INNFØRING I SPSS

Petter Laake
Avdeling for biostatistikk
Universitetet i Oslo

Denne innføringen er tenkt brukt gjennom medisinerstudiet. Den er skrevet til bruk for selvstudium, og den skal være så utfyllende at den kan leses uten lærerstøttet undervisning.

Statistikk undervises i SHBS-blokken i modul 1, men statistikk er nyttig gjennom hele studiet. Undervisningen i modul 1 inneholder bruk av programpakken SPSS. Så lenge statistikk er nyttig også senere, vil også SPSS være det. Det er derfor meningen at vi kan gå tilbake til denne innføringen etter hvert som det er behov for å gjøre en statistisk analyse.

Undervisningen i SHBS-blokken er lagt opp med mye bruk av SPSS. Der vil det refereres til hvilke kapitler i denne innføringen som må leses i forbindelse med forelesninger og gruppearbeid. Senere i studiet vil den måtte brukes til selvstudium

Bruker vi dette e-læringsopplegget bare til selvstudium, må vi selvfølgelig regne med å gå gjennom mer enn det som strengt tatt er nødvendig for den aktuelle analysen vi vil gjøre.

Denne innføringen er delt i 13 kapitler.

Innledning til statistiske analyser:

1. Komme i gang med SPSS
2. Innledning til SPSS
3. Innledning til statistiske analyser
4. Innlesing av data
5. Databearbeiding 1
6. Databearbeiding 2
7. Databearbeiding 3
8. Statistiske analyser via ordrefiler

Statistiske metoder:

9. Deskriptiv analyse
10. Diagrammer og plott
11. Univariable statistiske metoder
12. Multivariable statistiske metoder
13. Oppsummering av filtyper i SPSS

Hvert kapittel har underkapitler. For eksempel har kapittel 1 seks underkapitler. Men noen underkapitler har også under-underkapitler. I kapittel 5 har underkapittel 5.1 to under-underkapitler. Ved å gå inn på hvert kapittel i venstre spalte kan vi søke oss frem nedover i kapitlene og nedover i underkapitlene.

Skal vi utføre en statistisk analyse, må vi kjenne til hvordan vi kommer i gang med SPSS, hvordan vi leser inn fil og hvordan vi gjør enkle databearbeidinger. Skal vi kjenne hele gangen i en statistisk analyse, må vi derfor kjenne innholdet i kapitlene 1-8, før vi går til de statistiske metodene i kapitlene i kapittel 9-12. Men i mange tilfeller har vi lest en SPSS-fil som er klar til statistisk analyse, og vi behøver ikke kjenne annet enn innholdet i kapittel 9-12.

Uansett er det noen kapitler som er viktigere enn andre. Dette vil bli gjennomgått i undervisningen i SHBS-blokken.

Denne innføringen vil måtte endre seg over tid. Men den er også lett å oppdatere. Kommentarer, synspunkter, forslag til endringer mottas med takk og kan sendes Petter Laake petter.laake@medisin.uio.no.

Blindern, august 2016

Innhold

1. Komme i gang med SPSS

Læringsmål

I dette kapittelet skal vi lære hvordan vi kommer i gang med SPSS, og hvordan vi avslutter SPSS. For å komme i gang, er det det enkleste kanskje å laste ned programvaren SPSS fra UiOs nettsider til egen PC. Men det er også mulig å bruke SPSS på PC-stuene i Domus Medica, eller å bruke SPSS fra egen PC via såkalt UiO-kiosk.

For å gjøre statistisk analyse på filer som er tilrettelagt for oss må vi også lære å laste ned filer fra en nettside. Datafilene i SPSS er lagd med komma som desimalindikator. Vi må derfor også lære å skifte mellom punktum og komme som desimalindikator.

Når vi har gått gjennom kapittel 1, er vi klare til å gå inn i selve programpakken SPSS.

1.1 SPSS på egen PC

SPSS for Mac eller Windows kan installeres til egen maskin fra nettsiden

https://app.uio.no/programvare/

På listen over programmer går vi til SPSS for Mac eller SPSS for Windows, og vi følger instruksjonen for å installere på egen maskin. Dersom vi ikke får tilgang til installeringen, kan dette skyldes at vi ikke er innmeldt i en SPSS brukergruppe. Hvis dette skjer, må vi sende en e-post til programvare@usit.uio.no med opplysninger om UiO-brukernavn, at dere vil installere SPSS for Mac eller Windows for statistikkfaget i MED1100. Når vi har fått bekreftelse fra USIT om at dere er med i SPSS brukergruppe, kan dere installere SPSS slik som angitt over.

Hvis dere har problemer, kan dere henvende dere til orakel-hjelp@medisin.uio.no eller til foreleser i statistikk.

1.2 SPSS på PC-stue

I Domus Medica finnes det to slike PC-stuer. Der må vi logge oss på med UiO-brukernavn og passord. Da kommer det opp en skjerm, med ikonet for Windows i venstre hjørne på menylinjen nederst. Hvis vi klikker på den venstre musetasten på ikonet, kommer det opp en meny. Nederst i listen som da kommer opp ligger *All programs*. Vi klikker på den. Da kommer det opp en lang liste, som også inneholder en katalog som heter IBM SPSS Statistics. Når vi klikker på den, får vi opp innholdet i listen. Der ligger det to programmer, og vi klikker på IBM SPSS Statistics 22.

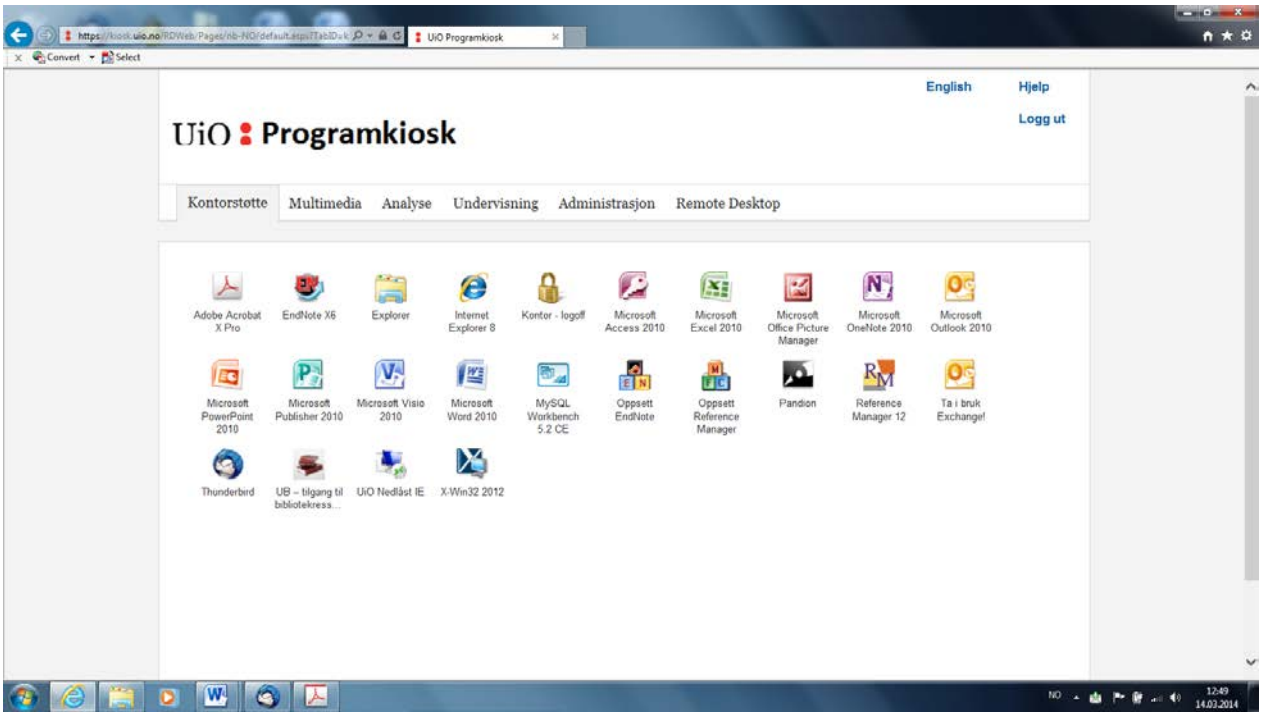
Hvis vi er nye brukere, vil det komme opp en eller to dialogbokser som vi klikker oss ut av med *OK*. Da ender vi i datavinduet til SPSS, med et forslag til å hente en datafil fra katalogen Dokumenter. Vi klikker oss ut av med *Cancel*, og får da opp et tomt dataark. Mer om de ulike vinduene i kapittel 2.

1.3 SPSS via UiOs programkiosk

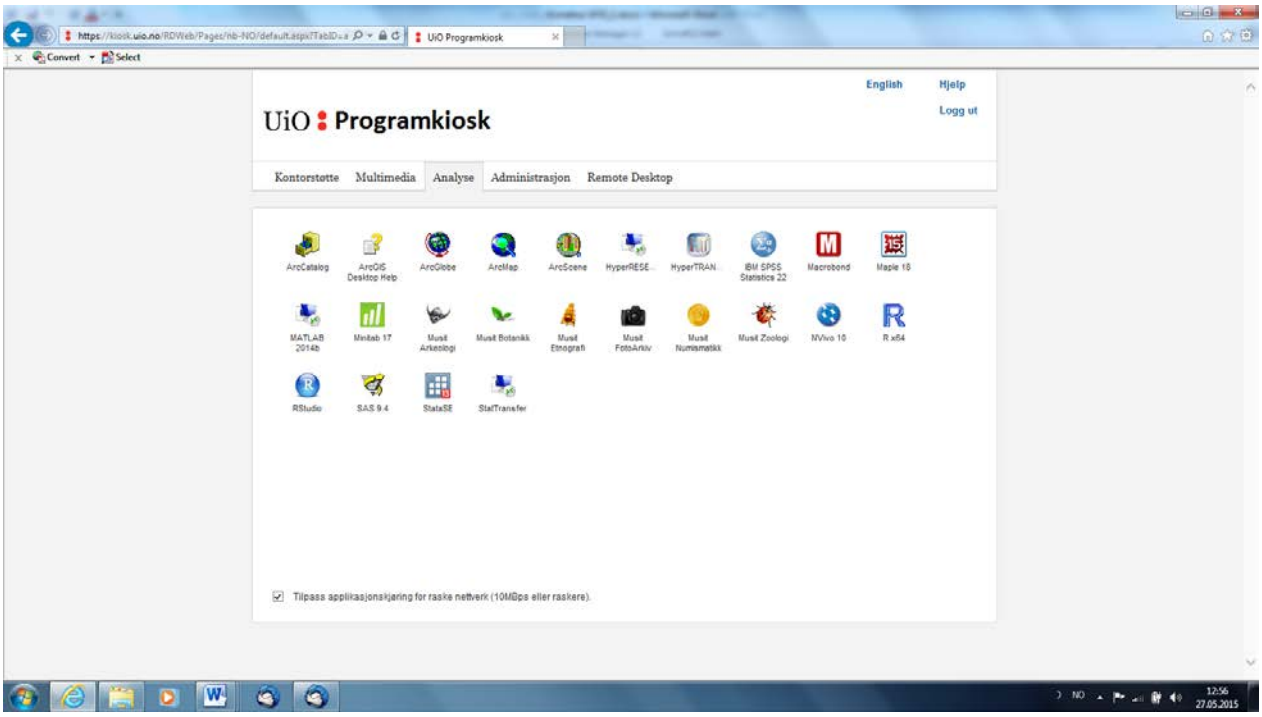
Vi går til Internet Explorer og skriver inn

www.kiosk.uio.no

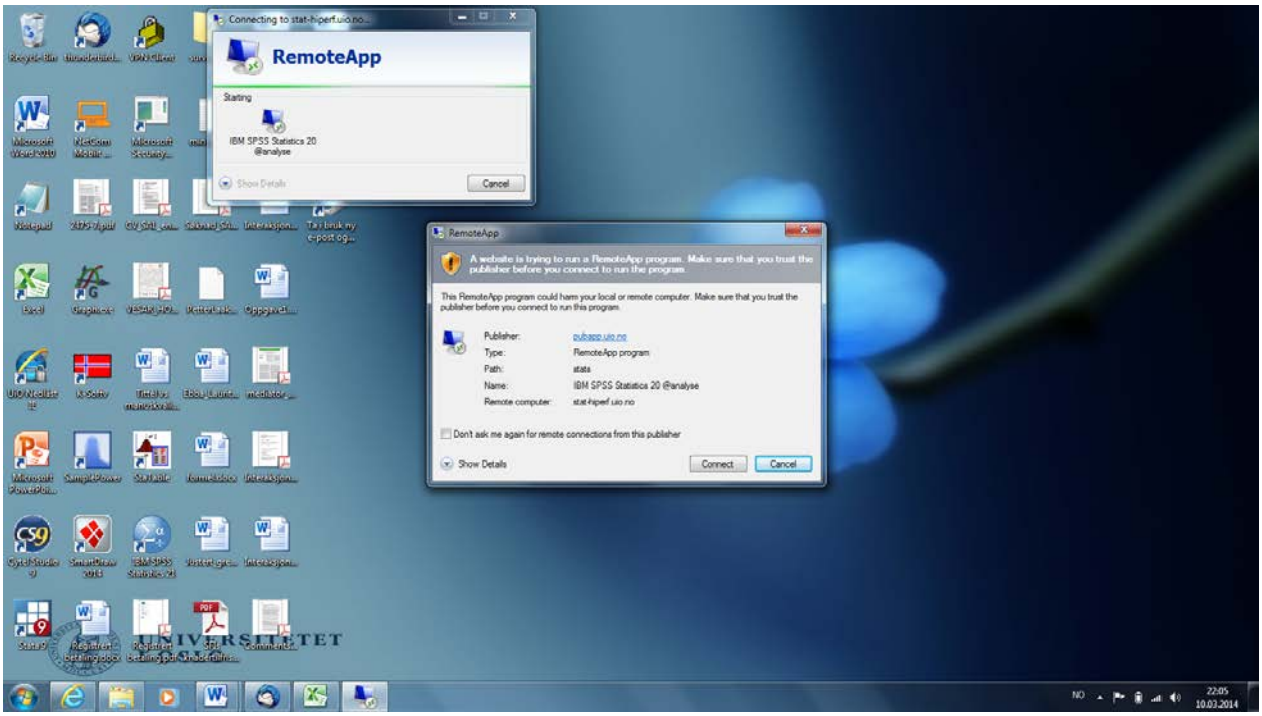
Der må vi logge inn med vårt UiO-brukernavn og passord. Da åpner UiOs programkiosk seg med følgende skjermbilde:



Her går vi *Analyse* i knapperekken, og får da følgende skjermbilde:



Her ser vi *IBM SPSS Statistics 22* i øverste linje. Vi dobbeltklikker på det ikonet. Dette kan ta litt tid! Men til slutt vil det åpne seg et nytt vindu, som under



Her klikker vi på *Connect* i det nederste vinduet. Da starter SPSS opp. Før vi kommer inn i SPSS sitt datavindu, må vi som nye bruker klikke *OK* på ett eller to vinduer som kommer opp.

Mer om de ulike vinduene i SPSS i kapittel 2.

1.4 Nedlasting av filer fra kurssidene til katalog på egen maskin

Vi går til hjemmesiden for e-læringsopplegget i SPSS: http://www.med-utv.uio.no/elaring/fag/med-statistikk/index.shtml

Nederst finner vi en oversikt over filene vi skal bruke i kurset. Vi velger ut den filen vi vil overføre til våre datamaskin. For å få nedlastet filen klikker vi på den høyre musetasten. Da kommer det opp en meny hvor vi velger *Save Target as.* Da åpner det seg et vindu, med forslag til hvor vi skal lagre filen. Vi velger en katalog som er dedikert til den analysen vi skal gjøre, eller til det kurset vi følger.

Som alltid: Det er viktig at vi har oversikt over hvor filene våre ligger!

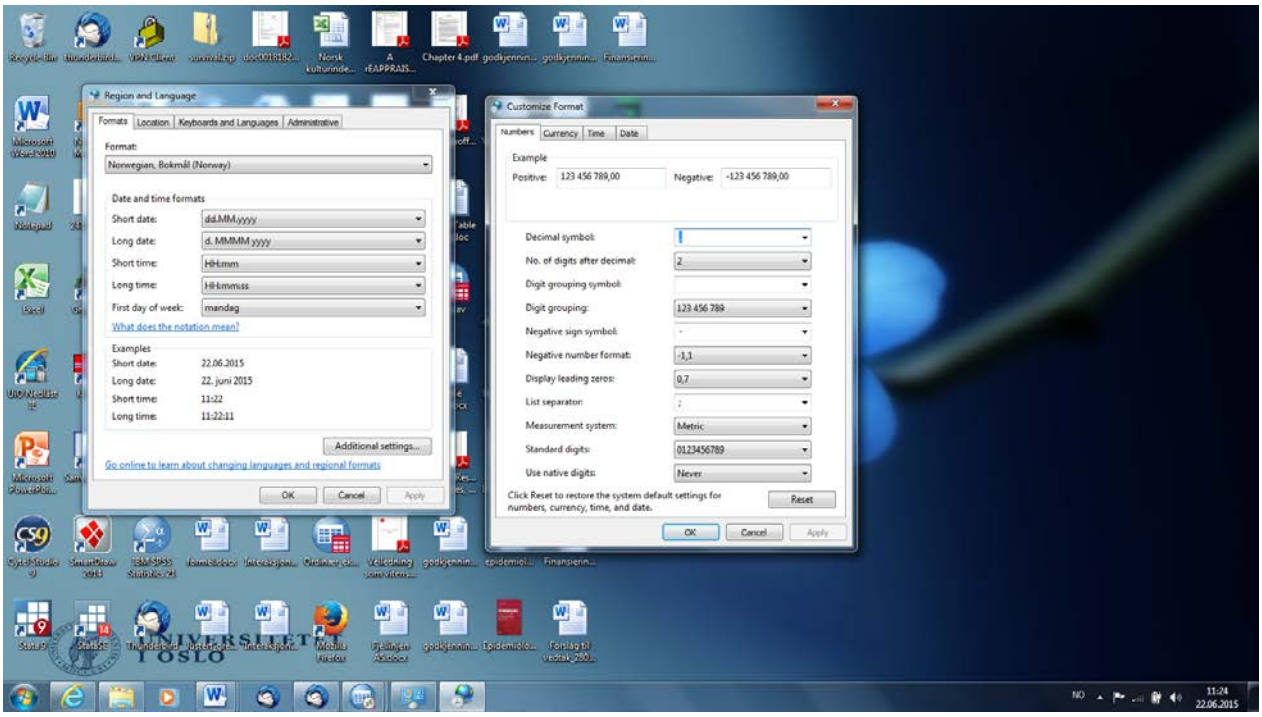
1.5 Desimalindikator

Desimalindikatoren kan være enten komma eller punktum. I norsk Windows settes denne til komma, men i engelsk Windows brukes vanligvis punktum. Dette er viktig å være klar over, siden data som legges inn med gal desimalindikator, ikke vil bli godtatt av SPSS.

Hvis vi kommer opp i situasjoner der de tallene vi vil legge inn i SPSS, ikke blir godtatt, skyldes dette vanligvis at desimalindikatoren er satt feil. At data ikke blir godtatt betyr at data som er tall (*Numeric*) blir oppfattet som bokstaver (*String*).

På alle datamaskiner som er installert på Universitetet, er desimalindikatoren satt til komma. Denne innføringen er skrevet for SPSS med desimalindikator komma. Alle de datafilene som lastes ned fra kursets hjemmeside, har komma som desimalindikator.

Dersom datamaskinen vi bruker er satt opp med punktum som desimalindikator, kan det for dette kurset være lurt å endre til komma. Valget av komma eller punktum som desimalindikator gjøres ikke i SPSS, men i Windows-hovedmenyen. Vi går da til Windows ikonet i nederste venstre hjørne. Den aktiviseres ved å venstreklikke med musen. Der går vi til *Control Panel/Region and Language* Der klikker vi på knappen nede i høyre hjørne, *Additional settings*. Den øverste boksen der gjelder *Decimal settings*. Der kan vi endre mellom komma og punktum.

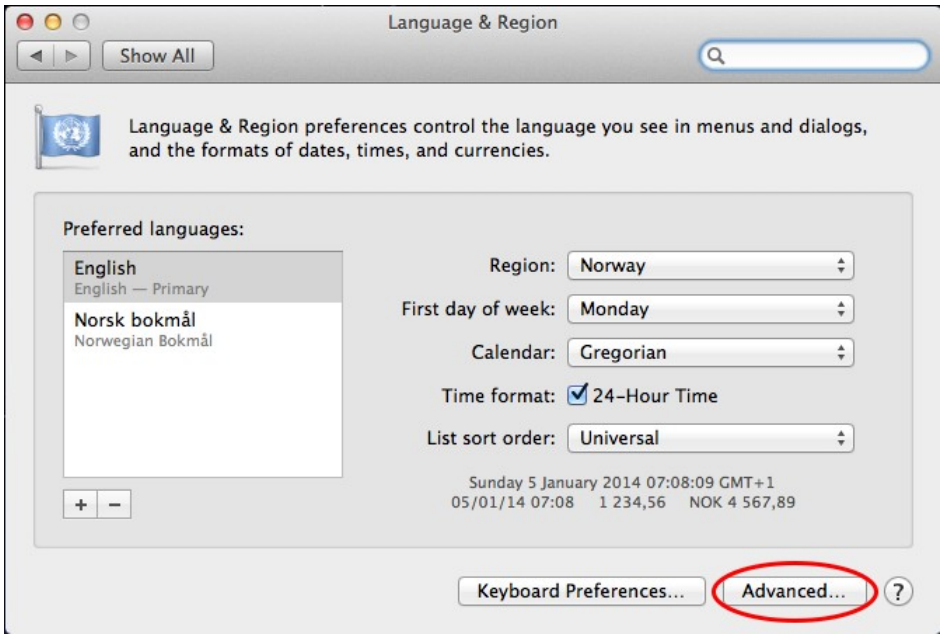


Forandring her blir først aktive når vi går inn i SPSS på nytt.

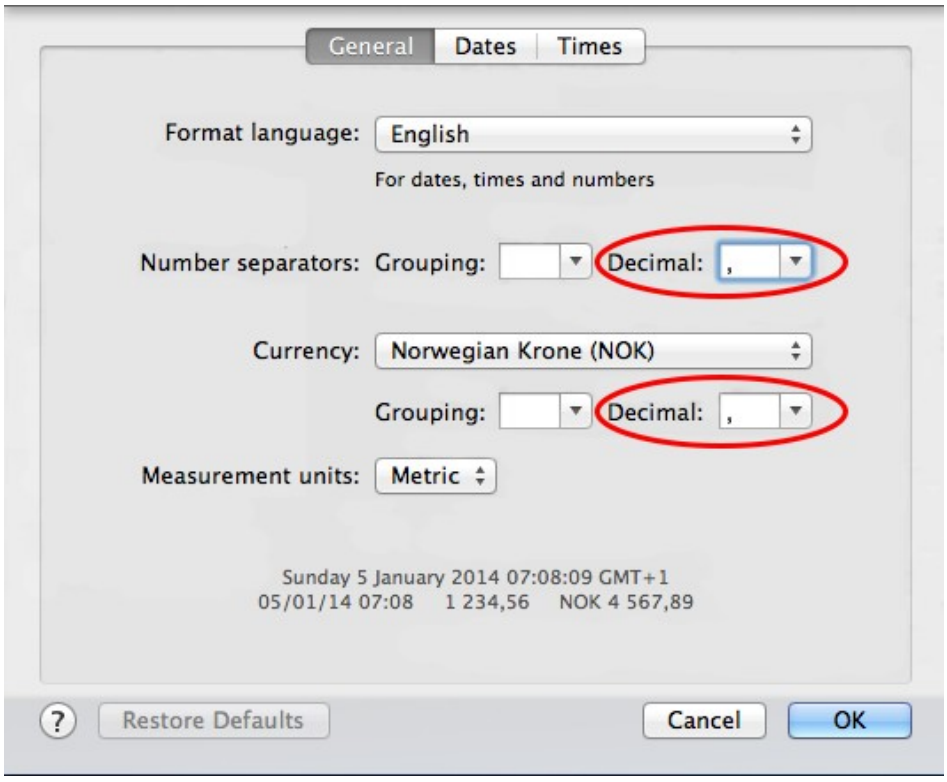
For å skifte desimaltegn fra punktum til komma på Mac som kjører Mac OS X 10.9.x, går vi først inn i System Preferences («Valg» heter det i norsk utgave).

I vinduet som dukker opp, velger vi *Language & Region*.

I det nye vinduet som da dukker opp, velger vi knappen merket med *Advanced*....



I det siste vinduet som dukker opp, velger vi fanen *General*, og setter komma begge steder som vist under. Til slutt trykker vi *OK*.



1.6 Hvordan gå ut av SPSS

Hvis vi vil gå rett ut av SPSS – uten at vi bevarer de filene vi har åpnet – klikker i den røde boksen med det hvite krysset øverst i høyre hjørne. Vi må da først klikke *Yes* som svar på det første spørsmålet og *No* på det andre, for også få lukket det andre vinduet som er åpnet.

Dersom vi vil bevare de filene vi har åpnet, må vi klikke på *File/Exit* i hovedmenyen. Da kommer det en del advarsler som det er viktig å forholde seg til. De gir en påminning om at vi bør lagre våre datafiler, resultater osv.

2. Innledning til SPSS

Læringsmål

I dette kapittelet skal vi gå gjennom menyene og de ulike vinduene som vi kan åpne i SPSS. SPSS har fire vinduer som vi må kjenne til: Datavinduet, utskriftsvinduet, ordrevinduet og grafikkvinduet. Vi legger dataene våre inn i datavinduet, og resultatene av analysene våre kommer i utskriftsvinduet. Plott og diagrammer legges også i utskriftsvinduet. Men skal vi editere på resultatene må vi bruke grafikkvinduet. Vi skal i hovedsak bruke menyene i SPSS for våre analyser. Men av og til er det fornuftig å lage ordrer som SPSS kan uføre for oss. Dette gjør vi i ordrevinduet.

Kjennskap til de forskjellige vinduene er helt avgjørende før vi kan analysere datafilene våre.

2.1. Hovedmenyene i SPSS

Øverst i alle SPSS-vinduene har vi 12 hovedmenyer. Hver av dem har undermenyer som framkommer når vi klikker på dem. Hovedmenyene varierer noe med hvilket vindu man er inne i. For datavinduet er det følgende hovedmenyer:

1. *FILE*: Benyttes for å utføre filoperasjoner, som å åpne eller lagre filer, innlesing av data fra datafiler vi har liggende. Når vi skal åpne en fil er det enklest hvis vi står i et vindu tilsvarende den filtypen vi vil åpne. Når vi skal lagre et vindu MÅ det vi skal lagre være i det aktive vinduet.

2. *EDIT*: Benyttes bl.a. for å flytte eller kopiere tall eller tekst fra data-, utskrifts- eller ordrevinduet.

3. *VIEW*: Benyttes til å bestemme en del egenskaper ved skjermbildet, som hvilke verktøylinjer som er synlig etc.

4. *DATA*: Benyttes bl.a. for å definere variable, velge ut spesielle undergrupper for videre analyse (*Select cases*) og til å flette data fra flere filer. Endringene som gjøres er midlertidige med mindre det gis ordre om noe annet.

5. *TRANSFORM*: Brukes til omkoding av variabler eller generering av nye variabler i arbeidsfilen.

6. *ANALYZE*: Brukes for å utføre ønsket statistisk analyse, som f. eks. t-tester, analyse av krysstabeller, regresjon osv.

7. *DIRECT MARKETING*: Ikke av interesse i dette kurset.et.
8. *GRAPHS*: Brukes til å generere grafer av ulike typer, slik som diagrammer og plott.

9. *UTILITIES*: Brukes til å få informasjon om alle variabler i arbeidsfilen, til å få opp en liste over alle SPSS-kommandoer og til å endre fonter.

10. *WINDOW*: Brukes til å manøvrere mellom de ulike vinduene (Data, Output og Syntax).

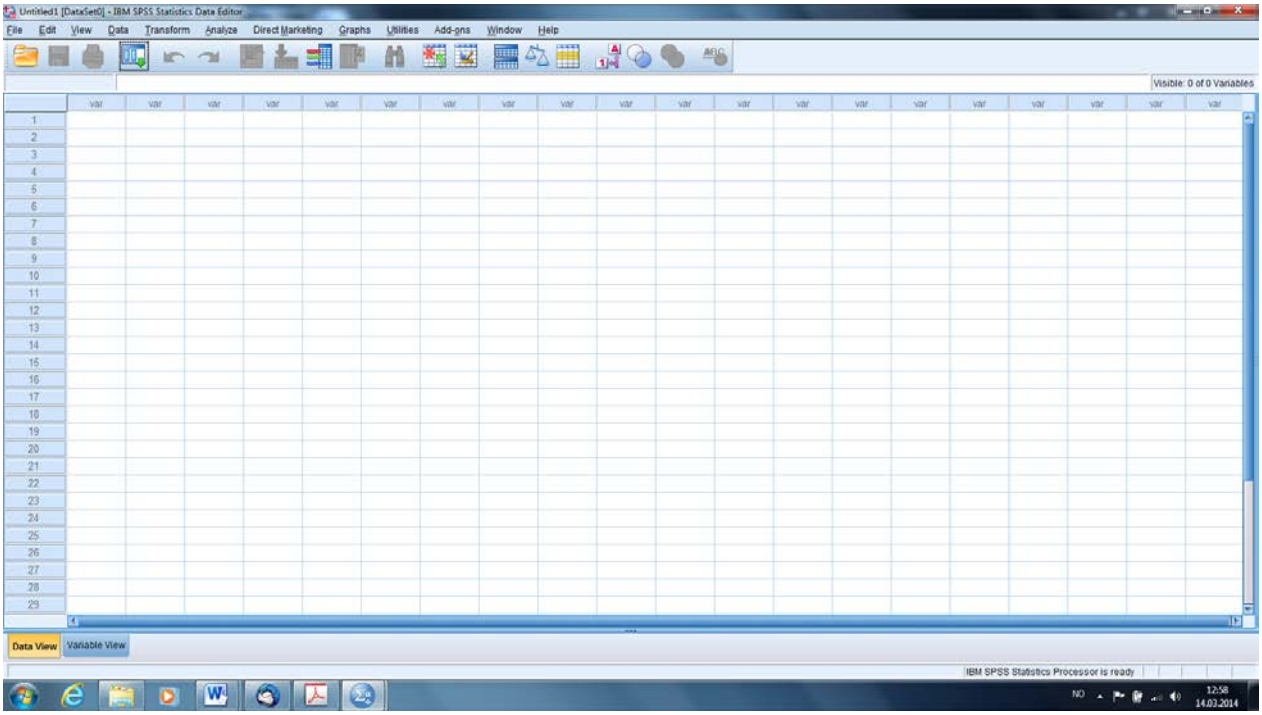
11. *ADD-ONS*: Ikke av interesse i dette kurset.

12. *HELP*: Åpner et hjelpe-vindu som inneholder informasjon om ulike SPSS-funksjoner.

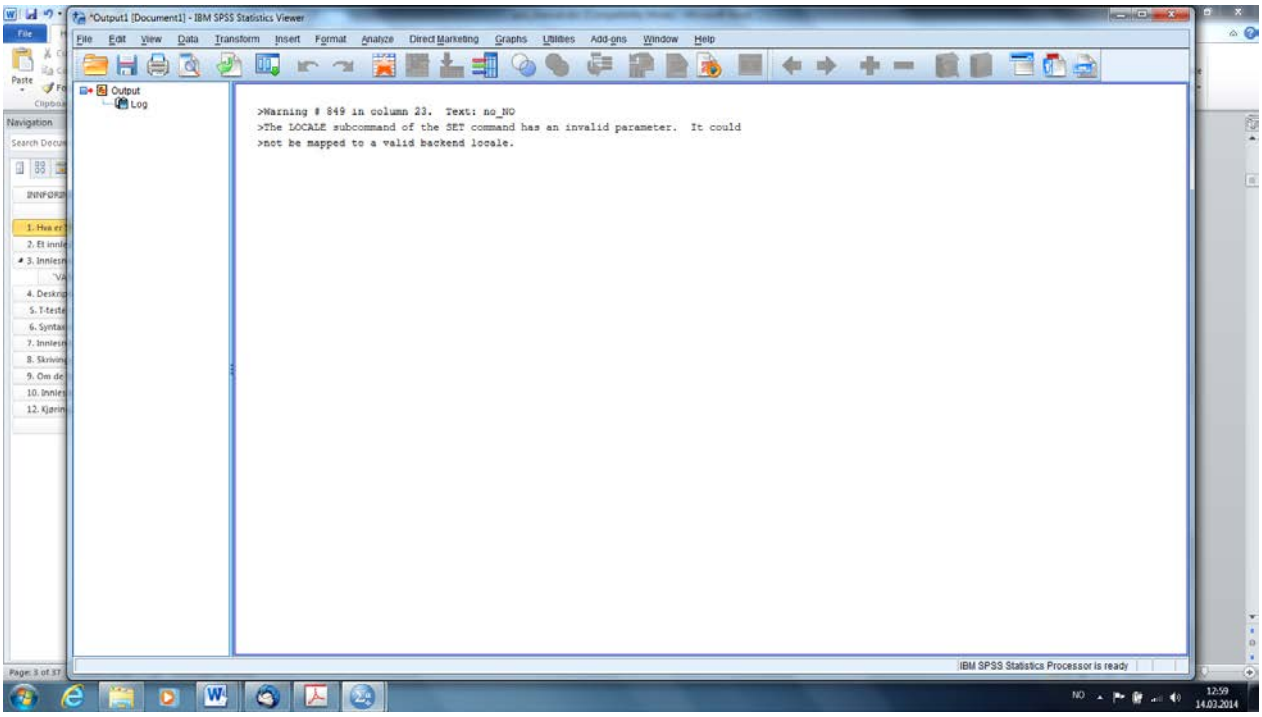
Menylinjen ligger rett under hovedmenyene. Ved hjelp av denne får vi raskt tilgang til en del viktige funksjoner. Hvilke som er til stede, er avhengig av det aktive vinduet. Forklaring på ikonene står nederst til venstre i vinduet når musepilen peker på ikonet.

