nivået hvis andelen defekte i utvalget $\hat{p}$ (= verdien av $\hat{P}$) enten er $\leq 0{,}048$ eller er $\geq 0{,}092$. Det vil igjen si at H_0 forkastes hvis $\hat{P} = \frac{X}{n} = \frac{X}{500} \leq 0{,}048$ dvs at $X \leq 0{,}048 \cdot 500 = 24$ eller $X \geq 0{,}092 \cdot 500 = 46$. Mao. hvis antall defekte i et utvalg på 500 er mindre enn eller lik 24 eller større enn eller lik 46 så er det grunn til å påstå at defektandelen har endret seg fra 0,07.

Anta at et tilfeldig utvalg på 500 viser 49 defekte. Siden $49 \geq 46$ så forkastes $H_0 : p = 0{,}07$ på 5 % -nivået. P-verdien for dette resultatet relativt til nullhypotesen og alternativet over er dermed

$$2P_{H_0}(X \geq 49) = 2(1 - P_{H_0}(X \leq 48)) = 2(1 - binomcdf(500, 0.07, 48)) = 0{,}0233(eksakt)$$

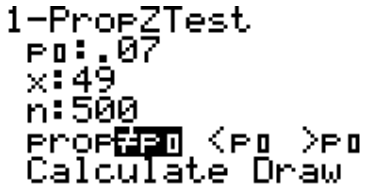
Bruker en alternativt normaltilnærmelsen har en

$$2P_{H_0}(X \geq 49) = 2Normalcdf(48.5, 10^{99}, 500 \cdot 0.07, \sqrt{500 \cdot 0.07 \cdot (1 - 0.07)}) = 0{,}0180$$

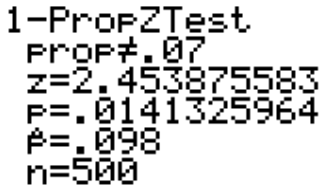
Bruker en nå kalkulatoren direkte finner en etter kommandoene

STAT
TESTS
5: 1-PropZTest…
ENTER

og etter å ha lagt inn de aktuelle tallene får en opp følgende bilde:



Velger en så kommandoen Calculate får en følgende bilde:



Hvis man ønsker å bruke MINITAB må en igjen gjøre bruk av kommandoene over. Velger så Summarized data og legger inn

Number of trials = 500

Number of events = 49

Deretter velger en Options og ber om

Test proportion	0,07
Alternative	not equal

og haker av i feltet hvor en skal avgjøre om en skal regne tilnærmet eller ikke (Use test and interval based on normal distribution or not). Eksakt utregning gir

Test and CI for One Proportion

```
Test of p = 0,07 vs p not = 0,07


                                                    Exact
Sample   X    N  Sample p         95% CI          P-Value
1       49  500  0,098000  (0,073383; 0,127489)    0,022
```

Tilnærmet utregning gir:

Test and CI for One Proportion

```
Test of p = 0,07 vs p not = 0,07


Sample   X    N  Sample p         95% CI          Z-Value  P-Value
1       49  500  0,098000  (0,071940; 0,124060)     2,45    0,014
```

En ser at begge utregningene stemmer godt med resultatene over.

En ser at det eksakte 95% konfidensintervall for den nye defektandelen går fra og med 7,3% til og med 12,7%. Kontroller selv at dette stemmer.

14. Inferens knyttet til to andeler.

Anta at vi nå har to populasjoner som begge antas å være binomisk fordelt med parametre henholdsvis p_1 og p_2. Anta vi er interessert i å finne ut om det er noen forskjell mellom disse andelene og hva denne forskjellen eventuelt er. La A være det kjennetegnet som vi interesserer oss for.

Anta at det tas to uavhengige tilfeldige utvalg på henholdsvis n_1 og n_2 fra de to populasjonene og la X_1 og X_2 være antall individer med kjennetegnet A (henholdsvis) i de to utvalgene. Dermed kan en definere de to estimatorene $\hat{P}_1$ og $\hat{P}_2$ for henholdvis p_1 og p_2 ved

$$\hat{P}_1 = \frac{X_1}{n_1} \text{ og } \hat{P}_2 = \frac{X_2}{n_2}$$

Det kan da vises at

$$E(\hat{P}_1 - \hat{P}_2) = p_1 - p_2,$$

at standardavviket til $\hat{P}_1 - \hat{P}_2$ er gitt ved

$$SE_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

og dermed at

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$$

Ved hjelp av denne kan en som tidligere utlede følgende $(1-\alpha)100\%$ konfidensintervall for $p_1 - p_2$ (Prøv selv. Vink: Erstatt $\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$ med $\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$ og ta utgangspunkt i at $P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1-\alpha$)

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Eks. Anta at to uavhengige utvalg av menn og kvinner på henholdsvis n_1=520 og n_2 = 480 viste at blant mennene var det x_1= 202 som var tilhengere av EU medlemskap og blant kvinnene var x_2 = 156 tilhengere. Finn et 95% konfidensintervall for differansen mellom andelen menn (= p_1) og andelen kvinner (= p_2) som er tilhengere av norsk EU medlemskap. En har nå at estimater for henholdsvis p_1 og p_2 er gitt ved

$$\hat{p}_1 = \frac{x_1}{n_1} = \frac{202}{520} \text{ og } \hat{p}_2 = \frac{x_2}{n_2} = \frac{156}{480}$$

dermed har en følgende utgangspunkt for å lage et 95% konfidensintervall for $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \pm z_{0,025}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} =$$

$$\frac{202}{520} - \frac{156}{480} \pm 1{,}96\sqrt{\frac{\frac{202}{520}(1-\frac{202}{520})}{520} + \frac{\frac{156}{480}(1-\frac{156}{480})}{480}} =$$

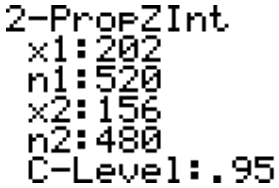
$$0{,}063 \pm 0{,}059$$

Dermed vil det endelige konfidensintervallet gå fra og med 0,004 og til og med 0,122.

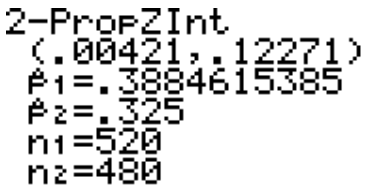
Velger en nå å bruke kalkulatoren direkte gjennom kommandoene

STAT
TESTS
B: 2-PropZInt…
ENTER

og så legge inn de observerte verdiene får en følgende bilde:



Velger en så kommandoen Calculate får en opp følgende resultat:



som bekrefter beregningene over.

Hvis man ønsker å bruke MINITAB må en (som før) bruke følgende kommandoer:

Stat
Basic Statistics
2 P 2 Proportion

Her velger en Summarized data og legger inn

	Trials	Events
First	520	202
Second	480	156

Går så inn på Options og velger 95% Confidence level. En får da følgende resultat (bruker ikke sammenslått estimat (gjelder kun hypoteseprøving)):

Test and CI for Two Proportions

```
Sample    X    N  Sample p
1       202  520  0,388462
2       156  480  0,325000


Difference = p (1) - p (2)
Estimate for difference:  0,0634615
95% CI for difference:  (0,00421108; 0,122712)
```

som igjen gir en bekreftelse av resultatene over. Jeg har her valgt å ikke ta med den delen av utskriften som har med hypoteseprøving å gjøre. Det skal vi nå straks behandle.

Hypoteseprøving for to andeler.

Anta som over at vi har to populasjoner som begge antas å være binomisk fordelt med parametre henholdsvis p_1 og p_2. La A være det kjennetegnet som vi interesserer oss for.

Anta at det tas to uavhengige tilfeldige utvalg på henholdsvis n_1 og n_2 fra de to populasjonene og la

$$Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

som over.

Anta at vi nå ønsker å teste

$$H_0 : p_1 = p_2$$

Vi skal som vanlig se på tre forskjellige alternative hypoteser: $p_1 \geq p_2$, $p_1 \leq p_2$ og $p_1 \neq p_2$. Som testobservator skal vi nå bruke

$$Z = \frac{(\hat{P}_1 - \hat{P}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}}$$

idet

$$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} = p(1-p)(\frac{1}{n_1} + \frac{1}{n_2})$$

der

idet $p_1 = p_2 = p$ (=den felles verdien under H_0)

Nå er imidlertid p ukjent, og må estimeres. Det gjøres ved hjelp av

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{\textit{antall med egenskapen A totalt i de to utvalgene}}{\textit{totalt antall utvalgte}}$$

I. Vi ser som vanlig først på testing av

$$H_0 : p_1 = p_2 \text{ mot } H_A : p_1 > p_2$$

Ved klassisk testing så forkastes H_0 på nivået α hvis

$$z \geq z_\alpha \text{ (=kritisk verdi på nivået } \alpha)$$

eller hvis P-verdien

$$P_{H_0}(Z \geq z)$$

blir tilstrekkelig liten. z er som vanlig den observerte verdien av *Z*.

La oss nå se på eksempelet over hvor det ble tatt to uavhengige utvalg av menn og kvinner på henholdsvis n_1=520 og n_2 = 480 som viste at blant mennene var det x_1 = 202 som var tilhengere av EU medlemskap og blant kvinnene var x_2 = 156 tilhengere. Test nå på 5%-nivået om andelen menn (= p_1) og andelen kvinner (= p_2) som er tilhengere av norsk EU medlemskap er like eller om det er slik at andelen menn er større. En har nå at estimat for p (den felles verdien av p_1 og p_2) er gitt ved

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{202 + 156}{520 + 480} = \frac{358}{1000} = 0{,}358$$

Dermed blir

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \frac{\frac{202}{520} - \frac{156}{480}}{\sqrt{0{,}358(1-0{,}358)(\frac{1}{520} + \frac{1}{480})}} = 2{,}09$$

Konklusjonen blir dermed at

$$H_0 : p_1 = p_2$$

forkastes til fordel for

$$H_A : p_1 > p_2$$

på 5%-nivået fordi

$$z = 2{,}09 \geq 1{,}645 = z_{0{,}05}$$

Det betyr mao. at vi med en sannsynlighet på høyst 0,05 for å ta feil påstår at andelen menn for EU i populasjonen er større en andelen kvinner.

Et litt sterkere resultat får en ved å beregne signifikanssannsynligheten (P-verdien) ved

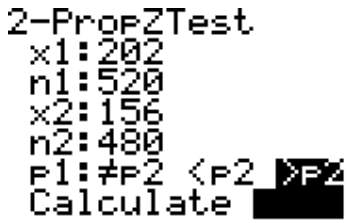
$$P_{H_0}(Z \geq 2{,}09) = Normalcdf(2.09, 10^{99}) = 0{,}018$$

Mao. vi får samme konklusjon, men risikoen for å ta feil er nå 0,018. Det eneste vi kan si ved hjelp av klassisk testing er at risikoen for å ta feil er mindre enn 0,05. Ved å beregne p-verdien har vi altså funnet et (tilnærmet) mål på sannsynligheten for å trekke gal konklusjoon.

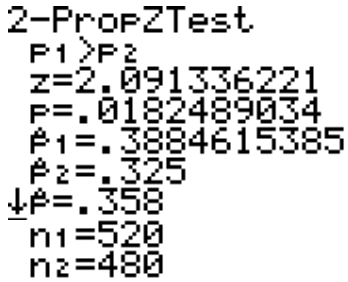
Bruker en nå kalkulatoren direkte finner en etter kommandoene

STAT
TESTS
6: 2-PropZTest…
ENTER

og etter å ha lagt inn de aktuelle tallene og valgt alternativ hypotese $H_A : p_1 > p_2$ får en opp følgende bilde



Velger en nå kommandoen Calculate får en opp følgende bilde:



og ser at man får samme konklusjon som foran.

Hvis man ønsker å bruke MINITAB må en (som før) bruke følgende kommandoer:

Stat
Basic Statistics
2 P 2 Proportion

Her velger en Summarized data og legger inn

	Trials	Events
First	520	202
Second	480	156

Går så inn på Options og velger alternativet ”greater than” og 95% Confidence level. En får da følgende resultat etter å ha valgt pooled estimate (sammenslått estimat)

Test and CI for Two Proportions

```
Sample    X    N  Sample p
1       202  520  0,388462
2       156  480  0,325000


Difference = p (1) - p (2)
Estimate for difference:  0,0634615
95% lower bound for difference:  0,0137370
Test for difference = 0 (vs > 0):  Z = 2,09
P-Value = 0,018
```

Dette er igjen en bekreftelse på beregningene over. Dessuten ser en at nedre grense i et 95% ensidig konfidensintervall er 0,0137. Hva betyr dette?

I et $(1-\alpha)100\%$ ensidig konfidensintervall utledes den nedre grensen av

$$\hat{p}_1 - \hat{p}_2 - z_\alpha \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Et 95% ensidig konfidensintervall har da nedre grense:

$$\frac{202}{520} - \frac{156}{480} - 1{,}645\sqrt{\frac{\frac{202}{520}(1-\frac{202}{520})}{520} + \frac{\frac{156}{480}(1-\frac{156}{480})}{480}} = 0{,}0635 - 0{,}0497 = 0{,}0138$$

som stemmer ganske bra med MINITAB sin beregning.

II. Anta nå at vi skal teste

$$H_0 : p_1 = p_2 \text{ mot } H_A : p_1 < p_2$$

Ved klassisk testing så forkastes H_0 på nivået α hvis

$$z \le -z_\alpha \text{ (=kritisk verdi på nivået } \alpha)$$

eller hvis P-verdien

$$P_{H_0}(Z \le z)$$

blir tilstrekkelig liten. z er som vanlig den observerte verdien av *Z*.

Eks. Anta at man ønsker å teste om det er slik at andelen jenter som må få ekstra leseopplæring på et bestemt klassetrinn er mindre enn den tilsvarende andelen gutter. Mao. en ønsker å teste

$$H_0 : p_1 = p_2 \text{ mot } H_A : p_1 < p_2$$

der p_1 er andelen jenter som må få ekstra leseopplæring og p_2 er den tilsvarende andelen gutter. I to tilfeldige utvalg på henholdsvis n_1=385 jenter og n_2 = 405 gutter fant man at blant jentene var det x_1 = 30 som trengte ekstra leseopplæring og blant guttene var det x_2 = 45. Tester nå på 5%-nivået om andelen jenter (= p_1) er lik andelen gutter (= p_2) som må få ekstra leseopplæring eller om det er slik at andelen jenter er mindre.

Nå forkastes $H_0 : p_1 = p_2$ på 5%-nivået hvis $z \le -z_{0,05} = -1{,}645.$ En har nå et estimat for *p* (den felles verdien av p_1 og p_2) gitt ved

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{30 + 45}{385 + 405} = \frac{75}{790} = 0{,}095$$

Da finner en følgende verdi av *Z*:

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{\frac{30}{385} - \frac{45}{405}}{\sqrt{0{,}095(1-0{,}095)(\frac{1}{385} + \frac{1}{405})}} = -1{,}590$$

Konklusjonen blir dermed at $H_0 : p_1 = p_2$ ikke kan forkastes på 5%-nivået idet *z* ikke har blitt mindre enn -1,645.

Beregner en alternativt P-verdien finner en

$$P_{H_0}(Z \le -1{,}590) = Normalcdf(-10^{99}, -1{,}590) = 0{,}0559$$

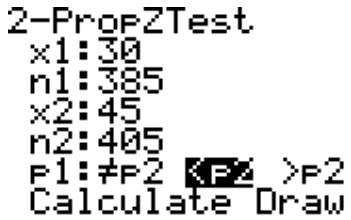
Her ser at hvis H_0 forkastes så er sannsynligheten for å gjøre feil (feilaktig forkastning) 0,0559 hvilket er større enn (så vidt) nivået på 0.05 (den øvre grensen for å gjøre feil). Men der er all grunn til å følge opp dette med ytterligere undersøkelser.

Bruker en nå kalkulatoren direkte finner en etter kommandoene

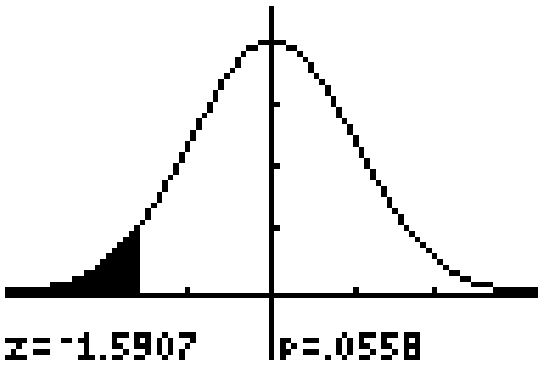
STAT
TESTS

6: 2-PropZTest…
ENTER

og etter å ha lagt inn de aktuelle tallene og valgt alternativ hypotese $H_A : p_1 < p_2$ får en opp følgende bilde:



Velger en så kommandoen Draw istedenfor Calculate ser en følgende:



Dette bekrefter beregningene over.

Hvis man ønsker å bruke MINITAB må en (som før) bruke følgende kommandoer:

Stat
Basic Statistics
2 P 2 Proportion

Her velger en Summarized data og legger inn

	Trials	Events
First	385	30
Second	405	45

Går så inn på Options og velger alternativet ”less than” og 95% Confidence level (dette siste må gjøres selv om kun ønsker hypoteseprøving. Hvis man ikke skriver noe i dette feltet velger allikevel MINITAB 95%). En får da følgende resultat etter å ha valgt pooled estimate (sammenslått estimat)

Test and CI for Two Proportions

```
Sample   X    N  Sample p
1       30  385  0,077922
2       45  405  0,111111


Difference = p (1) - p (2)
Estimate for difference:  -0,0331890
95% upper bound for difference:  0,000938775
Test for difference = 0 (vs < 0):  Z = -1,59
P-Value = 0,056
```

Legg merke til at nedre konfidensgrense er praktisk talt lik 0 (0,0009….) hvilket også bekrefter at resultatene nesten var signifikante (jfr. P-verdien på 0,056). Prøv selv å beregne denne nedre grensen)

III. Til sist skal vi se på testing av

$$H_0 : p_1 = p_2 \text{ mot } H_A : p_1 \neq p_2$$

Ved klassisk testing så forkastes H_0 på nivået α hvis

$$z \leq -z_{\alpha/2} \text{ eller } z \geq z_{\alpha/2} \text{ (=kritiske verdier på nivået } \alpha)$$

eller hvis P-verdien

$$2\,P_{H_0}(Z \leq |z|)$$

blir tilstrekkelig liten. $|z|$ er som tallverdien av den observerte verdien av *Z*.