2.2. Startvinduet

Når vi starter opp SPSS, møtes vi av det følgende skjermbilde:

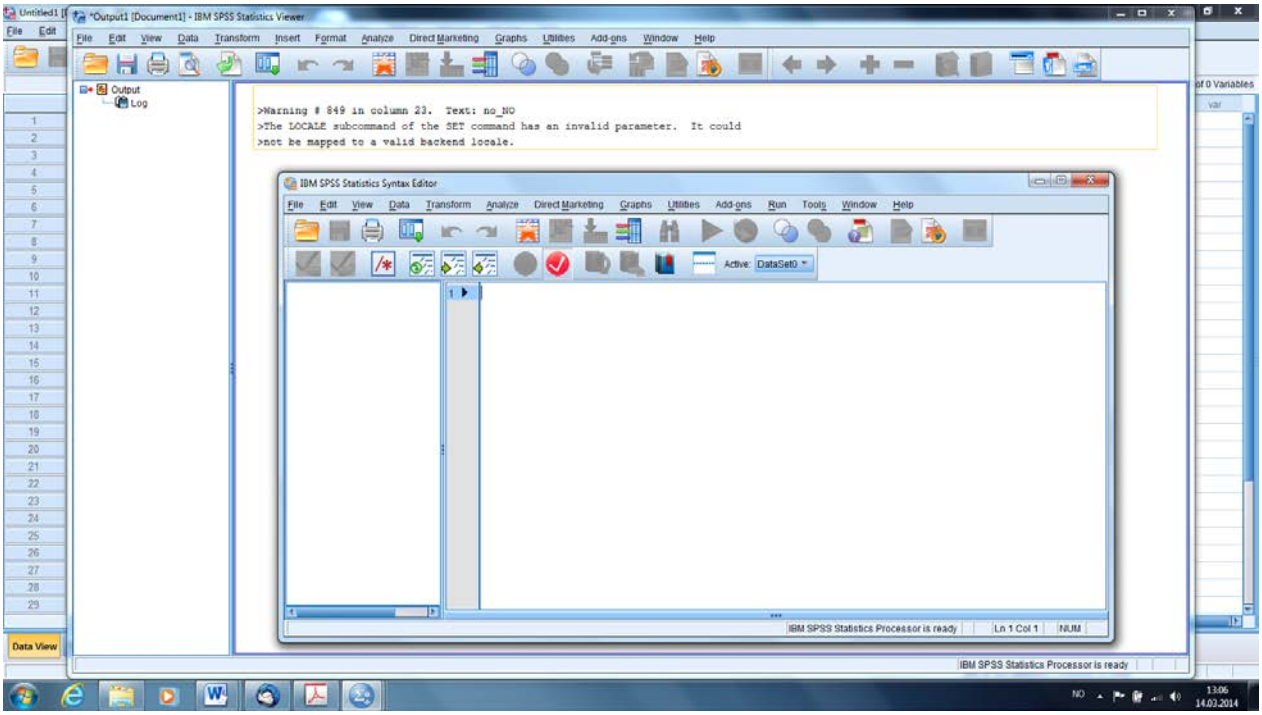


Dette er datavinduet. Men legg merke til at det også har åpnet seg et utskiftsvindu (Output):



Det står ikke mye her, men i dette vinduet blir utskriftene av de analysene vi gjør, lagt. Legg merke til at vi hvis går ned til ikonet for SPSS, som nå ligger nederst på skjermen, kan vi veksle mellom datavinduet og utskriftsvinduet.

SPSS har et tredje vindu som vi må kjenne til – ordrevinduet (Syntax). La oss åpne et slikt vindu. Det gjør vi ved å gå til *File/New/Syntax.* Når vi klikker på den, vil det åpne seg et nytt vindu, og skjermen ser da slik ut:



Nå har vi tre vinduer tilgjengelig i SPSS: Datavinduet, utskriftsvinduet og ordrevinduet. I tillegg har vi et grafikkvindu. Når vi lager grafiske fremstillinger, kan disse editeres i et grafikkvindu.

Vi kan klikke oss mellom de fire vinduene ved å gå ned på SPSS-ikonet, og velge hvilket vindu vi vil være i.

SPSS-vinduer

SPSS har fire vinduer som er spesielt viktige og nyttige:

1. Datavinduet (Data): Dette kommer vi automatisk inn i ved oppstart av programmet. Det brukes til å legge inn og editere data.
2. Utskriftsvinduet (Output): Her skrives resultatene av de statistiske analysene ut. Til venstre i vinduet er det en innholdsfortegnelse som gir oversikt og gjør det lett å bevege seg raskt mellom utskriftene fra de forskjellige analysene.
3. Ordrevinduet (Syntax): Her kan vi skrive direkte i SPSS språk (i motsetning til å bygge dem opp via meny-valg) og hvor kommandoene kan lagres til senere bruk.
4. Grafikkvinduet (Graph): Her kan vi editere diagrammer og plott som vi har lagd i analysene våre.

2.3 Datavinduet

Når vi har startet opp SPSS, kommer vi inn i datavinduet. Der ligger det linjer og kolonner. Linjene er nummert 1, 2, 3 osv. I et tomt data-vindu er variabelnavnene (var00001, var00002, osv.) dimmet som tegn på at de kun representerer potensielle variable.

Innlegging av data i en celle gjøres ganske enkelt ved å merke den aktuelle cellen ved. et museklikk, eller vi kan bevege seg rundt ved piltastene. Innholdet i den cellen vi arbeider med legges automatisk i øverste linje i den såkalte *Cell editor*. Etterat den ønskede verdien er lagt inn, lagres den ved å trykke *Enter* eller en piltast. Når data legges inn i en ny kolonne, etableres det en reell variabel, og navnet i øverste rekke er ikke lenger dimmet (grått).

Legg merke til at det nederst på dataarket er to knapper med navnene *Data View* og *Variable View*. Når vi er inne i dataarket, ser vi at *Data View* er aktiv. Vi kan aktivisere *Variable View* på flere måter. Vi kan gå over til *Variable View* ved å trykke på denne knappen i dataarket. Da kommer vi over til *Variable View* som gir en oversikt over variablene som er lagt inn. Vi kan da legge inn variabelnavnet ved å skrive det virkelige variabelnavnet over der det nå står var00001. Når vi så går tilbake til *Data View*, ser vi at variabelnavnet nå er skiftet over til det nye. Vi kan også komme over fra *Data View* ved å dobbeltklikke på variabelnavnet (var00001 osv.). Da kan vi på samme måte som over legge inn det ønskede navnet på variabelen.

Blanke celler uten datainnhold blir konvertert til *Missing value*, og for numeriske variabler er det et punktum. Data kan kopieres og flyttes ved de vanlige kommandoene *Copy* og *Paste*, enten ved å markere de som skal flyttes eller kopiers, klikke *Ctrl-c* og klikke på *Ctrl-v* når de skal flyttes eller kopieres. Dette kan også også gjøres ved først å markere det som skal

kopieres eller flyttes, og så gå inn i *Edit* på hovedmenyen og klippe på *Copy* og *Paste* på vanlig måte. Hele kolonner eller rekker kan behandles på tilsvarende måte.

Det er viktig å merke seg at målinger for hvert individ (personer, forsøksdyr osv.) legges nedover i linjene. Variablene som vi måler for hvert individ (f.eks. måleverdien på den variabelen vi er interessert, sammen med andre variabler, som alder, kjønn osv.) legger vi i kolonnene.

2.4 Utskriftsvinduet

Her legges alle resultatene fra de databearbeidingene eller analysene vi gjør. Det er her vi leser ut det vi er interessert i fra analysene våre. Vi kan kopiere resultatene i utskriftsfilen inn i en Word-fil, slik at vi får pene manuskripter som også kan gå inn i notater eller rapporter. Vi kommer tilbake til hvordan vi gjør det i kapittel 5.

2.5 Ordrevinduet

Vi kan gjøre analyser i SPSS enten ved å lage kommandoene via dialogboksene. Det er det enkleste dersom vi bare skal gjøre én og én analyse. Men av og til vil også gjøre en rekke tilsvarende analyser på den samme datafilen. Da kan det være enklere å gjøre analysene via ordrevinduet. Ofte er det også lurt å gjemme på de kommandoene vi har brukt til analysene våre. Det gjør vi ved å lagre kommandoene i ordrefiler. Vi skal se nærmere på dette i kapittel 8. Der vil vi bruke kommandoen *Paste* fra dialogboksen for å få skrevet kommandoene ned i ordrevinduet. Deretter vil vi vise hvordan vi gjør analyser fra ordrevinduet.

2.6 Grafikkvinduet

Dersom vi kjører en prosedyre som genererer et diagram eller et plott, blir den lagt sammen med andre resultater i Outputvinduet. Hvis den skal redigeres dobbeltklikker vi på diagrammet eller plottet og kommer inn i Grafikkvinduet. Vi kan også lage grafer direkte fra hovedmeny GRAPHS. Vi kommer tilbake til dette senere i denne innføringen.

2.7. Statistisk analyse utført ved menyer

Når vi skal utføre en statistisk analyse i SPSS må vi først ha lest inn ett datasett. Vi skal lese inn vårt første datasett i kapittel 4. Når vi har et datasett lagt inn i datavinduet velger vi statistisk prosedyre fra hovedmenyen *Analyze*. Her kan vi velge *Descriptive Statistics* og deretter *Descriptives* fra undermenyen. Vi kommer da inn i en dialogboks hvor vi kan flytte de variablene som skal med i analysen fra boksen til venstre over i boksen til høyre. Variablene merkes ved hjelp av et museklikk og flyttes ved å trykke på pilen i midten. Gjør vi en feil eller ønsker å utføre en analyse med andre variable, kan variablen(e) flyttes tilbake på tilsvarende måte.

Nederst i dialogboksen ligger en knapperekke som har følgende funksjoner:

OK	Utfører analysen, og resultatene legges i et Output-vindu.
Paste	Kommandoene som svarer til valgene vi har gjort i dialogboksen, legges over i et ordrevindu der de kan lagres og evt. modifiseres før analysen utføres (se kapittel 8).
Reset	Sletter alle endringene vi har gjort i dialogboksen og underdialogboksene.
Cancel	Sletter endringer og lukker dialogboksen.
Help	Gir brukeren et hjelpevindu med informasjon om dialogboksen.

Til høyre i dialogboksen ligger det også en knapperekke, som vil variere med hvilken dialogboks vi er i. Den kan for eksempel inneholde:

Options	Her kan vi spesifisere andre kommandoer i tillegg til dem som er gitt i dialogboken
Statistics	Her kan vi angi hvilke statistiske beregninger ønsker vi å få gjort, i tillegg til dem som er gitt i dialogboksen

Hvis vi i *Analyze/DescriptiveStatistics/Descriptives* går inn i underdialogboksen *Options*, kan vi her angi at vi ønsker å få beregnet varians (*Variance*) og summen (*Sum*). Hvis vi velger *Analyze/DescriptiveStatistics/Frequencies* finner vi underdialogboksene *Statistics*, *Charts* og *Format*. Åpner vi *Statistics* får vi muligheter til å spesifisere angivelser av persentiler, spredningsmål som standardavvik, sentral tendens som gjennomsnitt og median.

De tre knappene i underdialogboksen har følgende funksjoner:

Continue	Lagrer endringene og bringer brukeren tilbake til hoveddialogboksen.
Cancel	Sletter endringer og bringer brukeren tilbake til hoveddialogboksen.
Help	Gir brukeren et hjelpevindu med informasjon om subdialogboksen.

Legg merke til at innstillingene i dialogboksene blir stående også etter at det er utført analyser så lenge vi holder på i den samme SPSS-datafilen. Når vi klikker på *OK* eller *Paste*, lagres innstillingene. Åpner vi en ny datafil, settes innstillingene tilbake til «null», den såkalte «default» innstillingen.

2.8 Lagring av datafiler, utskriftsfiler, ordefiler og diagrammer/plott

Alle datafiler, resultater og order kan lagres ved å gå til *File/ Save as* for nye filer som ikke har fått navn enda. Vi må angi filnavn (*File name*) og katalog (*Look in*:). Ethvert filnavn består av selve navnet (før punktum) og ekstensjonen (etter punktum). SPSS velger selv ekstensjon etter hva slags type fil som skal lagres. Vi endrer IKKE på ekstensjonen. Er det en datafil, skal ekstensjonen være *sav*. Med denne ekstensjonen vet både vi og SPSS at vi har å gjøre med en datafil i SPSS-format.

Merk at det bare er filen i det vinduet vi står i som lagres. Står vi altså i datavinduet, er det datafilen som lagres. Da må vi angi hvilken katalog filen skal legges i, og hvilket filnavn den skal ha. Siden det er en datafil, velger SPSS selv ekstensjonen *sav*. Velger vi for eksempel vårt eget *navn* som filnavn, vil datafilen bli hetende *navn.sav*.

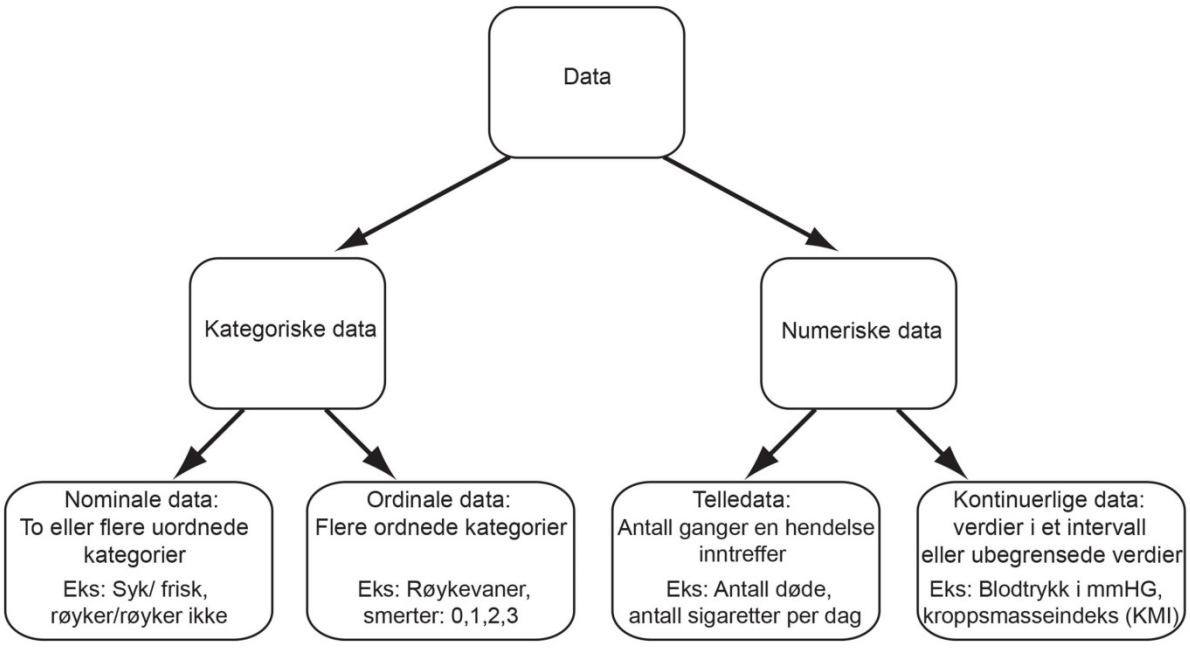
3 Innledning til statistiske analyser

Læringsmål

Før vi går inn i de statistiske analysene er det viktig å være klar over at det ligger noen hovedprinsipper til grunn for hvilke metoder som skal velges. Det er en sammenheng mellom datatyper og statistiske metoder. Vi vil derfor i dette kapittelet først gå gjennom hvilke datatyper vi har og så vil vi gå vi en oversikt over sammenhengen mellom datatyper og statistiske analysemetoder.

3.1 Datatyper

I medisin er vi interesserte i beregninger av størrelser som gjennomsnitt, variasjon og sammenhenger. Hvordan slike størrelser skal beregnes, er avhengig av hva slags type data som inngår i analysen, og det er viktig å uttrykke størrelsene på måleskalaer som er egnet nettopp for dem. Siden beregningene er numeriske (tallmessige), er det naturlig å måle på numeriske måleskalaer. Presentasjonen av data og statistiske analyser vil avhenge av måleskalaen. Vi skiller vanligvis mellom kategoriske og numeriske måleskalaer, se figuren nedenfor. Numeriske måledata kalles oftest kontinuerlige data, slik at det er like vanlig skille mellom kategoriske og kontinuerlige data.



En nominal variabel inndeles i ulike kategorier, og det er ingen naturlig ordning mellom kategoriene. Eksempler på slike variabler kan være sykdom og røyking. Her er sykdom inndelt i kategoriene friske og syke, og røyking er inndelt i kategoriene ikke-røykere og røykere. I statistiske analyser det vanlig å angi referansekategorien med verdien 0 og den andre kategorien med 1. Dette betyr at den kategorien vi ønsker å sammenligne med, gir vi verdien 0 og den andre får da verdien 1. Siden vi alltid ønsker å sammenligne de syke med de friske, gir vi de friske verdien 0 og de syke verdien 1. Tilsvarende gir vi røykerne verdien 1 og ikke-røykerne verdien 0, siden vi ønsker å sammenligne røykerne med ikke-røykerne.

En ordinal variabel deles inn i bestemte kategorier, på en slik måte at det finnes en underliggende ordning (avstandsmål) mellom kategoriene. Et eksempel på en ordinal variabel kan være røykevaner, med kategoriene ikke-røyker (med kategoriverdien 0), tidligere røyker (kategoriverdien 1), røyker av og til (kategoriverdien 2) og røyker daglig (kategoriverdien 3). Antall ganger en hendelse inntreffer, er ofte av interesse. Slike data kalles telledata. Eksempler kan være antall døde eller syke i en bestemt tidsperiode.

Kontinuerlige variabler karakteriseres ved at vi kan måle avstand mellom de ulike verdiene. Eksempler på dette kan være vekt, konsentrasjon, blodtrykk og alder, med verdier på variablene angitt direkte ved de observerte verdiene.

I SPSS er det viktig å ha klart for seg hva slags type type vi skal analysere. Valg av hvordan vi presenterer data (ved frekvensoversikter, histogrammer eller gjennomsnitt og median) avhenger av om vi har kategoriske eller kontinuerlige data.

3.2 Statistiske analyser

Presentasjonen av resultatene av de statistiske analysene er selvfølgelig helt avgjørende for at vi kan få forklart funnene våre. Det er tre hovedresultater som vi alltid skal presentere:

Presentasjon av statistiske hovedresultater

- Effektestimatet
- Usikkerheten til effektestimatet, i form av 95 % konfidensintervall
- Resultater fra testingen av én eller flere statistiske hypoteser, i form av p-verdier

Effektestimatet er det tallet som angir selve resultatet fra analysen. Vi kaller det effektestimatet, siden det er et beregnet tall (et estimat) på den effekten vi er interessert i. Effekten kan være uttrykt som differansen i to gjennomsnitt, en differanse mellom to risikoer (sannsynligheter), en relativ risiko eller en regresjonskoeffisient.

Alle effektestimater har en usikkerhet eller variasjon knyttet til seg. Et effektestimat med en liten usikkert kan vi feste mer lit til enn et estimat med stor usikkerhet. Det er viktig å få presentert denne usikkerheten. Det gjør vi ved et konfidensintervall. Et 95% konfidensintervall dekker populasjonsverdien av effektestimat med en sannsynlighet på 95%.

Sentralt i all forskning står hypotesetesting. Den statistiske null hypotesen er den vi ønsker å forkaste, slik at vi kan godta alternativ hypotesen. Verdien som angis av nullhypotesen kaller vi null-verdien. Ofte er null verdien lik 0, slik som når effektestimatet angir differansen

mellom to gjennomsnitt eller en regresjonskoeffisient. Men merk at null verdien også kan være lik 1, slik som for relativ risiko eller odds ratio.

En statistisk hypotese testes ved en test størrelse. P-verdien angir sannsynligheten for at denne test størrelsen er lik eller større enn den verdien vi har observert. Dersom p-verdien er < 0.05, sier vi at vi har et statistisk signifikant resultat, og null hypotesen forkastes.

Valg av statistiske metoder og test størrelse henger selvfølgelig nøye sammen med valg av effektmålet vi bruker, som igjen henger nøye sammen med målenivået på variablene.

Nedenfor gir vi en oversikt over hvilke effektmål og hva slags statistiske tester vi bruker, avhengig av datatype. Vi vil komme tilbake til de enkelte metodene i undervisningen i statistikk.

I statistiske analyser skiller vi mellom avhengig variabel og forklaringsvariabel. Den avhengige variabelen er den variabelen vi skal forklare. Forklaringsvariabelen(e) er den eller de variablene vi forklarer den avhengige variabelen med. I analyser av kontinuerlige data med to eller flere grupper, vil selve målevariabelen være den avhengige variabelen, og variabelen som angir gruppene vil være forklaringsvariabelen.

I SHBS-blokken i modul 1 vil vi gå gjennom de metodene som er angitt med *kursiv*.

Type data	**Effektmål**	**Uavhengige utvalg**	**Parede data**
Kontinuerlige data			
2 eller flere grupper	*Differanse i gjennomsnitt*/Differanse i medianer	*t-test*/ANOVA/ Wilcoxon-Mann-Whitney-test/KruskalWallis-test	*Paret t-test*/Wilcoxons test for parsammenligning/ Friedmans test
Kontinuerlig forklaringsvariabel	*Regresjonskoeffisient*	*Regresjonsanalyse*	Repeterte målinger
Nominale data			
2 eller flere grupper	*Risiko differanse/ Relativ risiko/Odds ratio*	*Kji-kvadrat-test*	McNemars test
Kontinuerlig forklaringsvariabel	Odds ratio	Logistisk regresjon	Betinget logistisk regresjon
Ordinale data			
2 eller flere grupper	Differanse i medianer	Wilcoxon-Mann-Whitney-test/ Kruskall-Wallis-test	Wilcoxons test for parsammenligning/ Friedmans test
Overlevelsesdata			

To eller flere grupper eller kontinuerlig forklaringsvariabel	*Hazard ratio*	*Kaplan-Meier plot, Cox regression*

4 Datafiler i SPSS

Læringsmål

I SPSS kan vi bruke datafiler som er lagd på flere ulike måter. De mest vanlige er

1. Lager datafilen ved å legge inn data i dataarket.
2. Bruke datafiler i SPSS-format, som er lagd ferdig for oss
3. Hente inn data som er i dataformat, såkalt ASCII-format, og lage SPSS-filer av dem.
4. Hente inn data fra andre formater, slik som Excel.

I kapittel 1 så vi hvordan vi kunne laste ned filer fra kursets hjemmeside til våre egen datamaskin. I dette kapittelet skal vi lære hvordan vi lager en SPSS-fil fra data som vi selv har hentet inn, har funnet i en lærebok, eller på annen måte har fått tak i (kapittel 4.1). Deretter skal vi lære hvordan vi skal få lagt inn i SPSS en datafil som allerede er lagd ferdig for oss i riktig SPSS-format (kapittel 4.2). Dette er det mest vanlige formatet. Deretter vil vi i kapittel 4.3 vise hvordan vi kan lage en SPSS-fil fra en datafil, i såkalt ASCII-format. Til slutt vil vi kapittel 4.4 vise hvordan vi gjør det samme med data fra en Excel-fil.

4.1 Data via dataarket. Eksempel: altman.dat

Vi skal starte med et enkelt eksempel hentet fra side 193 i boka til D.G.Altman: Practical statistics for medical research. Altman sammenlikner energiforbruket hos slanke (lean) og overvektige (obese). Vi skal legge disse dataene inn i SPSS og beregne gjennomsnitt for hver av gruppene. Dataene er som følger:

slanke	overvektige
6.13	8.79
7.05	9.19
7.48	9.21
7.48	9.68
7.53	9.69
7.58	9.97
7.90	11.51
8.08	11.85
8.09	12.79
8.11	
8.40	
10.15	
10.88	

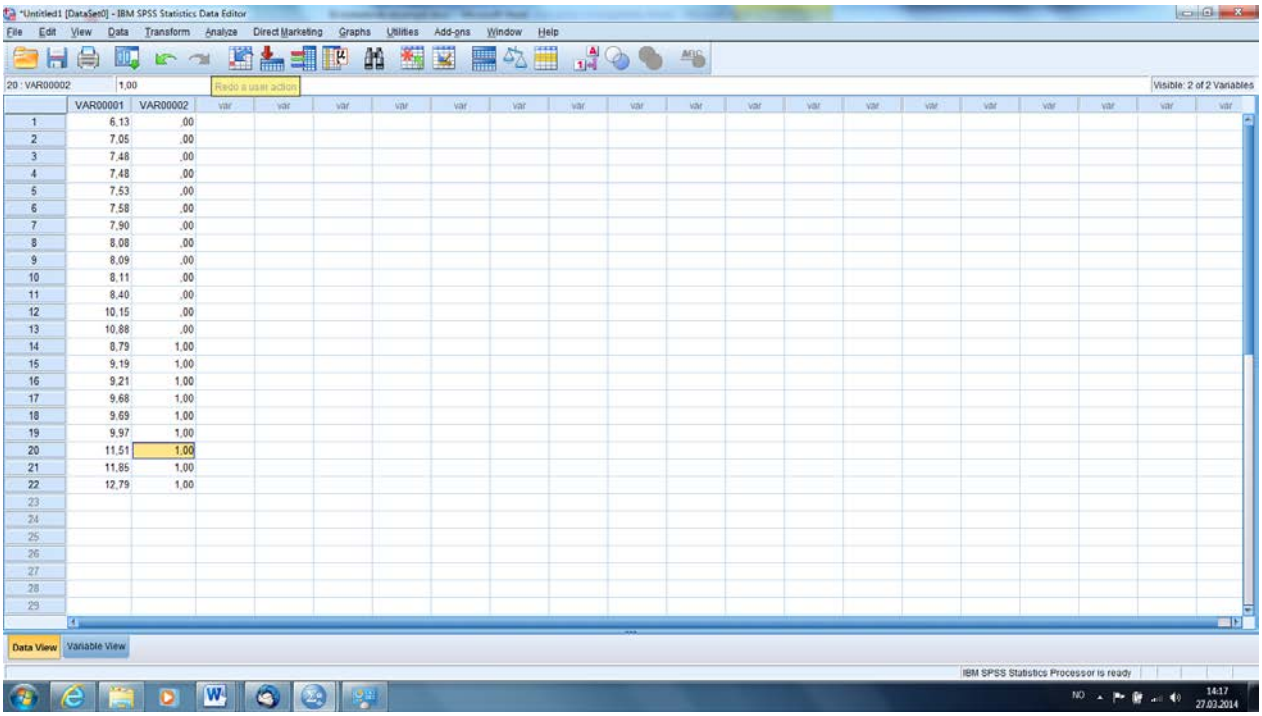
Mean	8.066	10.298	
SD	1.238		1.398

Vi går nå inn i et tomt dataark i datavinduet. Hvis vi allerede har noen data liggende, kan vi alltid åpne et nytt datavindu ved å gå til *File/New/Data*. Da får vi opp et tomt dataark.

Vi er nå klare til å legge data inn i dataarket. Men her merker vi oss at vi data for 22 personer (13 slanke og 9 overvektige). **Siden hver person skal leses inn med sine data linje for linje, kan vi ikke legge inn disse dataene i to kolonner som vist over**.

Energiforbruket for alle individene legges inn i første kolonne, dvs til sammen 13 (slanke) + 9 (overvektige) = 22 tall. For å kunne skille de to gruppene fra hverandre må vi også legge inn data i andre kolonne. Vi legger inn 0 for de slanke og 1 for de overvektige. SPSS har gitt variablene navnene var00001 og var00002.

Dataarket ser da slik ut



Når vi nå har lest inn altman-dataene, må vi huske å legge filen ned som en SPSS-fil i katalogen vår. Vi går da til *File/Save as*. Der velger vi katalogen vi skal legge filen i. Under *File name*: velger vi **altman**. Merk at SPSS foreslår av filnavnet skal slutte på sav. Det er viktig at vi beholder den ekstensjonen, siden vi da sikrer oss at SPSS kjenner igjen denne filen som en SPSS-fil. Når vi har gjort det, trykker vi på *Save*. Da ligger filen som en SPSS-fil med navnet **altman.sav**, og vi kan senere hente den opp som det.

4.2 Data i SPSS-format. Eksempel: lowbwt.sav

I forrige kapittel så vi på hvordan vi kunne lese inn data direkte i dataarket. Svært ofte er filene som vi skal bruke allerede ferdig lagde i det så kalte SPSS-formatet. Da kan vi hente filen direkte inn i SPSS. Da ligger variabelnavn osv. klart, uten at vi trenger å lage dem, slik vi gjorde i kapittel 4.1. SPSS-filer ligger der som en datafil med ekstensjon **sav** som vi kan hente direkte inn i SPSS. Det gjør vi ved *File/Open*. Da kommer vi inn i et Open Data vindu. I *Look in*: endrer vi til den katalogen der datafilen vår ligger. Når vi finner den SPSS-filen vi skal ha, markerer vi den ved å høyre-museklikke på den, og deretter klikke på *Open*.

I dette eksempelet skal vi bruke data som kommer fra LOW BIRTH WEIGHT studien som har blitt foretatt i Massachusetts, USA. Dataene ligger tilgjengelige på hjemmesiden for kurset, men datafilen er opprinnelige hentet fra hjemmesiden til University of Massachusetts at Amherst:

http://www.umass.edu/statdata/statdata/stat-desc.html

Der ligger også en beskrivelse av formålet med denne studien:

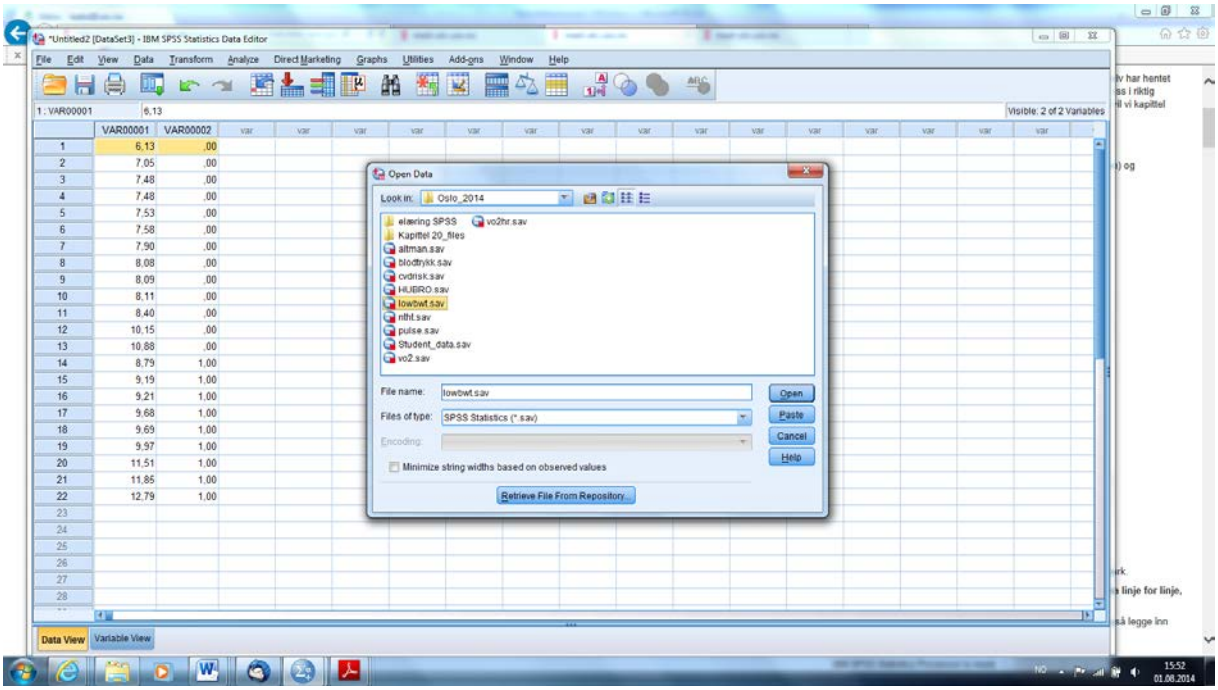
> The goal of this study was to identify risk factors associated with giving birth to a low birth weight baby (weighing less than 2500 grams). Data were collected on 189 women, 59 of which had low birth weight babies and 130 of which had normal birth weight babies. Four variables which were thought to be of importance were age, weight of the subject at her last menstrual period, race, and the number of physician visits during the first trimester of pregnancy.