Eks. Anta at vi ønsker å teste om det er noen forskjell på andelen personer i Oslo og i Bergen som følger med fast på en bestemt TV-serie. Siden vi ikke har data fra noen tidligere undersøkelser og vi ikke vet noe om tendenser i den ene eller andre retningen må vi her teste:

$$H_0 : p_1 = p_2 \text{ mot } H_A : p_1 \neq p_2$$

Anta at det ble foretatt to uavhengige tifeldige utvalg på henholdsvis n_1=854 i Bergen og n_2 = 923 i Oslo og at man fant at det i utvalget i Bergen var x_1 = 83 og at det i Oslo var x_2 = 95 som fulgte med på TV-programmet.

Nå forkastes $H_0 : p_1 = p_2$ på 5%-nivået hvis $z \leq -z_{0,05} = -1{,}645$ eller $z \geq z_{0,05} = 1{,}645$.

En har nå et estimat for *p* (den felles verdien av p_1 og p_2) gitt ved

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{83 + 95}{854 + 923} = \frac{178}{1777} = 0{,}100$$

Da finner en følgende verdi av *Z*:

$$z = \frac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = \frac{\frac{83}{854} - \frac{95}{923}}{\sqrt{0{,}100(1-0{,}100)(\frac{1}{854} + \frac{1}{923})}} = -0{,}403$$

Konklusjonen blir dermed at $H_0 : p_1 = p_2$ ikke kan forkastes på 5%-nivået idet *z* verken har blitt mindre enn -1,645 eller større enn 1,645. Mao. det tyder ikke på at $p_1 \neq p_2$.

Beregner en alternativt P-verdien finner en

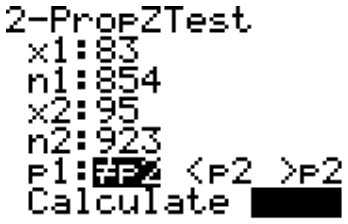
$$2\,P_{H_0}(Z \geq |-0{,}403|) = 2\,Normalcdf(0.403, 10^{99}) = 2 \cdot 0{,}343 = 0{,}686$$

Herav ser at H_0 slett ikke kan forkastes på 5%-nivået og ikke en gang på 34%-nivået. Det er mao. ganske sterke signaler om at $H_0 : p_1 = p_2$ er riktig (les: ikke bør forkastes)

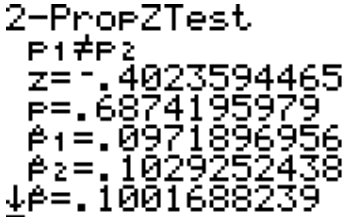
Bruker en nå kalkulatoren direkte finner en etter kommandoene

STAT
TESTS
6: 2-PropZTest…
ENTER

og etter å ha lagt inn de aktuelle tallene og valgt alternativ hypotese $H_A : p_1 \neq p_2$
får en opp følgende bilde:



Velger en kommandoen Calculate får en opp følgende bilde:



```
n1=854
n2=923
```

Som stemmer bra med resultatene over.

Hvis man ønsker å bruke MINITAB må en (som før) bruke følgende kommandoer:

Stat
Basic Statistics
2 P 2 Proportion

Her velger en Summarized data og legger inn

	Trials	Events
First	854	83
Second	923	95

Går så inn på Options og velger alternativet "less than" og 95% Confidence level (dette siste må gjøres selv om en kun ønsker hypoteseprøving. Hvis man ikke skriver noe i dette feltet velger allikevel MINITAB 95%). En får da følgende resultat etter å ha valgt pooled estimate (sammenslått estimat)

Test and CI for Two Proportions

```
Sample   X    N  Sample p
1       83  854  0,097190
2       95  923  0,102925

Difference = p (1) - p (2)
Estimate for difference:  -0,00573555
95% CI for difference:  (-0,0336455; 0,0221744)
Test for difference = 0 (vs not = 0):
Z = -0,40  P-Value = 0,687
```

som igjen stemmer med beregningene over. Her ser en dessuten at 95% konfidensintervall for $p_1 - p_2$ går fra -0,033 til 0,022. Vis dette resultatet og forklar hvordan du kan bruke dette i forbindelse med hypoteseprøvingen.

Mer om konfidensintervaller for en og to andeler.

En andel.

Eks.(Se Moore & McCabe s 573)

Anta at vi man knipser en mynt $n = 3$ ganger og får $x = 3$ kron. Basert på dette vil estimatet for kron være $\hat{p} = \frac{x}{n} = \frac{3}{3} = 1$ og et estimat for standardavviket være

$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{1(1-1)}{3}} = 0$. Det betyr at vi estimerer sannsynligheten for kron ved myntkast til 1 med et standardavvik på 0 (mao. ingen usikkerhet ved vårt estimat). Dette er ikke fornuftig. For å korrigere for slike situasjoner kan man gjøre bruk av følgende ide.

Lat som man har 4 tilleggsobservasjoner hvorav 2 er suksesser og 2 er fiaskoer. En estimator for *p* er da

$$\widetilde{P} = \frac{X+2}{n+4}$$

Denne estimatoren ble først foreslått av Edwin Bidwell Wilson i 1927 og kalles derfor for Wilson estimatoren. Fordi $\widetilde{p}$ (det nye estimatet for *p*) er basert på en populasjonsstørrelsen (n+4) så har en tilnærmet at

$$Z = \frac{\widetilde{P} - p}{\sqrt{\frac{p(1-p)}{n+4}}} \sim N(0,1)$$

Dermed blir $(1-\alpha)100\%$ konfidensintervall på formen

$$\widetilde{p} \pm z_{\alpha/2}\sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{n+4}}$$

som ikke vil avvike så mye fra konfidensintervallet på side 119

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

når *n* er stor. I eksempelet på side 120 hvor *p* er andelen stemmeberettigede i Norge som er for norsk medlemskap i EU fant vi et 95% konfidensintervall for *p* ved

$$\hat{p} \pm 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{495}{1200} \pm 1{,}96\sqrt{\frac{\frac{495}{1200}(1-\frac{405}{1200})}{1200}} = 0{,}413 \pm 0{,}028$$

Lager en nå isteden konfidensintervallet basert på Wilsonestimatoren finner en

$$\widetilde{p} \pm z_{\alpha/2}\sqrt{\frac{\widetilde{p}(1-\widetilde{p})}{n+4}} = \frac{497}{1204} \pm 1{,}96\sqrt{\frac{\frac{497}{1204}(1-\frac{497}{1204})}{1204}} = 0{,}413 \pm 0{,}028$$

M.a.o med 3 desimalers nøyaktighet er det ingen forskjell på disse to intervallene. Hvis derimot *n* ikke er stor så vil Wilsonintervallet gi et mer pålitelig forslag enn det vanlige konfidensintervallet.

Eks. Anta for eksempel at at det kun var 120 personer med i utvalget og at man da fant at det var 50 av disse som var for EU-medlemskap. 95% konfidensintervall ved de to metodene ville da gitt følgende resultater:

$$\hat{p} \pm 1{,}96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \frac{50}{120} \pm 1{,}96\sqrt{\frac{\frac{50}{120}(1-\frac{50}{120})}{120}} = 0{,}417 \pm 0{,}088 \text{ (direkte)}$$

og

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}} = \frac{52}{124} \pm 1{,}96\sqrt{\frac{\frac{52}{124}(1-\frac{52}{124})}{124}} = 0{,}419 \pm 0{,}087 \text{ (Wilson)}$$

Her ser en mao. en viss forskjell, men den er ikke dramatisk stor. Legg imidlertid merke til at $\pm$ leddet har økt ganske kraftig fra konfidensintervallene basert på et utvalg på *n* = 1200. Hva skyldes dette?

To andeler.

På side 130 utledet vi følgende $(1-\alpha)100\%$ konfidensintervall for $p_1 - p_2$:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Hvis man igjen bruker Wilsonteknikken foran og nå justerer resultatene med 2 suksesser og 2 fiaskoer på følgende måte: Legg til 1 suksess og 1 fiasko i hvert av utvalgene. Wilsonestimatorene for henholdvis p_1 og p_2 er da:

$$\tilde{P}_1 = \frac{X_1 + 1}{n_1 + 2} \text{ og } \tilde{P}_2 = \frac{X_2 + 1}{n_2 + 2}$$

og herav har følgende estimator for p_1 - p_2:

$$\tilde{P}_1 - \tilde{P}_2 = \frac{X_1 + 1}{n_1 + 2} - \frac{X_2 + 1}{n_2 + 2}$$

Standardavviket til $\tilde{P}_1 - \tilde{P}_2$ er da tilnærmet gitt ved

$$\sigma_{\tilde{P}_1-\tilde{P}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1+2}+\frac{p_2(1-p_2)}{n_2+2}}$$

Dermed kan en utlede følgende justerte $(1-\alpha)100\%$ konfidensintervall for $p_1 - p_2$:

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2}+\frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}$$

Går en nå tilbake til eksempelet knyttet til ekstra leseopplæring hvor en hadde to tilfeldige utvalg på henholdsvis n_1=385 jenter og n_2 = 405 gutter og fant at blant jentene var det x_1 = 30 som trengte ekstra leseopplæring og blant guttene var det x_2 = 45. 95% konfidensintervall for differansen mellom andelen gutter og andelen jenter som trenger ekstra leseopplæring blir da:

$$\hat{p}_1 - \hat{p}_2 \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} =$$

$$\frac{45}{405}-\frac{30}{385} \pm 1.96\sqrt{\frac{\frac{45}{405}(1-\frac{45}{405})}{405}+\frac{\frac{30}{385}(1-\frac{30}{385})}{385}} = 0{,}033 \pm 0{,}041$$

Velger en nå isteden å bruke Wilsonmetoden får en følgende 95%-konfidensintervall:

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2}+\frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}} =$$

$$\frac{47}{407}-\frac{32}{387} \pm 1.96\sqrt{\frac{\frac{47}{407}(1-\frac{47}{407})}{407}+\frac{\frac{32}{387}(1-\frac{32}{387})}{387}} = 0{,}033 \pm 0{,}041$$

Altså nok en gang ingen forskjell på de to metodene fordi utvalgene er forholdsvis store.

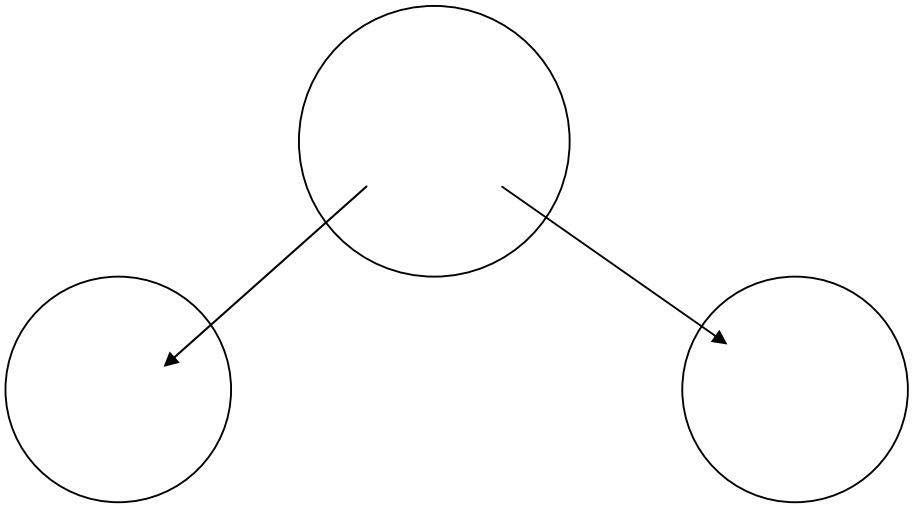
15. Ikkeparametriske tester.

Vi skal nå se på noen forskjellige tester som brukes når man skal sammenlikne to metoder (behandlinger) A og B. Tradisjonell metode (A) er den metoden som har vært praktisert "opp til i dag", mens ny metode (B) er den metode man nå ønsker å prøve ut mot tradisjonell metode og som man i de fleste sammenhenger har håp om at skal være bedre. Dette er testmetoder blant annet er mye brukt innenfor medisin, men som også kan brukes innenfor skoleforskning.

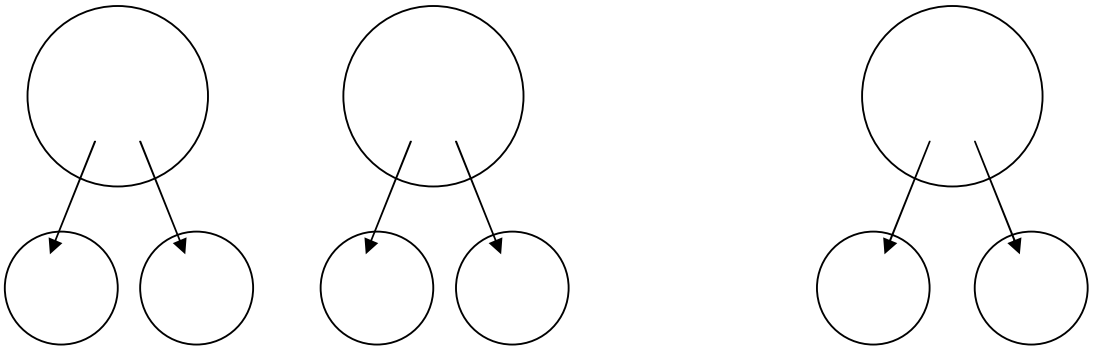
Gruppen av forsøksobjekter som skal være med i undersøkelsen bør være så homogene (like) som mulig. Dette gjøres ved at man lar en ekspertgruppe bestemme hvem som skal

være med i utvalget. Det betyr at det ikke er noen tilfeldighetsmekanisme så langt i forsøket. For å få inn dette momentet og dermed kunne bruke sannsynlighetsregningen skal vi se på to metoder: Fullstendig randomisering (tilfeldiggjøring) og randomisering innen blokker.

Fullstendig randomisering: Fra gruppen av forsøksobjekter trekker en tilfeldig ut (for eksempel ved loddtrekning) de som skal være behandlingsobjekter, dvs de som skal "behandles" med tradisjonell metode. De resterende blir såkalte kontrollobjekter.



Tilfeldiggjøring innen blokker: Forsøksobjektene deles først (etter skjønn) inn i flere grupper (også kalt blokker) på en slik måte at forsøksobjektene i hver gruppe blir så homogene som mulig. Deretter trekker en tilfeldig ut innen hver blokk de som skal være behandlingsobjekter. De øvrige blir kontrollobjekter.



Vi skal bare betrakte den situasjonen hvor hver blokk består av 2 forsøksobjekter slik at hver blokk gir et behandlingobjekt og et kontrollobjekt, og vi snakker da om såkalte parvise sammenlikninger.

Vi skal nå først se på to tester som brukes når man har fullstendig randomisering.

Fisher-Irwin-testen.

Eks. Anta at N = 25 karaktermessig like elever skal få opplæring i et fremmedspråk. Det kan gjennomføres ved to metoder: Vanlig klasseromsundervisning og ved et opplegg hvor elevene i stor utstrekning skal arbeide på egenhånd i grupper, men søke veiledning hos læreren når det er påkrevet. Det kan se ut som elever ved skolen gjennom gruppearbeid ikke greier å arbeide like effektivt og dermed ikke oppnår så gode resultater på tester. Man bestemmer seg for å trekke tilfeldig ut M = 13 elever (behandlingsobjektene) av de 25 som alle får undervisning ved "gruppearbeidsmetoden". De øvrige N-M = 12 elevene får tradisjonell undervisning. Anta at elevene etter 14 dager får en relativt vanskelig test hvor resultatet blir som følger:

Resultat / Metode	Godkjent	Ikke godkjent	SUM
Tradisjonell metode (klasseromsund.)	8	4	12
Ny metode (selvstudium)	6	7	13
SUM	14	11	25

Man ønsker nå å teste

H_0 : Det er ingen forskjell på de to undervisningsmetodene.

mot

H_A : Gruppearbeidsmetoden gir dårligere resultat enn tradisjonell metode.

Som testobservator skal vi nå bruke

X = Antall behandlingsobjekter (får ny metode) som får ikke godkjent

Små verdier av X er signifikante. Hvis H_0 er riktig så er X hypergeometrisk fordelt med N (= det totale antall forsøksobjekter) = 25, M (= antall spesielle elementer, dvs antall elever som får ikke godkjent) = 11 og n (= antall i utvalget, dvs antall elever som blir trukket ut til å være behandlingsobjekter) = 13

Nå observeres $X = 4$. P-verdien blir da:

$$P(X \le 4) = P(X = 0) + P(X = 1) + \cdots P(X = 4) =$$

$$= \frac{\binom{11}{0}\binom{14}{13}}{\binom{25}{13}} + \frac{\binom{11}{1}\binom{14}{12}}{\binom{25}{13}} + \cdots + \frac{\binom{11}{4}\binom{14}{9}}{\binom{25}{13}} = 0{,}1628 > 0{,}05$$

(kontroller beregningene selv)
Mao. H_0 kan ikke forkastes. Det er mao. ingen grunn til å påstå at gruppearbeidsmetoden gir dårligere resultat enn tradisjonell metode.

Den nærmeste testen i MINITAB er den såkalte Mann-Whitney-testen som tester om medianene i de to utvalgene er like. Dvs. at i denne situsjonen testes det:

Wilcoxontesten for to utvalg.

Anta nå at vi har samme situasjon som i eksempelet foran under Fisher-Irwintesten, men at vi nå i tillegg har poengene som den enkelte elev har fått på testen (Poeng er tildelt fra 0 til 100. For å bestå prøven må en ha minst 50 poeng)

Tradisjonell metode: 34, 50, 68, 88, 45, 80, 78, 14, 38, 74, 41, 61

Ny metode : 43, 63, 50, 90, 32, 18, 29, 31, 47, 56, 57, 12, 71

Igjen så tester vi:

H_0 : Det er ingen forskjell på de to undervisningsmetodene.

mot

H_A : Gruppearbeidsmetoden gir dårligere resultat enn tradisjonell metode.