Dataene som ligger på denne filen er:

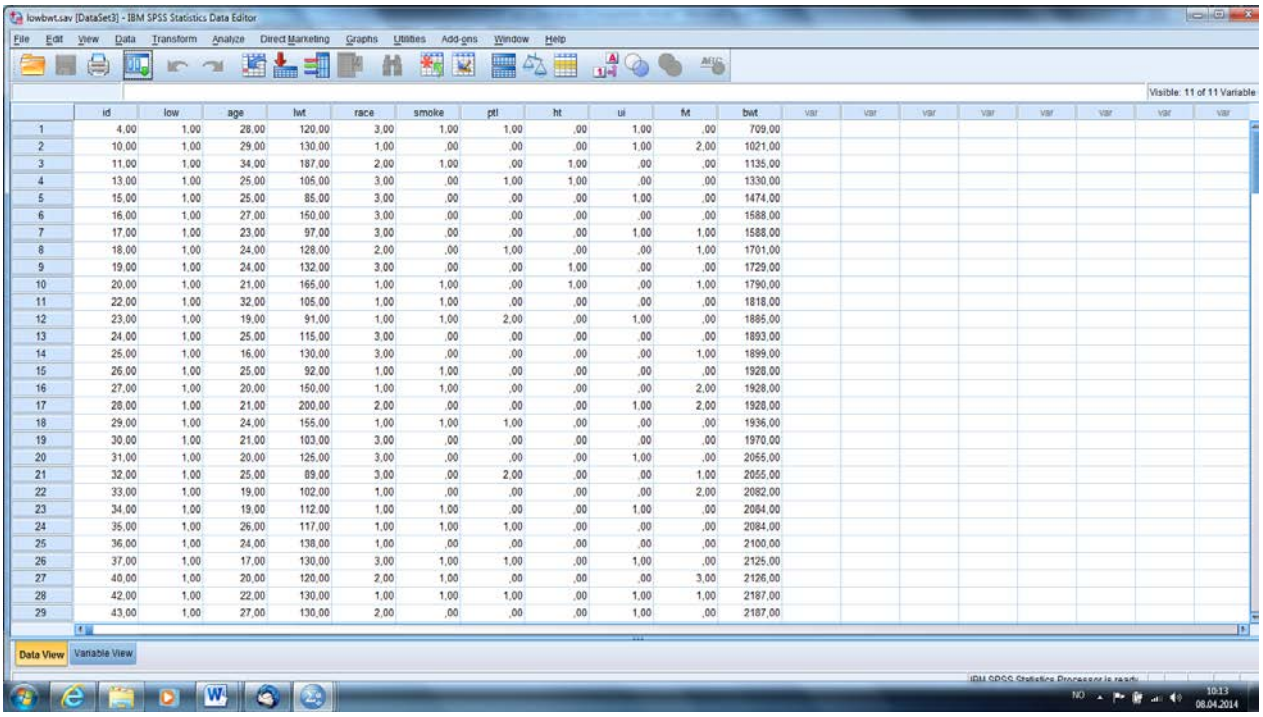
1	Id	ID
2	Low Birthweight (1=BWT<2500g, 0=BWT>2500g)	LOW
3	Age of the mother	AGE
4	Weight in pounds at last menstrual period	LWT
5	Race (1=white, 2=Black, 3=Other)	RAC
6	Smoking status (1=current, 0=not smoking during pregnancy)	SMOKE
7	History of premature labor (0,1,2...,)	PTL
8	History of hypertension (1=yes, 0=no)	HT
9	Uterine irratibility (1=yes, 0=no)	UI

10	Number of first trimester visits (0,1,2,3...,)	FVT
11	Birthweight in grams	BWT

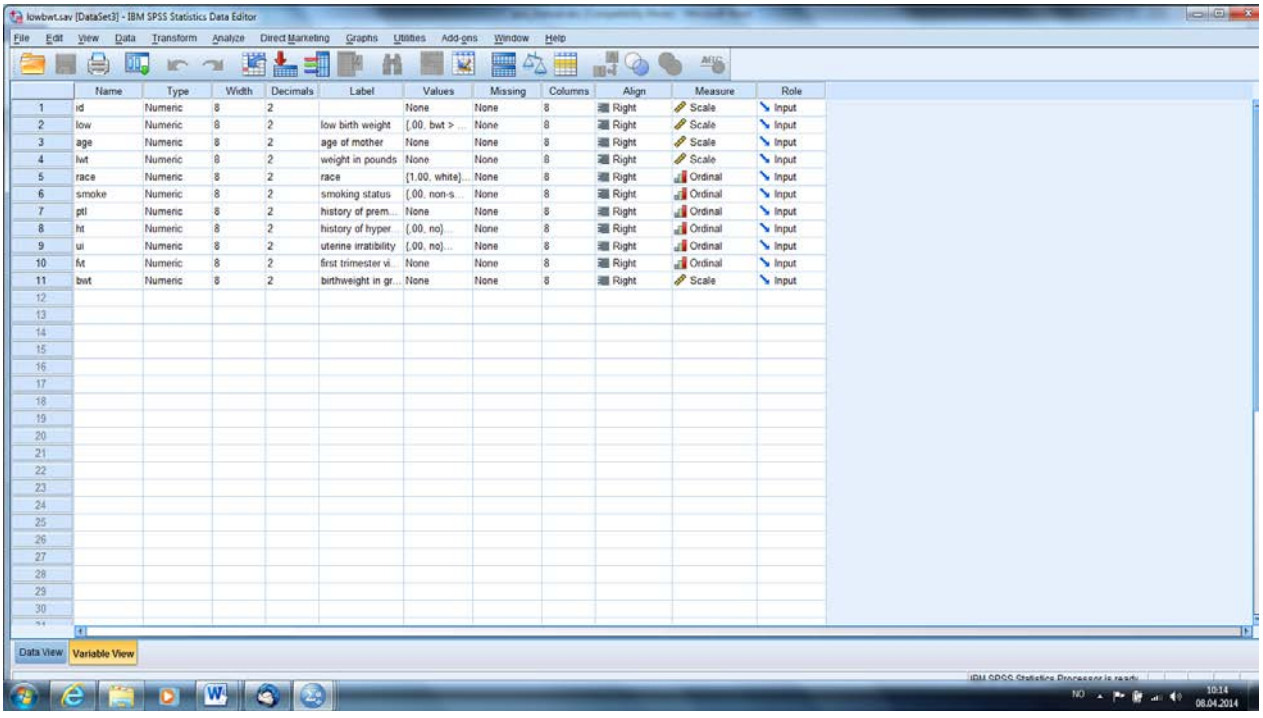
For å få lastet denne filen ned på vår PC, gjør vi slik vi lærte i kapittel 1. Vi legger filen ned i en katalog der vi har SPSS-filer for dette kurset. Når filen ligger der er vi klare til å hente den inn i dataarket vårt. Da ser SPSS-vinduet vårt slik ut:



Da legges dataene inn i dataarket, og vi kommer automatisk inn i datavinduet. Det ser nå slik ut:



og variabelvinduet ser slik ut:



4. 3 Data i ASCII-format. Eksempel: pulse.dat

Det finnes i dag et utall forskjellige dataprogrammer som alle lagrer data på ulike måter. SPSS lagrer også sine data på en særegen måte som bare SPSS benytter. Filer på dette formatet vi kaller vi SPSS-filer.

Imidlertid finnes det heldigvis en internasjonalt akseptert amerikansk standard for datafiler. Den standarden heter ASCII (American Standard Code for Information Interchange). Alle viktige programmer for statistisk analyse kan lese og skrive ASCII-datafiler og således kommunisere med hverandre. Vi skal nå se på hvordan SPSS kan lese ASCII data-filer. Vi skal også navnsette de forskjellige variablene i datasettet vårt og foreta enkel deskriptiv statistikk. Filen vi skal analysere heter **pulse.dat**. Den ligger på kursets hjemmeside. Start med å laste den ned til in katalog på PC-en din.

Dataene på denne filen kommer fra en undersøkelse av mannlige og kvinnelige studenter hvor vi målte pulsen deres før (PULSE1) og etter (PULSE2) en intervensjon, som her var løping. Halvparten av studentene løp under denne intervensjonen (RAN=1), den andre halvparten var i ro (RAN=2). I tillegg ble kjønn, høyde, vekt og røykevaner registrert. Datafilen **pulse.dat** inneholder bare tall skrevet i ASCII-format. Den består av 92 linjer svarende til 92 individer. For hvert individ er det 8 variabler bortover. Variabelnavn og fortolkning framgår av tabellen under.

Variable No.	Description	Name
1	First pulse rate	PULSE1
2	Second pulse rate	PULSE2
3	Running (1 = ran in place. 2 = did not run)	RAN
4	Smoking (1 = smokes regularly. 2 = does not smoke regularly)	SMO
5	Sex (1 = male. 2 = female)	SEX
6	Height in inches	HEI

De 10 første og de 5 siste enhetene på datafilen ser slik ut:

```
64  88 1 2 1 66,00 140 2
58  70 1 2 1 72,00 145 2
```

```
62  76 1 1 1 73,50 160 3
66  78 1 1 1 73,00 190 1
64  80 1 2 1 69,00 155 2
74  84 1 2 1 73,00 165 1
84  84 9 2 1 72,00 150 3
68  72 1 2 1 74,00 190 2
62  75 1 2 1 72,00 195 2
76 118 1 2 1 71,00 138 2
,
,
,
90  92 2 1 2 64,00 125 1
78  80 2 2 2 68,00 133 1
68  68 2 2 2 62,00 110 2
86  84 2 2 2 67,00 150 3
76  76 2 2 9 61,75 108 2
```

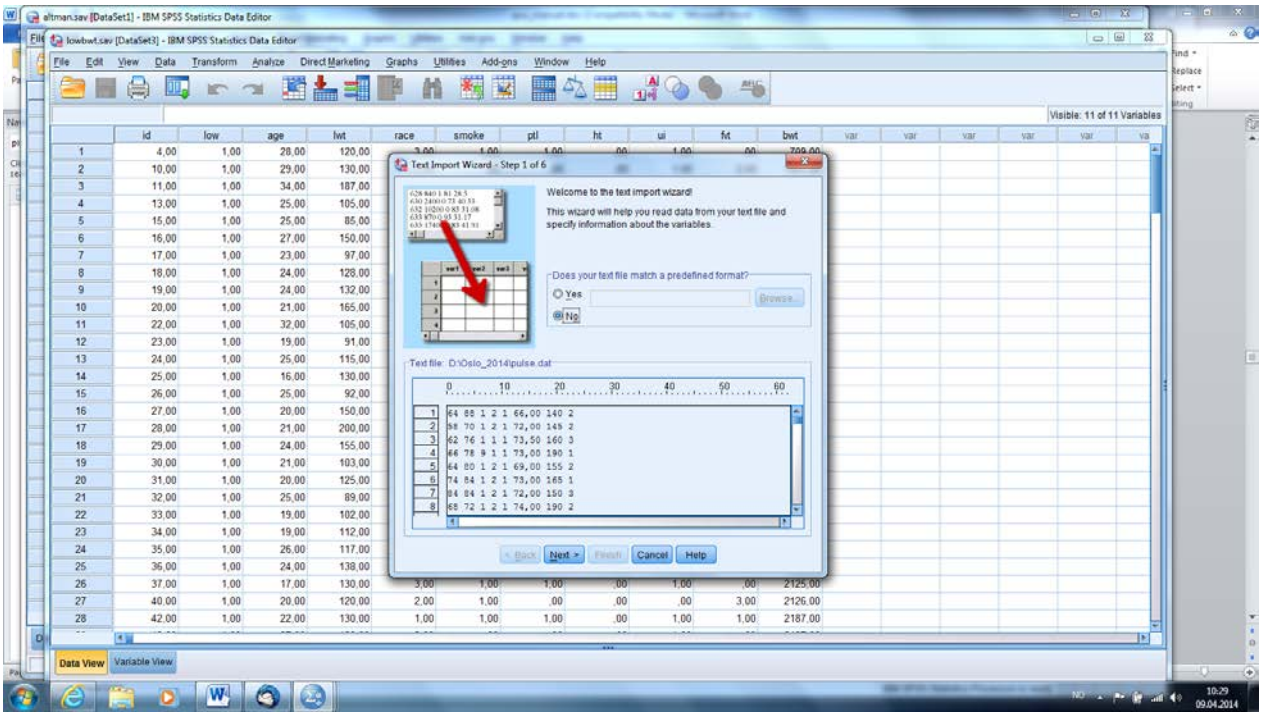
Legg merke til at desimalpunktum i denne filen er satt til komma. Det skyldes at Windows er satt opp med komma som desimalpunktum. Dersom Windows er satt opp med komma, må vi endre alle kommaene til punktumer i datafilen, før vi laster den inn i SPSS.

Denne filen ligger i såkalt **fritt format**. Det er et standard format for ASCII-filer. Det betyr at hver variabel er skilt med (minst) et **blankt tegn.** Det spiller ingen rolle om det er flere blanke tegn mellom variablene, og det spiller ingen rolle at lengden på hver datalinje er den samme. Det motsatte av fritt format er **fast format**. Da må dataene ligge på samme måte i alle datalinjene. Det går vi ikke nærmere inn på her.

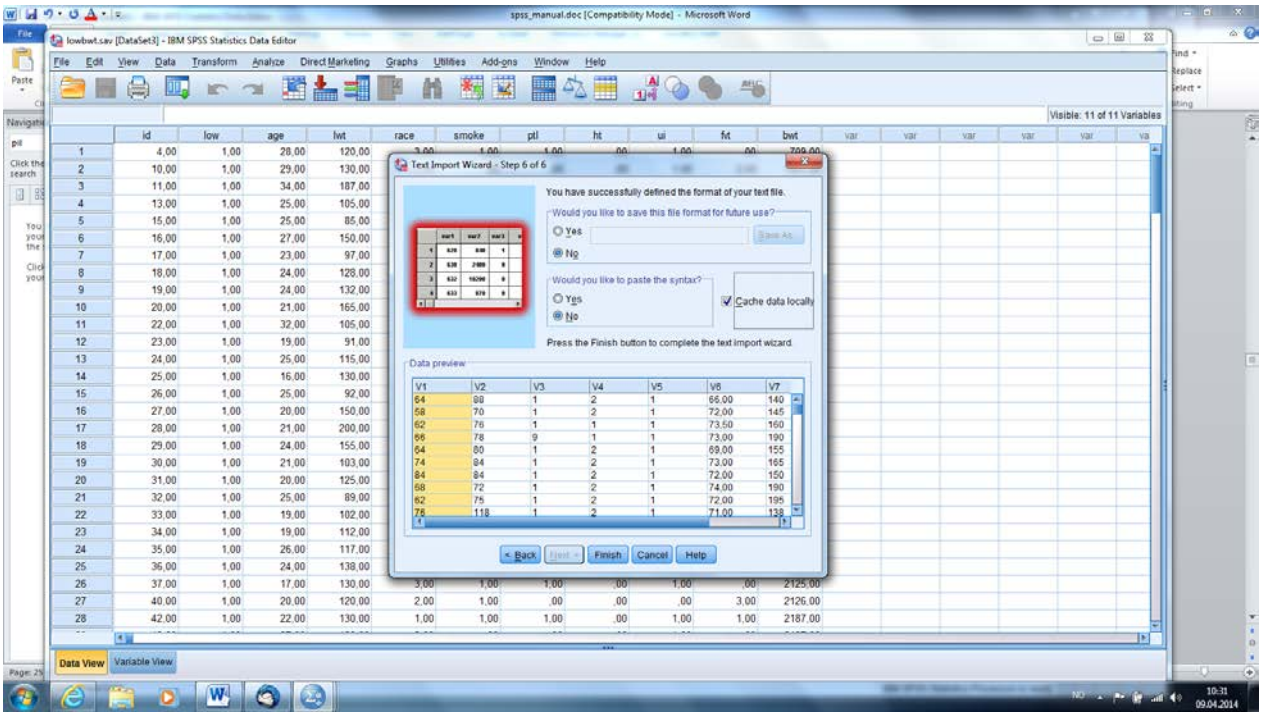
Legg også merke til at hver person (individ) ligger på én datalinje.

Vi går inn i SPSS og kommer igjen til et tomt datavindu. For å lese ASCII-filen **pulse.dat** inn i datavinduet klikker vi på *File/Read Text Data.* Vi går da til den katalogen vi har lagt datafilene våre, og velger da å ta inn datafilen **pulse.dat** derfra.

Vi kommer da inn i SPSS *Text Wizard Import* og må gå gjennom en meny i 6 trinn. Startmenyen ser slik ut:

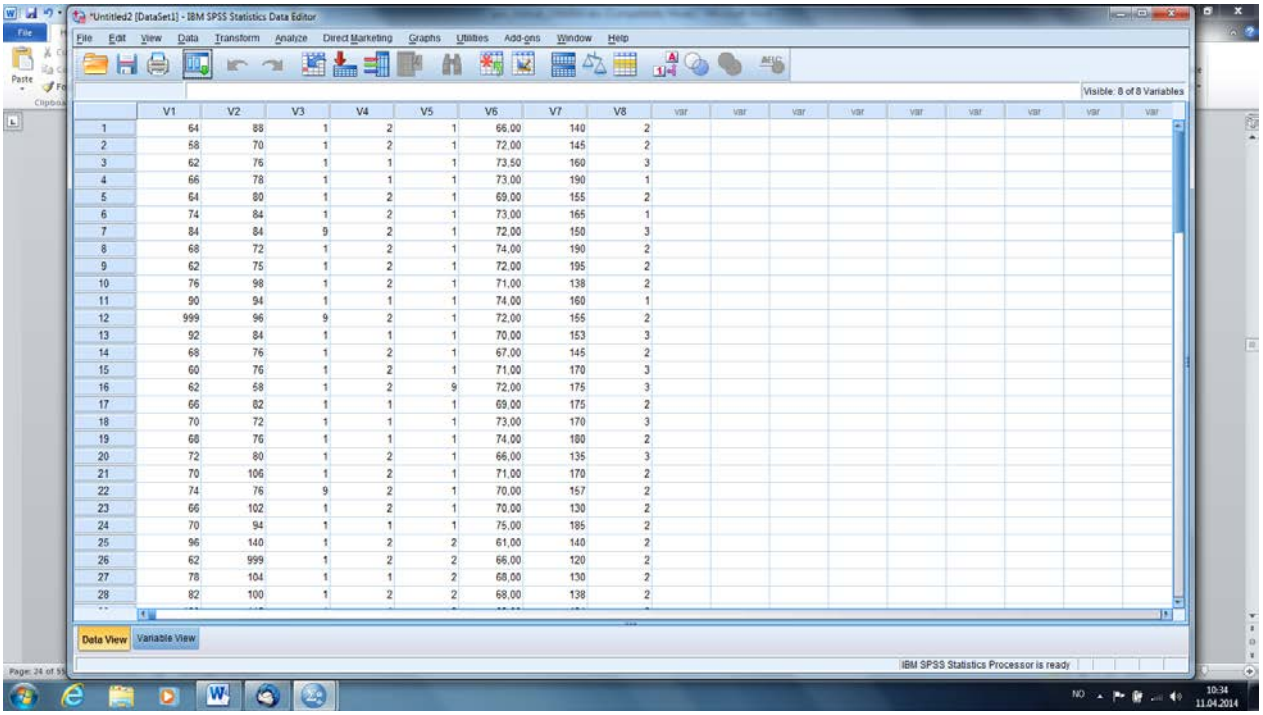


I alle trinnene velger vi de opsjonene som er satt opp av SPSS. På trinn 2 blir vi spurt om *Does your text file match a predefined format?*. Siden vår datafil er i fritt format, lar vi markeringen stå på *No*. Vi går trinn 3 via *Next*. Her blir vi spurt om *How are your variables arranged?* Vi velger *Delimited* siden dataene er delt med blanke. På neste trinn blir vi spurt om to ting: *How are your cases represented?*. Siden hver person ligger på én linje, velger vi også her øverste alternativ. Deretter blir vi spurt om *How many cases do you want to import*. Vi skal importere alle dataene. På neste side blir vi spurt om *Which delimiters appear between the variables*. I fritt format er det *Space* (blank). Når vi kommer til siste trinn velger vi V1 som variabelnavn. Vi har altså valgt *Next* hele veien inntil vi avslutter med *Finish*. Da ser dialogboksen slik ut:



Vi ser at dataene våre ligger i et lite vindu i dialogboksen, men etter at vi er ferdige, må vi likevel sikre oss at innlesingen av datafilen har gått riktig for seg. Vi går inn *Data View* og ser at dataene ligger riktig i kolonner og søyler. Har dette blitt galt, må vi gå tilbake og sjekke at selve datafilen vår er på riktig format. Hvis vi ser at vi har gjort noe galt, må vi starte på nytt med innlesingen av data.

I dataarket ser dataene (når de er riktige) slik ut:



Legg merke til at det i variabelen V1 og V2 ligger 999 flere steder. Dette er én av kodene som brukes for manglende opplysninger, eller *Missing values*. I V3 ligger det også flere steder 9. Andre koder som brukes er for eksempel 99. Men siden 99 er en gyldig verdi for variablene V1 og V2, er det her brukt 999. For V3 er heller ikke 9 en gyldig kode. SPSS tror i utgangspunktet at dette er en gyldig numerisk verdi. Vi må derfor fortelle SPSS at det er den ikke, den er bare en indikator på at det er en *Missing value*. Hvordan vi gjør det skal vi se på i kapittel 5.5, men først skal vi navngi variablene, og gi dem riktige koder.

Vi går over i *Variable View*. Vi ser at alle variablene er blitt numeriske, og at alle dataene er riktig lest inn. Legg spesielt merke til at alle dataene er markert som numeriske. Dersom vi ikke har riktig desimalpunktum, f.eks. punktum istedenfor komma, vil SPSS oppfatte variabelen med desimaltegn, som en tekstvariabel (*String*). Da kan vi ikke bruke SPSS til gjøre statistiske analyser på denne variabelen.

Når vi nå har lest inn datafilen vår, passer vi på å legge den ned som en SPSS-fil i katalogen vår. Vi går da til *File/Save as*. Der velger vi katalogen vi skal legge filen i. Under *File name*: velger vi **pulse**. Merk at SPSS foreslår av filnavnet skal slutte på sav. Det er viktig at vi beholder den ekstensjonen, siden vi da sikrer oss at SPSS kjenner igjen denne filen som en SPSS-fil. Når vi har gjort det, trykker vi på *Save*. Da har vi lagret filen **pulse.sav**.

Vi skal i kapittel 5 gå gjennom hvordan vi legger inn variabelnavn (*Variable name*), et mer utfyllende variabel variabelnavn (*Variabel label*), koder for verdiene på kategoriske variabler (*Value label*) og håndterer manglende verdier (*Missing value*).

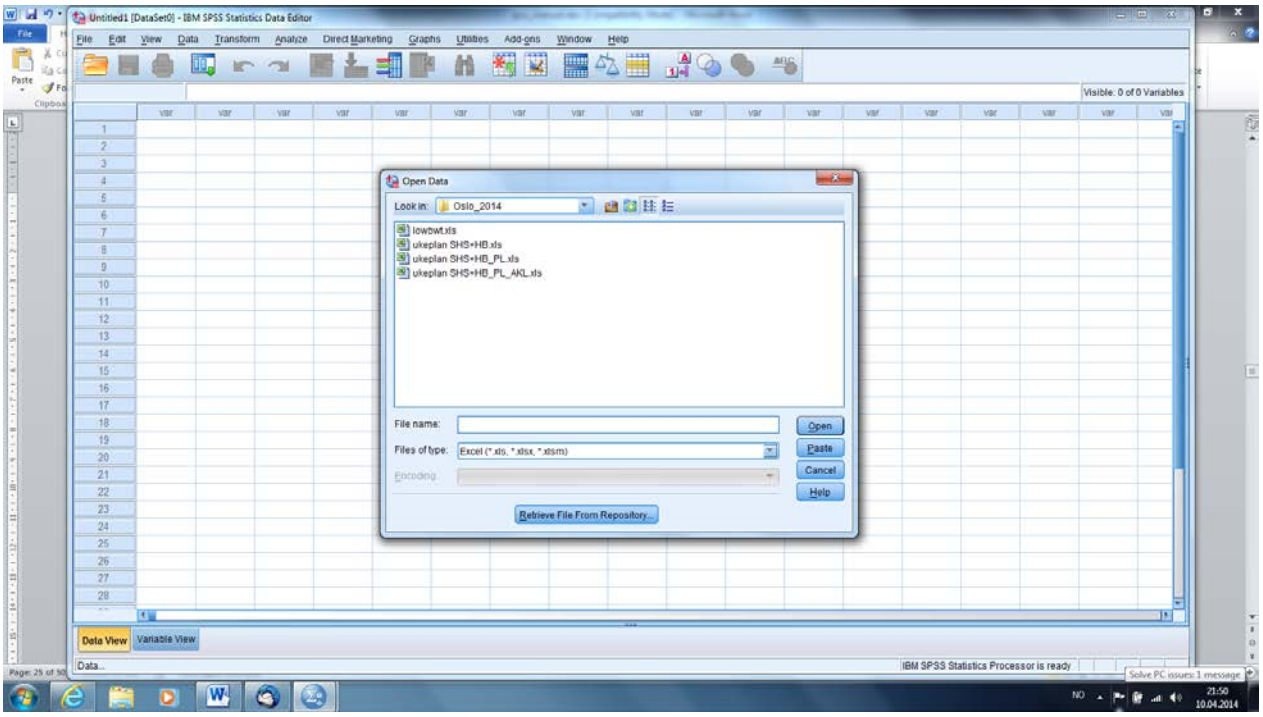
4.4 Innlesing av data fra Excel. Eksempel: lowbwt.xls

Data i Excel-format kan enkelt leses inn i SPSS. Det gjør vi ved å gå til *File/Open*. Under *File Type*: endrer vi fra standardvalget (SPSS) til Excel. Vi ser at det ligger en mengde andre filformater, som det kan være aktuelt å overføre fra (Lotus, SAS, Stata etc.), men for oss er Excel det mest vanlige formatet

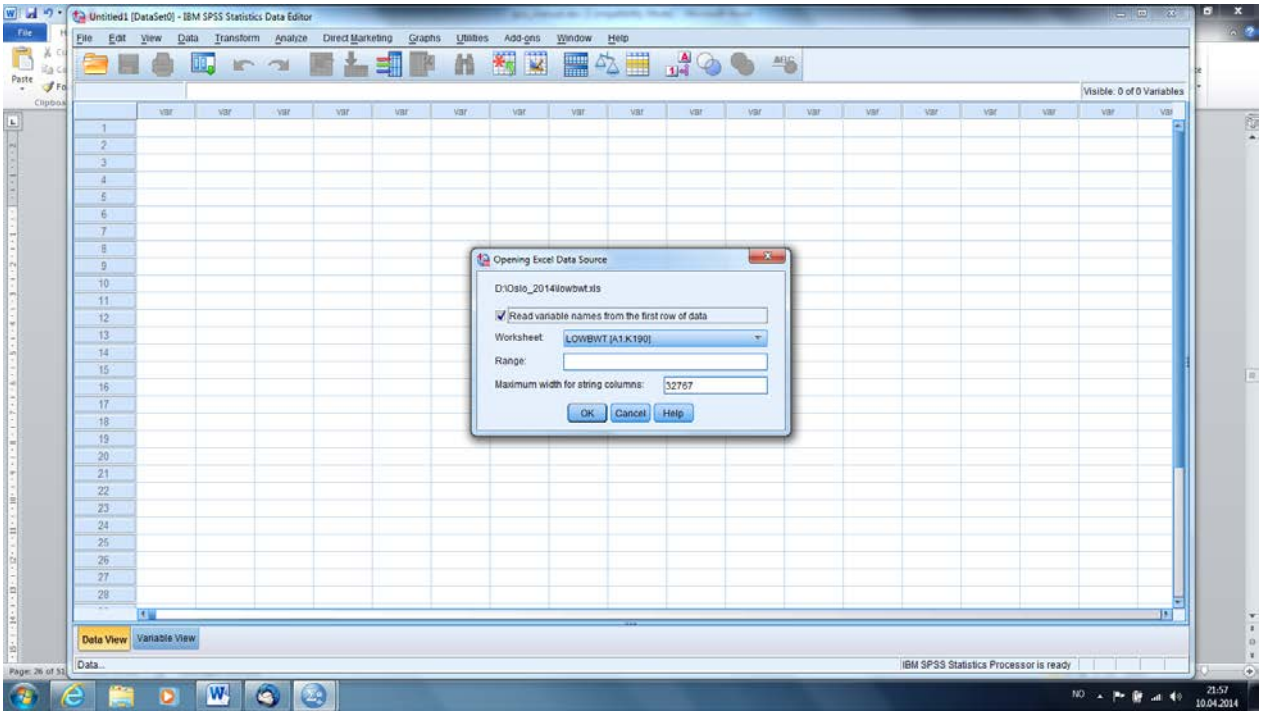
Vi går til den katalogen der filen vår ligger og klikker på den. Den kommer da inn i *File Name*: og vi klikker på *Open*. Da åpner Excel-filen seg i SPSS datavinduet. Hvis variabelnavnene ligger på Excel-filen, vil de også følge med over i SPSS.

Dataene fra Low Birth Weight studien ligger også på Excel-format. Den ligger på hjemmesiden for kurset, og må først legges ned til den katalogen vi har for våre kursfiler. Vi går da til *File/Open* og endrer filtypen til Excel.

Når vi går til riktig katalog, vil vi finne filen **lowbwt.xls**. Da ser dialogboksen våre slik ut:

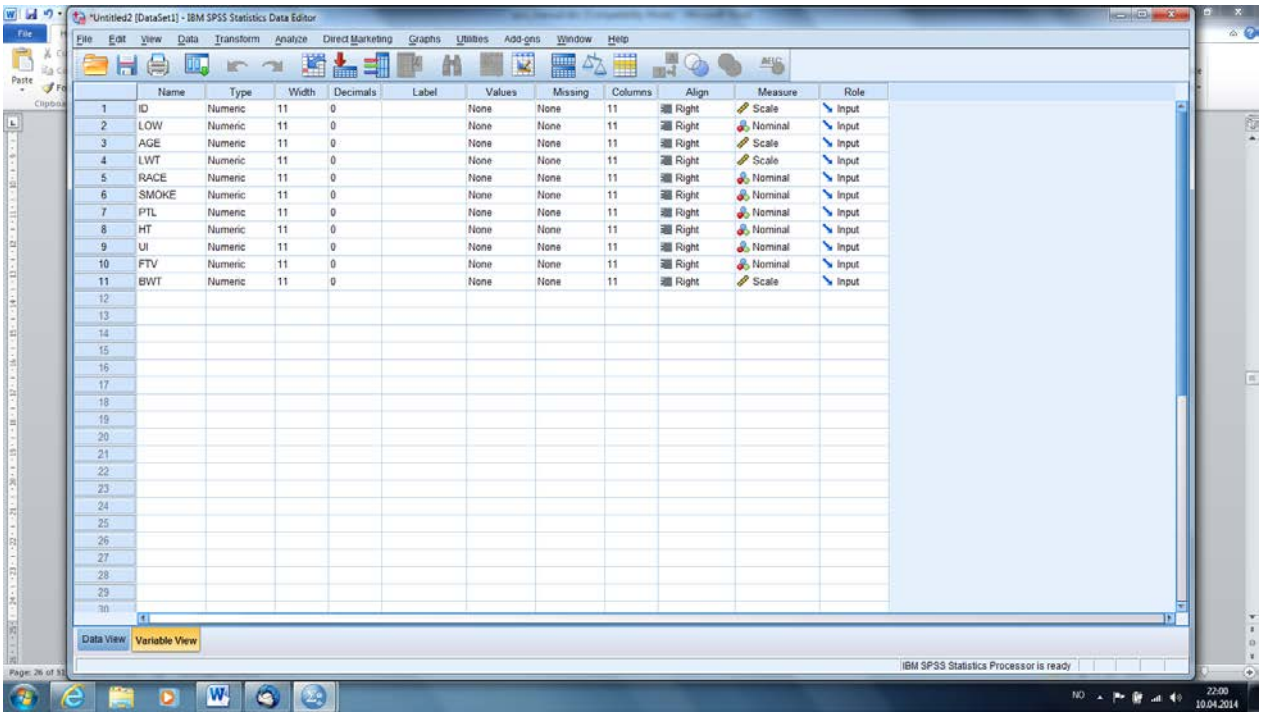


Når vi klikker på *Open*, åpner det seg en ny dialogboks, som vist under:



Vi klikker her på *OK*, og vi får dataene lagt inn i dataarket i SPSS. Legg merke til at alle variabelnavnene følger med, siden de også lå på Excel-filen.

Vi går til *Variable View*, og ser følgende skjermbilde:



Vi ser at alle variablene ligger riktig, men vi mangler *Variable label* og *Value label*. Det er typisk når vi overfører data fra et annet filformat. I slike tilfeller må vi legge inn *Variable label* og *Value label* i SPSS dataarket, som vi skal se på i neste kapittel.

5 Databearbeiding 1

Læringsmål

I dette kapittelet skal vi lære hvordan vi gir variablene navn (*Variable name*), variabelbeskrivelser (*Variable label*) og hvordan vi lager verdikoder (*Value label*) til variablene på dataarket. SPSS bruker var00001, var00002, var00003 osv. som variabelnavn, og det er ikke særlig hensiktsmessig når vi skal gjøre analyser. Det er også viktig at vi legger inn verdikoder for kategoriske variabler siden vi for senere bruk av datafilen da har kodene «hengt» på variablene, slik at vi ikke går surr i kodene.

I dette kapittelet skal vi også gjøre vår første statistiske analyse. Til slutt i dette kapittelet skal vi vise hvordan vi kan overføre utskrifter fra SPSS analyser til en Word-fil. Det er nyttig når vi skal skrive rapporter eller artikler.

5.1 Variable name

Det er helt nødvendig at vi gir variablene navn. SPSS bruker var00001, var00002, var00003 osv. som variabelnavn, og det er selvfølgelig helt umulig å holde orden på, spesielt når vi flere SPSS-filer åpne eller liggende på PC'en. Vi lager gjerne litt korte variabelnavn, siden vi har

muligheten til å gi en mer utfyllende beskrivelse under *Label*. SPSS tillater variabelnavn med æ, ø og å, men ikke mellomrom.

Vi får tilgang til *Name, Label* osv. ved å gå fra *Data View* over til *Variable View*. Begge disse ligger nederst på SPSS-arket. Når vi er i *Variable View* har vi en menylinje som inneholder informasjon om navn, antall desimaler etc. La oss se litt nærmere på den: Den inneholder

Name Type Width Decimals Label Values Missing Columns Align Measure Role

Name angir navnet, i kortform.

Type angir typen på variabel. Her er det Numeric som er viktig for oss. Vi vil bruke data som er numeriske. Merk at dersom vi leser inn data med galt desimaltegn, vil denne variablen settes som String. Da må vi endre desimaltegnet vårt før vi leser inn data på nytt.

Width angir antall posisjoner som dataene opptar.

Decimals angir antall posisjoner etter desimaltegnet

(Variable) Label angir variabelnavnet mer utfyllende enn det som står i Name. Det er dette som fremkommer i utskriften for denne variabelen. Vi kommer tilbake til denne i neste avsnitt.

Values angir navnet på kategoriene i en kategorisk variabel. Vi kommer også tilbake til dette i nest avsnitt.

Missing angir hvilke verdier som er nummeriske verdier for uoppgitt (missing). Vi kommer tilbake til dette

Columns angir antall posisjoner som dataene opptar.

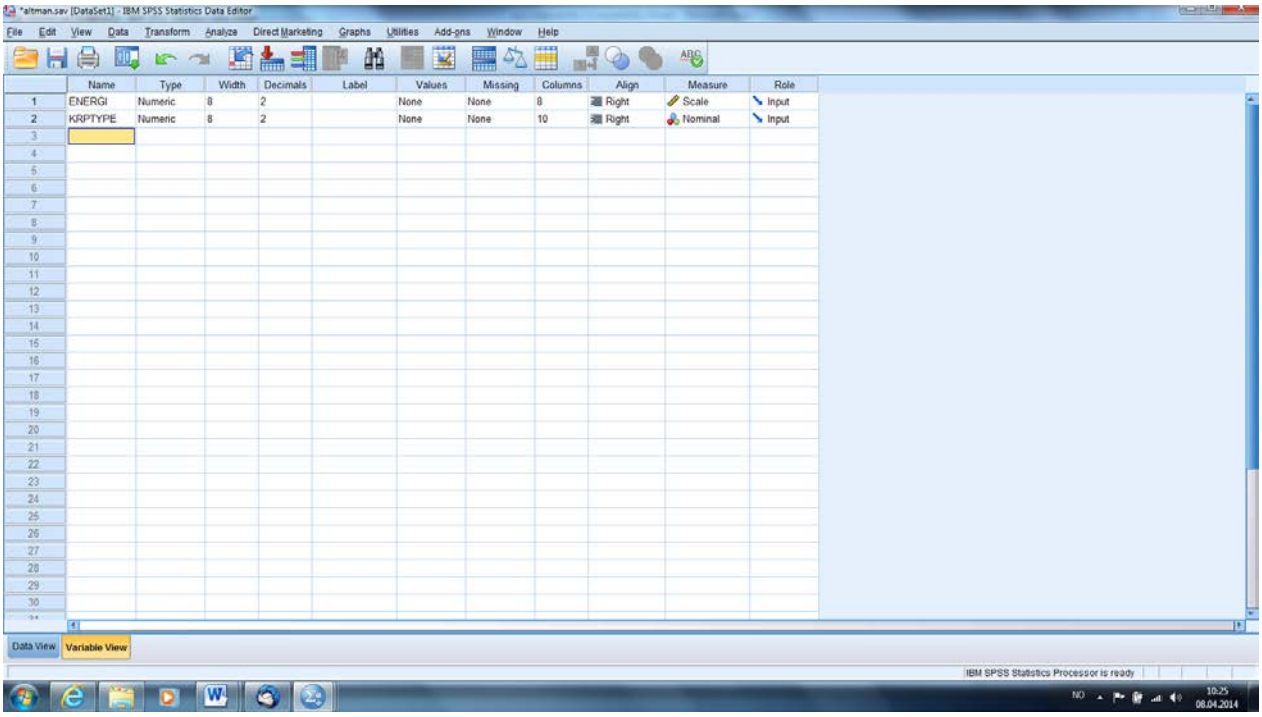
Align angir om dataene er høyrejuster (Right) eller venstrejustert (Left). Nummeriske data er høyrejustert.

Scale angir hva slagt type data vi har. I SPSS kan vi angir tre typer data, Scale angir at det en kontinuerlig variabel, ordinal angir at det er ordinal variabel og nominal at det er en kategorivariabel. I kapittel 3 gikk vi gjennom de forskjellige datatypene.

5.1.1 Eksempel: altman.sav

Vi går til SPSS-filen som vi har fra **altman-datene** som vi lagde i kapittel 4. Hvis disse dataene ikke ligger på dataarket vårt, må vi hente filen **altman.sav** fra den katalogen der vi lagret den. Vi kan nå enten dobbeltklikke på toppen av variabelens kolonne eller ved å klikke på *Variable View* nederst på dataarket. I begge tilfeller kommer vi inn i en oversikt over variablene, med en rekke kolonner. For dette eksemplet går vi bare gå inn i kolonnen med *Name*. Der skriver vi inn navnene på variablene var00001 og var00002. Dette betyr at vi skriver inn ENERGI for energivariabelen og KRPTYPE for kroppstypevariabelen for de to

variablene under *Name*. Vi skal komme tilbake til de andre kolonnene i *Variable View* etterhvert. Da vil *Varible View* se slik ut:

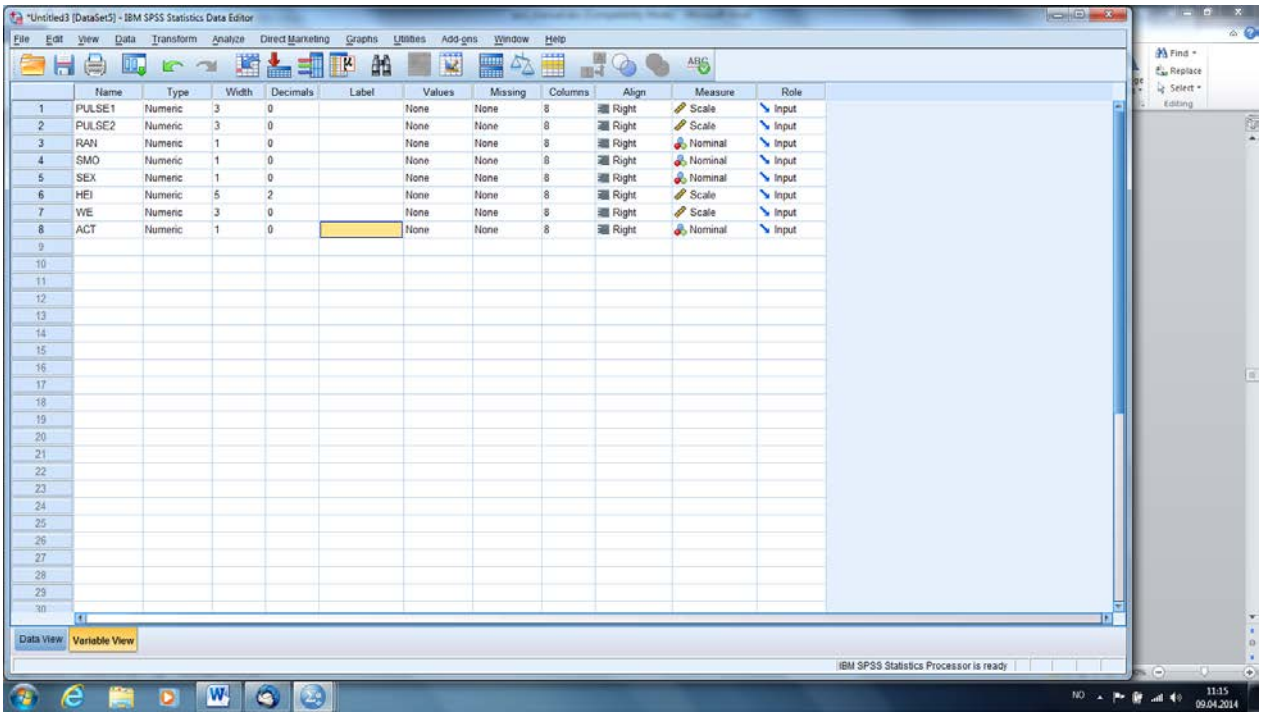


5.1.2 Eksempel: pulse.sav

I kapittel 4 leste vi inn datafilen **pulse.dat**, men vi lagret den som en SPSS-fil med navnet **pulse.sav**. Det er denne vi skal bruke nå. Når vi skal legge inn variabelnavn på denne datafilen, bruker vi dem som står i beskrivelsen av datafilen i kapittel 4, dvs. vi skriver inn PULSE1, PULSE2 etc. Vi forandrer på navnene ved å klikke på *Variable View* nederst på dataarket. Da kommer vi inn i en oversikt over alle variablene, med navn osv. Her skriver vi nå inn PULSE1 istedenfor V1 og PULSE2 for V2. Slik fortsetter vi gjennom alle variablene. Husk at alle variablene må legges inn i den rekkefølgen de ligger i på datafilen, og som da er oppgitt i listen over. SPSS foreslår i utgangspunktet (default) at vi har numeriske variable. Det stemmer for alle våre variable.

Vi ser også på *Type*, som ligger ved siden av *Name*, og undersøker om type variabel er blitt riktig. Den skal for alle variable være *Numeric*. Dersom vi går inn på feltet med *Numeric*, ser vi at det blir åpnet et lite grått område med prikker inne i. Dersom vi klikker på dette feltet kommer det opp hvilke alternativer vi har for datatyper. Gjør vi det, ser vi de ulike alternativene, med antall posisjoner og antall desimaler angitt.

Når vi er ferdige med variabelnavnene, ser *Variable View* slik ut:



5.2 Variable label

Vanligvis vil *Variable name* gi en kort beskrivelse av variabelen. *Label* gir oss muligheten til å angi en lengre beskrivelse enn den som fremkommer fra første kolonne, *Name*. Variabelnavnene blir derfor ofte litt ufullstendige. Når vi er i *Variable View*, har en menylinje som ser slik ut

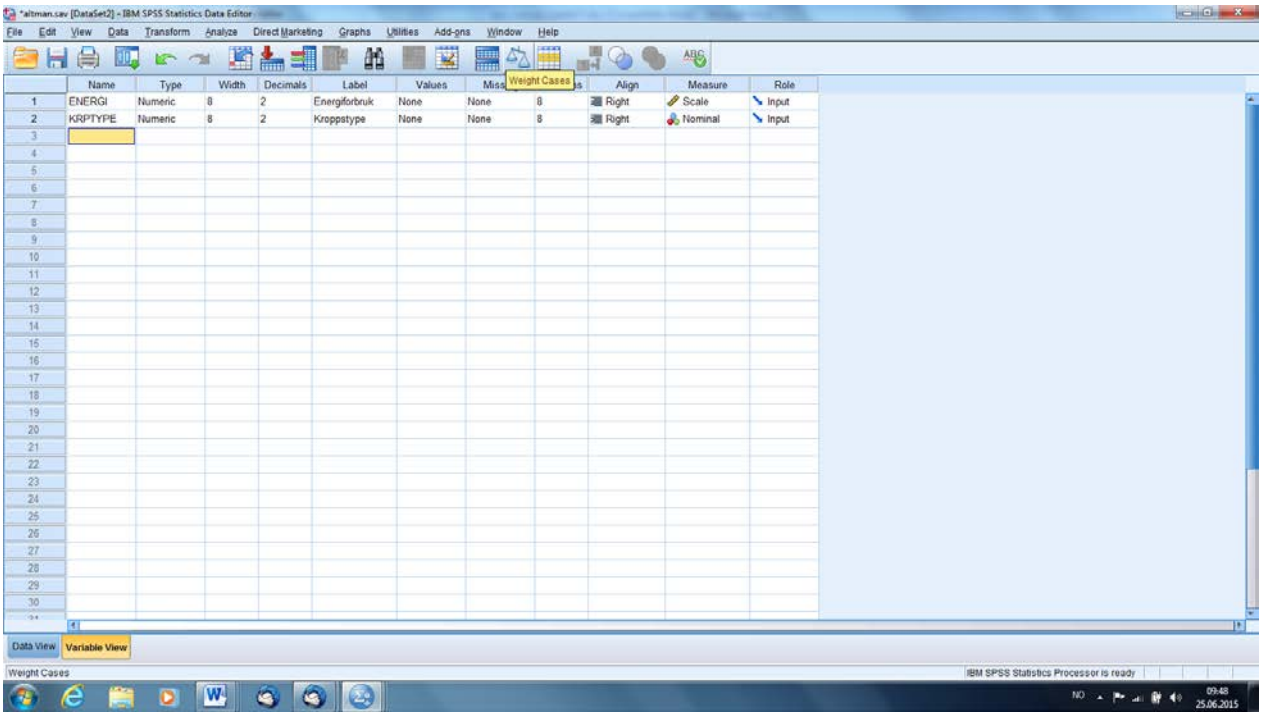
Name Type Width Decimals Label Values Missing Columns Align Measure Role

For å lage skikkelige variabelnavn går vi videre bort til *Label*. *Label* gir oss mulighet til å gi en nærmere beskrivelse av variabelen med flere enn 8 tegn, også æ/ø/å kan brukes. Denne må ikke forveksles med *Values* som ligger deretter.

5.2.1 Eksempel: altman.sav

I eksempelet med altman-dataene kan det være nyttig å beskrive ENERGI ved at det er energiforbruket. Vi skriver derfor inn Energiforbruk i *Label* for variabelen ENERGI. Tilsvarende skriver vi inn Kroppstype for variabelen KRPTYPE.

Da ser *Variable View* slik ut:

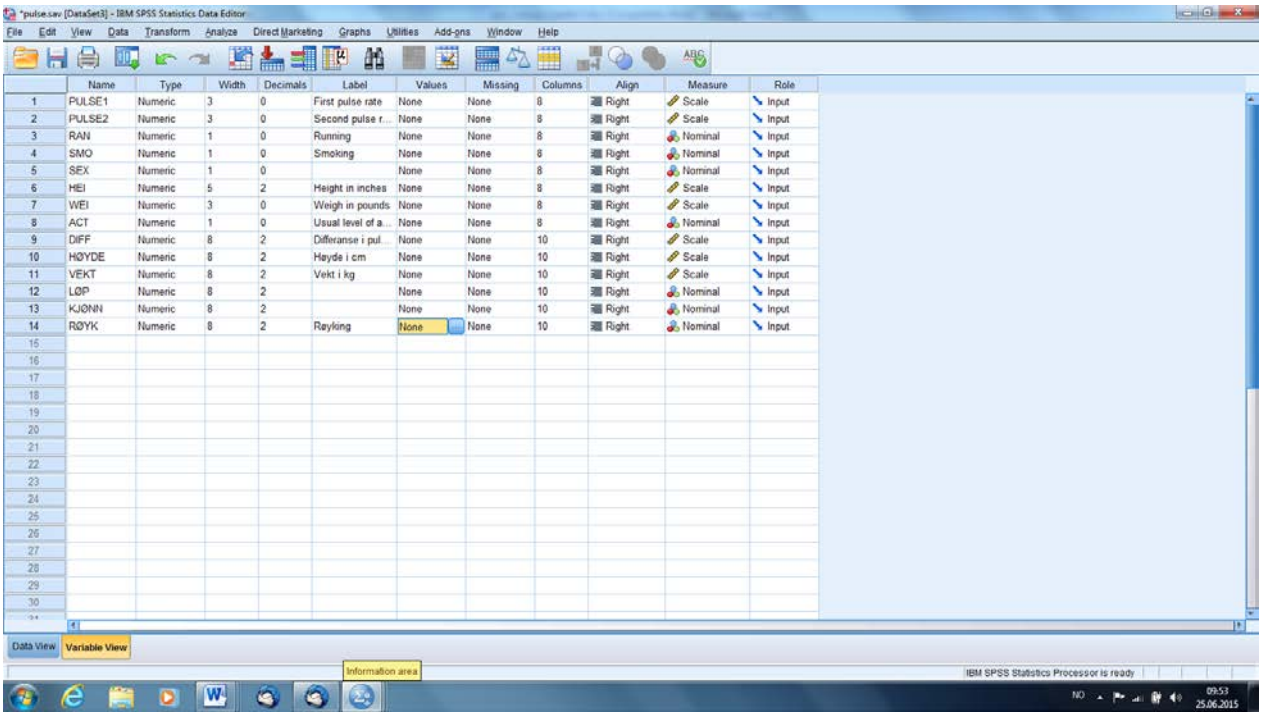


5.2.2 Eksempel: pulse.sav

Vi velger å lage *Label* for alle variablene våre, og følger den oversikten vi har fra selve beskrivelse av PULSE.DAT

Vi starter med å gå til feltet *Label* for variablen HEI. Her skriver vi inn Height in inches. Tilsvarende gjør vi for WEI der vi skriver inn Weight in pounds. For variablen ACT skriver vi inn Usual physical activity.

Da ser *Variable View* slik ut:



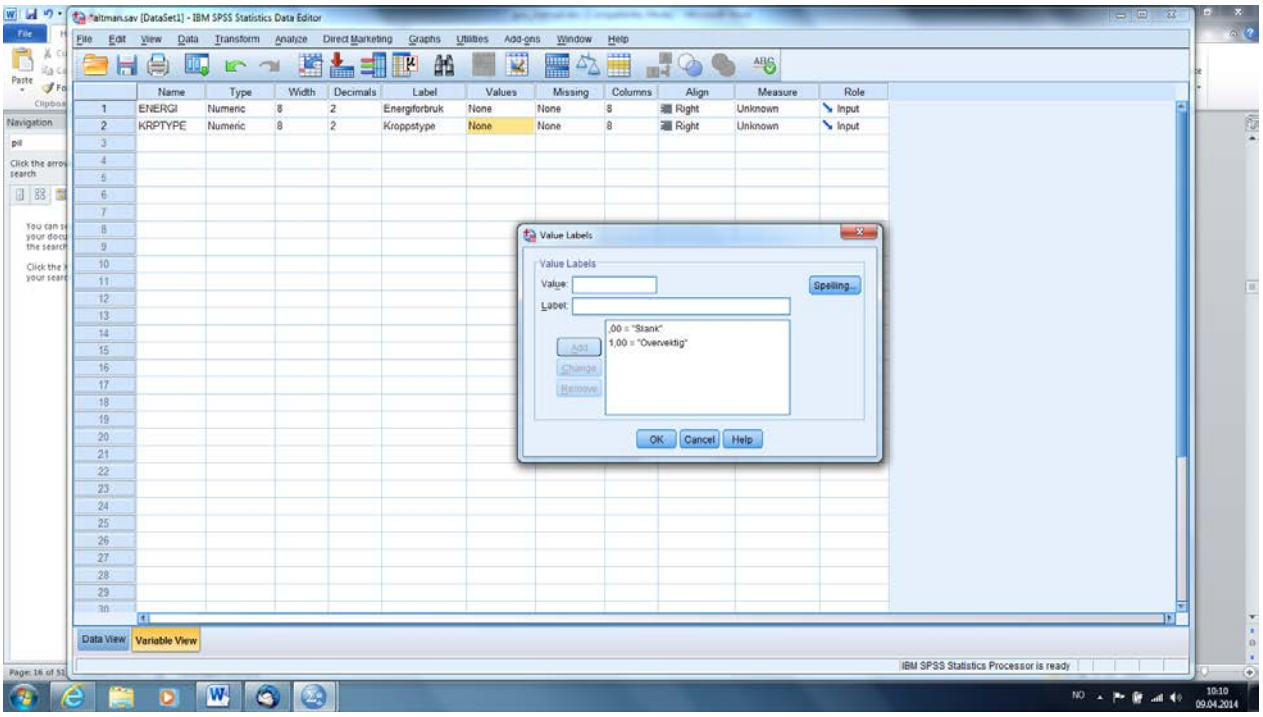
5.3 Value label

Variablene kan være kontinuerlige eller kategoriske. For de kategoriske variablene er det viktig at vi angir hva de forskjellige kodene angir. Vi bruker alltid numeriske koder, slik som 0, 1, 2 osv. Men vi glemmer fort hva de numeriske kodene betyr, og det er derfor viktig at vi angir hva de ulike numeriske kodene angir. Det mest vanlige eksemplet er at vi bruker kodene 0 og 1 for henholdsvis kvinner og menn. Da må vi angi til SPSS at 0 betyr kvinne og 1 betyr mann. Ved å gjøre dette via *Values* vil SPSS oppgi disse verdikodene når vi får utskrifter fra analysene våre. Merk at vi er i *Variable View*, får vi Value Labels ved å gå til *Values*. *Label* bruker vil til å angi en beskrivelse av selve variabelen. Den ligger på linjen med

Name Type Width Decimals Label Values Missing Columns Align Measure Role

5.3.1 Eksempel: altman.sav

I SPSS-filen **altman.sav** har variabelen KRPTYPE to verdier, 0 for de slanke og 1 for de overvektige. Vi må angi til SPSS at det er er slik de definert. Det gjør vi via *Values.*Vi går da til *Values* med pilindikatoren vår, og klikker i Value-feltet for variabelen KRPTYPE. Da åpner det seg et farget felt med prikker i (…). Vi klikker inn i det feltet. Da åpner det seg enda en ny dialogboks som heter *Value Labels*. I value skriver vi inn 0 og i *Label* skriver vi inn Slank. Da åpner det seg en ny knappetast, *Add*. Den klikker vi på. Etter det går vi tilbake til value og skriver inn 1, og i *Label* skriver vi inn Overvektig. Da ser dialogboksen i SPSS slik ut



Vi klikker på *OK*, og kodene blir lagt til på variabelen KRPTYPE.

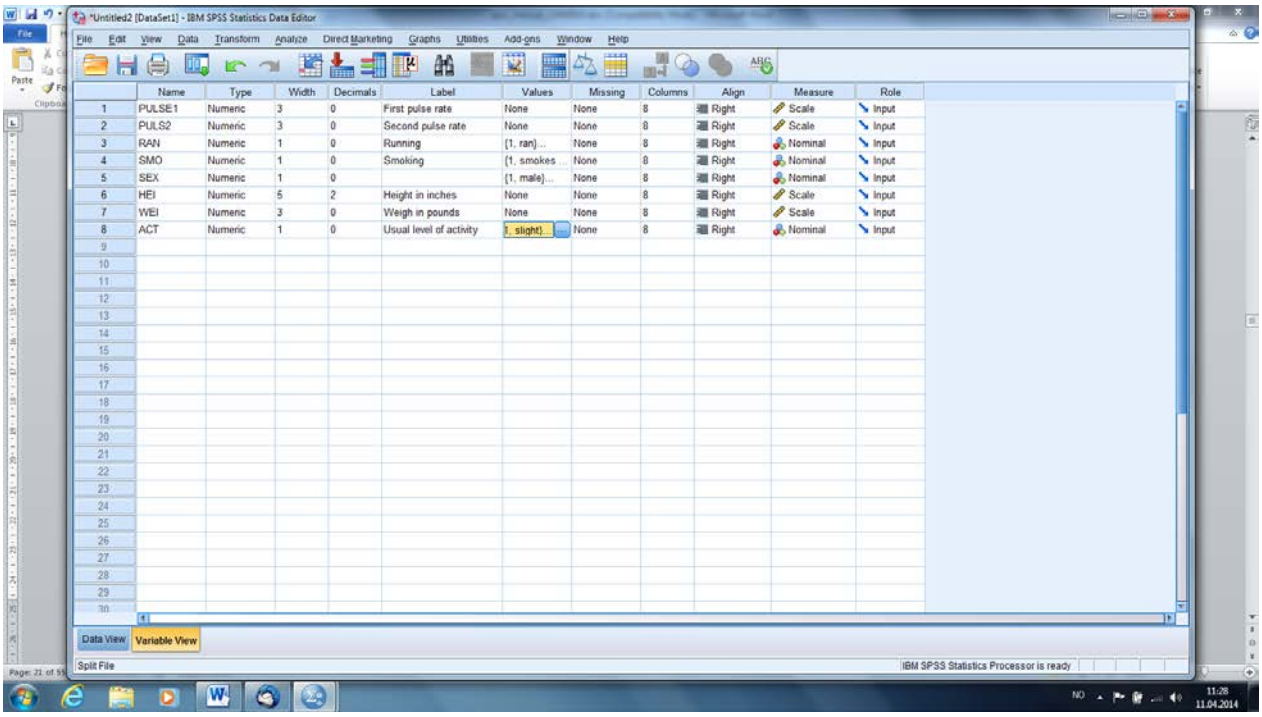
5.3.2 Eksempel: pulse.sav

Values gir oss tilsvarende mulighet til å fortelle programmet hva enkelte verdier på en variabel betyr, f.eks. at SEX = 1 betyr mann, og SEX = 2 betyr kvinne. For kontinuerlige variabler som PULSE1 trenger vi ikke noen *Value label*, men for kategorivariablene (variabler som antar noen distinkte verdier) som SEX, RAN, SMO og ACT er det nyttig med *Value label*.

Values gir oss altså en mulighet til å fylle ut forklaringer på de ulike verdiene på hver variabel. Dette gjøres ved å gå inn i feltet *Values* for den variabelen vi er interessert i. Da kommer det opp et grått felt som vi klikker på. Her kommer det så opp en ny undermeny til å definere de ulike *Values*. Etter innskriving av en *Value* og den tilsvarende *Value label* må vi trykke på *Add* for å få det registrert.

Vi gir nå utfyllende navn til alle variablene i samsvar med tabellen først i dette kapittelet. For å gjøre dette for variabelen SEX, må vi da gå inn i linjen med variabelen SEX og bort til kolonnen med *Values*. Vi klikker da på det grå feltet som kommer frem. Da kommer undermenyen frem. Vi legger så inn 1 på *Values* og skriver inn male på *Value label*. Deretter må vi trykke på Add. Så går vi opp til *Values* og skriver inn 2 og female på *Value label*. Deretter må vi først trykke på *Add*, så kan vi trykke på *OK*. Da ser vi at kodene er lagt inn i variabelarket. Slik fortsetter vi med de andre variablene. F.eks. kan vi gå til variabelen ACT og legge inn 1 på *Values* og skrive inn slight på *Value label*, så klikke på *Add*, deretter 2 og skrive moderate, så *Add*, og til slutt 3 på *Values* og skrive inn a lot på *Value label*, så Add. Til slutt trykker vi på *OK*. På denne måten kan vi ta å legge *Value label* på alle de variablene vi ønsker.

Etter dette ser dataarket, under *Variable View* slik ut:



5.4 Den første statistisk analysen. Eksempel: altman.sav

Når vi er i dataarket med filen **altman**, er vi klare til å gjøre våre første dataanalyser. I denne omgang skal vi bare lage en frekvensoversikt for variabelen KRPTYPE og en beregning av gjennomsnitt og standardavviket for ENERGI. Vi starter med frekvensfordelingen for KRPTYPE. Da går vi til *Analyze/Descriptive statistics/Frequencies.* Da kommer vi inn i en dialogboks. Der trekker vi KRPTYPE over i boksen *Variable(s).*

Da får vi følgende resultat:

Kroppstype

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Slank	13	59,1	59,1	59,1
	Overvektig	9	40,9	40,9	100,0
	Total	22	100,0	100,0	

Vi ser – som vi vet – at vi har 13 slanke og 9 overvektige i datafilen vår.

For å beregne gjennomsnittet og standardavviket, slik det er presentert i Altmans bok, klikker vi på *Analyze/Descriptive statistics/Descriptives.* Da kommer vi inn i en underdialogboks, og her må vi velge hvilke variabler vi vil ha et gjennomsnitt for fra boksen til venstre, og flytte dem over i boksen til høyre ved å klikke på pilen i midten. Vi velger ENERGI som markeres til venstre og flyttes til høyre med pilen. Analysen utføres når vi trykker på *OK*. Vi kommer

nå automatisk over i Output-vinduet hvor vi ser resultatet. Her får vi vite middelverdi og standardavvik av ENERGI for alle individene samlet. Da får vi følgende resultat:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
ENERGI	22	6,13	12,79	8,9791	1,69750
Valid N (listwise)	22				

Her ser vi at gjennomsnittet av energiforbruket er 8.98, med maksimum lik 12.79 og minimum like 6.13. Vi ser også at vi får beregnet standardavviket (Std. Deviation). Hva standardavviket uttrykker og hvordan vi skal fortolke resultatet kommer vil tilbake til senere i dette kurset.

For å få vite verdiene for gruppene enkeltvis, kan vi velge *Analyze/Descriptive/Explore.* Da kommer vi inn i en dialogboks, og her trekker vi ENERGI over i *Dependent List* og KRPTYPE over i *Factor List*. For å komme ut av denne dialogboksen klikker vi på *Continue*. Til slutt klikker vi på *OK*-knappen for å få analysen utført. Her er det veldig mye utskrift. Mye av dette kommer vi tilbake til. Men i dette innledende eksempelet viser vi bare utskriften av den deskriptive analysen:

Descriptives

	KRPTYPE			Statistic	Std. Error
ENERGI	,00	Mean		8,0662	,34338
		95% Confidence Interval for Mean	Lower Bound	7,3180	
			Upper Bound	8,8143	
		5% Trimmed Mean		8,0174	
		Median		7,9000	
		Variance		1,533	
		Std. Deviation		1,23808	
		Minimum		6,13	
		Maximum		10,88	
		Range		4,75	
		Interquartile Range		,77	
		Skewness		1,161	,616
		Kurtosis		1,768	1,191
	1,00	Mean		10,2978	,46596
		95% Confidence Interval for Mean	Lower Bound	9,2233	
			Upper Bound	11,3723	
		5% Trimmed Mean		10,2431	
		Median		9,6900	
		Variance		1,954	
		Std. Deviation		1,39787	
		Minimum		8,79	
		Maximum		12,79	
		Range		4,00	
		Interquartile Range		2,48	
		Skewness		,849	,717
		Kurtosis		-,719	1,400

I SPSS har vi nå to åpne vinduer. I *IBM SPSS Statistics Data Editor* har vi liggende selve dataene våre. I *IBM SPSS Statistics Viewer* har vi liggende utskriftene våre. Det er nå viktig at vi lagrer disse to filene for senere bruk. Da kan vi hente frem datafilen til senere analyser, og vi kan hente frem utskriftsfilen for å se på de analysene som vi har gjort til nå.

5.5 Missing values

I (praktisk talt) alle undersøkelser vil vi ha manglende opplysninger for noen variabler. I kliniske studier kan dette skyldes at forsøkspersonene ikke møter opp til undersøkelser, eller at man ikke klarer å få gjennomført én eller flere av de undersøkelsene som inngår i studien. I epidemiologiske studier kan personer nekte å svare på enkelte spørsmål i undersøkelsen.

Når vi registrerer data inn på SPSS dataarket, vil vi angi manglende opplysninger med kodene 9, 99, 999 osv. avhengig av hvilke koder som er gyldige koder for de enkelte variablene.Hvis vi ikke angir i SPSS at dette er koder for manglende data, vil SPSS selvfølgelig regne dem

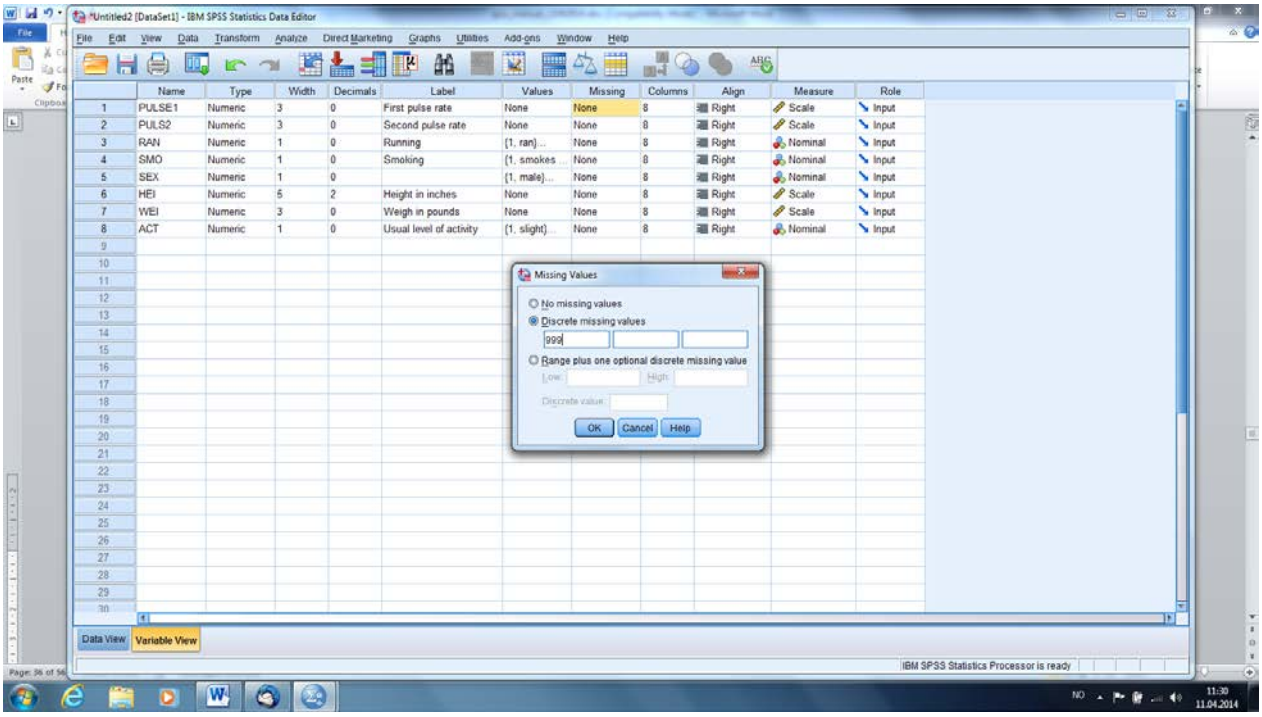
som gyldig koder, og bruke dem i de statistiske analysene. Det må vi selvfølgelig forhindre. Når vi er i *Variable View*, angir vi kodene for manglende data ved å gå til *Missing* på linjen med

Name Type Width Decimals Label Values Missing Columns Align Measure Role

5.5.1 Eksempel: pulse.sav

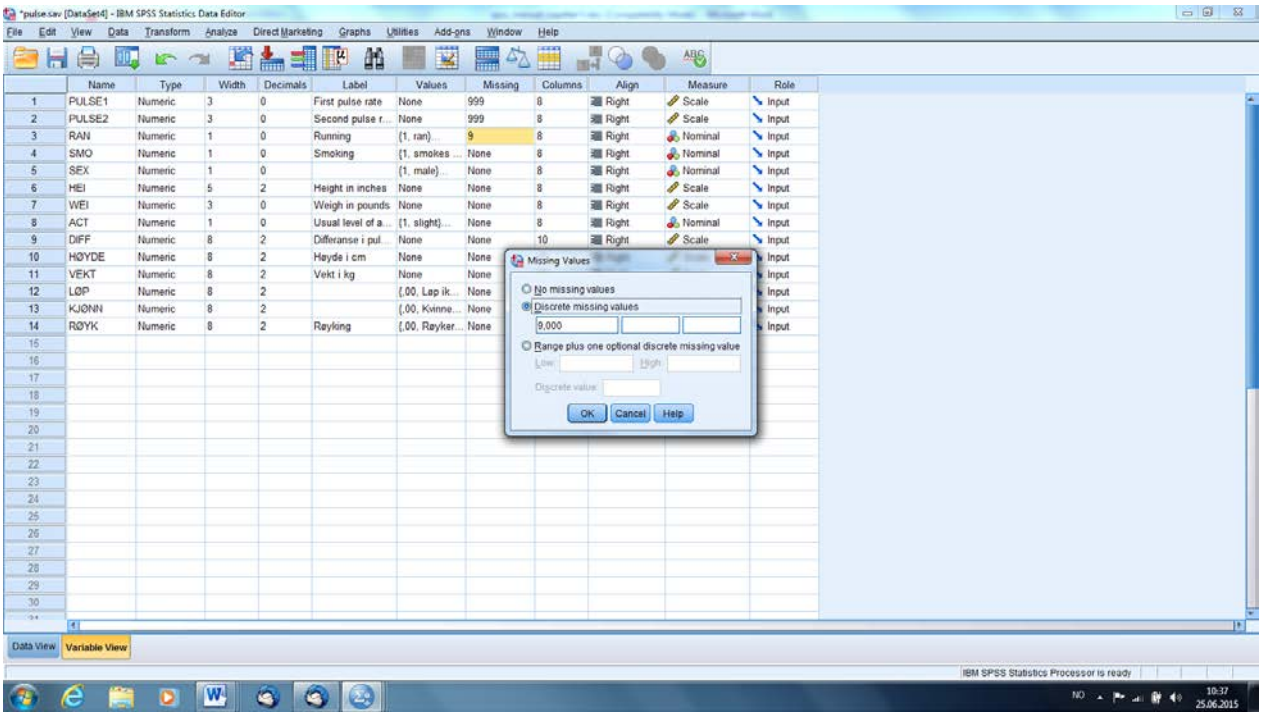
Vi går til datasettet **pulse.sav**. I vårt data materiale er 9 koden for *Missing* for variabelen RAN og 999 er koden for *Missing* for PULSE1 og PULSE2.