Men nå rangordner vi dataene fra den minste poengsummen og opp til den største og tildeler rang fra 1 og opp til 25.

	12	14	18	29	31	32	34	38	41	43	45	47	50	50	56	57	61
Rang	**1**	2	**3**	**4**	**5**	**6**	7	8	9	**10**	11	**12**	**13,5**	13,5	**15**	**16**	17

63	68	71	74	78	80	88	90
18	19	**20**	21	22	23	24	**25**

Når flere observasjoner er like så tildeles tallene gjennomsnittsrangen av de rangtallene de skulle hatt om de hadde vært forskjellige. Mao hadde de to observasjonene på 50 vært

forskjellige (for eksempel 50 og 52) skulle de hatt rang henholdsvis 13 og 14, dvs til sammen 27 i rang. Siden disse nå er like får de isteden begge rangen $\frac{13+14}{2} = 13{,}5$ (mao. fortsatt 13,5+13,5 = 27 til sammen i rang.

Som testobservator skal vi nå bruke den såkalte Wilcoconobservatoren W so er gitt ved

W = rangsummen til behandlingsobjektene

Når H_0 er riktig så kan det vises at W er tilnærmet normalfordelt med

$$E(W) = \frac{n(N+1)}{2} \quad \text{og} \quad Var(W) = \frac{n(N-n)(N+1)}{12}$$

Nå er små verdier av W signifikante. Tabellen over gir nå følgende verdi av W (se **rangene** med **uthevet skrift**)

w= 1+3+4+5+6+10+12+13,5+15+16+18+20+25 = 148,5

Dermed blir P-verdien

$$P(W \le 148{,}5) \approx P(Z \le \frac{148{,}5 + 0.5 - \frac{13 \cdot 26}{2}}{\sqrt{\frac{13 \cdot 12 \cdot 26}{12}}}) = P(Z \le -1{,}09) = 0{,}138$$

En ser mao. fortsatt at det ikke er noen grunn til å forkaste H_0.

Velger en nå bruke MINITAB finner ved kommandoene

Statistics
Nonparametrics
Mann-Whitney

den såkalte Mann-Whitney-testen som tester på medianene, mao.

$H_0 : \eta_1 = \eta_2$ (medianene i de to gruppene er like)

mot

$H_A : \eta_1 > \eta_2$ (medianen i kontrollgruppa er større enn medianen i behandlingsgruppa)

Dette blir en litt annen test enn Wilcoxontesten, men som imidlertid også baserer seg på rangering av dataene.

En finner her ved først å legge inn dataene i to kolonner, og så velge konfidenskoeffisient på 95% (må velges selv om en her kun ønsker å drive hypoteseprøving) og så velge alternativet $\eta_1 > \eta_2$. Resultatet blir da:

Mann-Whitney Test and CI: C4; C5

```
     N  Median
C4  12   55,50
C5  10   39,50


Point estimate for ETA1-ETA2 is 13,50
95,6 Percent CI for ETA1-ETA2 is (-12,00;33,01)
W = 157,0
Test of ETA1 = ETA2 vs ETA1 > ETA2 is significant at 0,111
```

Herav ser en altså at vi får en P-verdi på 0,111 som er litt mindre enn den den tilnærmede P-verdien på 0,135 som vi fant foran.

Fortegnstesten.

Anta at 10 personer skal utføre en arbeidoperasjon 2 ganger; 1 gang med tradisjonell metode og 1 gang med en ny og forhåpentligvis bedre metode. Det trekkes lodd for hver person om hviken motode som skal brukes først. Her vil mao. hver person utgjøre ” 2 forsøksobjekter”. Den enkelte er behandlingsobjekt med ny metode og kontrollobjekt med tradisjonell metode.
Anta at resultatet av forsøket ble (tiden på operasjonen ble målt i sekuder)

Person	1	2	3	4	5	6	7	8	9	10
Tradisj. metode	48	53	52	48	55	62	54	86	40	71
Ny metode	45	42	58	46	50	56	47	50	45	63
Differens	+	+	-	+	+	+	+	+	-	+

Der man har satt et ”+” tegn hvis tiden med tradisjonell metode hos en person er lenger enn med ny metode. Hvis ikke har man satt et ”-” tegn.

Vi tester nå som før:

$$H_0 : \text{Det er ingen forskjell på de to metodene.}$$

mot

$$H_A : \text{Ny metode gir bedre resultat enn tradisjonell metode.}$$

So m testobservator skal vi nå bruke:

X = antall ganger (av de 10) at ny metode var best
= antall ”+” tegn i de 10 forsøkene (herav fortegnstesten)

Når H_0 er riktig så er X binomisk fordelt med n =10 og p = 0,5. Grunnen til at p = 0,5 er at det under H_0 (ingen forskjell på de to metodene) er like sannsynlig at det blir + som -. Nå observeres X = 8 og siden store verdier av testobservatoeren er signifikante blir P-verdien

$$P_{H_0}(X \geq 8) = 0{,}0547 > 0{,}05$$

Konklusjonen blir dermed at H_0 (så vidt) ikke forkastes på 5%-nivået. Fortegnstesten kalles også ofte bare tegntesten (eng. signtest)

Velger en nå bruke MINITAB finner ved kommandoene

Statistics
Nonparametrics
1-Sampl. Sign…

etter at man har lagt inn tallene i kolonne C1 og C2 og beregnet C3 = C1-C2, og til slutt bedt om at det testes på tallene i C3:

Sign Test for Median: C3

```
Sign test of median =  0,00000 versus > 0,00000

     N  Below  Equal  Above       P  Median
C3  10      2      0      8  0,0547   4,000
```

En ser at en får nøyaktig samme P-verdi som over. Hva betyr det for øvrig at man i MINITAB tester på medianen?

Fortegnstesten er en enkel, men ikke så veldig god test idet den kaster bort en del informasjon. Om differensen i tid er +1 eller +36 vektlegges like mye. En test som tar bedre vare på tallmaterialet er den såkalte Wilcoxontesten for parvise sammenlikninger.

Wilcoxontesten for parvise sammenlikninger.

Hvis vi i eksempelet under fortegnstesten også hadde regnet ut selve tidsdifferansen ville man tatt mye bedre vare på tallmaterialet. Ved bare å tildele + eller – kaster man bort mye informasjon. Hvis man i tillegg hadde rangordnet disse differansene i henhold til deres absoluttverdi ville en fått:

Person	1	2	3	4	5	6	7	8	9	10
Tradisj. metode	48	53	52	48	55	62	54	86	40	71
Ny metode	45	42	58	46	50	56	47	50	45	63
Differens	3	11	-6	2	5	6	7	36	-5	8
Rang	2	9	**5,5**	1	3,5	5,5	7	10	**3,5**	8

Rangering (etter absoluttverdi) av differansene: 2, 3, 5, -5, -6, 6, 7, 8, 11, 36
Disse skulle (hvis de alle hadde vært forskjellige) ha rangtallene: 1, 2, 3, 4, 5, 6, 7, 8, 9 og 10. Men siden noen av observasjonene er like så tildeles gjennomsnittsrangene. Nå tester en som over:

H_0 : Det er ingen forskjell på de to metodene.

mot

H_A : Ny metode gir bedre resultat enn tradisjonell metode.

Som testobservator skal vi nå bruke

V = Rangsummen til de negative differansene

En finner her $V = 5{,}5+3{,}5 = 9$

Nå er små verdier av V signifikante . Når H_0 er riktig kan det vises at V tilnærmet er normalfordelt med :

$$E(V) = \frac{n(n+1)}{4} \quad og \quad Var(V) = \frac{n(n+1)(2n+1)}{24}$$

der n = antall forsøksobjekter.

P-verdien blir da:

$$P_{H_0}(V \le 9) = P(Z \le \frac{9 + 0{,}5 - \frac{10 \cdot 11}{4}}{\sqrt{\frac{10 \cdot 11 \cdot 21}{24}}}) = P(Z \le -1{,}83) =$$

$$Normal(-10^{99}, -1.83) = 0{,}0336 < 0{,}05$$

Konklusjon: H_0 forkastes nå på 5%- nivået. Det tyder mao. på at ny metode er bedre (raskere) enn tradisjonell metode.

Egentlig burde man her hatt en 0,25 –istedenfor en 0,5- korreksjon. Hvorfor? Gjennomfør testingen med 0,25-korreksjon.

Hvis man nå ønsker å bruke MINITAB så beregner en først differansene mellom de 2 kollonene og utfører så en Wilcoxontest på den nye kolonnen. En bruker da kommandoene:

Statistics
Nonparametrics
Mann-Whitney

(som gir den såkalte Mann-Whitney-testen som tester på medianene)

Resultatet av dette blir:

Wilcoxon Signed Rank Test: C3

```
Test of median = 0,000000 versus median > 0,000000

           N
          for   Wilcoxon          Estimated
     N  Test  Statistic      P     Median
C3  10    10       46,0  0,033      5,000
```

En ser at dette stemmer med beregningen av P-verdien over med normaltilnærmelsen hvor vi fikk 0,0336.

16. Variansanalyse.

Vi skal nå se på en generalisering av to-utvalgs t-tester. ANOVA(**an**alysis **o**f **va**riance)-tester er en metode hvor man kan sammenlikne mer enn to populasjonsgjennomsnitt samtidig. Som kjent kan Students t-tester kan kun brukes for 1 og 2 gjennomsnitt. Forutsetningen er som for t-tester at man har normalfordelte populasjoner.
Helt generelt tenker vi oss nå at vi har ***k* normalfordelte populasjoner** hvor populasjon 1 har et gjennomsnitt på μ_1 og en varians på σ^2, hvor populasjon 2 har et gjennomsnitt på μ_2 og en varians på σ^2, og …….., hvor populasjon *k* har et gjennomsnitt på μ_k og en varians på σ^2. De *k* populasjonene antas mao. alle å ha lik varians. En ønsker nå å teste om de *k* gjennomsnittene er like eller om det er slik at minst ett av de er forskjellige fra de øvrige. Man tester mao.

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$ (gjennomsnittet er likt i alle populasjonene)

mot

$$H_A : Minst\ en\ av\ \mu - ene\ er\ \ forskjellig\ \ fra\ de\ øvrige$$

Vi tenker oss nå at det tas et tilfeldig utvalg på *n* fra hver av de *k* populasjonene.
Som det ligger i navnet så baserer testen seg på variasjonen i tallmaterialet (og dermed på variansen: variasjon $= \sum (x - \bar{x})^2$ og varians $= \dfrac{\text{var}iasjon}{df}$ der *df* som før står for antall frihetsgrader).

Anta nå at x_{ij} er observasjon nr *j* i utvalg (gruppe) nr *i.*
En vil dermed få følgende datamatrise:

Resultatmatrise fra de *k* utvalgene
(de *nk* observasjonene)

Utvalg (gruppe) 1	Utvalg (gruppe) 2		Utvalg (gruppe) k
x_{11}	x_{21}	….	x_{k1}
x_{12}	x_{22}	….	x_{k2}
⋮ ⋮	⋮ ⋮	⋮ ⋮	⋮ ⋮
x_{1n}	x_{2n}	….	x_{kn}
$\bar{x}_1$	$\bar{x}_2$		$\bar{x}_k$

der de *k* gjennomsnittene $\bar{x}_1$, $\bar{x}_2$, ….., $\bar{x}_k$ er henholdsvis gjennomsnittet i utvalg (gruppe)1, utvalg 2, ….., utvalg *k*.

Legg merke til at en her har omvendt notasjon av hva som er vanlig når det gjelder matriseregning hvor x_{ij} betegner element i rad nr i og kolonne nr j idet en her har at x_{ij} er observasjon nr *j* i utvalg (gruppe) nr *i.*

Vi får også bruk for gjennomsnittet av alle de *nk* observasjonene $\bar{x}$, som ofte kalles det store gjennomsnittet. En har mao.

$$\bar{x} = \frac{1}{nk}\sum_{i,j} x_{ij}$$

Som testobservator skal man nå bruke

$$F = \frac{Variansen\ mellom\ gruppene(utvalgene)}{Variansen\ innenfor\ gruppene(utvalgene)}$$

Hvis mange av $\mu - ene$ er forskjellige så vil variasjonen (og dermed variansen) mellom gruppene bli stor, og dette fører igjen til at *F* vil bli stor. Mao. jo større *F* blir jo mer sannsynlig er det at H_A er riktig dvs.: $Minst\ en\ av\ \mu - ene\ er\ forskjellig\ fra\ de\ øvrige$ og dermed at det er grunn til å forkaste $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$. Mao. H_0 forkastes hvis *F* $\geq k$ (=kritisk verdi). Legg merke til at *F* ikke kan bli negativ. Hvorfor er dette tilfellet?

Nå er variasjonen mellom gruppene gitt ved

$$SSTr = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{x}_i - \bar{x})^2 = n\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2$$

forutsatt at det er like mange observasjoner i hvert utvalg. Dette er imidlertid ofte ikke tilfelle og da er SST gitt ved

$$SSTr = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\bar{x}_i - \bar{x})^2 = \sum_{i=1}^{k} n_i(\bar{x}_i - \bar{x})^2$$

der $n_1, n_2, \ldots, n_k$ er antall observasjoner i de forskjellige utvalgene.
En ser at *SST* sammenlikner gjennomsnittet i hvert utvalg med det totale gjennomsnittet.

Variansen mellom gruppene er dermed

$$MSTr = \frac{SSTr}{Antall\ \ frihetsgrader} = \frac{SSTr}{k-1} = \frac{n\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2}{k-1}$$

forutsatt at det er like mange observasjoner i hver gruppe.

Analogt er nå variasjonen innen gruppene gitt ved

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2$$

forutsatt at alle utvalgene er like store. Her ser en at hver eneste observasjon blir sammenliknet med sitt gruppegjennomsnitt idet hvis man skriver ut dobbeltsummen får en:

$$\begin{aligned} SSE &= \sum_{i=1}^{k}[(x_{i1} - \bar{x}_i)^2 + (x_{i2} - \bar{x}_i)^2 + \cdots + (x_{in} - \bar{x}_i)^2] = \\ &= (x_{11} - \bar{x}_1)^2 + (x_{12} - \bar{x}_1)^2 + \cdots + (x_{1n} - \bar{x}_1)^2 \\ &+ (x_{21} - \bar{x}_2)^2 + (x_{22} - \bar{x}_2)^2 + \cdots + (x_{2n} - \bar{x}_2)^2 \\ &+ \cdots \\ &\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ &+ (x_{k1} - \bar{x}_k)^2 + (x_{k2} - \bar{x}_k)^2 + \cdots + (x_{kn} - \bar{x}_k)^2 \end{aligned}$$

Hvis utvalgene ikke er like store så bruker en isteden

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

Variansen innen gruppene er dermed

$$MSE = \frac{SSE}{antall\ \ frihetsgrader} = \frac{SSE}{k(n-1)} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2}{k(n-1)}$$

forutsatt at det er like mange observasjoner i hver gruppe.

En ser av utrykket for *SSE* over (utskrevet) at denne ikke påvirkes i samme grad som *SST* når minst et av populasjonsgjennomsnittene er forskjellig fra de øvrige. Dermed blir

$$F = \frac{Variansen\ mellom\ gruppene(utval\,gene)}{Variansen\ innenfor\ gruppene(utval\,gene)} = \frac{MST}{MSE}$$

også stor når minst et av populasjonsgjennomsnittene er forskjellig fra de øvrige. Når $H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ er riktig kan det vises at *F* er såkalt *Fisherfordelt* med (*k*-1) og (*n-k*) frihetsgrader. Det betyr mao. at H_0 forkastes på nivået α hvis

$$F \geq f_{(k-1),(n-k),\alpha}$$

der den kritiske verdien $f_{(k-1),(n-k),\alpha}$ er den såkalte α -fraktilen i *Fisherfordelingen* med (*k*-1) og (*n-k*) frihetsgrader.

Det er vanlig å sette alle disse summene og frihetsgradberegningene inn i en såkalt ANOVA-tabell. Den ser ut som følger:

Variasjonskilde	*Kvadratsum*	*Frihetsgrader*	*Varians*	*Verdi av F*	*P-verdi*
Mellom grupper	*SSTr*	*k*-1	*MSTr*	$f = \frac{MSTr}{MSE}$	$P_{H_0}(F \geq f)$
Innen grupper	*SSE*	*k*(*n*-1)	*MSE*		
Total	*SST*	*kn*-1			

I tillegg til summene og frihetsgradene som er nevnt foran er det også satt inn *SST* som som står for totalvariasjonen (**S**um**S**quared**T**otal). Denne er gitt ved

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x})^2$$

hvis alle utvalgene er like store, og

$$SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x})^2$$

Det kan vises (matematisk) at

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{x}_i-\bar{x})^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij}-\bar{x}_i)^2$$

eller mao. at

$$SST = SSTr + SSE$$

Dvs. at totalvarisjonen er lik summen av variasjonen mellom utvalgene og variasjonen innen utvalgene. Dette er nyttig når man skal regne med papir og blyant, fordi man da kan kontrollere om summene er regnet riktig ut. Dette er jo imidlertid helt uinteressant når man enten bruker MINITAB eller kalkulator.

Eks. Anta at en person har tatt tiden hun bruker på å kjøre til jobben på de forskjellige ukedagene. Hun har valgt å ta 6 *tilfeldig* tider på hver ukedag. (Hvordan kan hun gjøre dette). Anta at resultatet av undersøkelsen ble:

Tid i minutter:

Mandag	Tirsdag	Onsdag	Torsdag	Fredag
45	41	46	42	41
53	43	42	45	39
48	42	40	47	40
50	47	45	44	42
47	42	44	48	38
51	43	41	44	40

Herav finner en de forskjellige ukedagenes gjennomsnittskjøretider:

$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$	$\bar{x}_5$
49	43	43	45	40

Målingene viser en litt høyere gjennomsnittstid på mandag enn på de øvrige ukedagene. Dessuten kan det se ut som om gjennomsnittstiden på fredag er litt lavere enn de andre dagene.