Vi legger inn denne informasjonen ved å gå til kolonnen med *Missing* for hver av disse variablene. Vi starter med PULSE1. Ved å klikke på det grå feltet, kommer det opp en underdialog. Her velger vi *Discrete missing value*, DISCRETE MISSING VALUE og legger inn 999 som *Discrete missing value*. Dette skrives inn i første boksen. Da ser skjermbildet vårt slik ut:

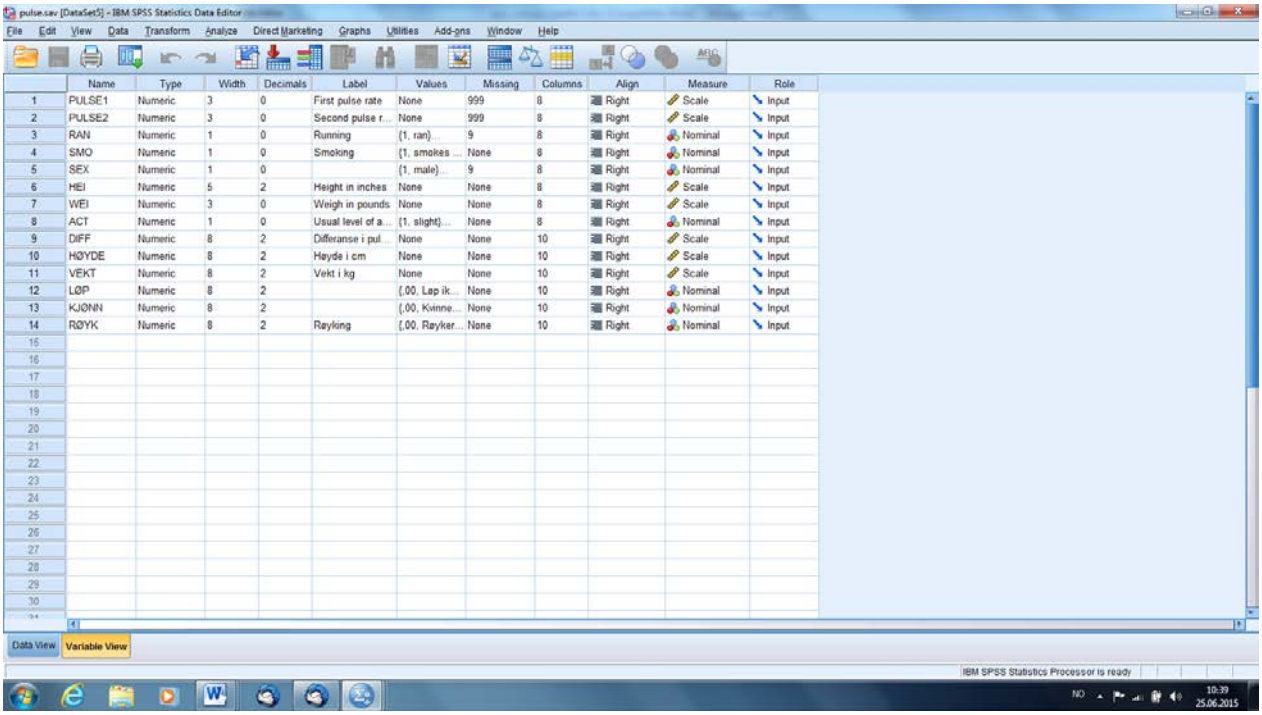


Deretter avslutter vi med *OK*. Denne prosedyren gjentar vi for PULSE2.

For RAN legger vi inn 9 som *Discrete missing value.* Da ser skjermbildet vårt slik ut:



Vi gjør tilsvarende for SEX, der også koden for *Missing* er 9. Når vi er ferdige med alle de fire variablene som har manglende opplysninger, ser skjermbildet for *Variable View* slik ut:



Da er vi ferdige med å definere filen **pulse.sav**. . Siden vi har gjort endringer på filen, gjenstår det å lagre filen. Det gjør vi på vanlig måte ved å gå til *File/Save As* og gi filen navnet **pulse.sav**.

5.6 Den andre statistiske analysen. Eksempel: pulse.sav

Vi lar **pulse.sav** være vår aktive fil. Vi kan nå lage en liten statistisk analyse av variabelen RAN. Dette er en kategorisk variabel, og vi fremstiller fordelingen til den gjennom en frekvensoversikt. Vi går da til *Analyze/Descriptive Statistics/Frequencies* og legger RAN over i *Variable(s).* Da får vi følgende utskrift:

Running

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	ran	32	34,8	36,8	36,8
	did not run	55	59,8	63,2	100,0
	Total	87	94,6	100,0	
Missing	9	5	5,4		
Total		92	100,0		

Legg merke til at vi har fått angitt at koden 9 er Missing, og at da totalt er 87 gyldige data for denne variablen. Mker at vi her skal bruke prosentfordelingen som er gitt under Valid Percent, siden det er den som regner ut prosentfordelingen UTEN Missing values. Vi kommer tilbake til dette når vi begynner på de statistiske analysene i kapitttel 19.

5.7 Lagring av filer. Eksempel: altman.sav og pulse.sav

Vi skal nå lagre filene med altman-dataene og pulse-dataene. Det er nødvendig siden vi har lagt på mye ny informasjon knyttet til dataene. Vi starter med altman-dataene, og passer på at vi har den som vår aktive fil, dvs. at det er den vi ser på skjermen. På hovedmenylinjen til SPSS velger vi nå *File/Save as*. Vi bruker fortsatt navnet **altman.sav.** Den nye versjonen av **altman.sav** inneholder ikke bare talldataene våre, men også de navnene og de kodene vi har gitt dem.

Når det er gjort gjør vi utskriftsvinduet til vårt aktive vindu. Igjen så går vi til *File/Save as*. Da åpner det seg en ny dialogboks, denne gangen for utskriftsfilen. Legg merke til at SPSS nå foreslår av filnavnet skal ende på **spv.** Vi beholder den ekstensjonen og velger å kalle filen for **altman1.spv.** Da husker vi at dette er den første utskriftfilen til analysene på altman-filen.

Vi gjør tilsvarende for pulse-dataene våre. Vi passer på at pulse-dataene er våre aktive data. Vi går da til *File/Save as* og legger denne filen ned i den katalogen vi har våre kursfiler, velger **pulse**under *File Name*: og klikker på *OK*. Da har vi fått lagd en oppdatert versjon av SPSS-filen **pulse.sav**.

Hensikten med å lage disse filene til SPSS-filer er at vi vi ved senere bruk kan hente frem dataene som analysefiler. Når vi har lagt på variabelnavn (Variable name), variabelbeskrivelser (Variabel label) og verdikoder (Value label) blir disse liggende på variablene i det vi har lagt dem ned på våre PC som en SPSS-fil.

Når disse filene er lagt trygt ned i en katalog på vår datamaskin, kan vi evt. gå ut av SPSS, med *File/Exit*. Hvis vi ikke har lagt ned filene til vårt eget område, gir nå SPSS oss en advarsel om å gjøre det.

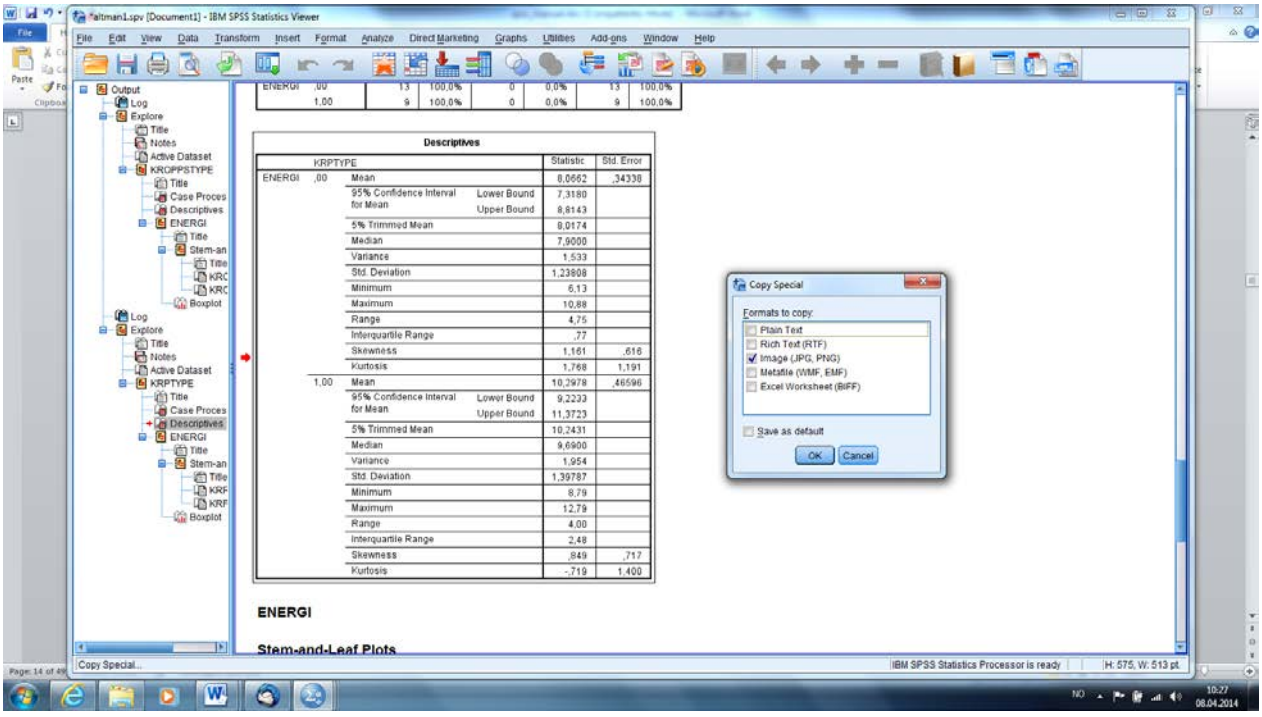
Men vi kan gå videre med å åpne en ny datafil. Hvis vi har **altman.sav** åpen i ett vindu, vil vi nå få åpnet den andre filen i et annet vindu.

5.8 Overføring av utskrifter fra SPSS til Word. Eksempel: altman.sav

Det er nyttig å kunne overføre utskrifter fra analyser fra SPSS utskriftsfilen til en Word-fil. Det gjør vi på følgende måte. Når vi er inne i en SPSS utskriftsfilen, klikker vi må det bildet vi skal ha overført. Så går vi opp på menyen *Edit/Copy Special*. Da kommer det opp et nytt dialogvindu, *Copy Special* vinduet. Der klikker vi av på *Image (JPG, PNG)*, og deretter *OK*.

Deretter går vi over i Word. Der vi hvor vi ønsker å ha bildet kopiert inn, legger vi markøren. Da trykker vi på *Ctrl/v* som er tastene for å kopiere inn fra minnet. Da kommer bildet fint inn i Word-filen.

Vi går over til utskriftsfilen fra analysen vi gjorde for **altman.sav.** Når utskriften er inne i SPSS, går vi ned til utskriften for *Descriptives*. Så går vi til menyen *Edit/Copy Special*. I dialogvinduet *Copy Special* vinduet klikker vi av på Image (JPG, PNG), og deretter *OK*. Da har vi følgende bilde på skjermen vår:



Da går vi over til Word-filen vår. Ved å trykke *Ctrl/v*, får vi lagt følgende utskrift inn i tekstfilen vår:

Descriptives

	KRPTYPE			Statistic	Std. Error
ENERGI	,00	Mean		8,0662	,34338
		95% Confidence Interval for Mean	Lower Bound	7,3180	
			Upper Bound	8,8143	
		5% Trimmed Mean		8,0174	
		Median		7,9000	
		Variance		1,533	
		Std. Deviation		1,23808	
		Minimum		6,13	
		Maximum		10,88	
		Range		4,75	
		Interquartile Range		,77	
		Skewness		1,161	,616
		Kurtosis		1,768	1,191
	1,00	Mean		10,2978	,46596
		95% Confidence Interval for Mean	Lower Bound	9,2233	
			Upper Bound	11,3723	
		5% Trimmed Mean		10,2431	
		Median		9,6900	
		Variance		1,954	
		Std. Deviation		1,39787	
		Minimum		8,79	
		Maximum		12,79	
		Range		4,00	
		Interquartile Range		2,48	
		Skewness		,849	,717
		Kurtosis		-,719	1,400

6 Databearbeiding 2

Læringsmål

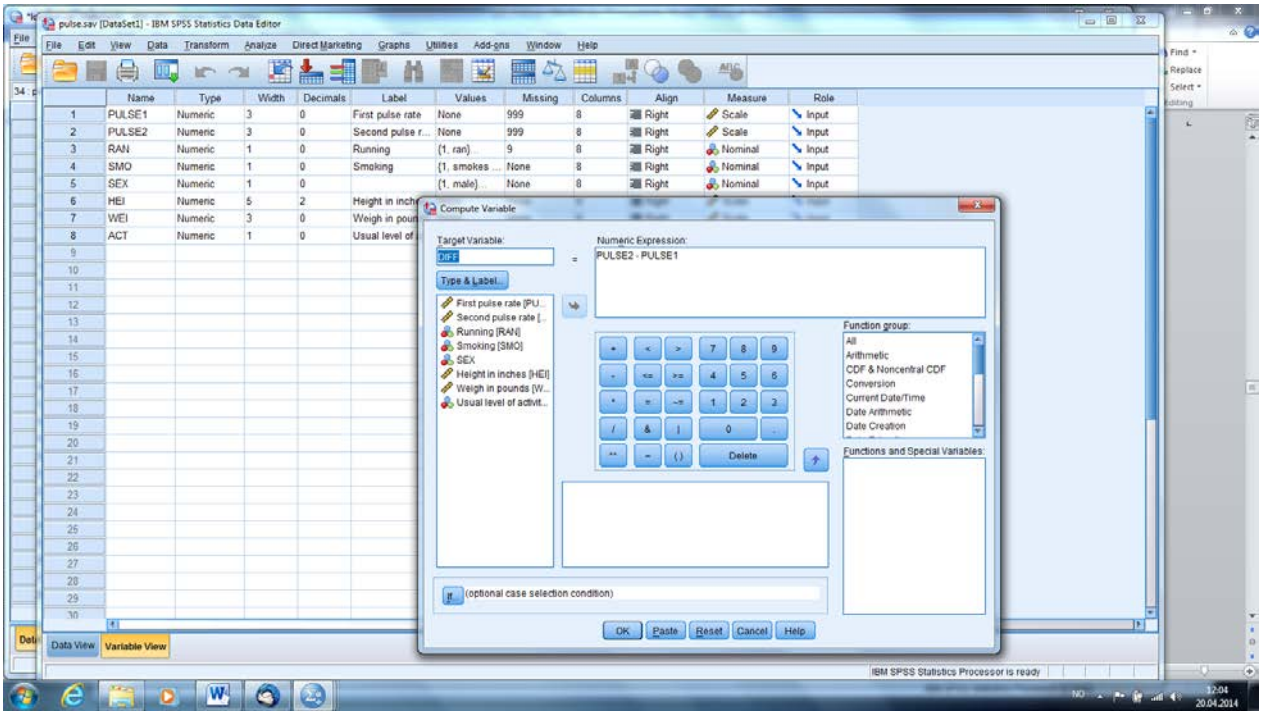
De viktigste funksjonene innenfor databearbeiding er å lage variabler som er transformert av andre variabler eller omkodet fra andre. Vi skal i dette kapittelet konsentrere oss om kommandoene *Compute* som lager en transformert variabel, fra andre variabler, via en gitt funksjon. Vi skal også se på kommandoen *Recode* som gir oss mulighetene for å omkode verdier på variablene.

De kommandoene vi skal bruke i dette kapittelet finner vi under *Transform/Compute*, *Transform/Recode into Same Variables* og *Transform/Recode into Different Variables* Merk at i *Transform/Compute* brukes tegnet * for multiplikasjon og ** for eksponering. Divisjonstegnet er /. Under *Function Group* ligger en rekke funksjoner som kan brukes.

6.1 Compute. Eksempel: pulse.sav

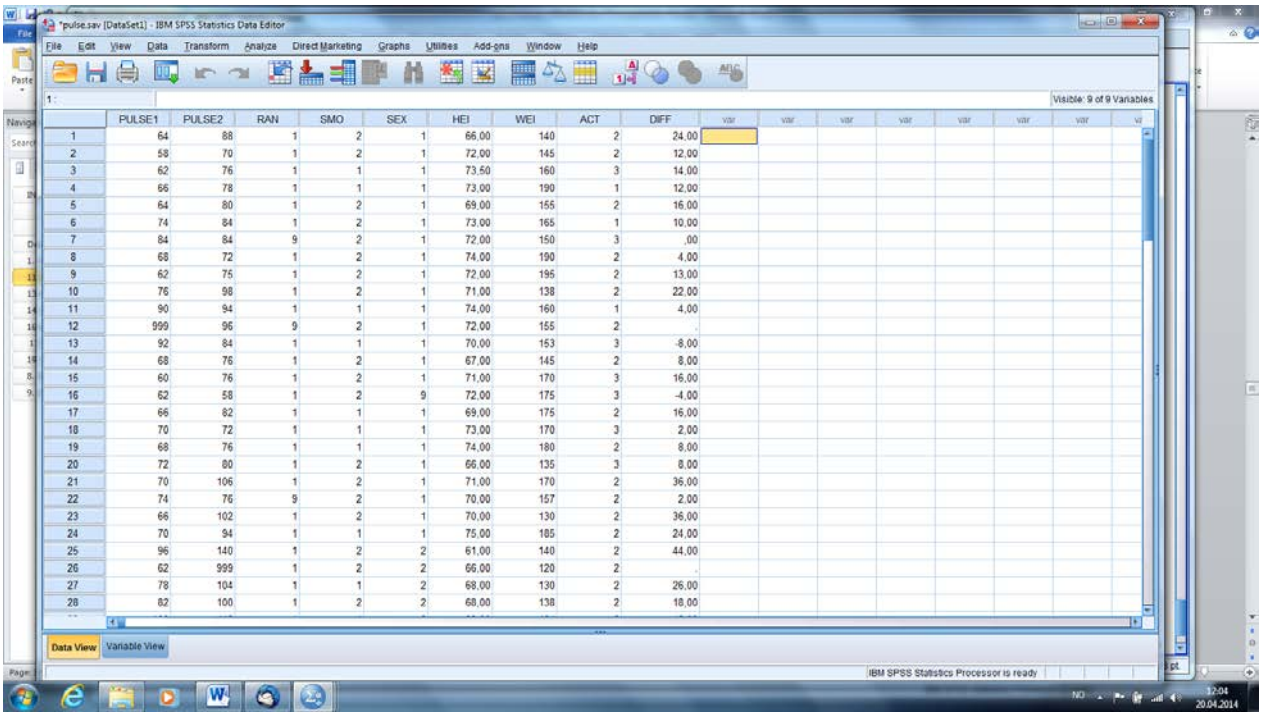
Vi skal bruke dataene fra **pulse.sav** for å vise hvordan vi enkelt kan avlede nye variabler og foreta omkodinger av dem vi allerede har. Dataene ligger lagret som en SPSS datafil med navnet **pulse.sav**. Hvis den ikke ligger på dataarket, henter vi denfrem ved å klikke på *File/Open* og oppgi navnet.

For senere analyser skal vi beregne differansen mellom PULSE2 og PULSE2. Ved bruk av dialogboksen innenfor hovedmenyknapp *Transform/Compute* får vi dette til. Ved å klikke på *Transform/Compute* kommer vi inn i en nye dialogboks. Vi skal lage en ny variabel DIFF som skal være differansen mellom PULSE2 og PULSE1. Da skriver vi inn DIFF som Target variable. Så merker vi av PULSE1 og trekker den over i boksen til høyre. Vi skriver inn så inn – (minustegnet), enten ved å skrive det fra tastaturet eller å bruke tegnet i dialogboksen. Så trekker vi over PULSE1 i boksen til høyre. Da ser dialogboksen vår slik ut:



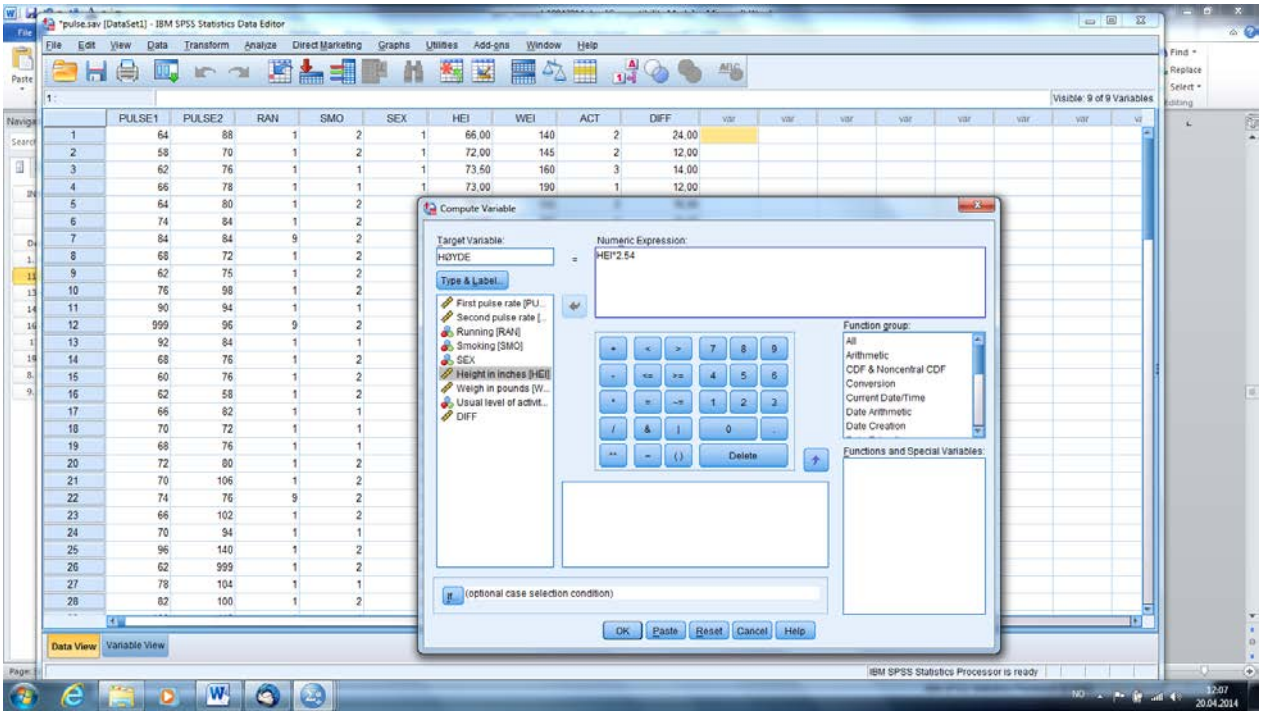
Vi klikker på *OK*. Da ender vi i utskriftsvinduet, der det er angitt at vi har fått beregnet DIFF. Vi går over i dataarket, med *Data View*, og ser at variabelen DIFF er lagt inn som ny variabel, med sine beregnede verdier.

Legg merke til at det ligger et punktum der det er *Missing values*. Dette er SPSS sin interne kode for manglende data. Datafilen ser da slik ut:



Vi skal fortsette å bruke **pulse.sav**. Vi skal nå omkode WEI (vekt i pounds) til VEKT (i kg) og HEI (høyde i inches) til HØYDE (høyde i cm). Omregningsfaktoren er 1 inch = 2.54 cm, og 1 pound = 0.45 kg.

Igjen bruker vi *Transform/Compute* og kommer inn i dialogboksen for omkoding. Der skriver vi inn HØYDE som *Target variable*. Så merker vi av HEI og den over i boksen til høyre. Vi skriver inn *2.54 rett etter HEI. Da ser dialogboksen ut som under:



Merk at vi her bruker punktum som desimalpunktum her, siden vi her inne i SPSS. Velger vi galt desimaltegn, vil SPSS si i fra om dette, og vi kan bare skifte i formelen i *Numeric Expression*. Vi klikker på *OK*. Da ender vi i utskriftvinduet, der det er angitt at vi har fått beregnet HØYDE. Vi går over i dataarket, med *Data View*, og ser at variabelen HØYDE er lagt inn som ny variabel, med sine beregnede verdier. Merk at på dataarket ligger dataene for HØYDE med komma som desimalpunktum.

Vi gjør det samme med å beregne VEKT i kg. Da skriver vi inn VEKT som *Target variable*, og merker WEI og flytter den over i boksen til høyre. Vi skriver inn *0.45, siden omregningsfaktoren mellom pound og kg er 0.45 og trykker *OK*. Når vi går til datavinduet, ser vi at VEKT lagt til etter HØYDE.

Når vi har gjort disse transformasjonene, er det viktig at vi får lagt disse til filen vår. Vi går derfor til *File/Save as*, og velger fortsatt navnet **pulse.sav** og legger den katalogen der den allerede ligger. Når vi klikker på *Save*, får vi beskjed om at filen allerede finnes og om vi vil erstatte (replace) den. Det vil vi ja til (yes).

6.2 Compute. Eksempel: lowbwt.sav

I datafilen **lowbwt.sav** ligger mors vekt, LWT, også angitt i pund. Barnets vekt er angitt i gram. Vi henter frem **lowbwt.sav**, og transformer på samme måte som i forrige eksempel. Vi skal omregne LWT til kg. Da skriver vi inn LWTKG som *Target variable*, og merker LWT og flytter den over i boksen til høyre. Vi skriver inn *0.45, siden dette er omregningsfaktoren mellom pound og kg er 0.45 og trykker *OK*. Når vi går til datavinduet, ser vi at LWTKG er lagt til som siste variabel.

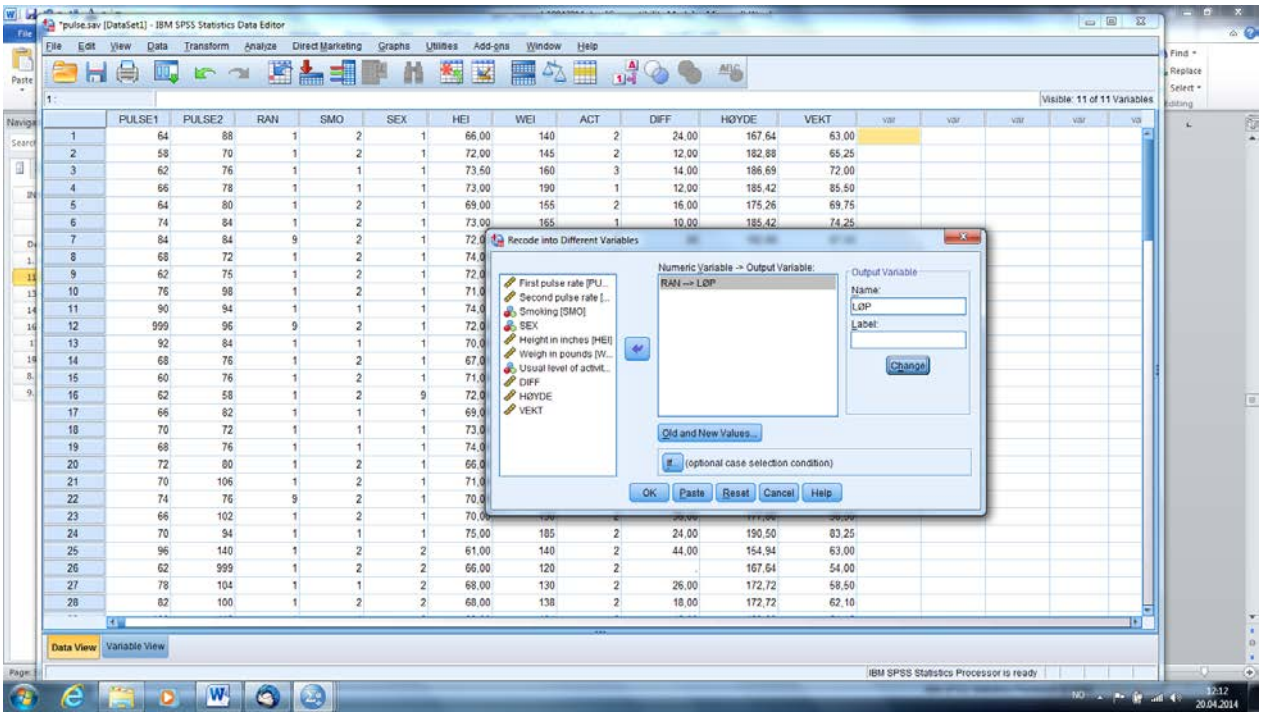
6.3 Recode. Eksempel: pulse.sav

Som vi allerede har nevnt i kapittel 4 om datatyper er det i statistiske analyser hensiktsmessig å la dikotome variable ha verdiene 0 og 1. Vi skal derfor omkode variablene RAN, SMO og SEX på den måten. Da kan vi:

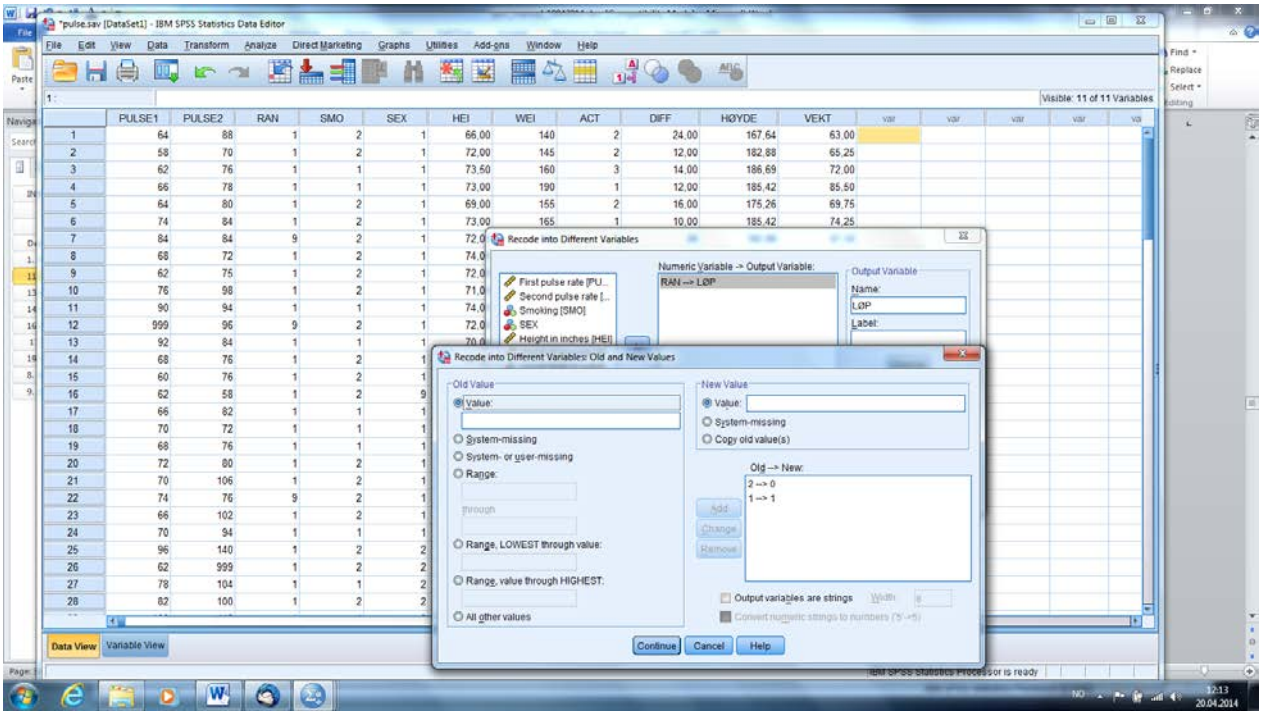
1. enten lage nye variabler som er kodet slik vi vil og beholde de gamle uforandret
2. endre kodingen av den gamle variabelen

SPSS gjør slike omkodinger gjennom kodeordet RECODE (tilfelle 1) og RECODE INTO (tilfelle2). Det er mest hensiktsmessig å bruke metode 2, siden vi da beholder den gamle variable, samtidig som vi får lagd de nye. Vi vil derfor holde oss til denne metoden. Vi skal nå omkode RAN til LØP, SEX til KJØNN og SMO til RØYK.