Vi ønsker nå å teste :

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

dvs at gjennomsnittlig kjøretid (i populasjonene) er like, mot

$$H_A : \textit{Minst en av } \mu_1, \mu_2, \mu_3, \mu_4, \mu_5 \text{ er forskjellige fra de øvrige}$$

Hva forstår du med gjennomsnittlig populasjonskjøretid i denne sammenhengen?

Vi må nå regne ut de forskjellige *SS*-summene, deretter de tilhørende variansene og tilslutt sette opp selve ANOVA-tabellen som brukes til å trekke konklusjon. Spørsmålet vi skal

besvare er mao.: kan vi på bakgrunn av de innsamlede dataene forkaste påstanden om at gjennomsnitlig kjøretid (i det lange løp) for denne personen er lik på de forskjellige ukedagene slik at de avvikene hun har observert på de forskjellige ukedagene kun skyldes tilfeldigheter.

Vi finner nå først

$$\bar{x} = \frac{1}{kn}\sum_{i,j} x_{ij} = \frac{49+43+43+45+40}{5} = 44$$

Forklar hvorfor dette stemmer. Hvordan kan du kontrollere disse beregningene?

Deretter finner man

$$SSTr = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{x}_i - \bar{x})^2 = n\sum_{i=1}^{k}(\bar{x}_i - \bar{x})^2 =$$
$$6[(49-44)^2 + (43-44)^2 + \cdots + (40-44)^2] = 264$$

og

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2 =$$
$$= [(45-49)^2 + (53-49)^2 + \cdots + (51-49)^2$$
$$(41-43)^2 + (43-43)^2 + \cdots + (43-43)^2$$
$$+ \quad \cdots\cdots\cdots\cdots \quad +$$
$$(41-40)^2 + (39-40)^2 + \cdots + (40-40)^2]$$
$$= 134$$

Kontroller utregningene (fint utgangspunkt for hoderegning) over. Beregn så *SST* og kontroller at

$$SST = SSTr + SSE$$

Dermed finner en følgende ANOVA-tabell:

Variasjons-kilde	*Kvadrat-sum*	*Frihets-grader*	*Varians*	*Verdi av F*	*P-verdi*
Mellom grupper	*SST r* = 264	*k*-1 = 4	*MSTr* = 66	$f = \frac{MSTr}{MSE} = 13{,}095.$	$P_{H_0}(F \geq 13{,}095)$ =6,956E^-6 =0,000006956
Innen grupper	*SSE* = 126	*k*(*n-1)* = 25	*MSE* =5,04		
Total	*SST* = 390	*kn*-1 =29			

Der P-verdien er regnet ut ved hjelp av kalkulatoren med følgende kommandoer:

1-
2ND
VARS
9: Fcdf(
ENTER

Ved nå å legge inn grensene 0 og 13,095, samt frihetsgradene 4 og 25 finner en nå arealet under F-fordelingen mellom 12,31 og uendelig til $1{,}146 \cdot 10^{-5}$ (dvs. 0,000011..).
TI-84 viser nå:

```
1-Fcdf(0,13.095,
4,25)
   6.95585497E-6
```

Det betyr at $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ forkastes på 0,002% nivået og det tyder da på at minst et av populasjonsgjennomsnittene er forskjellige fra de øvrige.

En kan også gjennomføre hele variansanalysen på kalkulatoren. Da må en først legge inn tallene (dataene fra de 5 forskjellige ukedagene) i 5 forskjellige lister $L_1, L_2, \cdots, L_5$. Dette gjøres ved kommandoene

STAT
EDIT
ENTER

Når dette er gjort så bruker en kommandoene

STAT
TESTS
F:ANOVA
ENTER

Legger så inn $L_1, L_2, \cdots, L_5$ slik at kalkulatoren viser ANOVA(L_1, L_2, L_3, L_4, L_5). Trykker en deretter ENTER får en følgende skjermbilde:

```
One-way ANOVA
 F=13.92857143
 p=4.1775385E-6
 Factor
  df=4
  SS=280.8
↓ MS=70.2
 Error
  df=25
  SS=126
  MS=5.04
_Sxp=2.24499443
```

Dette ser en stemmer godt overens med den *F*- og *P*-verdien en har fått over.

Velger en nå å bruke MINITAB direkte, får en ved hjelp av følgende kommandoer:

STATISTICS
ANOVA
One-Way (unstacked)

One-way ANOVA: Mandag; Tirsdag; Onsdag; Torsdag; Fredag

```
Source  DF      SS     MS      F      P
Factor   4  264,00  66,00  13,10  0,000
Error   25  126,00   5,04
Total   29  390,00

S = 2,245   R-Sq = 67,69%   R-Sq(adj) = 62,52%


                          Individual 95% CIs For Mean Based on
                          Pooled StDev
Level    N    Mean  StDev    -+---------+---------+---------+--------
Mandag   6  49,000  2,898                            (----*----)
Tirsdag  6  43,000  2,098           (-----*----)
Onsdag   6  43,000  2,366           (-----*----)
Torsdag  6  45,000  2,191                 (-----*----)
Fredag   6  40,000  1,414    (----*-----)
                             -+---------+---------+---------+--------
                          38,5      42,0      45,5      49,0

Pooled StDev = 2,245
```

En ser at dette (selvfølgelig) også stemmer med beregningene over

Når dataene er gitt som her med en kolonne for mandagsdatene, en kolonne for tirsdagdataene, osv må en velge alternativet One-Way (unstacked). Hvis dataene (kjøretidene) alle sammen lå i en kolonne og dagene mandag (for eksempel =1), tirsdag (for eksempel =2) osv … fredag (for eksempel =5) lå i en annen kolonne skulle en ha brukt kommandoen One-Way.

Forutsetningen for å bruke enveisvariansanalyse er som nevnt tidligere at populasjonene er normalfordelte med gjennomsnitt $\mu_1, \mu_2, \mu_3, ..., \mu_k$ og med ukjent varians (men lik varians) σ^2 i alle populasjonene. Se på i hvilken utstrekning du mener dette er oppfyllt. (Vink: Se på utvalgsvariansene)

Hva betyr det at 95% konfidensintervallene er individuelle? Beregn ved hjelp av andre kommandoer i MINITAB et av disse konfidensintervallene og se om du får samme svar.

Hvis det hadde vært slik at det også hadde vært 6 forskjellige sjafører involvert kunne en ha gjennomført en såkalt 2-veis variansanalyse restvariansen *MSE* ville nå blitt mindre fordi en nå får to *SSE*- summer som i sum er mindre enn den opprinnelige *SSE*. Vi skal ikke komme inn på dette i denne boka, men kun gjøre oppmerksom på at hvis man har tilleggsopplysninger om variabel til så bør denne være med i analysen i dette blir en bedre analyse (tar bedre vare på tallmaterialet)

17. Regresjon og variansanalyse.

På side 30 så vi på den enkle regresjonsmodellen

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

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. $X = x$. 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 først ønsker å teste

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

mot

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

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 MINITAB finner vi nå (se dataene side 36) ved hjelp av kommandoene

STAT
REGRESSION
REGRESSION
Responce C2 (y-verdiene), Predictors C1 (x-verdiene)
OK

Regression Analysis: C2 versus C1

```
The regression equation is
C2 = - 67,5 + 0,816 C1


Predictor     Coef  SE Coef      T      P
Constant   -67,48    44,77  -1,51  0,170
C1         0,8162   0,2558   3,19  0,013


S = 7,38448   R-Sq = 56,0%   R-Sq(adj) = 50,5%
```

Herav ser en at t-verdien blir 3,19 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 MINITAB tester tosidig, dvs.

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

og dermed blir P-verdien (pga. symmetri)

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

Legg også merke til at man får en såkalt variansanalyse utskrift (uansett om man ønsker det eller ikke). Dette er en annen måte å gjennomføre regresjonsanalyse på. Legg imidlertid merke til at P-verdien er den samme som under regresjonsanalysen.

Nå kan det vises (se side 42) hvis man i hvert eneste punkt kvadrerer og summerer avvikene over at

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

dvs. at

Total variasjon = Forklart variasjon + Uforklart variasjon

eller at

$$SST = SSR + SSE$$

Bruker en tallmaterialet side 36 finner en

x_i	160	165	189	187	182	168	181	170	174	172
y_i	70	62	91	82	75	59	83	67	78	85
$\hat{y}_i =$ -67,5+0,816x	63,1	67,1	86,7	85,1	81,0	69,6	80,2	71,2	74,5	72,9

og herav

$$SST = \sum_i (y_i - \bar{y})^2 = \left[\sum_i y_i^2 - n\bar{y}^2 \right] = \left[57542 - 10 \cdot (\frac{752}{10})^2 \right] = 110{,}177... = 991{,}59$$

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2 = [\ (63{,}1\text{-}75{,}2)^2 + \ldots\ldots + (72{,}9\text{-}75{,}2)^2\ ] = 555{,}10$$

$$SSE = \sum_i (y_i - \hat{y}_i)^2 = [\ (70\text{-}63{,}1)^2 + \ldots\ldots + (85\text{-}72{,}9)^2\ ] = 436{,}28$$

En ser at en får bekreftet påstanden om at

$$SST = SSR + SSE$$

Den lille forskjellen mellom venstresiden og høyresiden skyldes avrundinger. (SST = 991,51 og $SSR + SSE = 991{,}38$)

Bruker en så MINITAB og bruker Variansanalyse til å gjennomføre regresjonsanalyse (mao. en bruker variansene i regresjon til å danne F) får en (se siste del av MINITAB – utskriften)

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	555,36	555,36	10,18	0,013
Residual Error	8	436,24	54,53		
Total	9	991,60			

Mao. nøyaktig samme resultat som ved t-testen over (P-verdi = 0,013). Legg merke til at $t^2 = 3{,}19^2 = 10{,}18 = F$. Dette er et resultat som kan vises helt generelt.

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 . Oftest er denne verdien lik 0, fordi man ønsker å teste om det er noen sammenheng mellom variablene X og Y

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

mot

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

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 29, og legger dette inn i MINITAB og bruker kommandoene

STAT
BASIC STATISTICS
CORRELATION

og legger inn tallene, gir dette følgende utskrift:

Correlations: C1; C2

Pearson correlation of C1 and C2 = 0,748
P-Value = 0,013

Det betyr mao. at korrelasjonskoeffesienten mellom X og Y er 0,748, som vi har funnet tidligere, og at testing av

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

mot

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

gir en signifikanssannsynlighet på 0,013 som er presis samme signifikanssannsynlighet som den man fant på side … hvor man testet

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

mot

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

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 ? og ?)

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

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. $(s_x > 0 \; og \; s_y > 0)$

18. Multippel regresjon.

Foran i heftet (side 30) betraktet vi den enkle regresjonsmodellen

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

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. $X = x$ 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. I denne situasjonen er det variabelen X (alene) som skal forklare variasjonen i Y. I de fleste situasjoner (i virkelighetens verden) vil det som regel være flere forklaringsvariable knyttet til en variabel.

Vi skal nå anta at vi har k forklaringsvariable $X_1, X_2, \ldots, X_k$ slik at den statistiske modellen for multippel regresjon nå blir

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots \beta_k x_{ik} + \varepsilon_i, \;\; i = 1, 2, \ldots, n$$

der $\varepsilon_i \sim N(0, \sigma)$ for $i = 1, 2, \ldots, n$ antas å være uavhengige. Det betyr at

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

som igjen kalles for populasjonsregresjonslikningen. Denne estimeres så ved hjelp av den såkalte utvalgsregresjonslikningen

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

som igjen beregnes ved minste kvadraters metode ved hjelp av et utvalg av nk observasjoner , dvs det er tatt n observasjoner i hver av de k underpopulasjonene: X_1 er knyttet til populasjon 1, X_2 er knyttet til poulasjon 2,…., og X_k er knyttet til populasjon k. De (k +1) ukjente estimatene $b_0, b_1, b_2, \cdots, b_k$ finner en som tidligere ved å minimere de kvadrerte avvikene

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

Dette leder til de såkalte (k+1) normallikningene:

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

der $\sum_i x_1 = \sum_i x_{i1} =$ summen av de *n* observasjonene fra populasjon 1 , $\sum_i y = \sum_i y_i =$ summene de *n* observerte *y*- verdiene, osv……., $\sum_i x_k y = \sum_i x_{ik} y_i =$ produktsummen av de *n* observasjonene fra populasjon *k* og de *n* observerte *y*- verdiene.
Eks. Anta at man ønsker å sjekke sammenhengen mellom variablene pris (=*Y*), kilometerstand (= X_1) og alder (= X_2) på bilmerket Toyazda MZW i et bestemt distrikt på Østlandet. Anta at man tar et tilfeldig utvalg på *n* = 10 biler og finner følgende:

Km-stand x_1 (i 1000 km)	Alder x_2 (i mnd)	Pris y (i 100000kr.)
17	14	2,7
23	28	2,5
45	50	2,1
28	18	2,6
67	63	2,0
9	10	2,8
52	36	2,3
20	12	2,7
31	25	2,4
40	30	2,4

Vi skal nå gjøre bruk av den lineære regresjonsmodellen

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

Det vil egentlig si at

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \ i = 1, 2, ..., 10$$

Estimat for $\beta_0, \beta_1 \ og \ \beta_2$ gitt ved henholdsvis $b_0, b_1 \ og \ b_2$ gir dermed den estimerte regresjonslikningen

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

Likningsystemet over antar nå formen

$$\begin{aligned} b_0 \cdot n \ &+ \ b_1 \sum_i x_1 \ + \ b_2 \sum_i x_2 \ &= \sum_i y \\ b_0 \sum_i x_1 + b_1 \sum_i x_1^2 \ &+ b_2 \sum_i x_1 x_2 \ &= \sum_i x_1 y \\ b_0 \sum_i x_2 + b_1 \sum_i x_2 x_1 &+ b_2 \sum_i x_2^2 \ &= \sum_i x_2 y \end{aligned}$$

Vi må mao. bestemme 8 summer (hvorfor ikke flere?) , deretter sette opp de 3 likningene med de 3 ukjente og så løse disse.

Nå finner en (legg inn de tre kolonnene i liste 1, 2 og 3 på kalkulatoren)

$$\sum_i x_1 = 17 + 23 + \cdots + 31 + 40 = 332$$
$$\sum_i x_1^2 = 17^2 + 23^2 + \cdots + 31^2 + 40^2 = 13864$$
$$\sum_i x_2 = 14 + 28 + \cdots + 25 + 30 = 286$$
$$\sum_i x_2^2 = 14^2 + 28^2 + \cdots + 25^2 + 30^2 = 10834$$
$$\sum_i x_1 x_2 = 17 \cdot 14 + 23 \cdot 28 + \cdots + 31 \cdot 25 + 40 \cdot 30 = 12034$$
$$\sum_i y = 2{,}7 + 2{,}5 + \cdots + 2{,}4 + 2{,}4 = 24{,}5$$
$$\sum_i x_1 y = 17 \cdot 2{,}7 + 23 \cdot 2{,}5 + \cdots + 31 \cdot 2{,}4 + 40 \cdot 2{,}4 = 773{,}9$$
$$\sum_i x_2 y = 14 \cdot 2{,}7 + 28 \cdot 2{,}5 + \cdots + 25 \cdot 2{,}4 + 30 \cdot 2{,}4 = 660{,}8$$

Med $n = 10$ finner en da følgende likningssystem:

$$\begin{aligned} 10b_0 + 332b_1 + 286b_2 &= 24{,}5 \\ 332b_0 + 13864b_1 + 12034b_2 &= 773{,}9 \\ 286b_0 + 12034b_1 + 10834b_2 &= 660{,}8 \end{aligned}$$

Løser en dette ved hjelp av kalkulatoren eller ved papir og blyant finner en

$$b_0 = 2{,}8988,\ b_1 = -0{,}0032 \text{ og } b_2 = -0{,}1194$$

For de som har tatt kurset i lineæralgebra har en

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} 10 & 332 & 286 \\ 332 & 13864 & 12034 \\ 286 & 12034 & 10834 \end{bmatrix}^{-1} \cdot \begin{bmatrix} 24{,}5 \\ 773{,}9 \\ 660{,}8 \end{bmatrix} = \begin{bmatrix} 2{,}8988 \\ -0{,}0032 \\ -0{,}0119 \end{bmatrix}$$

Det betyr mao. at den estimerte regresjonslikningen (eller utvalgsregresjonslikningen) blir

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 = 2{,}8988 - 0{,}0032 x_1 - 0{,}0119 x_2$$

For de som ikke har dette kurset må de selv velge hvilken måte de ønsker å løse likningssystemet på. Det greieste er kanskje gjennom innsettingsmetoden å skaffe seg 2 likninger med 2 ukjente og derfra løse disse enten med innsetingsmetoden en gang til eller bruke addisjonsmetoden.

Det betyr for eksempel at en Toyazda MZW som har kjørt 0 km og er 0 mnd. gammel skal koste 289880 kroner (Nå er ny pris på denne modellen 290000 kroner. Hvordan vil du forklare at vi nå sier at den skal koste 289880 kroner?)
Hvis den har kjørt 25000 km og er 18 mnd gammel skal den koste (estimert pris)

$$\hat{y} = 2{,}8988 - 0{,}0032 \cdot 25 - 0{,}0119 \cdot 18 = 2{,}6046$$

mao. den vil koste 260460 kroner i følge modellen.

Bruker en nå MINITAB med kommandoene

Stat
Regression
Regression

og så legger inn dataene får en følgende utskrift:

Regression Analysis: Pris versus Km; Alder

The regression equation is
Pris = 2,90 - 0,00336 Km - 0,0118 Alder

Predictor	Coef	SE Coef	T	P
Constant	2,89905	0,03880	74,72	0,000
Km	-0,003361	0,002717	-1,24	0,256
Alder	-0,011799	0,002808	-4,20	0,004

S = 0,0553563 R-Sq = 96,6% R-Sq(adj) = 95,6%

En ser at dette stemmer ganske bra med beregningene over. Vi skal nå se på de øvrige utregningene i MINITAB og hva de brukes til.