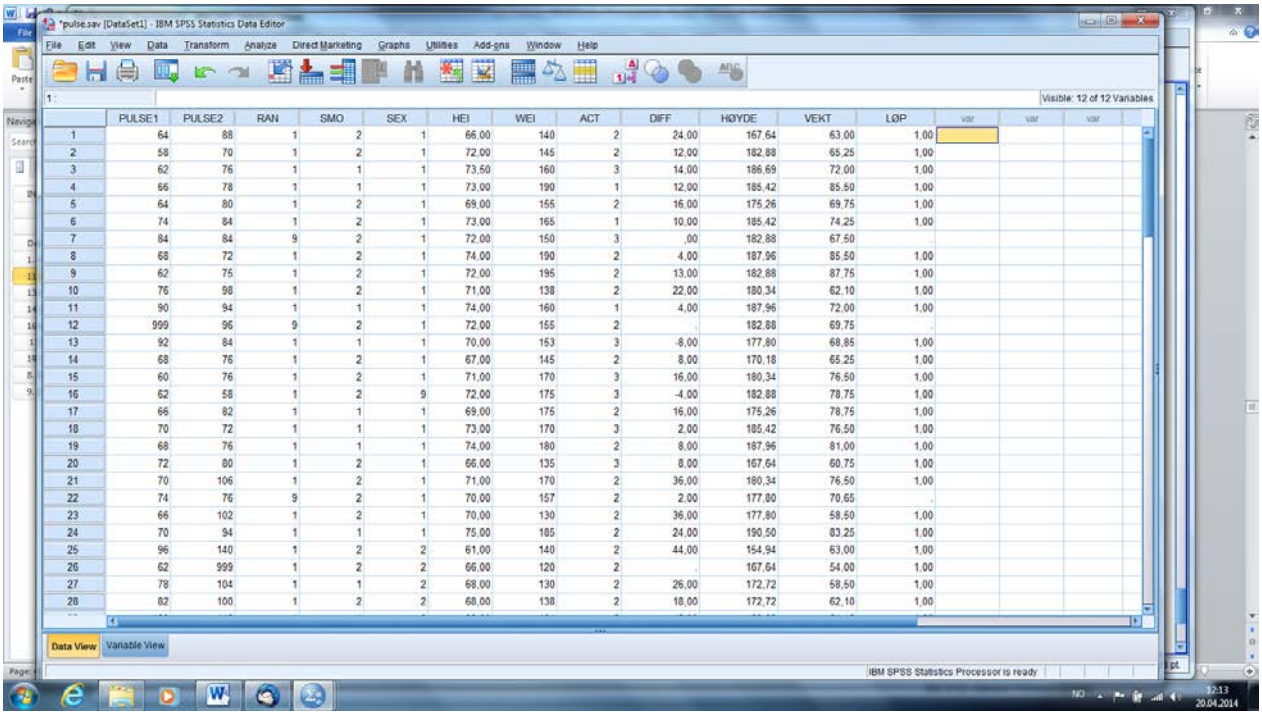
Vi starter med RAN. Vi går inn i *Transfrom/Recode into Different Variables*. Vi trekker RAN over i boksen i midten. Da åpner det seg en ny boks på høyre side, med overskriften Output. Der skriver vi inn LØP, og skifter ved *Change*. Da ser dialogboksen vår slik ut:



Vi er ennå ikke ferdig. Vi må da gå ned til *Old and New Values*. Vi klikker på den. Da kommer vi inn i en ny dialogboks. Her er det to bokser vi skal bruke, boksen med *Old Values* og boksen med *New Values*. I boksen med *Old Values* skriver vi inn 2 og i boksen med *New Values* skriver vi 0. For å få denne aktivisert må vi klikke på *Add*. Vi må også la koden 1 for RAN også være koden 1 for LØP. Det gjør vi enkelt ved å skrive 1 for *Old Values* og 1 for *New Values*. Når vi har klikket på *Add*, ser dialogboksen slik ut:

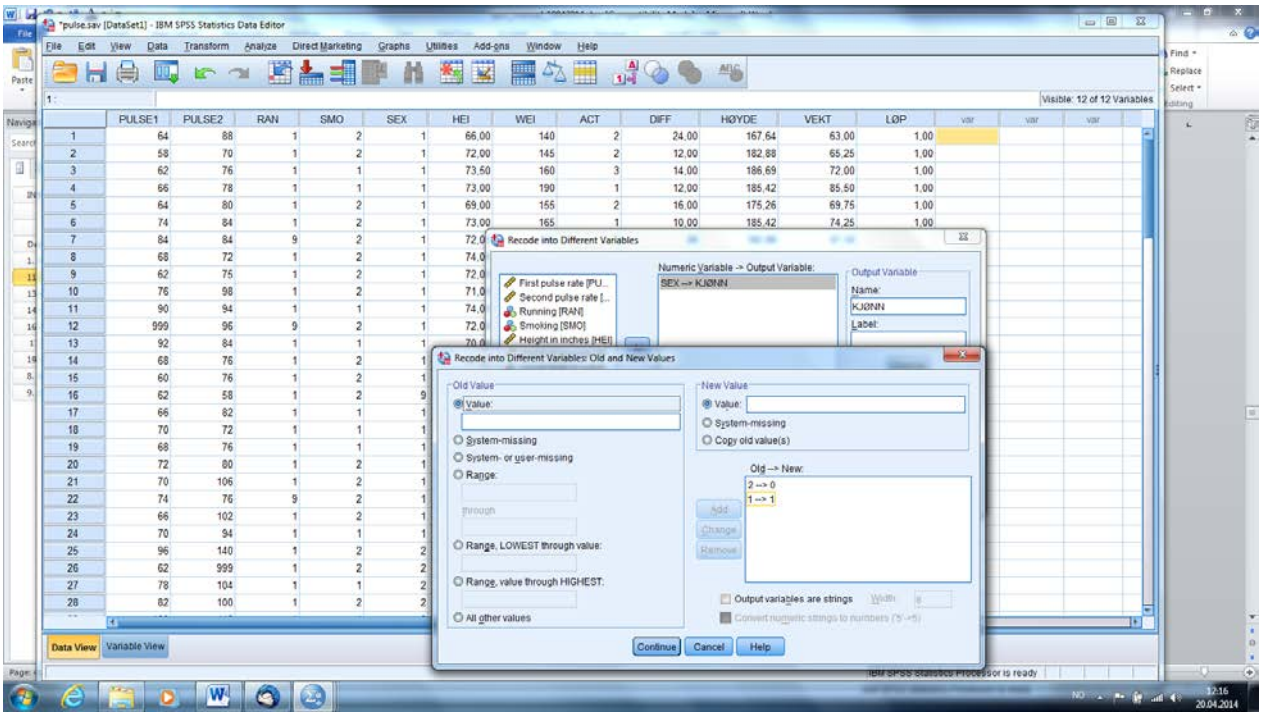


Da klikker vi på *Continue*, og kommer tilbake til den forrige dialogboksen. Der klikker vi på *OK*. Da ender vi i utskriftfilen. Går vi over til datafilen, ser vi at LØP har blitt lagt på til slutt. Da ser datafilen slik ut:



Legg merke til at omkodingen har gått riktig for seg, og at *Missing values* har blitt kodet med et punktum.

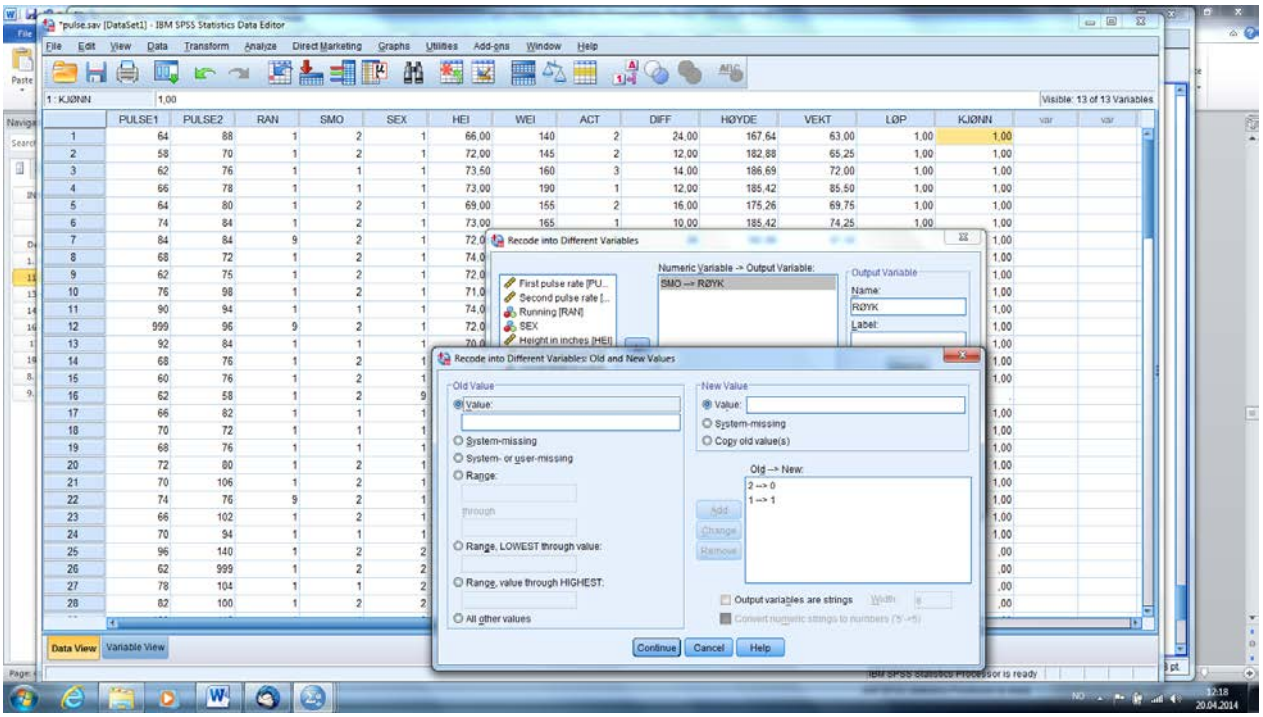
Vi fortsetter på samme måte med omkodingen fra SEX til KJØNN. Her skal vi også omkode fra kodene 1 og 2 for SEX til 1 og 0 for KJØNN. Vi går frem på akkurat samme måte som for omkodingen fra RAN til LØP, og vi gjentar den ikke her. Dialogboksen for omkodingen av SEX ser slik at før vi avslutter selve omkodingen:



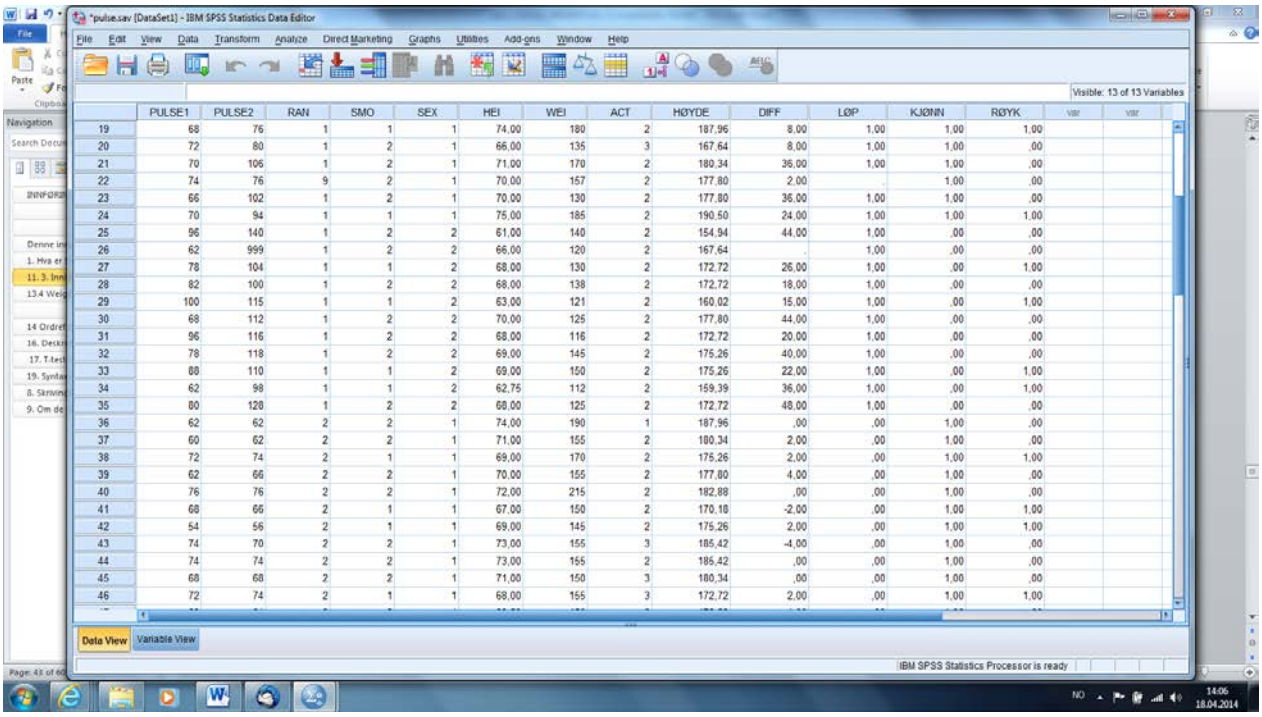
Legg merke til at SPSS «husker» de siste omkodingene som ligger i boksen *Old* → *New*.

Her klikker vi da på *Continue* på den siste dialogboksen og på *OK* på den første. Da får vi omkodingen utført.

Til slutt gjør vi sammen omkodingen for SMO til RØYK. Her blir dialogboksen til slutt seende slik ut:



For sikkerhets skyld går vi nå inn i dataarket og sjekker at omkodingene har gått riktig for seg. Dataarket vil da se slik ut:



Når vi lager en frekvensoversikt for variabelen LØP får vi følgende resultat:

LØP

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	55	59,8	63,2	63,2
	1,00	32	34,8	36,8	100,0
	Total	87	94,6	100,0	
Missing	System	5	5,4		
Total		92	100,0		

Vi merker oss at kodene 0 og 1 her ikke er angitt. Det skal vi se på i kapittel 6.6

Når vi nå har lagd våre omkodede variabler, sikrer vi oss at disse blir lagt på filen **pulse.sav**. Som vanlig går vi til *File/Save As* og lagrer filen som **pulse.sav** i den katalogen der vi har har våre kursfiler.

6.4 Recode. Eksempel: lowbwt.sav

Vi henter frem filen **lowbwt.sav**. Vi skal først omkode variablene PTL og FTV. Grunnen til dette finner vi i frekvensfordelingene til disse to variablene. Vi går da til *Analyze/Descriptive Statistics/Frequencies*. Der trekker vi over PTL og FTV over i boksen med *Variable(s)* og klikker på *OK*. Da får vi følgende utskrift:

history of premature labor

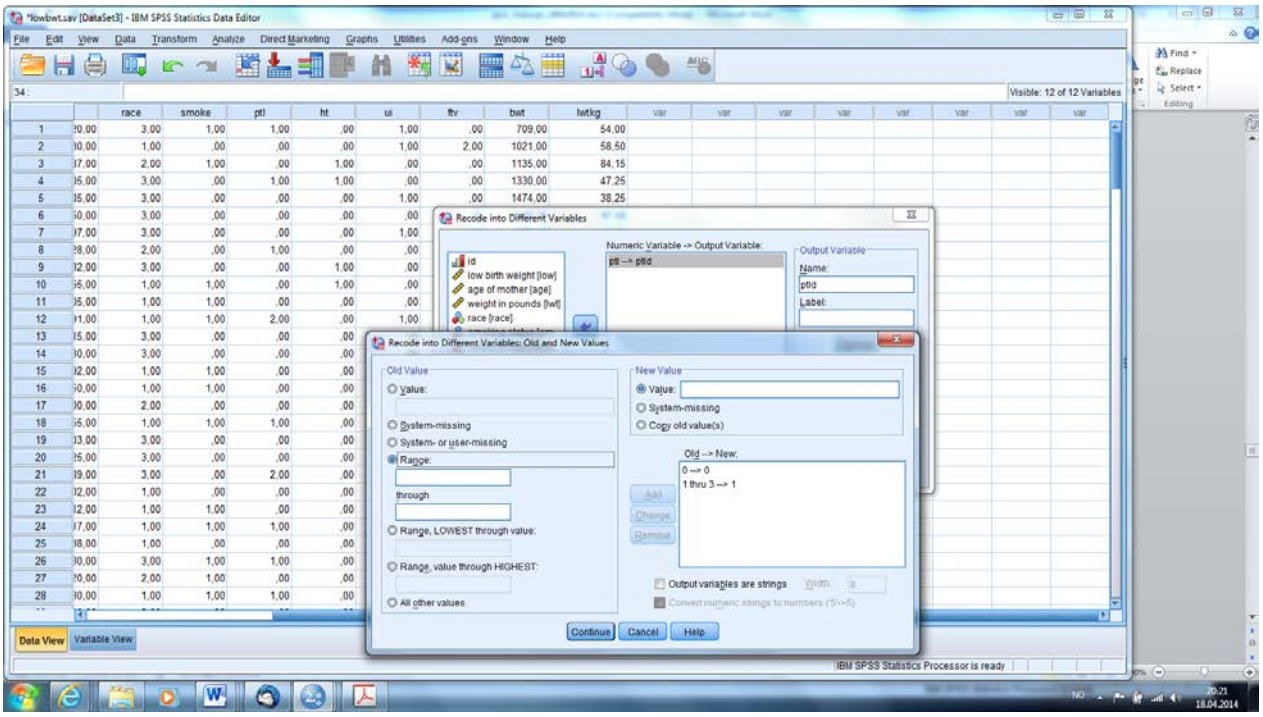
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	159	84,1	84,1	84,1
	1,00	24	12,7	12,7	96,8
	2,00	5	2,6	2,6	99,5
	3,00	1	,5	,5	100,0
	Total	189	100,0	100,0	

first trimester visits

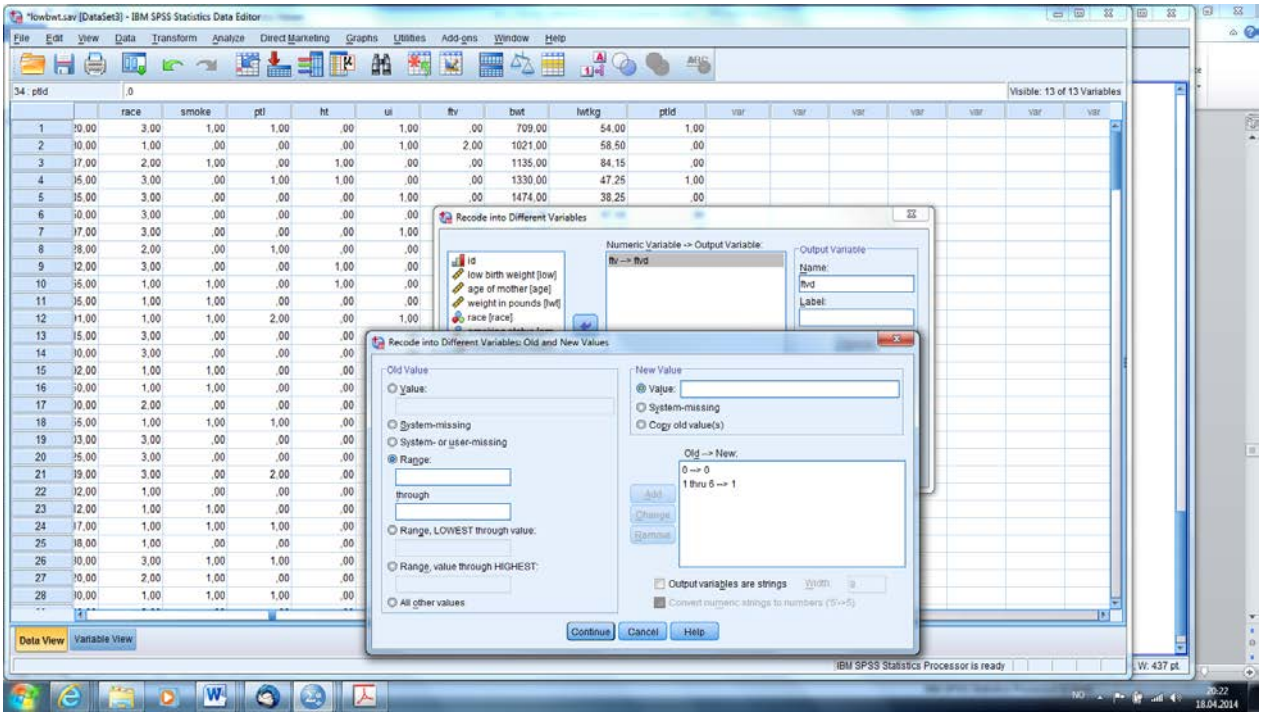
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	100	52,9	52,9	52,9
	1,00	47	24,9	24,9	77,8
	2,00	30	15,9	15,9	93,7
	3,00	7	3,7	3,7	97,4
	4,00	4	2,1	2,1	99,5
	6,00	1	,5	,5	100,0
	Total	189	100,0	100,0	

Her ser vi at det er få personer med verdier større enn 1 for PTL (5+1=6) og 42 for FTV (30+7+4+1=42). Vi bør ha grupper av en viss størrelse når vi skal gjøre analyse på dem. Dette henger igjen sammen med at usikkerheten i frekvensfordelingen blir stor for små grupper. Vi bør derfor lage oss større grupper ved å slå sammen kategoriene. Dette gjør vi for PTL og FTV ved at vi omkoder veridene 2 og 3 til verdien 1 for PTL, og 2, 3, 4 og 6 til verdien 1 for FTV. Dermed betyr koden 1 for PTL og FTV at vi har verdien 1 eller større for disse variablene.

For å gjøre dette går vi inn i *Transform/Recode into Different Variables*. Der trekker vi over PTL over i boksen i midten og angir at vi skal kalle den nye variabelen PTLD. Vi klikker på *Old and New Values*. Nå må vi angi at verdiene 2 og 3 skal bli omkodet til 1. Det gjør vi i *Old Values* ved å gå til *Range* og der angi at 1 til 3 skal få *New Value* 1. Så klikker vi på *Add*. Deretter må vi angi at koden 0 for PTL også skal være 0 for PTLD. Dette gjør ved å skrive inn 0 i *Old Value* og 0 i *New Value*, og deretter klikke på *Add*. Da vil dialogboksen vår se slik ut:



Ved å klikke oss gjennom *Continue* og *OK*, får vi utført omkodingene våre. Vi gjør det helt tilsvarende for transformasjonen fra FTV til FTVD. Men når vi er i dialogboksen for omkodingen til FTVD, må vi huske at vi i *Range* må angi at 1 til 6 skal omkodes til 1. For FTVD vil dialogboksen se slik ut:



Etter å klikket på *Continue* og *OK* blir også denne transformasjonen utført.

Det er en god vane å sjekke at slike omkodinger har gått riktig for seg. Det gjør vi ved å lage en frekvensfordeling for PTLD og FTVD. Vi går da inn *Analyze/Descriptive Statistics/Frequencies.* Der trekker vi over PTLD og FTVD over i boksen med *Variable(s)*

ptld

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	159	84,1	84,1	84,1
	1,00	30	15,9	15,9	100,0
	Total	189	100,0	100,0	

ftvd

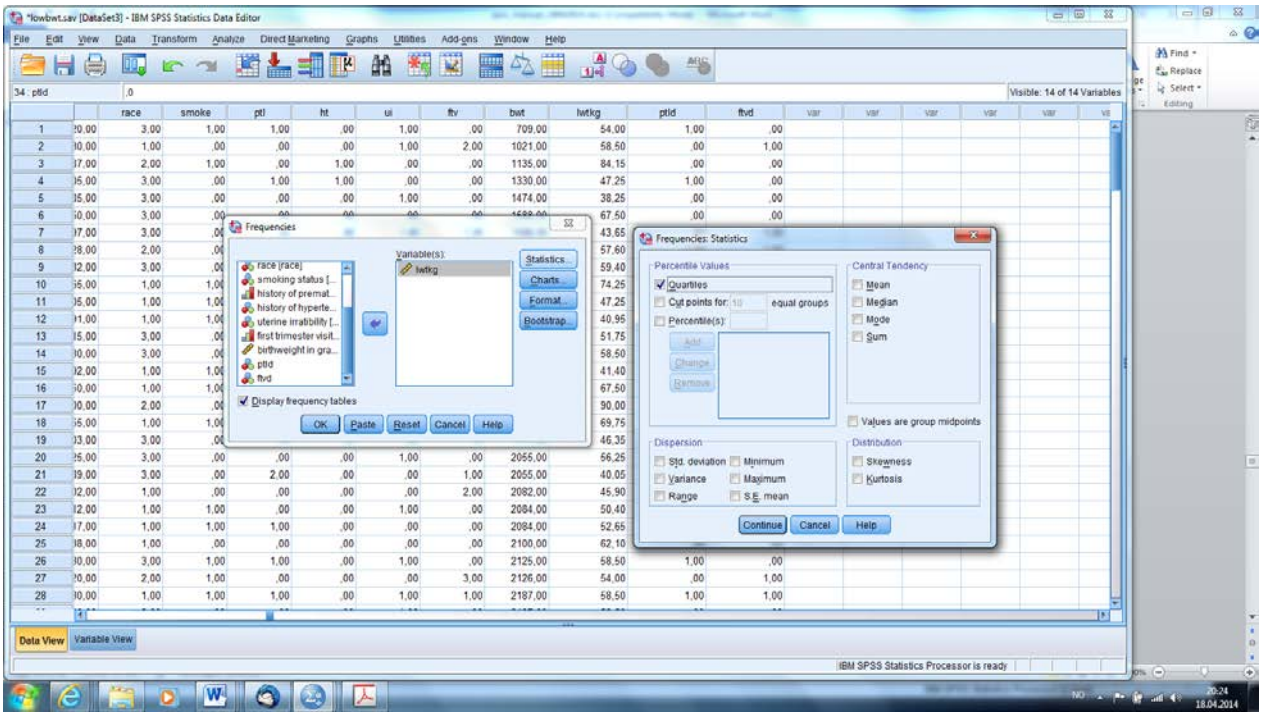
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	100	52,9	52,9	52,9
	1,00	89	47,1	47,1	100,0
	Total	189	100,0	100,0	

Vi ser at fordelingen for PTLD og FTVD har blitt riktige. Vi har det samme antall 0'ere for begge variablene, og de resterende har fått verdien 1, som da også har blitt riktig.

Legg merke til at det ikke ligger variable label eller value labels til disse to variablene. Slik vil det alltid være for transformerte variabler. Vi kommer tilbake til dette.

I en del situasjoner er det aktuelt å omkode en kontinuerlig variabel til en kategorivariabel. Dette er særlig aktuelt i regresjonsanalyser, når vi er interessert i om sammenhengen mellom den avhengige variabelen og en forklaringsvariabel virkelig er lineær. I slike situasjoner er det naturlig å dele variabelen oppi fire kategorier, etter kvartilene. Kvartilene deler variabelen i fire like store deler. Første kvartil vil da bestå av den fjerdedelen med de minste observasjonene, annen kvartil vil bestå observasjoner som faller ovenfor laveste kvartil, men nedenfor midten av observasjonene. Tredje kvartil vil ha de observasjonene som faller over midten, og er blant de tre fjerdedeler minste observasjonene. Øverste kvartil består av de 25% største observasjonene. Vi vil nå omkode variablene slik at observasjoner som faller i første kvartil får verdien 0, de som faller i annen kvartil får verdien 1, observasjonene i tredje kvartil verdien 3, og observasjonene i fjerde kvartil får verdien 4.

Vi skal anvende dette til variabelen LWTKG. Vi først finne observasjonene som gir oss inndelingen i kvartiler. For å finne disse går vi til *Analyze/Descriptive Statistics/Frequencies*. Her går vi videre til dialogboksen *Statistics*. Under *Percentile Values* klikker vi av på *Quartiles*. Da ser dialogboksene slik ut:



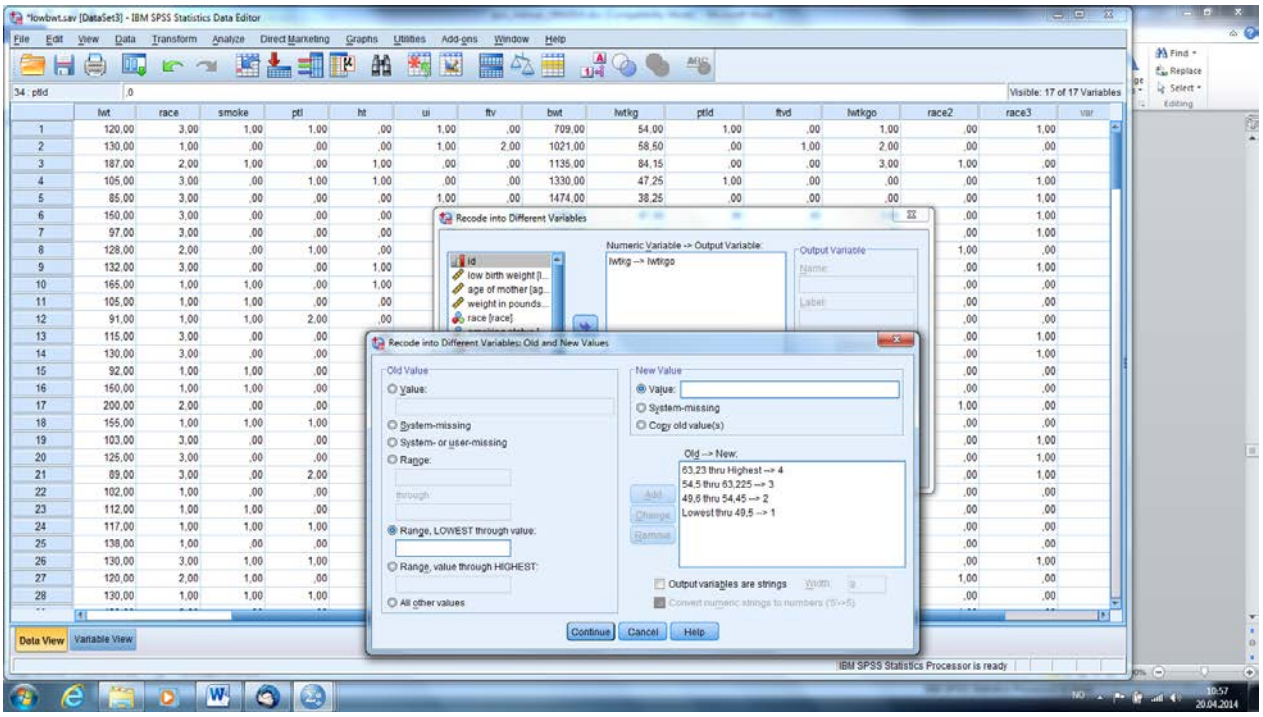
Ved å klikke på *Continue* og *OK*, får vi følgende resultater:

Statistics

lwtkg

N	Valid	189
	Missing	0
Percentiles	25	49,5000
	50	54,4500
	75	63,2250

Nå er vi klare til å omkode LWTKG til LWTKGO (O for å vise at dette er en ordinal variabel). Da går vi inn i *Transform/Recode into Different Variables*. Her trekker vi LWTKG over i boksen i midten, og skriver LWTKGO inn i boksen til høyre (*Output Variable*). Da må vi klikke på *Change*, for å få aktivisert denne transformasjonen. Deretter klikker vi på *Old and New Values*. Da kommer det opp en dialogboks hvor vi skal legge inn verdiene for de fire kvartilene. Vi legger først inn observasjonen i første kvartil. Observasjoner som er mindre eller lik 45.5 skal omkodes til 1. Vi skriver da inn *Range Lowest through value*: 49,5. Da går vi opp i boksen med *New Value* og skriver inn 1. Da er vi klar til å klikke på A*dd*. Så går vi videre til neste kvartil. Da klikker vi på *Range*, og her skriver vi inn 49.6 i øverste boks og 54,45 i nederste boks. Når det er gjort går til *New Value* og skriver inn 1, og klikker på *Add*. Da skal vi omkode for tredje kvartil. Vi holder oss til *Range* og skriver inn 54,5 i øverste boks og 63,225 i nederste boks, og skriver inn 3 i *New Value*. Vi må også klikke på *Add*. Til slutt skal vi ta for oss fjerde kvartil. Da går vi til *Range Value* through HIGHEST, og skriver inn 63,23, og skriver inn 4 i *New Value*. Da ser dialogboksen slik ut:



Da klikker vi på *Continue* og *OK*, og får utført transformasjonen. For å se at transformasjonen har gått riktig, lager vi en frekvensfordeling for LWTKGO. Det gjør vi ved *Analyze/Descriptive Statistics/Frequencies* og trekker over LWTKGO i boksen med *Variable(s).* Når vi klikker på *OK* får vi følgende utskrift:

lwtkgo

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1,00	53	28,0	28,0	28,0
	2,00	43	22,8	22,8	50,8
	3,00	46	24,3	24,3	75,1
	4,00	47	24,9	24,9	100,0
	Total	189	100,0	100,0	

Vi ser at fordelingen er tilnærmet riktig. Vi kan ikke forvente å få nøyaktig 25% av observasjonene i kvartil, siden vi har observasjoner som har like verdier. Fordelingen ser derfor riktig ut.

6.5 Recode til dummy-variabler. Eksempel: lowbwt.sav

Kategoriske variabler kan ha to eller flere verdier. Dersom variabelen har to verdier, gir vi dem verdiene 0 og 1, med 0 for referansekategorien og 1 for den kategorien vi skal sammenligne referansekategorien med. Vi gir altså de eksponerte, for eksempel røykerne, eller de syke verdien 1, og de ueksponerte eller friske verdien 0. Kategoriske variabler med to verdier er greie å bruke i statistiske analyser, siden de kan brukes direkte i analysene.

Slik er det ikke nødvendigvis for kategoriske variabler med flere verdier. I datafilen har vi to variabler som er flerkategoriske, nemlig RACE og LWTKGO. I noen statistiske analyser, slik som i variansanalyser (se kapittel 12.1), kan vi bruke de flerkategoriske variabler direkte. Men i regresjonsanalyser (se kapittel 12.2) kan vi ikke bruke de flerkategoriske variablene. Da må vi bruke så kalte dummy-variable.

Dummy-variabler er et sett av kategoriske variable med to kategorier. Disse skal sammenlignes med en referansekategori. En kategorisk variable med *k* kategorier trenger $k-1$ dummy-variabler. Vi ser på variabelen RACE som har tre kategorier. Vi lager en frekvensfordeling for RACE, ved *Analyze/Descriptive Statistics/Frequencies*, trekker RACE over i boksen i midten og klikker på *OK*. Da får vi følgende frekvensfordeling:

race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	white	96	50,8	50,8	50,8
	black	26	13,8	13,8	64,6
	other	67	35,4	35,4	100,0
	Total	189	100,0	100,0	

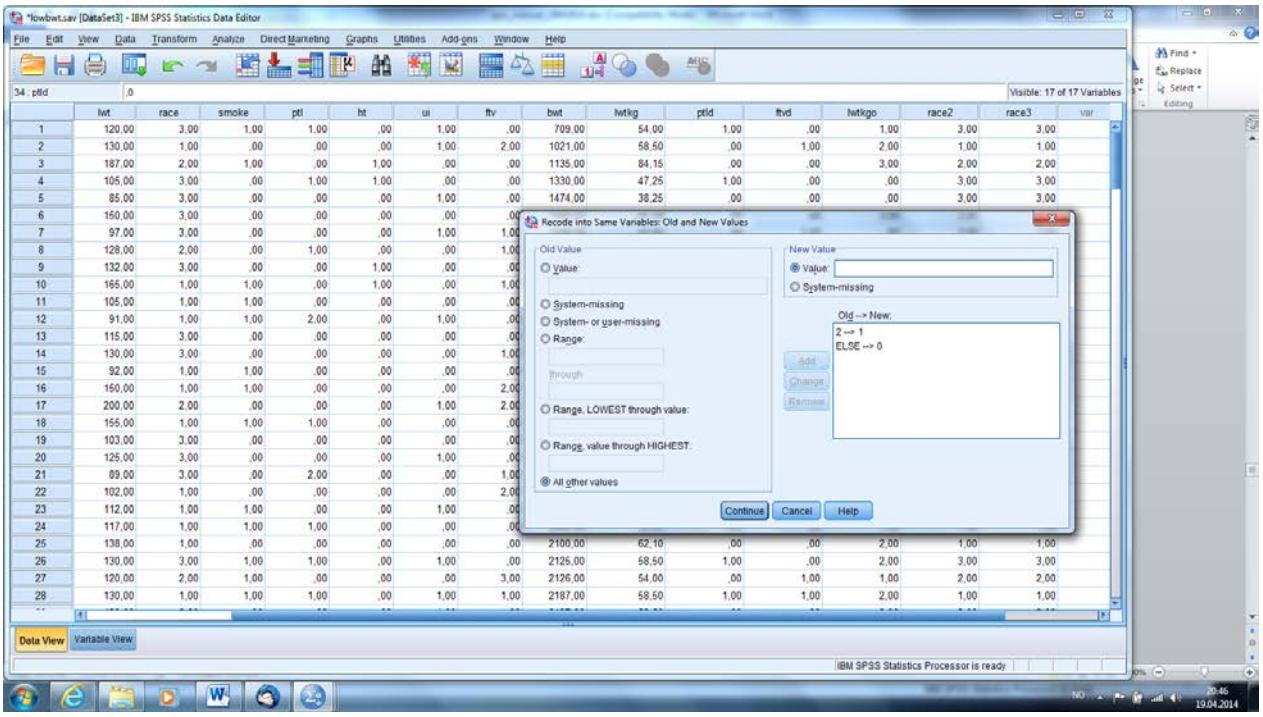
Vi har tre kategorier, og vi skal da lage to dummy-variabler. Her skal vi nå bruke WHITE som referansekategori. Vi skal nå lage to dummy-variabler RACE2 og RACE3, slik som angitt i tabellen nedenfor. Merk at begge variablene har to kategorier, 0 og 1.

RACE	RACE2	RACE3
WHITE	0	0
BLACK	1	0
OTHER	0	1

Her ser vi at RACE2 = 1 når RACE = BLACK, og RACE3 = 1 bare når RACE = OTHER. Altså vil RACE2 angi at etnisiteten er BLACK, mens RACE3 angir at etnisiteten er OTHER. Omvendt, når RACE = WHITE er RACE2 = RACE3 = 0. Når RACE = BLACK, er RACE2 = 1 og RACE3 = 0. Når RACE = OTHER, er RACE2 = 0 og RACE3 = 1. Vi ser altså at vi trenger de to dummy-variablene RACE2 og RACE3 for å beskrive de tre kategoriene i RACE.

Vi skal nå lage RACE2 og RACE3 i SPSS. Først beregner vi RACE2 og RACE3 som begge er identisk lik med RACE. Deretter omkoder vi disse to variablene. Først går vi inn i Transform/Compute, og skriver RACE2 i venstre vindu (*Target Variable*) og RACE i høyre vindu (*Numeric Expression*), og så klikker vi på *OK*. Vi gjør helt tilsvarende med RACE3. Da får vi to nye variabler lagt til på dataarket vårt. Foreløpig er de identiske med RACE.

Da er vi klare til å omkode RACE2 og RACE3. Vi tar først RACE2 og går til *Transform/Recode Into Same Variable*. Vi trekker RACE2 over i vinduet til høyre. Da åpner det seg en ny knapp med *Old and New Values*. Vi klikker på den. Vi skriver inn 2 i vinduet med *Old Values* og 1 i vinduet med *New Values*, og klikker på *Add*. Under Old Values går vi så ned til *All other values*, skriver inn 0 i vinduet *New Values* og klikker på *Add*. Da ser dialogboksen slik ut:



Da klikker vi på *Continue* og *OK*.

Vi gjør da det samme med RACE3. Her skal koden 3 kodes om til 1 og alle andre verdier skal settes til 0.

Når vi har gjort en omkoding, er det en god vane å sjekke at omkoding har gått riktig for seg. Det gjør vi ved lage en frekvensoversikt for den variabelen vi omkoder fra og de vi omkoder til. Vi går derfor inn i *Analyze/Descriptive Statistics/Frequencies* og trekker over RACE, RACE2 og RACE3. Da får vi følgende resultater:

race

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	white	96	50,8	50,8	50,8
	black	26	13,8	13,8	64,6
	other	67	35,4	35,4	100,0
	Total	189	100,0	100,0	

race2

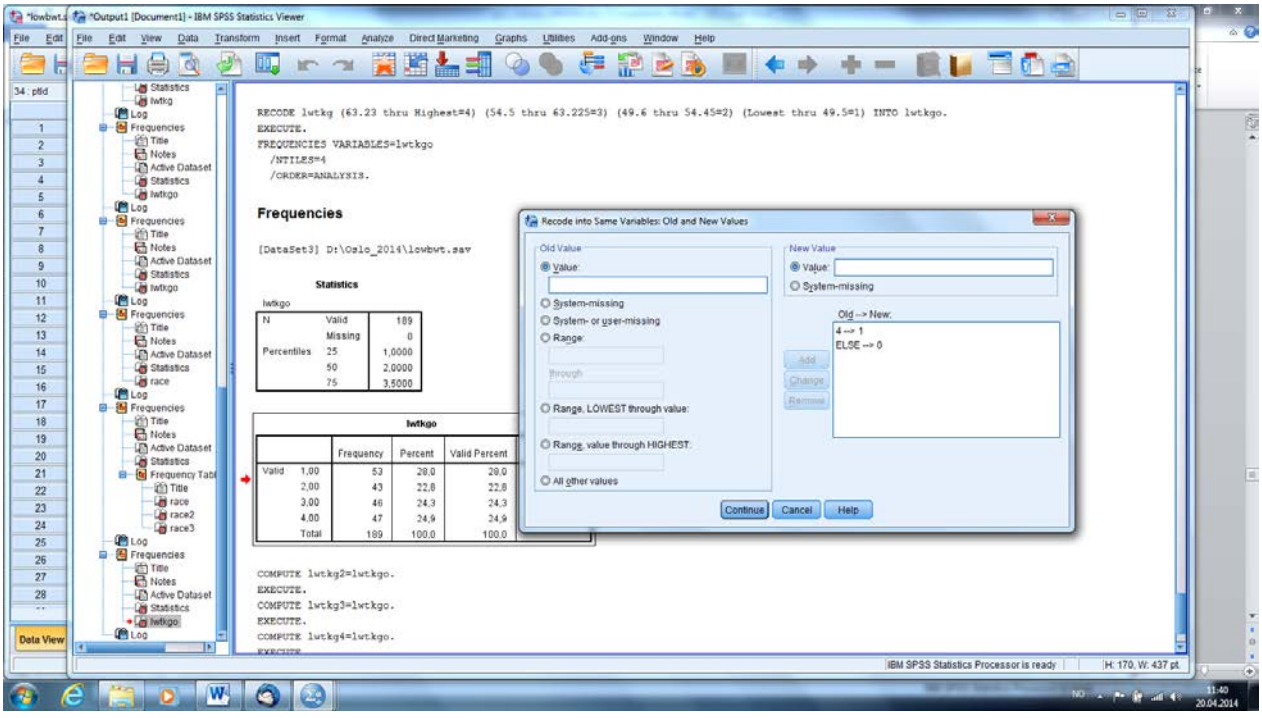
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	163	86,2	86,2	86,2
	1,00	26	13,8	13,8	100,0
	Total	189	100,0	100,0	

race3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	122	64,6	64,6	64,6
	1,00	67	35,4	35,4	100,0
	Total	189	100,0	100,0	

Her ser vi at RACE2 angir at RACE = BLACK og RACE2 angir at RAC E = OTHER. Omkodingen har blitt riktig.

På filen **lowbwt.sav** ligger også variablen LWTKGO som er en ordinal variabel. Denne variabelen kan også omkodes til dummy-variabler. Siden LWTKGO har fire kategorier, må vi lage tre dummy-variabler. Disse kaller vi LWTKG2, LWTKG3 og LWTKG4. Vi gjør dette på samme måte som for RACE2 og RACE3. Vi bruker LWTKGO = 1 som referanseverdi. LWTKG2 = 1 når LWTKGO = 2, LWTKG3 = 1 når LWTKGO = 3, og LWTKG4 = 1 når LWTKG = 4. Dette gjør vi ved å inn i *Transform/Compute* og lage tre kopier av LWTKGO, inn i LWTKG2, LWTKG3 og LWTKG4. Deretter går vi til *Transform/Recode into Same Variable* og rekoder LWTKG2, LWTKG3 og LWTKG4, slik vi gjorde med RACE2 og RACE3. Når vi gjør dette for LWTKG4 ser dialogboksen våre slik ut



Som for omkodingen for RACE lager vi nå en frekvensfordeling for LWTKG2, LWTKG3 og LWTKG4. Dette gjør vi ve å gå inn i *Analyze/Descriptive Statistics/Frequencies* og trekker over de tre variablene i vinduet til høyre. Da får vi følgende frekvensfordeling;

lwtkg2

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	146	77,2	77,2	77,2
	1,00	43	22,8	22,8	100,0
	Total	189	100,0	100,0	

lwtkg3

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	143	75,7	75,7	75,7
	1,00	46	24,3	24,3	100,0
	Total	189	100,0	100,0	

lwtkg4

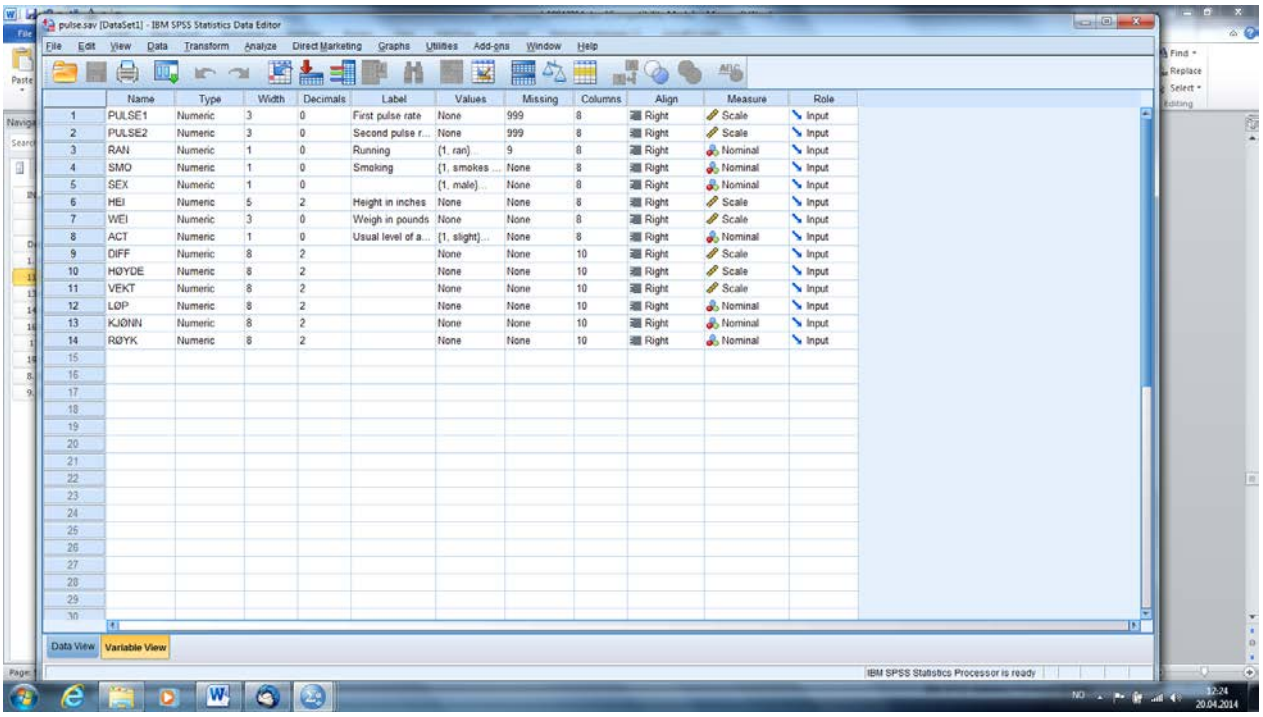
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	,00	142	75,1	75,1	75,1
	1,00	47	24,9	24,9	100,0
	Total	189	100,0	100,0	

Vi ser at disse frekvensfordelingene stemmer overens med den vi hadde for LWTKGO i kapittel 6.4. Vi sikrer oss nå at vi får lagt denne filen ned i katalogen vår. Det gjør vi på vanlig måte med *File/Save As*, velge navnet **lowbwt.sav** og legger den i riktig katalog. Vi vil få beskjed om at filen allerede eksisterer, og om vil overskrive den gamle filen. Vi svarer *Yes* til det.

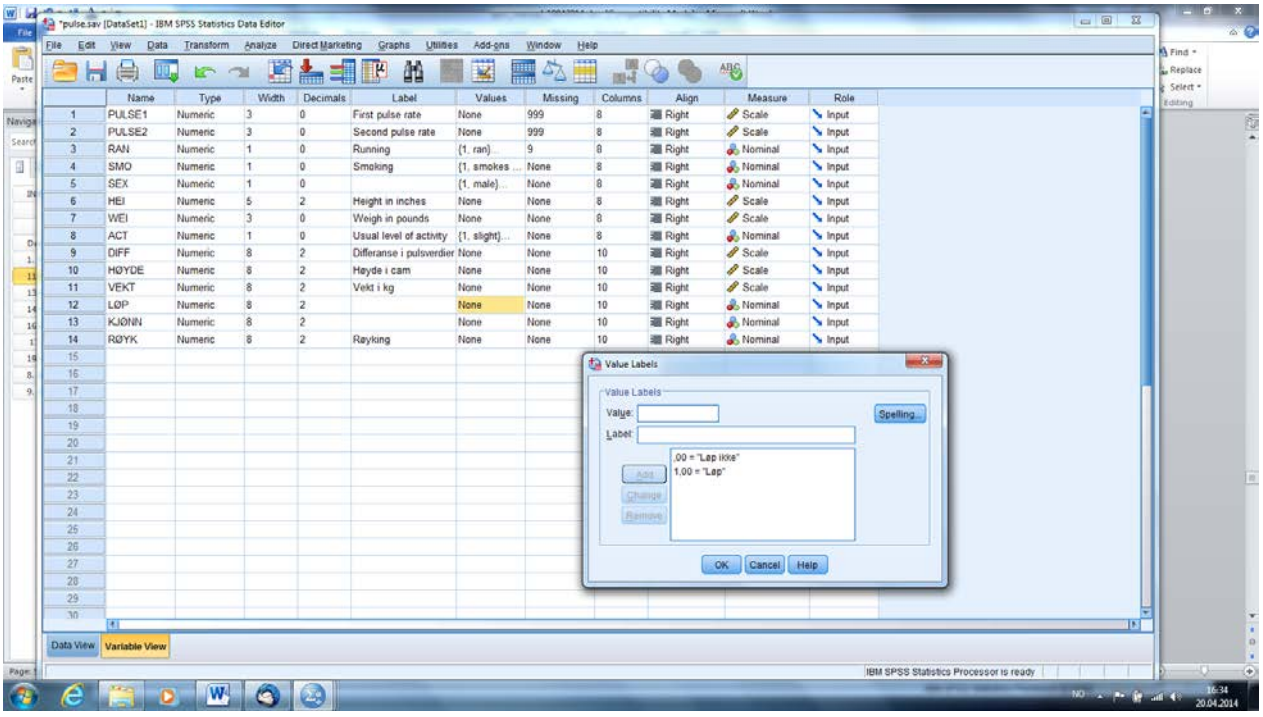
6.6 Variable label og Value label for omkodede variabler . Eksempel: pulse.sav

Når vi har gjort omkodinger må vi sikre oss at vi legger til *Variable labels* og *Value labels* for nye beregnede og omkodede variabler. Vi skal nedenfor gjøre dette for variablene på filen **pulse.sav**.

Vi går nå inn i datafilen **pulse.sav**. Når vi går inn på *Data View*, ser den slik ut:

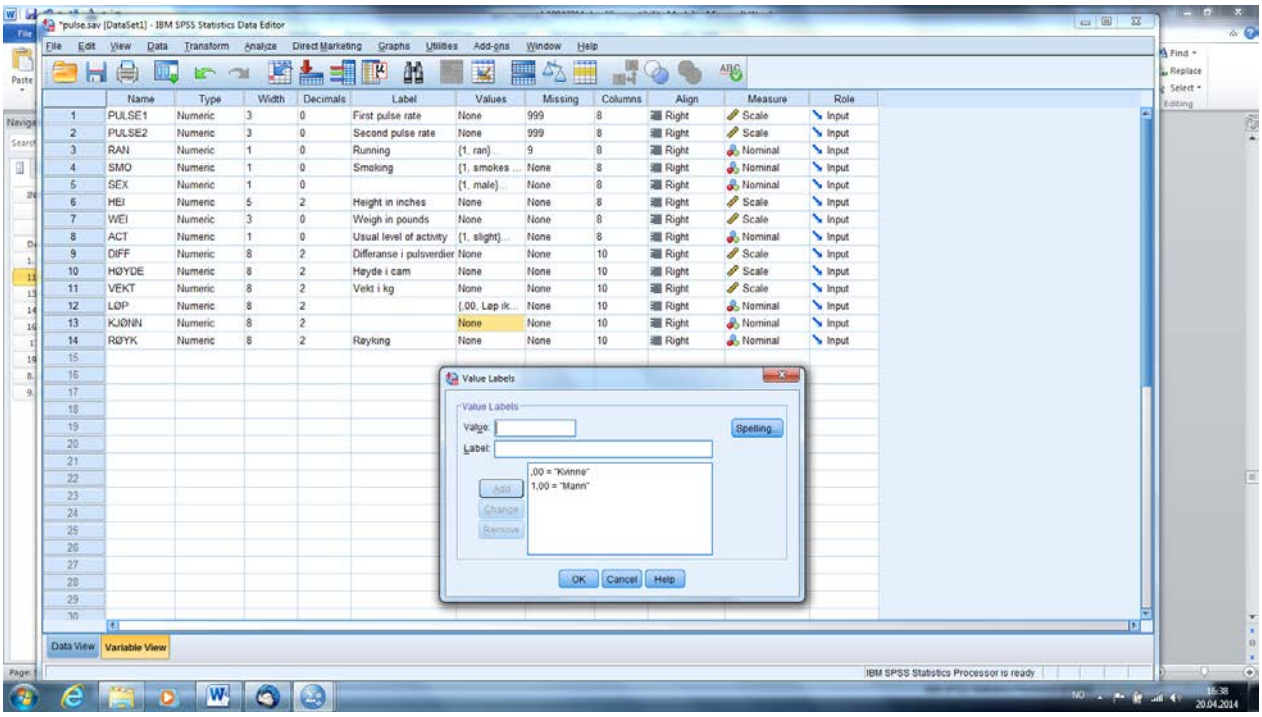


Vi ser at de omkodede variablene mangler *Variable label* og *Value label*. Vi starter med *Variable label* og for DIFF skriver vi inn Differanse i pulsverdier, for HØYDE skriver vi Høyde i cm, for VEKT Vekt i kg, og for RØYK skriver vi Røyking. Men viktigere er å få *Value label* på plass. Vi går til LØP og bort til *Values*. Der klikker vi i feltet og går inn feltet med prikker. Her skriver vi inn verdien 0 med label Løp ikke og 1 med label Løp. Etter hver av disse må vi huske å klikke på *Add*. Da ser dialogboksen vår slik ut:

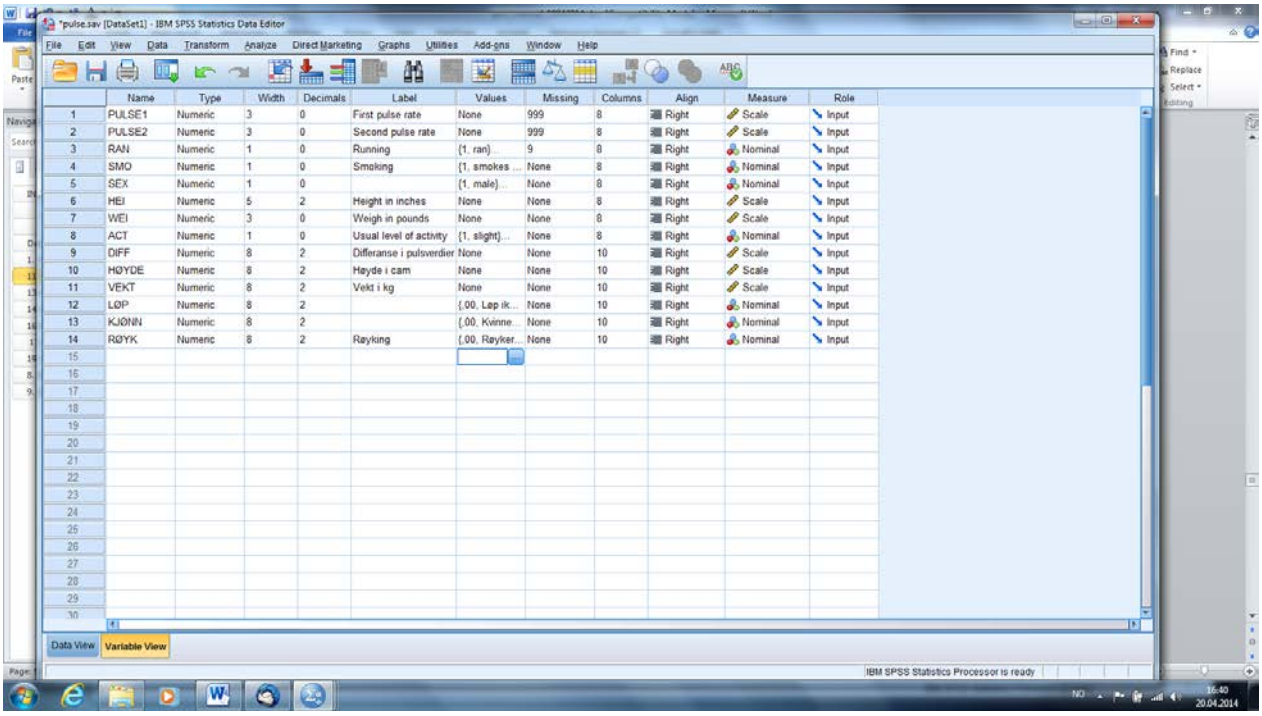


Vi gjør det tilsvarende med variablene KJØNN og RØYK. Husk at for alle variablene skal verdien 0 angir referansegruppen og 1 angir de som vi vil sammenligne med referansegruppen.

Merk at kvinner er referansegruppen, som da menn sammenlignes med, slik som i dialogboksen under:



Når vi er ferdige ser datafilen, med *Variable View*, slik ut:



Hvis vi nå lager en frekvensfordeling for LØP, ved *Analyze/Descriptive Statistics/Frequencies*, får vi følgende resultat:

LØP

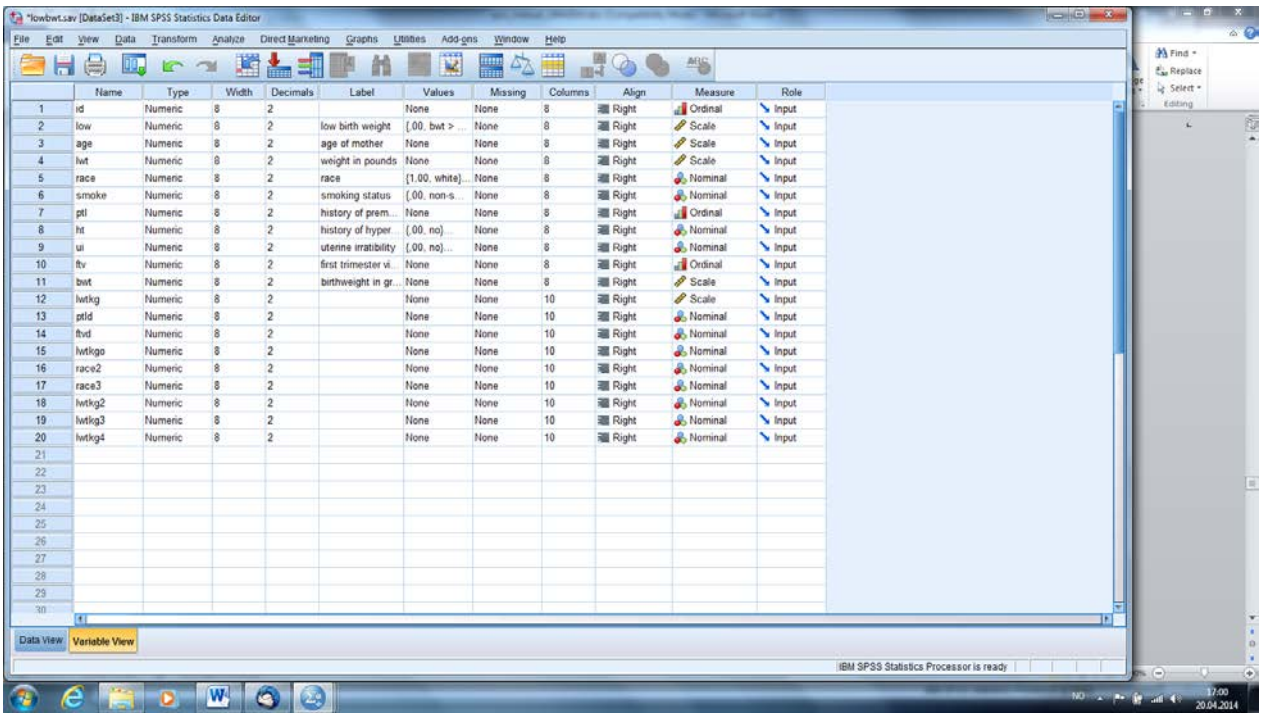
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Løp ikke	55	59,8	63,2	63,2
	Løp	32	34,8	36,8	100,0
	Total	87	94,6	100,0	
Missing	System	5	5,4		
Total		92	100,0		

Vi ser at kodene til variabelen blir lagt på i frekvensfordelingen. Dette er en stor fordel!

Når vi er ferdige med omkodingene må vi huske å legge filen ned i katalogen vår. Det gjør vi på vanlig måte med *File/Save As*, velger navnet **pulse.sav**, legger den i riktig katalog og svarer Y*es* til at vi vil overskrive den gamle filen, med samme navn.

6.7 Variable label og Value label for omkodete variabler. Eksempel: lowbwt.sav.

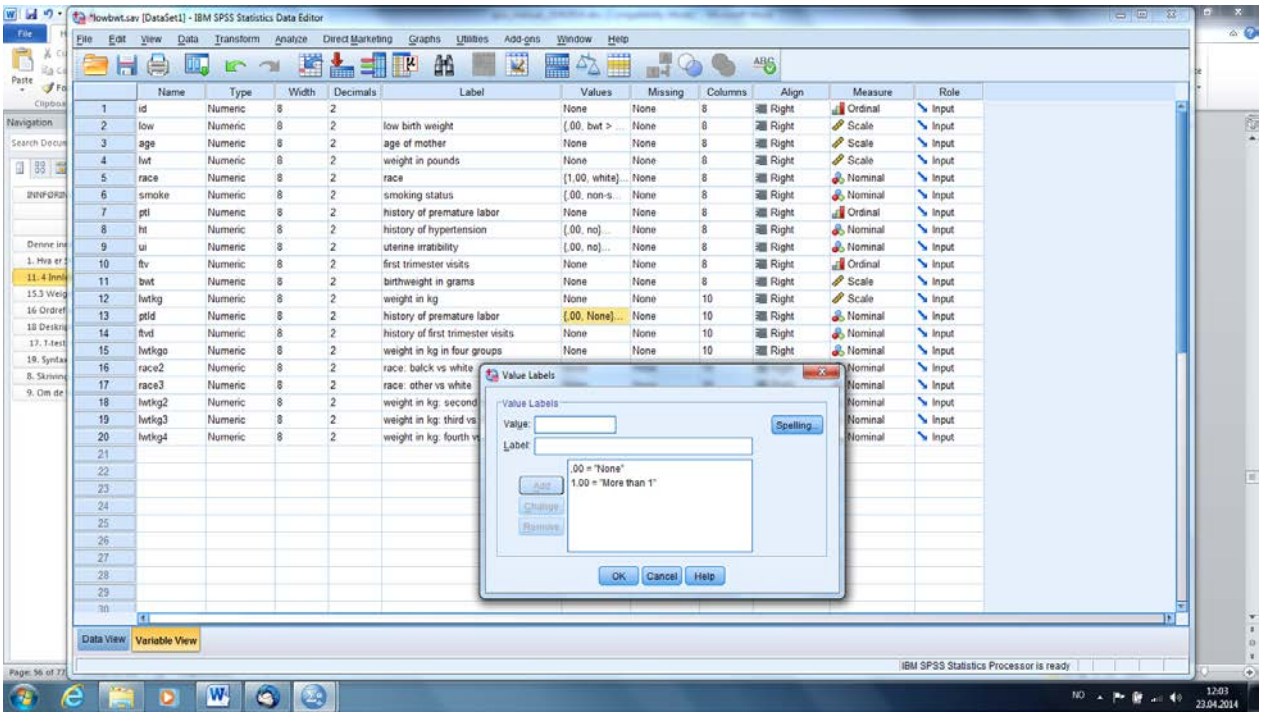
Vi skal nå gjøre det samme som vi gjorde for variablene på filen **pulse.sav** for variablene på **lowbwt.sav**. Vi henter da frem datafilen **lowbwt.sav**. Her er det gjort en rekke beregninger av nye variabler og omkodinger. Vi må få gitt *Variable label* og *Value label* til disse. Datafilen, i *Variable View*, ser slik ut:



Vi ser at variablene fra LWTKG til LWTKG4 hverken har *Variabel label* eller *Value label*. Det skal vi legge til nå. Vi tar første *Variable label*, og skriver inn under *Label*:

lwtkg: weight in kg
ptld: history of premature labor
ftvd: history of first trimester visists
lwtkgo: weight in kg in four groups
race2: race: black vs white
race3: race: other vs white
lwtkg2: weight in kg: second vs first quartile
lwtkg3: weight in kg: third vs first quartile
lwtkg2: weight in kg: fourth vs first quartile

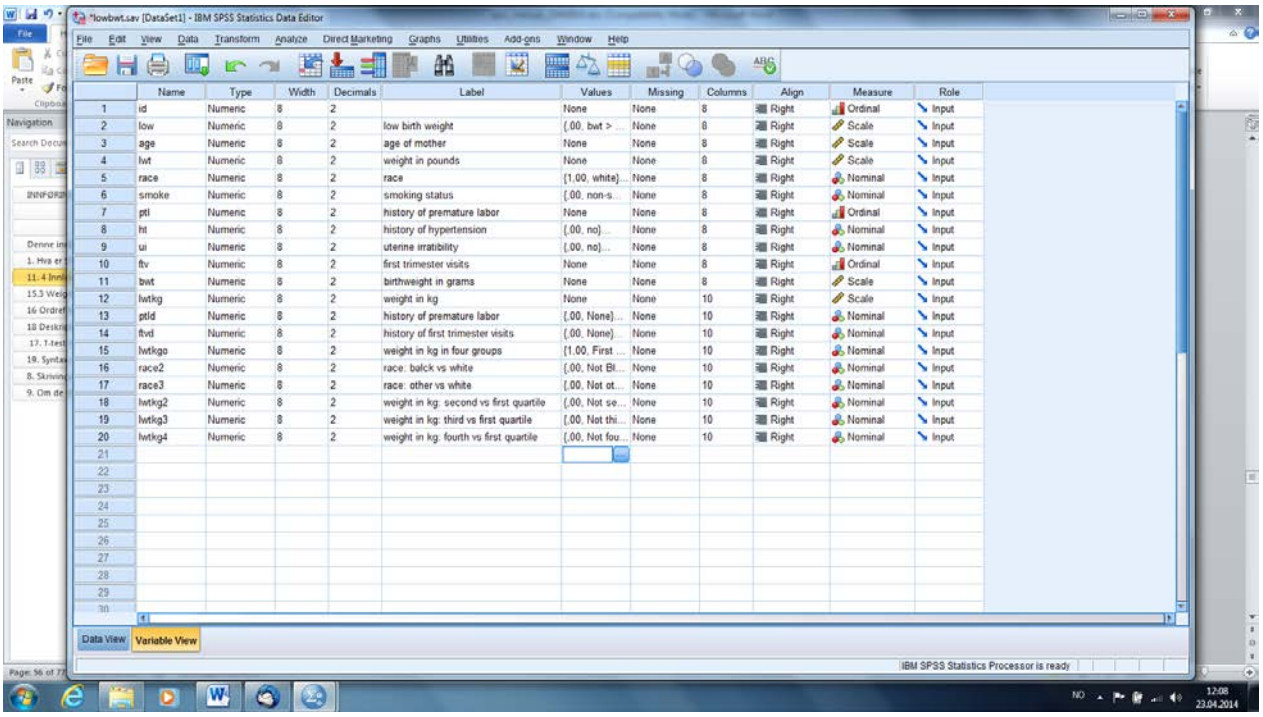
Når vi har dem på plass, går vil til *Value label*. Her er det variablene fra PTLD til LWTKG4 som må gis *Value labels*. Vi starter med variabelen PTLD og går til *Values*. Der skriver vi inn i *Value*: 0 og *Label*: None, og klikker på *Add*. Deretter går vi til *Value*: 1 og *Label*: More than 1 og klikker på *Add*. Da ser dialogboksen vår slik ut:



Da klikker vi på *OK*, og kodene blir lagt til. Vi går da gjennom de andre variablene og legger inn *Value Labels* som følger:

ftvd: 0: None 1: More than 1
lwtkgo: 1: First quartile 2: Second quartile 3: Third quartile 4: Fourth quartile
race2: 0: Not black 1: Black
race3: 0: Not other 1: Other
lwtkg2: 0: Not in second quartile 1: Second quartile
lwtkg3: 0: Not in third quartile 1: Third quartile
lwtkg4: 0: Not in first quartile: 1: Fourth quartile

Da vil datarket vårt se slik ut:



Vi kjører ut frekvensoversikten for LWTKG2, LWTKG3 og LWTKG4 på nytt og får:

weight in kg: second vs first quartile

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not second quartile	146	77,2	77,2	77,2
	Second quartile	43	22,8	22,8	100,0
	Total	189	100,0	100,0	

weight in kg: third vs first quartile

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not third quartile	143	75,7	75,7	75,7
	Third quartile	46	24,3	24,3	100,0
	Total	189	100,0	100,0	

weight in kg: fourth vs first quartile

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Not fourth quartile	142	75,1	75,1	75,1
	Fourth quartile	47	24,9	24,9	100,0
	Total	189	100,0	100,0	

Vi ser at tabellene er vesentlig mer informative enn dem vi lagde i 14.5, UTEN *Variable* og *Value Labels.*

Nå har vi foretatt en rekke endringer i filen **lowbwt.sav**. Vi må da sikre oss at vi lagrer file som nå har åpen i dataarket vårt. Vi gjør det på vanlig måte ved å gå til *File/Save As* og lagre filen under samme navn som vi har, nemlig **lowbwt.sav**.