I modellen over har vi sagt at feileddene $\varepsilon_i \sim N(0,\sigma)$. Mao. at feileddene er normalfordelte med en forventning på 0 og et standardavvik på σ. Det kan vises at σ^2 kan estimeres med s^2 der s^2 er gitt ved

$$s^2 = \frac{1}{n-k-1}\sum_i e_i^2 = \frac{1}{n-k-1}\sum_i (y_i - \hat{y}_i)^2$$

Bruker en nå dataene i tabellen over finner en

Km-stand x_1 (i 1000 km)	Alder x_2 (i mnd)	Pris y (i 100000kr)	Estimert verdi $\hat{y}$	Estimert avvik e_i
17	14	2,7	2,68	0,02
23	28	2,5	2,49	0,01
45	50	2,1	2,16	-0,04
28	18	2,6	2,59	0,01
67	63	2,0	1,93	0,07
9	10	2,8	2,75	0,05
52	36	2,3	2,30	0
20	12	2,7	2,69	0,01
31	25	2,4	2,50	-0,10
40	30	2,4	2,41	-0,01

Herav finner en så

$$s^2 = \frac{1}{n-k-1}\sum_i e_i^2 = \frac{1}{10-2-1}(0{,}02^2 + 0{,}01^2 + \cdots (-0{,}01)^2 + (-0{,}01)^2) = \frac{0{,}0218}{7} = 0{,}00311$$

og dermed tilslutt

$$s = 0{,}0558$$

som stemmer meget bra med MINITAB-utskriften hvor $s = 0{,}0553563$.

SE Coef som angir standardfeilen (avviket) til de estimerte koeffesientene b_0, b_1 og b_2 er en del vanskeligere å regne ut så vi skal ikke komme inn på det her.

For de som allikevel ønsker å komme til bunns i dette kan f.eks se etter i John E. Freunds bok: Mathematical Statistics with applications (7. ed) side 465. foklaringen her bygger en del på kunnskaper i lineær algebra.

t-verdiene finner en av

$$t = \frac{b_j - b_j^0}{SEb_j}$$

der b_j^0 er den verdien av parameteren som en tester ved nullhypotesen. Det mest vanlige er å teste

$$H_0 : \beta_j = b_j^0 = 0$$

og det er det som er gjort i MINITAB-analysen. Det betyr mao. at

$$t = \frac{b_j}{SEb_j}$$

Ser en nå på tallene tabellen over finner en

$$t_{Const} = \frac{2{,}89905}{0{,}03880} = 74{,}72$$

$$t_{Km} = \frac{-0{,}003361}{0{,}002717} = -1{,}24$$

$$t_{Alder} = \frac{-0{,}011799}{0{,}002808} = -4{,}20$$

som stemmer med tallene i tabellen. En finner da følgende P-verdier (ved hjelp av kalulatoren) når man tester

$$H_0 : \beta_j = b_j^0 = 0 \quad \text{mot} \quad H_1 : \beta_j \neq 0$$

For konstanten : $2P(t \geq 74{,}72) = 2{,}120 \cdot 10^{-11} = 0{,}000$

For antall kjørte km: $2P(t \leq -1{,}24) = 0{,}255$

For alder: $2P(t \leq -4{,}24) = 0{,}004$

Her er alle sannsynlighetene regnet ut ved hjelp av kalkulatoren med

$df = n\text{-}k\text{-}1 = 10\text{-}2\text{-}1 = 7$ frihetsgrader.

Kontroller selv at beregningene stemmer. En ser mao. at nullhypotese 1 og 3 forkastes, mens nullhypotese 2 ikke kan forkastes.

Hvis en isteden velger å gjennomføre variansanlyse i den multiple regresjonssituasjonen får en ved hjelp av MINITAB

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	2	0,60355	0,30177	98,48	0,000
Residual Error	7	0,02145	0,00306		
Total	9	0,62500			

Den generelle ANOVA-tabellen i multippel regresjon er som følger:

Kilde	Antall frihetsgrader *df*	Sum av kvadrater *SS*	Gj.sn. kvadrat summer *MS*	*F*
Modell (regresjon)	*p*	$SSM=\sum_i(\hat{y}_i-\bar{y})^2$	$MSM=\frac{SSM}{p}$	$\frac{MSM}{MSE}$
Feil (residualfeil)	*n-p*-1	$SSE=\sum_i(y_i-\hat{y}_i)^2$	$MSE=\frac{SSE}{n-p-1}$	
Total	*n*-1	$SST=\sum_i(y_i-\bar{y})^2$		

Sjekk selv om verdiene over i MINITAB-utskriften stemmer. Legg merke til at det også her er mulighet til å kontrollere noen av mellomregningene idet følgende gjelder:

1. *df.tot = df.modell + df.feil*
2. *SST = SSM + SSE*

Hva er det så man tester her? Jo i motsetning til over hvor vi har testet

$$H_0 : \beta_j = b_j^0 = 0 \quad \text{mot} \quad H_1 : \beta_j \neq 0$$

for j = 1, 2 og 3. Mao. 3 nullhypoteser testes hver for seg. Dermed gjelder resultatene også hver for seg. Med variansanalyse tester en derimot

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

mot

$$H_1 : Minst\ en\ av \beta_j - ene \neq 0$$

Det vil mao. si i eksempelet over at man tester samtidig

$$H_0 : \beta_1 = \beta_2 = 0$$

mot

$$H_1 : Minst\ en\ av\ \beta_1\ og\ \beta_2 \neq 0$$

Legg merke til at man her kun tester på koeffisientene til prediktorene (forklaringsvariablene) og på koeffisienten til konstantleddet.

19. Oppgaver

Oppgave 1

Beregn gjennomsnittet i tallmaterialene

a) 2, 5, 3, 6, 7, 4, 9, 11, 8, 13
b) 6, 15, 9, 18, 21, 12, 27, 33, 24, 39.
c) Ser du noen sammenheng mellom de to tallmaterialene og deres gjennomsnitt?

Oppgave 2

Beregn gjennomsnittet i tallmaterialene

a) 3, 4, 6, 5, 9, 2, 8, 3, 4, 3
b) 13, 14, 16, 15, 19, 12, 18, 13, 14, 13
c) Ser du noen sammenheng mellom de to tallmaterialene og deres gjennomsnitt?

Oppgave 3

Anta at et tallmateriale er gitt ved $x_i, i = 1, 2, \cdots, n$
Beregn gjennomsnittet i tallmaterialene

a) $y_i = kx_i \quad i = 1, 2, \cdots, n$ der k er en vilkårlig konstant
b) $z_i = x_i + c \quad i = 1, 2, \cdots, n$ der c er en vilkårlig konstant

Oppgave 4

a) Beregn typetallet i oppgave 1.2 a)
b) Beregn typetallet i oppgave 1.1 a)
c) Hva er typetallet i tallmaterialet
3, 4, 3, 8, 9, 5, 6, 5, 9, 11

Oppgave 5

Finn medianen i tallmaterialene

a) 3, 5, 2, 4,18, 6, 9
b) 3, 5, 2, 4,18, 6
c) Finn gjennomsnittet i b) og sammenlikn med medianen. Kommenter kort.

Oppgave 6

Beregn σ^2 i tallmaterialet 2, 3, 6, 4, 7, 3, 8, 9, 1, 5

a) ved hjelp av $\frac{1}{n}(\sum_{i=1}^{n} x_i^2) - \bar{x}^2$ og

b) ved hjelp av $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$

c) Beregn standardavviket σ

Oppgave 7

Beregn s^2 i tallmaterialet 2, 3, 6, 4, 7, 3, 8, 9, 1, 5

a) ved hjelp av $\frac{1}{n-1}(\sum_{i=1}^{n} x_i^2) - \frac{n}{n-1}\bar{x}^2$ og

b) ved hjelp av $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

c) Beregn standardavviket *s*.

d) Legg inn tallene i en liste på kalkulatoren og beregn σ og *s*. Kommenter kort forskjellene.

Oppgave 8

a) Vis at

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

b) Vis at

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

(Vink: $(a-b)^2 = a^2 - 2ab + b^2$, $\sum_i (a_i + b_i) = \sum_i a_i + \sum_i b_i$ og $\sum_i k a_i = k\sum_i a_i$ der *k* er en vilkårlig konstant)

Oppgave 9

Gitt de to tallmaterialene x_i : 2, 4, 5, 1, 5, 8, 7, 9, 3, 8, 9 og y_i :3, 4, 6, 7, 6, 8, 9, 8, 10, 7, 11

a) Beregn variasjonsbredden i de to tallmaterialene.
b) Finn variansene i de to tallmaterialene (betrakt dataene som utvalg)
c) Bestem første, andre og tredje kvartil.
d) Finn tilslutt kvartilbredden i de to tallmaterialene.

Oppgave 10

Gitt de to tallmaterialene x_i : 2, 4, 5, 1, 5, 8, 7, 9, 3, 8, 9 og y_i : 22, 24, 25, 21, 25, 28, 27, 29, 23, 28, 29. (begge utvalg)

a) Beregn variansen i de to tallmaterialene. Kommenter kort resultatene.
b) La tallmaterialet z_i være gitt ved $z_i = \frac{y_i}{10}$. Finn variansen i dette tallmaterialet. Kommenter kort resultatet.

Oppgave 11

a) Vis at hvis $y_i = x_i + k$ der k er en vilkårlig konstant så har tallmaterialene samme varians.
b) Vis at hvis $y_i = kx_i$ så er variansen til y-ene der k er en vilkårlig konstant så er variansen til y-ene, s_y^2, gitt ved

$$s_y^2 = k^2 s_x^2$$

der s_x^2 er variansen til x-ene.

Oppgave 12

Gitt tallmaterialet x_i: 3, 8, 6, 9, 3, 7, 12, 10, 4, 5

a) Finn første og tredje kvartil i tallmaterialet ved regning.
b) Legg tallene inn i en liste på kalkulatoren. Finn så første og tredje kvartil ved hjelp av kalkulatoren.
c) Legg så tallene inn i en kolonne i MINITAB og finn kvartilene.
d) Sammenlikn resultatene i a), b) og c) og kommenter eventuelle forskjeller.

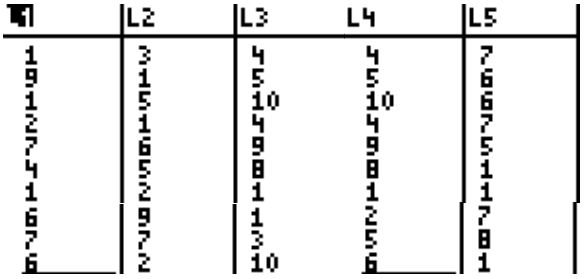
Oppgave 13

Gitt tallmaterialet x_i: 2, 3, 2, 2, 2, 4, 5, 5, 6, 7, 7, 3, 4, 5, 7, 1, 7, 8, 8, 15

a) Finn kvartilene og medianen.
b) Avgjør om observasjonen 15 er en outlier. Hvis så er tilfellet så fjern den fra datamengden.
c) Finn gjennomsnittet på enklest mulig måte.
d) Bestem standardavviket s_x og finn andelen observasjoner som faller innenfor $\bar{x} \pm s_x$

Oppgave 14

Gitt tallmateriallmaterialet x_i , i =1, 2, ….,50 som er framkommet ved å be kalkulatoren lage 50 tilfeldige heltall mellom 1 og 10 (randint(1,10,10) 5 ganger og lagre 10 og 10 tall i liste 1,…, liste 5.



a) Beregn gjennomsnittet og medianen i tallmaterialet. Kommenter kort resultatene.
b) Finn første og tredje kvartil og kvartilbredden
c) Finn 10% og 90% percentilene og herav 10-90 percentilbredden $P_{90} - P_{10}$

Oppgave 15

a) Simuler 250 tilfeldige trekninger fra en normalfordelt populasjon med gjennomsnitt på 180 og standardavvik på 7 ved hjelp av kalkulatoren (randNorm (180, 7, 250). Lagre disse i en tilfeldig liste og bruk så kalkulatoren til å gjennomføre 1-Var Stats på disse dataene

b) Gjør det samme i MINITAB og bruk kommandoen descriptive statistics.

Oppgave 16

Gitt tallmaterialet x_i: 2, 3, 2, 2, 2, 4, 5, 5, 6, 7, 7, 3, 4, 5, 7, 1, 7, 13, 14, 15

a) Beregn 3.ordensmomentet m_3 gitt ved

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

ved hjelp av kalkulatoren. Kommenter kort resultatet.

b) Beregn så g_1 og G_1 gitt ved henholdsvis

$$g_1 = \frac{m_3}{m_3^{3/2}} \quad \text{og} \quad G_1 = \frac{\sqrt{n(n-1)}}{n-2} g_1$$

c) Legg så dataene inn i MINITAB og beregn skjevheten (the skewnes)

Oppgave 17

Simuler et utvalg på 100 normalfordelte observasjoner på kalkulatoren fra en populasjon hvor $\mu = 50$ og $\sigma = 3$.

a) Beregn 4.ordensmomentet m_4 gitt ved

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

ved hjelp av kalkulatoren. Kommenter kort resultatet.

b) Beregn så g_2 og G_2 gitt ved henholdsvis

$$g_2 = \frac{m_4}{m_2^2} - 3 \quad \text{og} \quad G_2 = \frac{n-1}{(n-2)(n-3)}\left[(n+1)g_2 + 6\right]$$

c) Legg så dataene inn i MINITAB og beregn spissheten (the kurtosis) (Legg merke til forskjellen på utrykkene for g_1 og g_2. I det siste utrykket er brøken redusert med 3 for å kunne sammenlikne med normalfodelingen hvor spissheten er nøyaktig lik 3. Det betyr mao. at i en fordeling hvor spissheten er større enn hos normalfordelingen vil g_2 være større enn 0, i en fordeling hvor spissheten er mindre enn normalfordelingen vil g_2 være mindre enn 0.)

d) Hva bør g_1 omtrent være for dataene i denne oppgaven? Beregn g_1 og G_1 ved hjelp av kalkulatoren.

e) Beregn g_2 og G_2 i dataene i oppgave 1.14. Virker resultatene rimelige?

Oppgave 18

Anta at det ble avholdt prøvevalg på en ungdomsskole med 250 elever og at resultatet ble som følger:

Parti	Antall stemmer
Arbeiderpartiet	48
Fremskrittspartiet	52
Høyere	35
Kristelig Folkeparti	13
Rød Valgallianse	18
Senterpartiet	20
Sosialistisk Venstreparti	48
Venstre	16

a) Fremstill dataene grafisk ved hjelp av et stolpediagram (bar chart) og ved hjelp av et kakediagram (pie chart)
b) Sammenlikn grafisk den prosentvise oppslutningen til "regjeringspartiene" (AP, SV og SP) mot de andres.

Oppgave 19

I en klasse ble karakterene i avgangsprøven i norsk som følger:

Karakter	1	2	3	4	5	6
Antall	2	5	8	7	4	1

Fremstill dataene grafisk både ved hjelp av kalkulatoren og ved hjelp av MINITAB

Oppgave 20

En utdanningsstatistikk fra statistisk sentralbyrå for de nordiske landene for 2005* viste følgende antall (mht. høyeste fullførte utdanning)

Utdanning (begge kjønn)	Danmark	Finnland	Island	Norge	Sverige
Grunnskole-utd.	1342585	1293832	78900	941497	1604341
V.g. utd. og påbygging. til v.g.utd.	1658410	1598246	80400	1411322	3280466
Univers.- og høgsk.utd.	889492	1063631	37700	844425	1514181

Lag en passende grafisk framstilling av disse dataene. (* Tallene fra Island er fra 2004)

Oppgave 21

Statistisk årbok fra 2007 viser følgende oversikt over høyeste fullførte utdanning for personer på 16 år og over i perioden 1970-2005 (i prosent):

År	1970	1980	1990	2000	2001	2002	2003	2004	2005
Univ.og høgsk. nivå,kort (≤4år)	5,7	8,8	12,2	17,0	17,4	17,7	18,1	18,6	19,2

Lag en passende grafisk framstilling av disse dataene.

Oppgave 22

Inndel tallmaterialet i oppgave 1.16 i tre like klasser (samme klassevidde)

a) Beregn .det aritmetiske middeltall og medianen i tabellen
b) Finn første og tredje kvartil ved regning og ved hjelp av kalkulatoren
c) Tegn sumfordelingskurven og bruk denne til å lese av alle tre kvarilene.
d) Finn til sist de tre kvartilene ved hjelp av

$$Kvartil = L + \frac{Rest}{f} v$$

Oppgave 23

I en klasse på 28 elever er det 16 gutter. Tegn et kakediagram som illustrerer andelen gutter og andelen jenter i klassen

Oppgaver 24

På en skole er aldersfordelingen blant de 42 lærerne som følger:

23, 36, 56, 54, 60, 58, 45, 65, 67, 54, 24, 34, 37, 50, 28, 44, 41, 43, 59, 38, 39, 49, 51, 57, 34, 56, 50, 48, 50, 45, 63, 64, 65, 59, 58, 67, 39, 36, 35, 60, 53, 51

a) Tegn et stamme-blad diagram som illustrerer aldersfordelingen.
b) La MINITAB tegne et stamme-blad diagram
c) Tegn et histogram med klassevidder på 10 år.

Oppgave 25

I følge statistisk sentralbyrå så fordelte antall voksenopplæringskurs i perioden 1998 -2006 seg som følger:

År	1998	1999	2000	2001	2002	2003	2004	2005	2006
Kurs i alt	61428	61319	60326	54530	52181	50083	46925	42800	39088

Framstill tallmaterialet grafisk.

Oppgave 26

I følge statistisk sentralbyrå så var antall studenter i høyere utdanning i utlandet i perioden 1995-2005 som følger:

År	1995	2000	2005
Antall stud.	8596	13928	12343
Andel kvinner	51,1	56,7	57,6

a) Tegn et kurvediagram som viser utviklingen på antall kvinner og et som viser utviklingen på antall menn i den aktuelle perioden.
b) Tegn 3 kvadrater som viser utviklingen av andelen kvinner i høyere utdanning i utlandet. Velg side = 2 cm i 1995.

Oppgave 27

På en hukommelsesprøve ble det lest opp 25 ord som elevene skulle skrive ned flest mulig av etter at opplesningen var ferdig. Resultatet blant de 15 elevene ble som følger:

x_i: 12, 23, 10, 8, 4,13, 14,15, 9, 12, 11, 10, 11, 12, 7

a) Finn $x_{\min}, Q_1, M, Q_1$ *og* x_{maks} og boksplottet.
b) Legg dataene inn i en liste på kalkulatoren og tegn så det vanlige boksplottet og det modifiserte boksplottet.
c) Avgjør ved regning om verdien 23 er en outlier eller ikke. Sammenlikn med det du finner pkt b)
d) Finn $\bar{x}$ med og uten verdien 23.

Oppgave 28

Gitt tallmaterialet

$$x_i: 32, 35, 43, 39, 91.$$

Avgjør både ved regning og ved en grafisk framstilling om 91 er en outlier. Finn gjennomsnittet og medianen med og uten 91.

Oppgave 29

Undersøk om de 10 observasjonene

100,95; 99,64; 99,77; 103,94; 96,14;
99,11; 104,66; 100,85; 105,29; 108,14

kan være normalfordelt (eller ekvivalent kan sies å være et tilfeldig utvalg fra en normalfordelt populasjon) ved å tegne et normal kvantilplot.

Oppgave 30

Legg inn de 15 observasjonene gitt under på kalkulatoren og avgjør om de kan sies å være normalfordelt. Bruk kalkulatoren optimalt.

9, 8, 10, 15, 10, 10, 11, 8, 7, 11, 8, 9, 5,10, 11

Oppgave 31

a) Simuler 250 normalfordelte observasjoner med $\mu = 50$ *og* $\sigma = 3$ på kalkulatoren. Lagre observasjonene i en liste og tegn et sannsynlighetsplot.

b) Gjør så det samme i MINITAB og gjennomfør en normalitytest.(Vink: Velg Anderson-Darling)

Oppgave 32

A. Anta at man har målt høyden *X* til 200 tilfeldig valgte kvinnlige: studenter og at man fant:

Høyde (i cm)	[155,160)	[155,160)	[155,160)	[155,160)	[155,160)	[155,160)	[155,160)
Frekvens	6	20	42	60	44	18	10

a) Framstill dataene grafisk og finn $\bar{x}$.

b) Finn standardavviket s_x og andelen som faller innenfor intervallet $[\bar{x} - s_x, \bar{x} + s_x]$ ved regning.

c) Finn de relative frekvensene r_k og deretter de kumulerte relative frekvensene R_k. Tegn sumfordelingskurven og bruk denne til å kontrollere beregningen av andelen i b).

B. Man ønsker nå å kontrollere om dataene følger en normalfordeling med $\mu = 172{,}5 \; og \; \sigma = 7{,}0$. La variabelen $Y - N(172{,}5; 7{,}0)$.

a) Vis at da er $p_1 = P(Y < 160) = 0{,}0371$ og $p_2 = P(160 \le Y < 165) = 0{,}1049$. Beregn deretter $p_3 = P(165 \le Y < 170)$ og $p_4 = P(170 \le Y < 175)$. I resten av oppgaven kan du bruke (Skal **ikke** vises) at $p_5 = P(175 \le Y < 180) = 0{,}2185$, $p_6 = P(180 \le Y < 185) = 0{,}1049$ og $p_7 = P(Y \ge 185) = 0{,}0371$

c) La nå X = høyden på en tilfeldig valgt kvinnlig student. Test da med utgangspunkt i dataene i pkt. A om du vil forkaste

H_0 : X er normalfordelt med $\mu = 172{,}5 \; og \; \sigma = 7{,}0$

til fordel for

H_1 : X er <u>ikke</u> normalfordelt med $\mu = 172{,}5 \; og \; \sigma = 7{,}0$

Velg nivå på 5%.

d) Gjennomfør testingen ved en normalitetstest i MINITAB (ikke gitt til eksamen)

Oppgave 33

En kontrollprøve i matematikk gav følgende resultat

Poeng x_k	8-12	13-17	18-22	23-27	28-32	33-37	38-42	43-47	48-52
Frek-vens f_k	1	1	3	6	7	13	10	6	3

(60 poeng er maksimal poengsumm)

a) Framstill dataene grafisk.

b) Beregn det aritmetiske middeltallet $\bar{x}$ og medianen *M*. Kommenter kort eventuelle forskjeller.

c) Framstill sumfordelingskurven grafisk og finn andelen elever som faller mellom $\bar{x}$ og M både ved regning og ved avlesning.

d) Beregn standardavviket s_x.

Et mål på skjevheten (the skewness) er gitt ved

$$a_3 = \frac{\frac{1}{n}\sum_k (x_k - \bar{x})^2 f_k}{s_x^3}$$

(I en symmetrisk fordeling er $a_3 = 0$, dvs det er ingen skjevhet)

e) Beregn skjevheten i dette tallmaterialet.

Oppgave 34

Anta en har følgende 7 parobservasjoner $(x_i, y_i) \quad i = 1,2,...,7$

x	1	3	3	5	8	10	12
Y	3	5	6	7	10	13	16

a) Tegn spredningsdiagrammet.

b) Avsett punktet $(\bar{x}, \bar{y})$ i spredningsdiagrammet og tegn den rette linja som "passer best i punksvermen"

c) Finn likningen for denne rette linja. Denne linja er den såkalte regresjonslinja for *y* mhp. *x* funnet ved grafisk metode.

d) Anta lineær regresjonsmodell

$$y = \alpha + \beta x + \varepsilon \quad der\ \varepsilon\ er\ N(0, \sigma)$$

og sett så opp normallikningene til bestemmelse av *a* og *b* i den såkalte estimerte regresjonslikning

$$\hat{y} = a + bx$$

e) Løs normallikningene mhp. *a* og *b*. Tegn så denne minste kvadraters regresjonslikning inn i samme koordinatsystem som over.

f) Kontroller både ved hjelp av MINITAB og kalkulatoren at dine beregninger over stemmer.

g) Tegn residualplottet og sjekk om forutsetningen i regresjonsmodellen kan sies å være oppfylt eller ikke.

Oppgave 35

Ved en skole er det tilfeldig trukket ut 10 barn som alle har vært igjennom en matematikktest og en lesetest. Resultatet av undersøkelsen ble:

Barn	1	2	3	4	5	6	7	8	9	10
Poeng matematikktest x	5	13	23	20	36	40	54	60	29	45
Poeng lesetest y	8	15	22	31	40	45	50	55	32	45

a) Tegn spredningsdiagrammet og se om det er noen tendens i tallmaterialet.

b) Beregn $\bar{x}, \bar{y}, s_x$ *og* s_{xy}. Finn så minste kvadraters regresjonslinje og tegn den inn i spredningsdiagrammet. Estimer antall poeng en elev vil få på lesetesten gitt at vedkommende fikk 50 poeng på matematikktesten

c) Finn så s_y og bruk denne sammen med beregningene i punkt b) til å finne korrelasjonskoeffisienten mellom x og y. kommenter kort resultatet i lys av punkt a).

d) Bruk både kalkulatoren og MINITAB til å finne korrelasjons koeffisienten.

e) Skisser spredningsdiagrammet.

Oppgave 36

Anta en har følgende 8 parobservasjoner $(x_i, y_i) \quad i = 1,2,...,8$

x	1	3	3	5	9	10	12	13
y	3	5	6	7	6	5	4	2

a) Tegn spredningsdiagrammet.

b) Avsett punktet $(\bar{x}, \bar{y})$ i spredningsdiagrammet og tegn den kurven som "passer best i punksvermen". Angi den modellen du finner rimelig å bruke (foreslå et funksjonsuttrykk, $y = f(x)$, som du vil bruke)

c) Sett opp det antall likninger som du finner rimelig og løs så dette likningssettet (Vink: 3 likninger med 3 ukjente)

d) Bruk så kalkulatoren og MINITAB til å finne det estimerte regresjonsutrykket. Tegn inn denne kurven sammen med den kurven du tegnet i b) . Bruk både den kurven du fant i b) og den kurven du fant i c) til å estimere verdien av y når x er 7. Var du god til skissere i b)?

Oppgave 37

a) Finn totalvariasjonen, den forklarte variasjonen og den uforklarte variasjonen i oppgave 2.2.

b) Vis at totalvariasjonen er lik den forklarte variasjonen pluss den uforklarte variasjonen. Finn r^2 (the coeffesient of determination) og herav r.

Oppgave 38

Anta en har følgende 6 parobservasjoner $(x_i, y_i) \quad i = 1,2,...,6$

x	2	5	7	9	12	13
y	4	8	11	16	21	22

a) Sett opp normallikningene og bruk disse til å finne likningen for regresjonslikningen for y mhp. x. Tegn denne linja i et koordinatsystem.

b) Bruk kalkulatoren til å finne totalavviket, det forklarte avviket og det uforklarte avviket i alle punktene. Tegn inn disse avvikene for to av punktene.

c) Finn r^2. Hvordan vil du tolke denne verdien? Hva blir korrelasjonskoeffisienten mellom x og y ?

d) I multippel regresjon (som vi kommer tilbake til senere) skal bruke at den såkalte multiple korrelasjonskoeffisienten (ved mer enn to variable involvert) er det samme som korrelasjonskoeffesienten mellom y og $\hat{y}$. Sjekk om dette også er tilfellet i enkel regresjon.

Oppgave 39

Har vi blitt klokere og klokere? En undersøkelse basert på IQ-målinger av menn på sesjon i perioden 1960-1980 viste følgende sammenheng mellom x (tid) og y (IQ):

x (tid)	1960	1965	1970	1975	1980
y (IQ)	100,2	102,0	103,8	105,6	107,1

a) Framstill dataene i et spredningsdiagram og sett opp normallikningene til bestemmelse av regresjonslikningen for y mhp. x. $\hat{y} = a + bx$, ved å sette $x = 1$ i 1960, $x = 2$ i 1965, …, og $x = 5$ i 1980.

b) Bestem regresjonslikningen ved å løse normallikningene mhp. a og b. Angi et estimat intelligenskvotienten i år 2003.

Anta at populasjonsregresjonslikningen er gitt ved

$$\mu_y = \alpha + \beta x$$

c) Test da

$$H_0 : \beta = 1{,}7 \text{ mot } H_A : \beta > 1{,}7$$

på 5%-nivået.

Nå viste målingene fra 1985 og fram til år 2000 følgende sammenheng mellom x og y :

x (tid)	1985	1990	1995	2000
y (IQ)	108,0	108,8	109,8	109,9

d) Skisser nå spredningsdiagrammet for hele perioden 1960-2000. Bestem så ved hjelp av kalkulatoren en funksjon $\hat{y} = f(x)$ som best mulig beskriver det totale tendensen fra 1960 til 2000 og bruk denne til å estimere intelligenskvotienten (for unge menn på sesjon i Danmark) i år 2003.

Oppgave 40

I følge Statens Institutt for Folkehelse i Norge så utviklet influensaliknende sykdommer seg på følgende måte i siste del av 1995:

Uke nr.x	41	42	43	44	45
Antall rapporterte tilfeller y	2249	2569	2402	2877	3433

a) Tegn et spredningsdiagram for den gitte situasjonen og bestem likningen for regresjonslinjen for y med hensyn på x. Hvor mange rapporterte tilfeller vil du anslå for uke 50?

Du kan her bruke at $\sum x = 215$, $\sum y = 13530$, $\sum xy = 584466$ og $\sum x^2 = 9255$.

Nå utviklet influensaen seg videre som følger:

Uke nr. x	46	47	48	49
Antall rapporterte tilfeller y	4391	6743	10219	12828

b) Tegn et nytt spredningsdiagram for alle 9 ukene og vurder nå modellen i a).

Det blir nå foreslått å tilpasse en ny modell,

$$\hat{y} = A \cdot B^x \qquad (\text{dvs. } \hat{y} = (lnA) + (lnB)x = a + bx),$$

til dette tallmaterialet fra og med uke 41 til og med uke 49. Bestem nå a og b ved minste kvadraters metode og angi så et estimat for antall rapporterte tilfeller i uke 50 og i uke 2 (1996)

c) Finn så tilslutt A og B og sett opp den endelige modellen.

Nå ble det registrert 15684 tilfeller i uke 50 og 7342 i uke 2. sammenlign med dine estimater og kommenter kort.

Oppgave 41

For å undersøke om det er noen sammenheng mellom score (= poengsum) til eksamen i kvantitative metoder og antall timer man hadde studert (utenom forelesningene) en vanlig uke, intervjuet man 10 tilfeldig valgte studenter m.h.t.arbeidsvaner og fant blant annet:

Ant. timer man hadde studert pr. uke x	**Score y (maksimal poengsum=100)**
10	62
12	73
18	89
5	30
4	25
21	95
8	61
11	65
13	70
7	49

a) Tegn spredningsdiagrammet og sett opp normallikningene til bestemmelse av den estimerte regresjonslikningen for y mhp. x, $\hat{y} = a + bx$

b) Løs disse likningene mhp. a og b ved hjelp av kalkulatoren. Forklar kort framgangsmåten .

c) Beregn korrelasjonskoeffesienten mellom x og y . Vis de nødvendige mellomregningene.

d) Test

$$H_0 : \beta = 0 \text{ mot } H_A : \beta > 0$$

på 5%-nivået (β er stigningstallet for populasjonsregresjonslikningen $\alpha + \beta x$)

En MINITAB-utskrift av regresjonsanalysen viser:

Regression Analysis: Score versus Antall timer

```
The regression equation is
Score = 18,6 + 3,97 Antall timer


Predictor       Coef  SE Coef     T      P
Constant      18,617    5,223  3,56  0,007
Antall timer  3,9709   0,4333  9,16  0,000


S = 7,05227   R-Sq = 91,3%   R-Sq(adj) = 90,2%
```

Stemmer dine beregninger så langt med dette? Kommenter kort.

e) Angi et 95% konfidensintervall for β .

Test tilslutt

$$H_0 : \rho = 0 \quad \text{mot} \quad H_0 : \rho > 0$$

ved å angi P-verdien. (ρ er populasjonskorrelasjonskoeffisieneten)

Oppgave 42

Er det noen sammenheng mellom når på året man er født og hvorledes man scorer på en skrivetest? (maksimum score er 80) et tilfeldig utvalg på 12 elever fra samme klassetrinn viste følgende resultat:

Mnd. x	1	2	3	4	5	6	7	8	9	10	11	12
Score Y	71	66	72	70	68	61	65	67	59	63	59	52

a) En regresjonsanalyse med MINITAB viser:

Regression Analysis: Score versus Mnd

The regression equation is
Score = 73,3 - 1,36 Mnd

Predictor	Coef	SE Coef	T	P
Constant	73,258	2,100	34,89	0,000
Mnd	-1,3601	*	-4,77	*

S = * R-Sq = 69,5% R-Sq(adj) = 66,4%

Her er verdiene av SE Coef og P fjernet for prediktoren Mnd. Dessuten mangler verdien av s (Se *). Finn disse 3 verdiene ved å bruke kalkulatoren optimalt. Hvordan vil du tolke den verdien du finner for P? Forklar kort hvilke uttrykk som ligger til grunn for dine beregninger og skriv opp noen av leddene som inngår.

b) Den samme regresjonsanalysen (i MINITAB) viser også følgende variansanalyse:

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	*	264,55	22,73	*
Residual Error	10	*	*		
Total	11	380,92			

Her er to av SS-verdiene, en av MS-verdiene og P fjernet (Se *). Finn disse 4 verdiene ved å bruke kalkulatoren optimalt. Hvordan vil du tolke denne P-verdien du finner?.Gå forøvrig fram som i pkt a).

Oppgave 43

Dødeligheten y (antall dødsfall pr. 100000) av hjerteinfarkt blant menn i Norge i perioden 1945-1965 utviklet seg som følger:

Tid x	1945	1950	1955	1960	1965
Dødelighet y	500	550	625	740	780

a) Framstill dataene grafisk og tegn en rett linje gjennom punktet $(\bar{x}, \bar{y})$ slik at denne "passer best mulig" til dataene.

b) Bestem regresjonslinjen $\hat{y} = ax + b$ ved minste kvadraters metode. Anslå et estimat for dødeligheten i år 2000.

I perioden 1970-1990 var dødeligheten som følger:

Tid x	1970	1975	1980	1985	1990
Dødelighet y	800	760	720	700	675

c) Bestem nå en regresjonslikning $\hat{y} = f(x)$ for hele perioden 1945-1990 og bruk denne til å anslå dødligheten i år 2000. Sammenlign med resultateti b). Bruk kalkulatoren til å bestemme $f(x)$, men foklar kort hva du gjør.

Oppgave 44

En undersøkelse foretatt av NHO i perioden 1987-1993 viste følgende sammenheng mellom sykefraværet y (i %) for kvinner (arbeidere) og tiden x:

År	Sykefraværet i %
1987	13,2
1988	13,1
1989	12,0
1990	11,7
1991	11,0
1992	10,3
1993	10,1

a) Tegn et spredningsdiagram for tallene og bestem et estimat for regresjonslikningen for y med hensyn på x ved minste kvadraters metode, dvs. bestem $\hat{y} = ax + b$. Estimer sykefraværet i år 2000.

Nå var sykefraværet i perioden 1983-1986 og 1994-1996 som følger for den samme gruppen:

År	Sykefraværet i %
1983	11,7
1984	12,3
1985	13,2
1986	13,4
1994	9,3
1995	9,6
1996	10,0

b) Tegn nå et nytt spredningsdiagram for all 14 årene og vurder modellen i a) samt estimatet for år 2000.

d) Foreslå en ny modell på bakgrunn av alle dataene over og anslå nå sykefraværet i år 2000. Forklar kort hvorledes du går fram. (Vink: Du skal nå finne et eksplisitt funksjonsuttrykk som er best mulig tilpasset alle dataene 1983-1996. Kalkulatoren kan nå brukes optimalt, dvs at duikke trenger å vise noen mellomregning)

Oppgave 45

Den tiden som trengs for at en bil skal stoppe etter at sjåføren har oppdaget en situasjon som kan bli farlig er summen av reaksjonstiden (tiden som går før vedkommende tråkker på bremsepedalen)og tiden som vedkommende bremser. Anta at det er gjort 9 målinger på en lukket bane med samme bil og samme sjafør. Resultatet av førsøkene ble:

Hastighet v (i km/t)	30	40	50	60	70	80	90	100
Stopplengde s (i meter)	15	23	38	53	72	102	149	201

a) Tegn spredningsdiagrammet og tegn en kurve som passer godt til tallmaterialet. Angi en ikkelineær modell.

b) Bruk så kalkulatoeren optimalt til å bestemme utrykket for aktuelle modellen $\hat{y} = f(x)$

c) Bruk modellen til å estimere stopplengden hvis bilen skulle kjøre i 85 km/t og hvis den skulle kjøre i 105 km/t

d) Finn ved regning likningen for den kurven du har skissert i a) (Vink: Lag 3 likninger med 3 ukjente)

Oppgave 46

Målinger utført av Nasjonalt folkehelseinsttutt i perioden 1976-2001 ppå 40-42 år gamle menn i Oppland viste følgende sammenheng mellom tiden x og %-andelen y med kraftig overvekt (Bodymassindex = BMI ≥ 30):

x	1976	1981	1986	1991	1994	1997	2001
y	7,1	7,2	8,4	8,5	11,0	15,8	18,0

a) Tegn spredningsdiagrammet og finn regresjonslikningen $\hat{y} = b_0 + b_1 x$. Vis de nødvendige mellomregningene.

b) Test

$$H_0 : \beta_1 = 0 \quad \text{mot} \quad H_1 : \beta_1 \neq 0$$

i modellen $y_i = \beta_0 + \beta_1 x + \varepsilon_i$. Bruk $\alpha = 0{,}05$.

c) Bestem residualene y_{res} og tegn residualplottet. Bruk dette til vurdere modellen. Bruk kalkulatoren til å finne en bedre regresjonslikning (ikkelineær) enn den du fant i a).

Oppgave 47

Anta at man har et tilfeldig innledende utvalg på $n = 40$ tatt fra en normalfordelt populasjon hvor standardavviket $\sigma = 3{,}7$. I utvalget finner en gjennomsnittet $\bar{x} = 12{,}7$.

a) Sett opp et 95% og et 99% konfidensintervall for populasjonsgjennomsnittet μ.

b) Finn de samme to konfidensintervallene ved hjelp av kalkulatoren og ved hjelp av MINITAB.

Anta at man senere ønsker å gjennomføre et større forsøk hvor man på forhånd bestemmer seg for at bredden på det nye 95% konfidensintervallet skal være 1,0.

c) Bestem hvor stort utvalget skal være for oppfylle dette kravet. Hvis man istedenfor 95% hadde ønsket et 99% konfidensintervall med samme bredde hva måtte da utvalgsstørrelsen være.

Oppgave 48

En kjøttbedrift er flere ganger blitt tatt for ikke å oppgi korrekt vekt på sine 1 kgs pakninger med kabonadedeig. Bedriften påstår nå å ha ordnet opp i problemet. Det blir tatt et tilfeldig utvalg på 30 av en kontrollinstans. Man finner nå at gjennomsnittsvekten i dette utvalget er på 1,05 kg. Anta at man ut i fra tidligere målinger kan anta at pakkemaskinen pakker med en varians på $\sigma^2 = 0{,}04$

a) Sett opp et 95% konfidensintervall for (populasjons-) gjennomsnittsvekten for karbonadedeigpakkene etter justering. Ville du vært fornøyd med dette resultatet?

b) Hva bør gjennomsnittsvekten i et utvalg på 30 være hvis man ønsker at nedre grense i et 95% konfidensintervallet på 0,999?

Oppgave 49

Bedriften i oppgave 3.2 gjør ytterligere endringer og er nå ganske sikre på at gjennomsnittsvekten på pakningene er på mer enn 1,0 kg. Det tas nok et tilfeldig utvalg av pakker fra den løpende produksjonen, denne gangen på n =25.

a) Formuler en H_0 og en H_1 for den gitte situasjonen. Angi en test observator og avgjør om store eller små verdier av denne er signifkante.

b) Finn kritisk verdi på 5%-nivået og gjennonfør testingen ved hjelp av denne.

c) Bestem også *P*-verdien. Skisser også en normalfordeling og tegn inn de størrelsene du bruker i b) og c).

d) Sett opp 95% ensidig konfidensintervall i situasjonen over.

Oppgave 50

I den såkalte Osloundersøkelse 1 i 1972/1973 ble alle menn i Oslo i alderen 40-49 år invitert til undersøkelse av risikofaktorer for hjerte-og karsykdom. Man fant da blant annet en gjennomsnittsvekt på 77,3 kg. I år 2000 ble den såkalte Osloundersøkelsen 2 gjennomført og man fant da i et utvalg av menn i samme aldersgruppe en gjennomsnittsvekt på 81,7 kg. Man har hele tiden fra 1972/1973 og fram til i dag hatt mistanke om at vekten ville gå opp på grunn av en gradvis endret livsstil.

a) Formuler en H_0 og en H_1 for den gitte situasjonen. Angi en test observator og avgjør om store eller små verdier av denne er signifkante.

b) Gjennomfør testingen på 5%-nivået idet du kan bruke at populasjonsstandard-avviket hele tiden har vært 3,5 kg.

c) Kommenter eventuelle svakheter ved resultatene.

Oppgave 51

Er det slik at elever bruker mindre tid på lekser hjemme nå enn for 10 år siden. En større undersøkelse på en skole på Østlandet for 10 år siden blant alle 7-klassinger viste at de gjennomsnittlig brukte 5,8 timer pr uke på hjemmelekser.
Et tilfeldig utvalg på 15 syvende-klassinger viste nå følgnde tider:

3,4; 5,8; 12,8; 6,5; 8,2; 4,3; 5,7; 6,2; 3,8; 4,2; 5,9; 6,1; 2,8; 3,9 ; 5,2

a) Avgjør først om verdien 12,8 er en outlier. Hvis dette er tilfellet så fjern den fra tallmaterialet.

b) Formuler en H_0 og en H_1 for den gitte situasjonen. Angi en testobservator og avgjør om store eller små verdier av denne er signifkante.

c) Gjennomfør testingen på 5%-nivået

Oppgave 52

Har virkestoffet Phaseolus Vulgaris som man blant annet finner i stoffet Reduxin som er et ekstrakt fra hagebønner noen effekt på vektreduksjon? Ved Universitetssykehuset i Nord-Norge ble en gruppe overvektige mennesker gitt dette i 4 uker. Resultatet blant de 40 menneskene som deltok i undersøkelsen ble en gjennomsnittlig vekt reduksjon på 3,94 kg med et standardavvik på 1,5 kg.

a) Sett opp et 99% konfidensintervall for den gitte situasjonen.

I en kontrollgruppe på 35 personer var gjennomsnittlig vektreduksjon 0,77 kg. Standardavviket var her på 0,6 kg.

b) Test om det er noen signifaikant forskjell i vektreduksjon på de to gruppene på 5%-nivået.

Oppgave 53

Er det slik at man lettere får promille når man lager seg en drink med lettbrus enn med vanlig brus? Ved et sykeshus i Adeleide i Australia ble en gruppe studenter på 14 delt tilfeldig i to ved loddtrekning. De fikk all samme alkoholmengde i drinken, men den ene gruppa fikk sin alkoholmengde blandet med lettbrus mens den andre gruppa fikk sin drink blandet med vanlig brus. Undersøkelsen viste nå at en drink med lettbrus gjennomsnittlig brukte 21 minutter på å passere gjennom magen med et standardavvik på 1,8 minutter. En drink med sukkerhodig væske brukte gjennomsnittlig 36 minutter på å passere gjennom magen med et standardavvik på 1,9 minutter.

a) Sett opp 95% konfidensintervall for gjennomsnittlig "passeringstid" for de to gruppene. Avgjør om det er en signifikant forskjell.

Etter en stund viste det seg at gruppa som fikk drink med lettbrus hadde en gjennomsnittlig promille på 0,5 med et standardavvik på 0,09, mens den andre gruppa hadde en gjennomsnittlig promille på 0,3 med et standardavvik på 0,08.*

b) Test påstanden om at gjennomsnittlig promille med lettbrus målt en stund etter inntak blir mindre enn gjennomsnittlig promille med sukkerholdig brus. Bruk et nivå på 2,5%.

*Legene ved sykehuset mener at årsaken til dette fenomenet er at lettbrusen får (pga. manglende sukkerinnhold) alkoholen til å gli hurtigere gjennom fordøyelsessystemet og ut i blodårene

Oppgave 54

I det såkalte Betula-prosjektet hvor tannleger, psykologer og nevrologer i Stockholm, Umeå og Tromsø har samarbeidet for undersøke påstanden om at at eldre med egne tenner har bedre hukommelse enn dem uten. Det har vært med 2000 tilfeldig valgte eldre mennesker fra Umeå-området. Disse ble delt i to grupper ; de som hadde trukket mange tenner og de som hadde de fleste tennene inntakt. De har blant annet vært gitt en hukommelsesprøve som har gått på huske flest mulig av 30 tilfeldige oppleste ord. Anta at i gruppen på 1250 av personer med de fleste tennene inntakt husket man gjennomsnittlig 15,7 ord, mens det i den andre gruppen ble husket gjennomsnitlig 12,8 ord. Standardavviket i den første gruppa var 2,7, mens det var 3,2 iden andre gruppa.

a) Formuler en H_0 og en H_1 for den gitte situasjonen. Angi en test observator og avgjør om store eller små verdier av denne er signifkante.

b) Gjennomfør testingen både ved å finne kritsk verdi og ved å finne P-verdien.

c) Sett opp 95% konfidensintervall for differansen mellom gjennomsnittlig antall ord som huskes hos de med de fleste tennene inntakt og gjennomsnittlig antall ord hos de som har trukket mange tenner.

Oppgave 55

En gruppe på 10 barn fikk målt puls før og etter et visst TV-program som mener ikke passer for aldersgruppen (inneholder for mye vold). Anta resultatet av målingene ble:

Barn	Puls etter	Puls før
1	130	105
2	110	106
3	100	95
4	108	103
5	110	100
6	95	92
7	90	91
8	107	98
9	98	90
10	103	99

a) Formuler en H_0 og en H_1 for den gitte situasjonen. Angi en testobservator og dennes fordeling under H_0 ved å gjøre de nødvendige forutsetningene. Avgjør om store eller små verdier av testobservatoren er signifikante.

b) Gjennomfør testingen på 5%-nivået ved å finne kritisk verdi.

c) Det kan se ut som barn nr.1 er en "outlier". Avgjør om dette er tilfellet. Gjennomfør eventuelt (Hvis "outlier") testen på nytt ved å bruke kalkulatoren optimalt

Oppgave 56

Gjennomsnittsvektenfor et tilfeldig utvalg på 14 niårige jenter tatt i 1975 var 29,8 kg ved en skole på Østlandet. Standardavviket var 3,2 kg. I år 2000 viste et tilfeldig utvalg på 15 niårige jenter ved den samme skolen en gjennomsnittsvekt på 32,9 kg. Standardavviket var da 3,3 kg.

a) Avgjør ved hypoteseprøving om det grunn til å påstå at gjennomsnittsvekten på 9-årige jenter kan sies å ha økt ved skolen. Velg nivå på 5%. Anta at populasjonsstandardavvikene var like i 1975 og 2000.

b) Sett opp et 95% konfidensintervall for differansen mellom populasjons-gjennomsnittsvektene i 2000 og 1975.

Oppgave 57

I forbindelse med markedsføringen av lettproduktet EKSTRALETT reklameres det med at det på 100 g vare gjennomsnittlig skal være 5 g fett (dvs. at populasjonsgjennomsnittet $\mu = 5\ g$). Matvarekontrollen har imidlertid mistanke om at produktet gjennomsnittlig inneholder mer fett enn dette. Man tar derfor et tilfeldig utvalg på 11 EKSTRALETT-produkter og analyserer disse. Prøvene viste derfor følgende fettinnhold i gram:

$$x_i = 4, 7, 9, 5, 3, 8, 7, 9, 6, 5, 6$$

a) Formuler en nullhypotese og et alternativ for den gitte situasjonen. Angi testobservator og avgjør om store eller små verdier av denne er signifikante (dvs. gjør at nullhypotesen forkastes).

b) Finn kritisk verdi på 5%-nivået og gjennomfør hypotesetesingen.

Oppgave 58

På en hovedveg hvor det er satt opp skilt om radarovervåkning via foto- og radarutstyr, ønsker man å kontrollere om det er slik at det gjennomsnittlig kjøres fortere like etter passering av kontrollstedet enn rett før.
Anta at hastigheten blir målt for 10 tilfeldig valgte blister 50m før fotoboksen og 50m etter, og at man fant:

Bil nr	1	2	3	4	5	6	7	8	9	10
Hast.før	75	79	80	85	83	77	76	79	79	80
Hast.etter	81	78	89	90	91	92	86	82	79	76

a) Angi en nullhypotese H_0 og en alternativ hypotese H_1 for den gitte situasjonen. (Vink: Beregn differansene for hver bil og bruk en

b) Velg nivå på 5% og gjennomfør testingen.

c) Testingen kunne også vært gjennomført ved en ikkparametrisk testmetode. Gjennomfør testingen ved den alternative metoden. Velg også her nivå på 5%.

Oppgave 59

150 studenter har vært igjennom en matematikktest med henblikk på å kontrollere deres forkunskaper. Resultatene, i form av en poengsum *x* for hver person, fordelte seg som følger:

	Hyppighet	
Poeng *x*	Menn	Kvinner
0-9	2	0
10-19	3	1
20-29	7	3
30-39	15	10
40-49	28	23
50-59	14	25
60-69	8	7
70-79	3	1

a) Gi en grafisk framstilling av dataene for mennene og for kvinnene.

b) Beregn gjennomsnittet $\bar{x}$ og standardavviket s_x både for mennene og kvinnene.

c) Sett opp en taberll over de kumulative relative hyppighetene for både mennene og kvinnene. Tegn sumfordelingskurven for begge gruppene og finn medianene.

d) Ut fra tidligere prøver kan det se ut som kvinnene gjør det litt bedre enn mennene. Man ønsker derfor nå å teste

$$H_0 : \mu_{kvinner} = \mu_{menn} \quad \text{mot} \quad H_1 : \mu_{kvinner} > \mu_{menn}$$

Velg nivå på 5% og gjennomfør testingen.

Oppgave 60

Det blir påstått at unge som blir tilbudt alkohol hjemme drikker mer enn jevnaldrende som ikke blir tilbudt alkohol hjemme. For undersøke denne påstanden blir det tatt et utvalg på $n_1 = 10$ nittenåringer som blir tilbudt alkohol hjemme og et annet uvhengig

og tilfeldig utvalg på $n_2 = 10$ som ikke blir tilbudt alkohol hjemme. Resultatet av undersøkelsen blir:

Alkoholforbruket pr. år (omregnet i liter ren sprit)

Blir tilbudt alkohol hjemme	6,4	7,3	0	8,1	5,9	9,2	7,1	3,6	7,9	6,9
Blir ikke tilbudt alkohol hjemme	2,1	3,8	4,1	2,5	5,1	2,7	3,0	4,7	0	0

a) Sett opp 95% konfidensintervall for differansen mellom populasjonsgjennomsnittene i de to gruppene.

b) Test

$$H_0 : \mu_1 = \mu_2 \quad \text{mot} \quad H_0 : \mu_1 > \mu_2$$

Velg nivå på 5%.

d) Bruk en alternativ ikke parametrisk test til å løse punkt b).

Oppgave 61

Målinger ble foretatt av gripestyrken i venstre og høyre hånd for et tilfeldig utvalg på 10 kjevhendte personer.

Resultatet ble som følgende:

Gripestyrke i vestre	140	90	125	130	95	121	85	97	131	110
Gripestyrke i høyre	138	87	110	132	96	120	86	90	129	100

a) Gjennomfør en t-test på 5%-nivået for å finne ut om det kan sies at kjevhendete har større gripestyrke i venstre enn i høyre hånd.

b) Finn et 95% konfidensintervall for (forventningen til) differansen mellom gripestyrken i venstre og høyre hånd. ($(\mu_V - \mu_H)$)

d) Bruk en alternativ rangbasert test til å løse punkt a)

Oppgave 62

For å se om det er noen sammenheng mellom mening om EU-medlemskap og alder har man intervjuet 330 tilfeldig valgte personer og funnet:

Alder i år / Mening om EU-medl.sk.	20-34	35-49	50→
Nei til medlemskap	30	50	40
Ja til medllemskap	45	35	28
Vet ikke	25	31	46

a) Formuler en H_0 og en H_1 for den gitte situasjonen. Angi testobservator og avgjør når H_0 forkastes på 5%-nivået.

b) Gjennomfør testingen på 5%-nivået ved hjelp av a). Hva blir P-verdien i forsøket?

d) Gjennomfør testingen ved hjelp av MINITAB.

Oppgave 63

På spørsmålet:

Synes du at Norge bør ta imot betydelige mengder arbeidskrakt fra andra land for å fylle stillinger i helsevesenet og andre servicebransjer, synes du det bør importeres minst mulig arbeidskraft eller har du ingen mening om dette?

svarte et representativt utvalg på 600 norske kvinner og menn som følger: (V.G. 14. oktober 2000)

Kjønn / Mening	Mann	Kvinne
Betydelige mengder	175	130
Minst mulig	122	113
Ingen mening	33	27

Man ønsker nå å finne ut om det er noen sammenheng (avhengighet) mellom kjønn og mening om import av arbeidskraft.

a) Formuler en nullhypotese og et alternativ for den gitte situasjonen. Angi testobservator og avgjør om store eller små verdier av denne er signifikante.

b) Gjennomfør testingen på 5%- nivået.

c) Bruk kalkulatoren og MINITAB til å gjennomføre testingen.

Oppgave 64

I en undersøkelse foretatt av Norsk Gallup for VG, svarte 600 tilfeldig valgte norske menn og kvinner som følger på spørsmålet:

” *Hvis du reiser eller skulle reise med NSB, ville du føle deg helt trygg eller ville du føle deg litt utrygg for at sikkerheten er godt nok ivaretatt?”*

Kjønn / Mening	Mann	Kvinne
Helt trygg	225	153
Litt utrygg	87	135

Man lurer på om det på bakgrunn av undersøkelsen kan sies å være noen sammenheng mellom hva slags mening man har, og om man er mann eller kvinne.

a) Formuler en nullhypotese og et alternativ for den gitte situasjonen. Angi testobservator og avgjør om store eller små verdier av denne er signifikante.

b) Gjennomfør testingen på 5%-nivået ved å bruke kritisk verdi. Bruk så kalkulatoren til å finne P-verdien. Hvordan kan denne brukes til å gjennomføre testingen?

Oppgave 65

For å få kjennskap til om det er noen sammenheng mellom mening om bygging av gasskraftverk (med dagens teknologi) og kjønn så utførte MMI en gallup for Dagbladet i februar i år. På spørsmålet :

”*Er du for eller imot bygging av gasskraftverk med dagens teknologi?”*

Fikk man følgende resultat:

Kjønn / Mening	Menn	Kvinner
For	179	90
Imot	112	137
Vet ikke	81	195

a) Sett opp en nullhypotese og et alternativ for den gitte situasjonen. Angi testobservator og avgjør om store eller små verdier av denne er signifikante (gjør at nullhypotesen må forkastes)

b) Gjennomfør testingen på 5%-nivået på bakgrunn av tallene foran. Hvor signifikant er resultatet? (Vink: Hva blir P-verdien?)

e) Gjennomfør testingen ved hjelp av kalkulatoren.

e) Gjennomfør testingen ved hjelp av MINITAB.

Oppgave 66

I en kommune på Østlandet er det gjort en undersøkelse for å se om det er noen sammen heng mellom leseferdighet og kjønn. Resultatet av undersøkelsen blant 250 tilfeldig valgte 7.-klassinger ble:

Leseferdighet / Kjønn	Dårlig	Middels	God	Sum
Jenter	20		45	119
Gutter	38	56		131
Sum	58			250

a) Fyll ut tabellen og gi en grafisk framstilling av dataene. Formuler en nullhyptese og et alternativ for den gitte situasjonen.

b) Angi en testobservator og avgjør om store eller små verdier av denne er signifikante. Bestem kritisk verdi på 5%-nivået og gjennomfør testingen.

Oppgave 67

Utlånene på biblioteket på HIBU (avdeling Hønefoss) i en tilfeldig uke i 2007 var som følger :

Ukedag	Ma	Ti	On	To	Fr
Antall utlånte bøker	68	78	64	52	44

Test påstanden om at antall utlånte bøker fordeler seg likt gjennom uken. Formuler nullhypotese og alternativ hypotese, angi testobservator og gjennomfør testingen på 5%-nivået.

Oppgave 68

Anta at en terning kastes 120 ganger og at man observerer:

Antall øyne	1	2	3	4	5	6
Frekvens O_i	13	23	22	18	17	19

Gjennomfør en kjikvadrattest for å avgjøre om terningen kan sies å være rettferdig.

Oppgave 69

En produsent av fiskesnører har de siste årene importert råstoff fra et industrialisert land og observert at de i det lange løp har produsert 4 kvaliteter: D(=dårlig og må vrakes), C(=god), B(= bedre) og A (=best) med følgende fordeling:

Kvalitet	A	B	C	D
Fordeling i prosent	40	35	20	5

Nå har imidlertid prisen på råstoffet blitt så dyrt at produsenten bestemmer seg for å importere fra lavkostland til en mye lavere pris. Det bekymrer imidlertid ledelsen ved bedriften at råstoffet muligens holder mye dårligere kvalitet. Før de eventuelt bestemmer seg for å bytte leverandør gjennomfører de en prøveproduksjon av fiskesnører. Et tilfeldig utvalg på 200 fiskesnører produsert med det nye råmaterialet gir følgende fordeling av kvaliteter:

Kvalitet	A	B	C	D
Antall O_i	60	80	44	16

a) Formuler en nullhypotese og et alternativ for den gitte situasjonen. Bestem kritisk verdi på 1%-nivået.

b) Gjennomfør testingen både ved hjelp av kritisk verdi og ved å finne P-verdien.

Oppgave 70

Anta at man har kastet 5 mynter 120 ganger og observert antall krone og at resultatet ble:

Antall kron	0	1	2	3	4	5
Frekvens O_i	20	85	197	182	97	19

a) Sett opp sannsynlighetsmodellen til dette forsøket og finn de forventede verdiene E_i

b) Test påstanden om at modellen $X \sim bin(5, 0.5)$ er holdbar (dvs. at mynten er "OK" og avvikene skyldes kun tilfeldigheter) på 5%-nivået.

Oppgave 71

Anta at antall alvorlige trafikkuhell X pr uke på en bestemt sterkt traffikert veistrekning i en tilfeldig uke ble observert til:

Antall uhell	0	1	2	3	4	5	≥ 6
Frekvens O_i	7	27	28	30	15	11	5

a) Anta at $X \sim Poisson(\lambda)$. Angi et estimat $\hat{\lambda}$ for λ ved hjelp av dataene over.

b) Sjekk ved hjelp av en kjikvadrattest antagelsen om modellen over kan sies å være rimelig. Bruk 1% nivå.

Oppgave 72

Anta at man har målt høyden X på 150 tilfeldig valgte mannlige studenter på HIBU og funnet

Høyde i cm	Frekvens O_i
160-164	5
165-169	11
170-174	20
175-179	40
180-184	39
185-189	21
190-194	10
195-199	4

a) Finn $\bar{x}$ og s_x.

b) Anta at $X \sim N(\mu, \sigma)$. Angi et estimat for μ *og* σ. Finn så de tilsvarende forventede verdier E_i i de 8 klassene.

c) Test ved en kjikvadrattest om antagelsen i b) kan sies å være rimelig. Bruk et nivå på 5%.

d) Finn også P-verdien og fortell kort hva du kan slutte av den.

e) Bruk MINITAB og kalkulatoren til å gjennomføre en normalitetstest

Oppgave 73

29. april 2006 viste Dagbladets partibarometer (utført MMI) følgende oppslutning for de 7 største politiske partiene i Norge i april. Undersøklesen baserer seg på telefonintervju med 902 stemmeberettigede personer:

Parti	Frp.	Ap.	Høyre	SV	Krf.	Sent.p.	V
Oppslutning i %	30,3	28,3	13,7	9,7	5,6	5,2	4,9

a) Sett opp 95% konfidensintervall for Ap.s og for Frp.s oppslutning i april 06 basert på Dagbladets partibarometer.

b) Ved valget i 2005 hadde Ap. En oppslutning på 32,7%. Avgjør ved hypoteseprøving om Ap. Nå kan sies å ha fått en signifkant tilbakegang på 5%-nivået ved å teste

$$H_0 : p = 0{,}327 \quad \text{mot} \quad H_1 : p < 0{,}327$$

(Vink: Bruk normaltilnærmelsen)

d) Regn ut styrken for testen foran når p = 0,32; 0,30; 0,28; 0,26 og 0,24. Bruk gjerne kalkulatoren i dine beregninger. Skisser styrkefunksjonen.

Oppgave 74

I en meningsmåling offentliggjort i Aftenposten 16. mars 2003 kunne finne følgende statistikk basert på spørsmålet:

"*Dersom det var folkeavstemning om norsk medlemskap i EU i morgen, hva ville du da stemme?"*

Ja (i %)	58	67	63	62
Nei (i %)	42	33	37	38
Tidspkt.	Des.02	Jan.03	Feb.03	Mar.03

Den siste undersøkelsen er basert på 1000 intervjuer. 81% har avgitt svar om holdning til EU-medlemskap. Undersøkelsen er foretatt ved telefonintervju.

a) Angi et 95% konfidensintervall for andelen av de som har bestemt seg og vil stemme nei, og et 95% konfidensintervall for andelen av de som har bestemt seg og som vil stemme ja i mars 2003.

b) Framstill dataene grafisk. I pkt. a) er det ikke tatt hensyn til de 19% vet ikke stemmene. Hvordan vil disse påvirke det grafiske billedet og konfidensintervallene i a)? Gjør de nødvendige beregningene.

c) Bestem annenordens trendkurver $\hat{y} = ax^2 + bx + c$ både for Ja-siden og Nei-siden ved hjelp av kalkulatoren. Kommenter kort resultatene.

Oppgave 75

20% av alle bilførere bruker fortsatt mobiltelefonen mens de kjører bil, uten at de har "handsfree"-opplegg. Anta at det blir gjennomført en kampanje i media og at det samtidig blir gjort en rekke kontroller av politiet. Dette mener man vil få ned andelen p som feilaktig bruker mobiltelefonen mens de kjører bil. Anta at man ønsker å gjøre en innledende undersøkelse med 500 tilfeldig valgte sjafører.

a) Formuler en H_0 og en H_1 for den gitte situasjonen. Angi en testobservator og dennes fordeling under H_0. Bestem kritisk verdi k på 5%-nivået. (Vink: Bruk normaltilnærmelsen)

b) Av de 500 sjaførene brukte 90 mobiltelefonen feilaktig. Hvilken konlusjon vil du da trekke? Sett også opp et 95% konfidensintervall for p.

Anta at man i hovedundersøkelsen ønsker $P(\left|\hat{P} - p\right| < 0{,}02) \geq 0{,}95$ der $\hat{P}$ estimatoren for andelen *p*.

c) Hvor mange personer bør da være med i undersøkelsen for å oppfylle dette kravet?

Oppgave 76

En undersøkelse utført av MMI for Dagbladet i oktober 2000 basert på et tilfeldig utvalg på 927 stemmeberettigede viste blant annet følgende oppslutning for Ap., Frp., H. og Krf.

Parti	Ap.	Frp.	H	Krf.
Oppslutning	24,9	31,6	12,4	13,3

a) Sett opp 95% konfidensintervall for Ap.s og Frp.s oppslutning.

b) Er det grunn til å tro at Krf. Har større oppslutning i populasjonen enn H? Gjør de nødvendige beregningene.

c) Hva er grunnen til at ± leddet i konfidensintervallet (og dermed usikkerheten) er størst for Frp? Hva er den største verdien ± leddet kan ha for et slikt utvalg (dvs. med *n* =927)

d) Test nullhypotesen om at Høyre har en oppslutning på 14,3% (= Høyres oppslutning ved siste stortingsvalg) mot alternativet at de har fått en tilbakegang. Velg nivå på 5%.

Oppgave 77

Et firma som selger tannkremen SOLIBOKS har en markedsandel på 28%. De vil gjerne øke denne andelen og kjører derfor en reklamekampanje på TV 7. Dette mener de bestemt vil føre til økt markedsandel. Etter kampanjen tar de et tilfeldig utvalg på *n* = 1000 personer og intervjuer mht. hvilken tannkrem man bruker.

a) Formuler en nullhypotese og en alternativ hypotese for den gitte situasjonen. Angi testobservator og avgjøer om store eller små verdier av denne er signifikante.

b) Bestem kritisk verdi på 5%-nivået. Nå svarer 305 av de 1000 at de bruker SOLIBOKS. Hva betyr dette?

Anta at den nye markedsandelen er 30%.

c) Hva er da sannsynligheten for feilaktig å påstå at markedsandelen fortsatt er 28%? Hva er testens styrke i dette tilfellet?

Oppgave 78

Vil plaster med cellegift kunne forlenge livet til kreftpasienter?
I følge en artikkel i Dagbladet ble plasteret som kalles for Glidel og er utviklet av Guildford farmasøytiske firma i Baltimore, Maryland prøvd på 330 pasienter i USA og Skandinavia. Fordelen med denne metoden er at den har mindre bivirkninger enn vanlig cellegift. Resultatet av undersøkelsen ble.

Resultat / Behandlingsmetode	I live etter 1 år	Døde i løpet det første året	Sum
Behandlet med plaster	99	66	165
Behandlet med tradisjonell metode	33	132	165
Sum	132	198	330

Analyser dataene i tallmaterialet ved en Fisher-Irwintest. (Formuler en nullhypotese og et alternativ, angi testobservator og gjennomfør testingen.)

Oppgave 79

Anta man har følgende 6 observasjonspar av variabelene X (forklaringsvariabelen) og Y (responsvariabelen):

x	2	3	5	7	8	10
y	16	14	12	11	10	8

Anta at regresjonsmodellen

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \text{der} \quad \varepsilon_i \sim N(0,\sigma) \text{ og uavhengige } i=1, 2, 3,\ldots,n$$

kan brukes.

a) Bestem likningen for regresjonslinjen for y mhp. x ved minste kvadraters metode

b) Angi et estimat for σ

c) Angi estimat for standardfeilen til β_0 *og* β_1 og sett opp et 95% konfidensintervall for β_1.

d) Test nullhyoptesen om at $\beta_1 = 0$ mot alternativet $\beta_1 \neq 0$. Test også nullhypotesen at $\beta_1 = -1{,}8$ mot alternativet $\beta_1 < -1{,}8$

Oppgave 80

Anta man har følgende 5 observasjonspar av variabelene X og Y

x	3	4	5	7	8
y	10	6	9	8	10

a) Test

$$H_0 : \rho = 0 \quad \text{mot} \quad H_0 : \rho \neq 0$$

på 5% nivået ved å finne kritisk verdi.

b) Hva er p-verdien i testen over?

c) Test

$$H_0 : \rho = 0 \quad \text{mot} \quad H_0 : \rho \neq 0$$

i tallmaterialet i oppgave 79.

Oppgave 81

Et tilfeldig utvalg på ti 6.klassinger har blitt testet i matatematikk. De samme ti elevene har også besvart et spørreskjema knyttet til skole- og hjemmesituasjonen. Resultatet av undersøkelsen viste bl.a. følgende sammenheng mellom variablene

Y = poeng på matematikktesten (fra 0 til 100)
X_1 = mors utdanning (antall år ut over grunnskole)

Elev nr	1	2	3	4	5	6	7	8	9	10
x_1	8	10	5	0	5	0	3	7	3	6
y	75	80	73	65	75	69	73	76	69	72

Anta at du har en normal regresjonsmodell.

a) Tegn spredningsdiagrammet og finn den estimerte regresjonslikningen for y med hensyn på x_1.

$$\hat{y} = a + bx_1$$

(Vis mellomregningen)

b) Bruk kalkulatoren til å bestemme residualene y_{res} gitt ved

$$y_{res} = y_{obs} - y_{pred} \quad (= y_i - \hat{y}_i)$$

Forklar kort hvilke kommandoer du bruker.
Tegn residualplottet og kommenter (kort) dette.

c) Sett opp et 95% konfidensintervall for β.

I denne samme undersøklesen har man også spurt elevene om "hvorvidt de ønsker å lære matematikk, for senere å kunne bruke det i en jobb". Denne variabelen betegnes med X_2 og kan anta følgende verdier: 1 = svært uenig, 2 = meget uenig, 3 = uenig, 4 = verken uenig eller enig, 5 = enig, 6 = meget enig og 7 = svært enig. Resultatet av dette ble:

Elev nr	1	2	3	4	5	6	7	8	9	10
x_2	4	7	4	1	5	2	3	6	2	3

En MINITAB-utskrift av den multiple regresjonsanalysen med X_1 og X_2 som forklaringsvariable viser nå bl.a:

Regression Analysis: Poeng versus Mors utd.; Matem./jobb

```
The regression equation is
Poeng = 65,1 + 0,398 Mors utd. + 1,56 Matem./jobb


Predictor       Coef  SE Coef      T      P
Constant     65,0711   0,8710  74,70  0,000
Mors utd.     0,3978   0,2383   1,67  0,139
Matem./jobb   1,5565   0,4123   3,78  0,007


S = 1,16801   R-Sq = 94,1%   R-Sq(adj) = 92,4%
```

d) Bruk nå kalkulatoren til å beregne $\hat{y}$ i den multiple situasjonen og finn deretter den multiple korrelasjonskoeffesienten $R_{y\hat{y}}$ (= korrelasjons-koeffesienten mellom y og $\hat{y}$).
Bruk utskriften til å kontrollere om du har regnet riktig.

MINITAB-utskriften viser også

```
      Analysis of Variance

Source          DF       SS      MS      F      P
Regression       2  152,550  76,275  55,91  0,000
Residual Error   7    9,550   1,364
Total            9  162,100


Source       DF   Seq SS
Mors utd.     1  133,107
Matem./jobb   1   19,443
```

e) Vis ved hjelp av kalkulatoren at tallene over stemmer. Hva kan du slutte ut av denne utskriften.

Litteraturliste

(1) David S. Moore, George P. McCabe : Introduction to the Practice of Statistics (6.ed. 2008)

(2) Jostein Lillestøl: Sannsynlighetsregning og statistikk med anvendelser (5. utgave 1997)

(3) De Veaux, Velleman og Bock : Intro Stats (2. ed. 2006)

(4) Ajit C. Tamhane, Dorothy D. Dunlop : Statistics and Data Analysis, from Elementary to the Intermediate (1. ed. 2000)

(5) Larson, Farber : Elementary Statistics Picturing the World (4. ed. 2008)

(6) Alan Agresti, Christine Franklin : Statistics, The Art and Science of learning from Data. (2. ed. 2008)

(7) Miller og Miller : John E. Freunds Mathematical Statistics with Applications (7. ed. 2004)

(8) Michael Sullivan, III: Statistics, Informed Decisions using Data. (2.ed.2007)

(9) Gunnar G. Løvås : Statistikk for universitet og høgskoler (2. utgave 2005)

(10) Ronald J. Wonnacot, Thomas H. Wonnacot : Introductory Statistics (4.ed. 1985)

(11) Alan Agresti, Barbara Finley : Statistical Methods for the Social Sciences (2.ed. 1986)

(12) Ryan, Joiner : MINITAB Handbook (4.ed. 2001)

(13) Ruth Meyer, David Krueger : A MINITAB Guide to Statistics (2.ed. 2001)

(14) Texas Instruments: Bruksanvisning TI-84 Plus (2003)

Høgskolen i Buskerud
Postboks 235
3603 Kongsberg
Telefon: 32 86 95 00
Telefaks: 32 86 98 83

www.hibu.no

HØGSKOLEN
i Buskerud