7 Databearbeiding 3

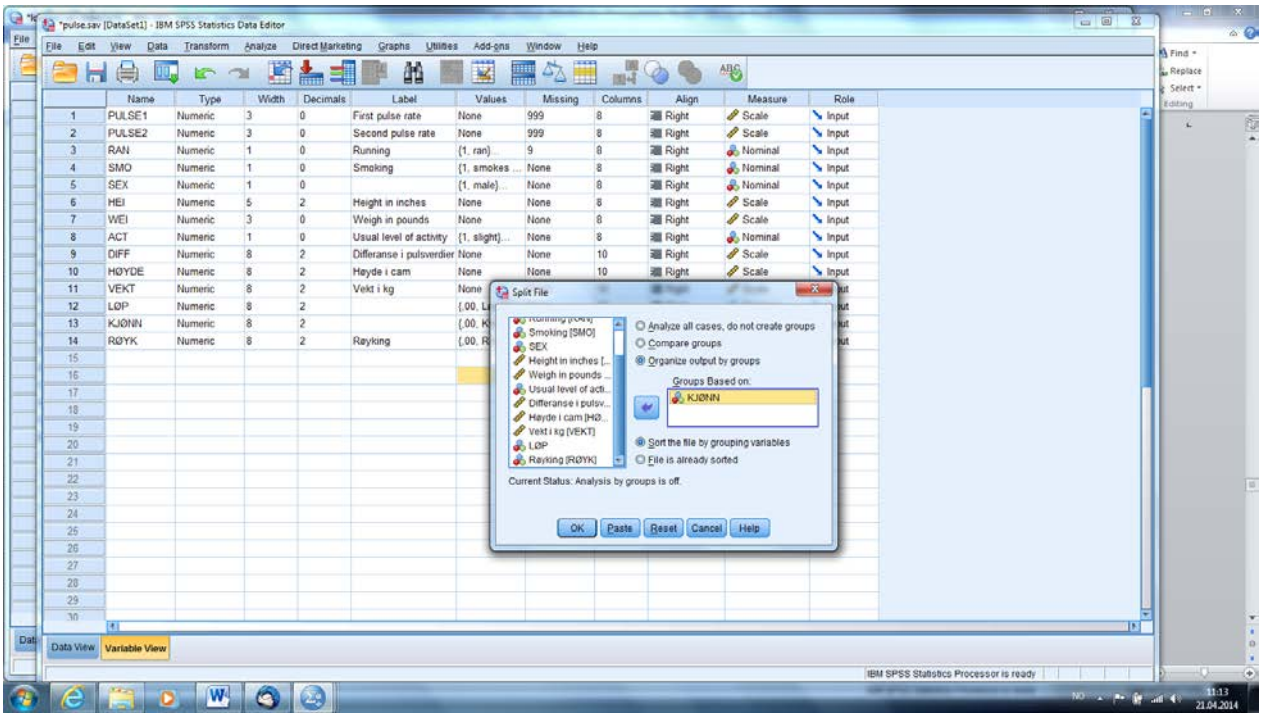
Læringsmål

Ofte er vi interessert i å undersøke undergrupper av individene i datavinduet vårt. Denne muligheten får vi under hovedmenyknappen *Data*. Der er det de to kommandoene *Split file* og *Select cases* som er aktuelle for oss.

SPSS har dataene liggende i linjer og variablene i kolonner. SPSS tror at hver datalinje representer ett individ. Men av og til har vi data presentert på aggregert form. Da må vi ha mulighet til å vekte opp data. Dette skal vi gjøre via *Weight cases*.

7.1 Split file. Eksempel: pulse.sav

Ofte vil vi presentere dataene oppdelt etter verdiene til en annen variabel. Vi vil for eksempel studere gjennomsnitt eller frekvensfordelingen etter kjønn. Da kan vi bruke Split file i SPSS. Vi går til datafilen **pulse.sav**. Der går vi til *Data/Split file*. Da åpner det seg en ny dialogboks. Her åpner det seg flere valgmuligheter. Vi skal velge *Organize output by groups* og vi klikker av på denne. Da kan vi trekke KJØNN over i vinduet til høyre. Det gir oss en dialogboks som under.



Vi klikker på *OK*. Da skjer det ingenting. Merk at denne funksjonen trer i funksjon når vi velger analyse. Da vil SPSS presentere resultatene fordelt etter kjønn.

Vi går derfor til *Analyze/Descriptive Statistics/Descriptives og* trekker DIFF over i vinduet som åpner seg i vinduet som åpner seg i dialogboksen. Vi klikker på *OK* og får vi følgende resultat:

Descriptive Statistics[a]

	N	Minimum	Maximum	Mean	Std. Deviation
Differanse i pulsverdier	32	-6,00	48,00	10,0312	16,41299
Valid N (listwise)	32				

a. KJØNN = Kvinne

Descriptive Statistics[a]

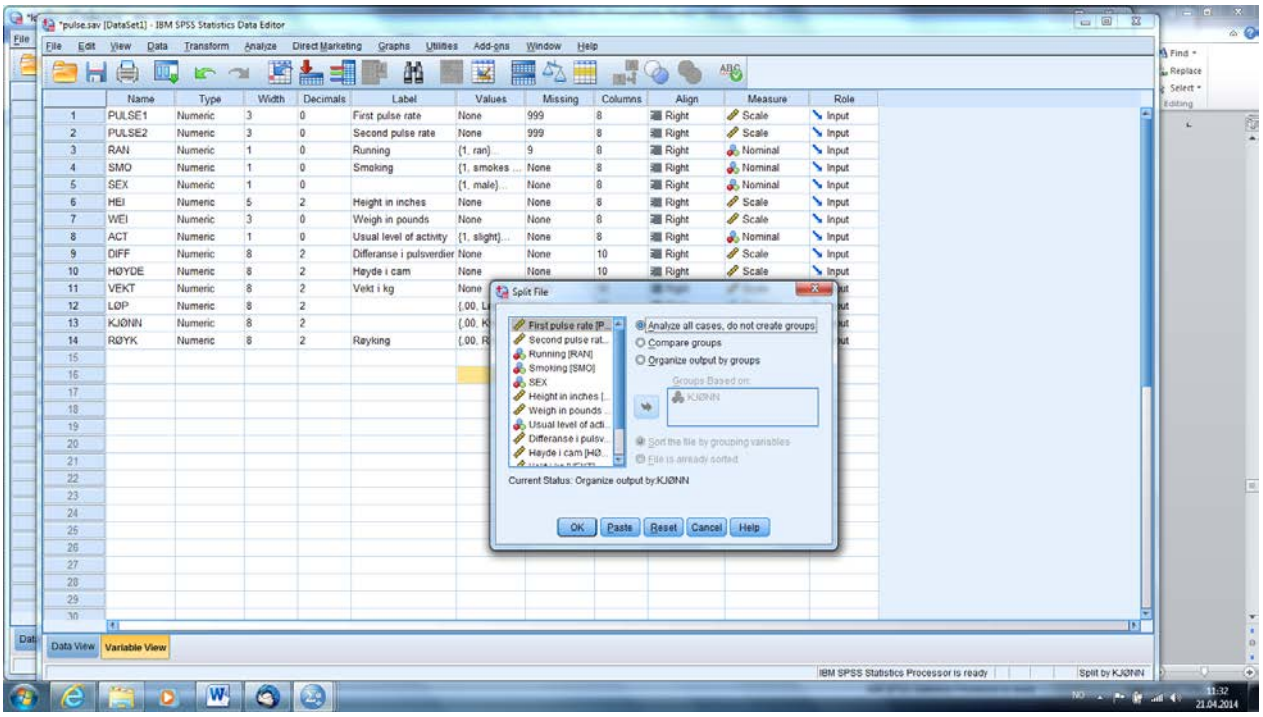
	N	Minimum	Maximum	Mean	Std. Deviation
Differanse i pulsverdier	53	-8,00	36,00	5,1887	9,80003
Valid N (listwise)	53				

a. KJØNN = Mann

Merk at vi først får utskrift for Kjønn = . Dette er for dem som har *Missing value* på KJØNN. Dem er vi selvfølgelig ikke interessert i verdien for, og vi ser også at gjennomsnittsverdien for DIFF ser rar ut.

Men deretter kommer verdiene for Kjønn = Kvinne og Kjønn = Menn, hver for seg. Vi ser at utskriften gir en fin oversikt over resultatene. Kvinner har en økning i pulsverdiene på 10.0, menn har en økning på 5.2.

Husk at når vi skal gå tilbake til å presentere resultatene samlet, må vi gå til *Data/Spilt file* og i dialogboksen velge *Analyze all cases do not create groups*, slik som vist under

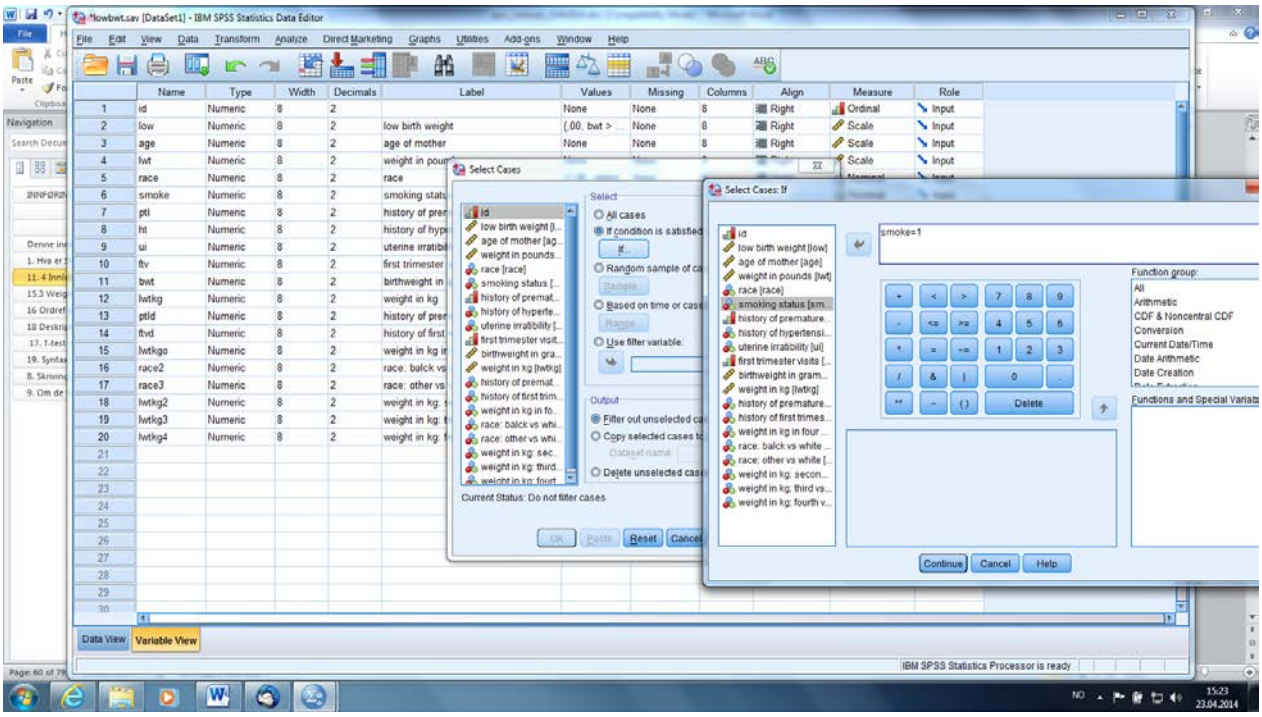


Når vi klikker på *OK*, blir senere resultater presentert for alle individer samlet.

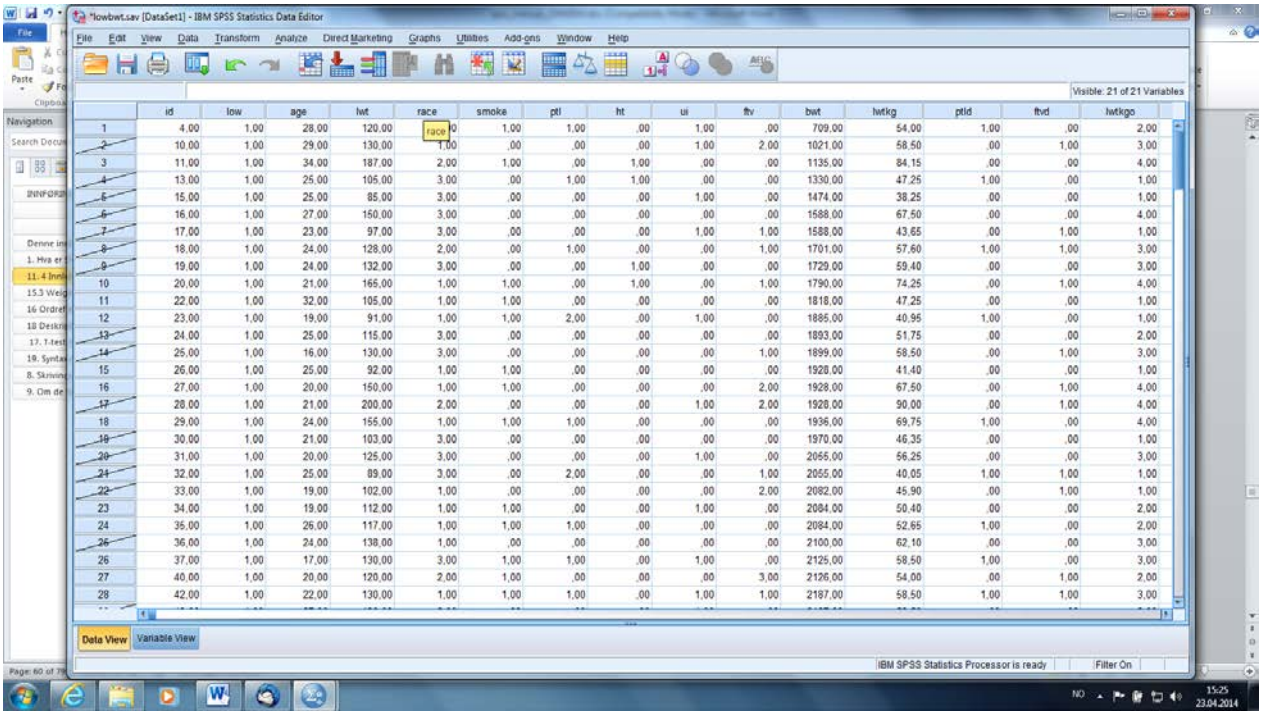
7.2 Select cases. Eksempel: lowbwt.sav

Når vi er interessert i flere analyser på undergrupper av individene i datavinduet vårt, er kommandoen *Select cases* mest aktuell for oss. Vi går inn i datafilen **lowbwt.sav**.

Anta at vi er interessert i å analysere røykende mødre. Vi velger *Data/ Select Cases* og ser at i denne dialogboksen i utgangspunktet er satt på *All cases* alle. Vi merker av på *If condition is satisfied* og klikker på *If*. Da åpner det seg en ny dialogboks nemlig *Select Cases: If*. Der må vi velge hvilke personer som skal være med. Vi trekker over SMOKE i boksen til høyre og skriver så = 1. Med dette valget får med de individene som har SMOKE = 1, altså røykerne. Da ser dialogboksene våre slik ut:



Vi klikker på *Continue* og *OK*, og får utført kommandoen. Men det skjer tilsynelatende ikke noe. Men la oss gå tilbake til dataarket, med *Data View*. Da ser datafilen vår slik ut:



Legg merke til at det nå står *Filter On* til høyre på nederste grå linjen. Men hvis vi nå ser i datavinduet vil vi se at mange av individene har en strek over sitt identifikasjonsnummer helt ut til venstre. Det er de som røyker som har fått en strek over seg. De er da filtrert ut.

La oss nå gjøre noen analyser på datafilen vår. Vi går inn i *Analyze/Descriptive Statistics*/*Frequencies* og trekker over LOW i vinduet. Denne variabelen angir om barnet hadde en fødselsvekt under grensen på 2500 gram. Da får vi følgende resultat:

low birth weight

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	bwt > 2500g	44	59,5	59,5	59,5
	bwt < 2500g	30	40,5	40,5	100,0
	Total	74	100,0	100,0	

Vi ser at det er 40.5% av barna født av røykende mødre som har en fødselsvekt under 2500 gram.

La oss se på gjennomsnitt og standardavvik blant disse barna. Da går vi til *Analyze/Descriptive Statistics/Descriptives* og trekker over BWT i vinduet. Når vi klikker på *OK*, får vi følgende resultat:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
birthweight in grams	74	709,00	4238,00	2773,2432	660,07517
Valid N (listwise)	74				

Vi ser at gjennomsnittet for barn født av røykende mødre er 2773 gram, standardavviket er 660 gram, og den minste og største fødselsvekten er henholdsvis 709 gram og 4238 gram.

Hvis vi skal gjøre mange analyser på en slik undergruppe kan vi lage en egen SPSS datafil (en ***.sav**-fil) for undermaterialet. Da går vi tilbake til *Data/Select cases*. I dialogboksen går vi nederst i Output: og klikker av på Da klikker vi på *Delete unselected cases*. Når vi klikker på *OK* og går tilbake til datavinduet, ser vi at alle enhetene som ble selektert ut, og dermed fikk en strek over seg, nå er tatt ut. Da kan vi lagre filen i datavinduet som en egen **sav**-fil. Men vi må selvfølgelig velge et nytt navn slik at vi bevarer hele materialet på den originale **sav**-filen.

Når vi har et filter på, dvs. velger ut en undergruppe av hele datafilen vår, setter SPSS opp en midlertidig variabel FILTER_$. Den er 0 for de individene som ikke er med og 1 for de som er med. Hvis vi ser helt til høyre i datavinduet vårt, vil vi se denne variabelen.

Når vi skal gå tilbake til hele datafilen, og gjøre nye analyser for alle enhetene, må vi gå tilbake til *Data/Select Cases*, og velge *All cases* i dialogboksen. Når vi klikker på *OK*, og går tilbake til datavinduet, ser vi at alle strekene over de selekterte enhetene er tatt bort.

Husk at dersom vi skulle gjøre en feil i en dataseleksjon, kan vi alltid gå tilbake og hente frem igjen den filen vi arbeider med. Det gjør vi via *File/Open/Data*. Men dette forutsetter selvfølgelig at vi ikke har overskrevet den opprinnelige filen med en datafil som inneholder de selekterte enhetene. Derfor er det svært viktig at vi alltid bruker NYE filnavn når vi lager filer som er lagd på undergrupper av dataene våre.

7.3 Weight cases. Eksempel: blodtrykk.sav

I mange tilfeller ønsker vi å analysere data som er presentert i bøker eller artikler. Dersom vi skal analysere kontinuerlige data, slik som energiforbruk i datasettet **altman.sav**, må vi selvfølgelig ha rådata, dvs. alle verdiene for alle enhetene. Disse må vi da legge inn i SPSS for videre analyser.

Men i mange tilfeller er vi interessert i å analysere data som er presentert i tabellform. Da er dataene aggregert i de cellene som utgjør kombinasjonene av de variablene vi studerer. Et eksempel på en slik tabell er tabellen nedenfor. Her ser vi på sammenhengen mellom behandling av aspirin mot blodtrykk. Alle i studien hadde høyt blodtrykk før inklusjon i studien. Antallet i cellene angir antallet som gikk ned til normalt blodtrykk etter behandling med aspirin og i antallet som gikk ned i kontrollgruppen. Alle i kontrollgruppen fikk placebo. Her vil vi altså sammenligne undersøke hvor mange som gikk ned fra høyt til normalt blodtrykk.

Vi ser at 30 av de 34 i aspirin-gruppen gikk ned til normalt blodtrykk, mens 20 av de 31 i placebo-gruppen gikk ned til normalt blodtrykk. Vi er interessert i å undersøke sammenhengen mellom behandling og blodtrykk. Hvordan vi skal analyse sammenhengen skal vi se på senere. Nå skal vi lære å lese inn dataene fra en tabell, ved å bruke *Weight Cases*.

I tabellen har vi to variabler:

Behandling, med kodene 0 = Placebo, 1 = Aspirin
Blodtrykk, med kodene 0 = Høyt blodtrykk og 1 = Normalt blodtrykk

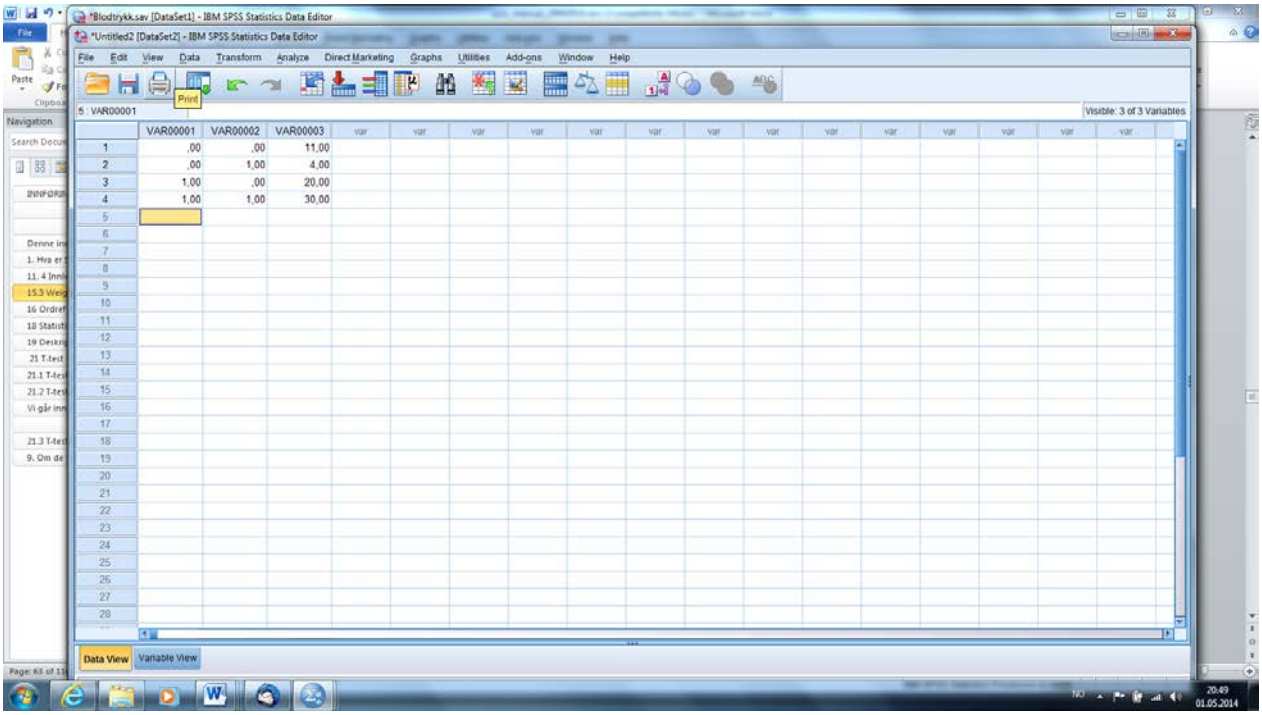
Merk at vi alltid bruker kodene 0 for referansekategorien, i denne sammenhengen de som er i placebogruppen og de med høyt blodtrykk. Da vil cellen i øverste venstre cellen være identifisert med 0, 0, siden Blodtrykk = 0 og Behandling = 0. Cellen i øverste høyre hjørne vil være 0, 1, siden Blodtrykk = 0 og Behandling = 1. Antallet i de to cellene er henholdsvis 11 og 4.

	Behandling		
Blodtrykk	Placebo	Aspirin	Totalt
Høyt blodtrykk	11	4	15
Normalt blodtrykk	20	30	50
Totalt	31	34	65

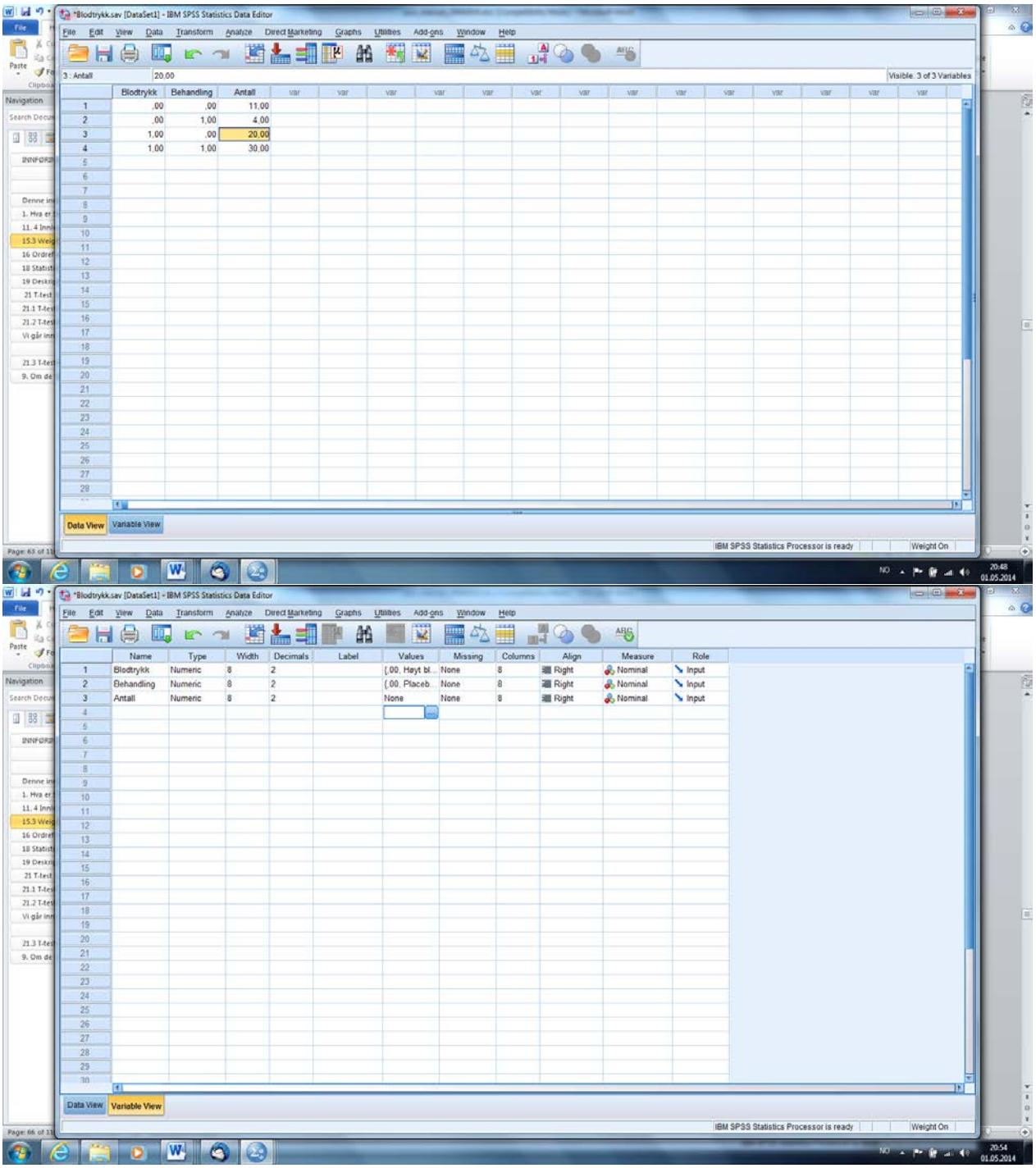
Når vi leser inn disse dataene, leser vi dem inn fra tabellen linjevis. Vi leser øverste linje først og deretter linjen under. Vi leser først inn hvilken celle vi er i, og så antallet i den cellen. Vi leser inn fire linjer, hver med tre kolonner:

0	0	11
0	1	4
1	0	20
0	0	30

Første linje angir øverste venstre celle, neste linje øverste høyre celle, og så gå vi videre til neste linje i tabellen. Når vi leser dette inn i SPSS, blir det slik:

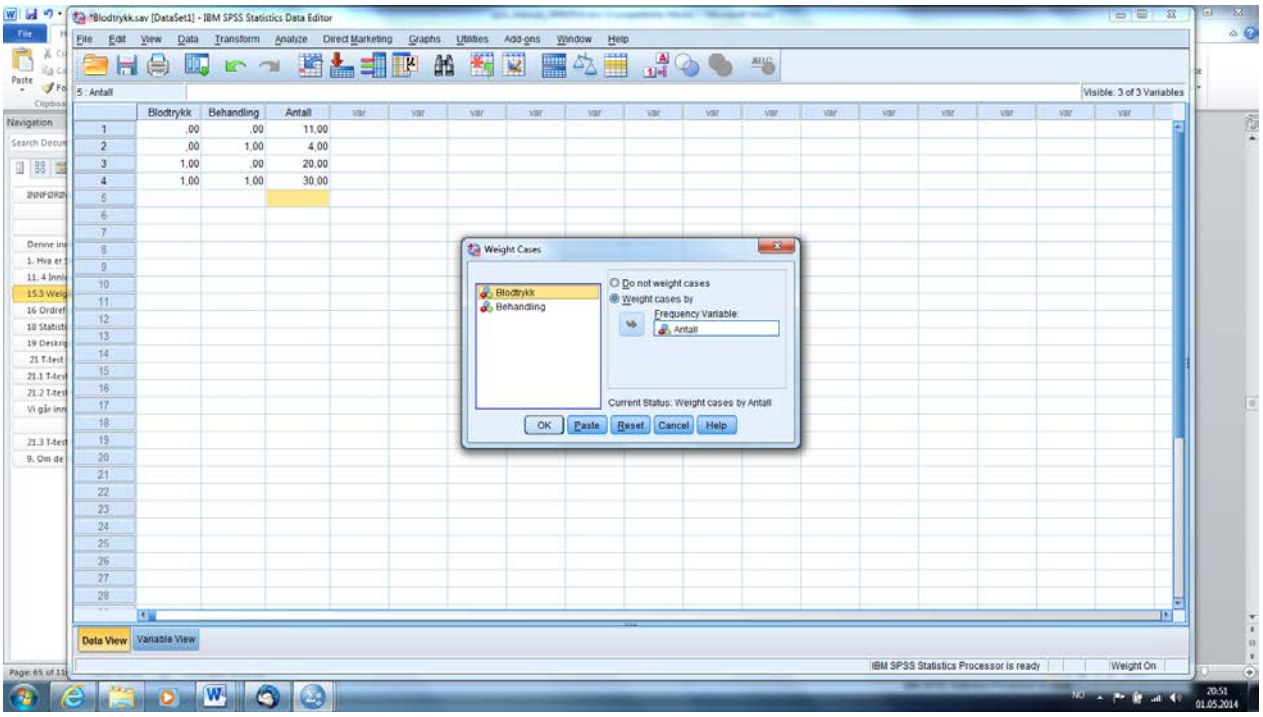


Så må vi navngi variablene. Siden vi leser inn tabellen linjevis, er første variabel Blodtrykk og annen variabel er Behandling. Siste variabel er Antall. Vi angir variabelnavnene, og går til *Value label*. For Blodtrykk angir vi 0 for Normalt blodtrykk og 1 for Høyt blodtrykk. For Behandling angir vi 0 for Placebo og 1 for Aspirin. Da blir dataarket, med *Data View* og *Variable View* slik:

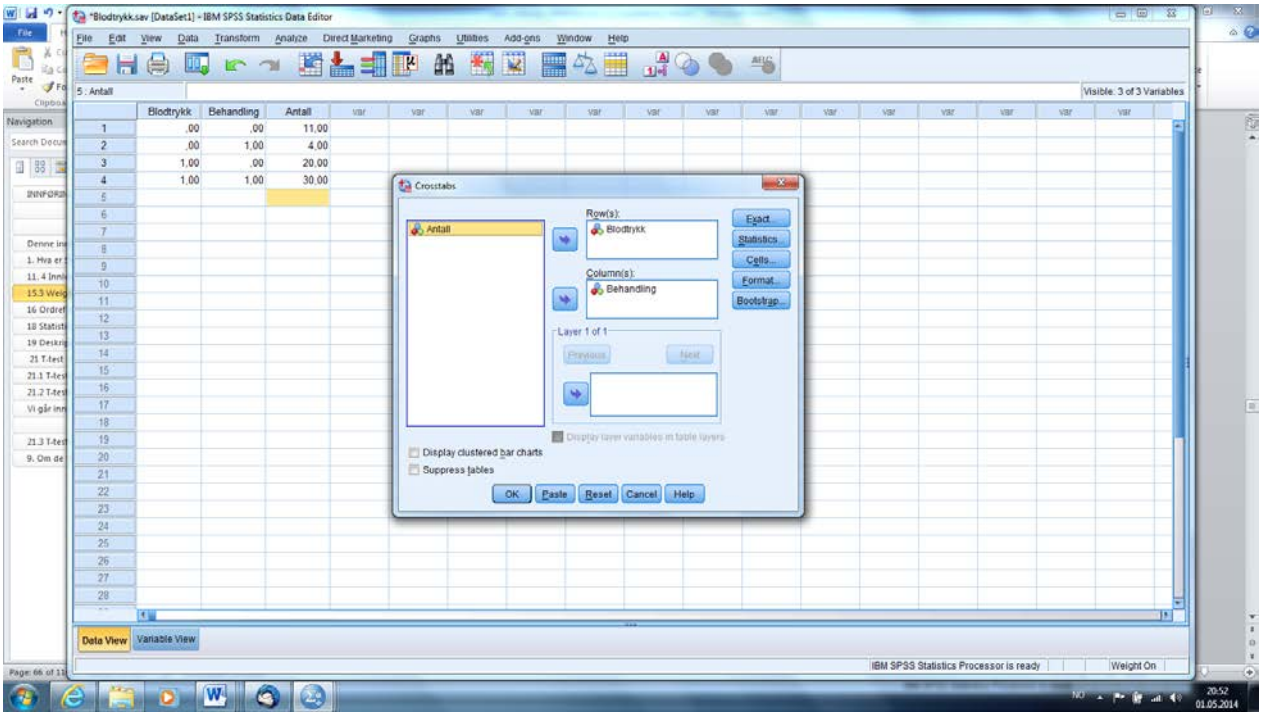


Nå er tiden inne for å legge denne filen ned i katalogen vår. Vi gir den navnet **blodtrykk.sav** og legger den i katalogen med de andre filene våre.

Nå tror SPSS at dette er en datafil med bare fire enheter. Vi må fortelle SPSS at dette er tall som skal vektes opp. Det gjør vi ved *Data/Weight cases*. Vi kommer inn i en dialogboks. Der må vi markere at vi skal *Weigh cases by*, og i vinduet som åpner seg, trekker vi over ANTALL.



Merk at det ikke skjer noe på dataarket vårt. Men SPSS er nå klar over at dette egentlig er 65 personer, som er fordelt i celler etter antallet vi har angitt. Nå kan vi da lage vår første krysstabell, mellom BLODTRYKK og BEHANDLING. Vi går inn *Analyze/Descriptive Statistics/Crosstabs*. I den dialogboksen som åpner seg legger vi over BLODTRYKK i *Rows* og BEHANDLING i *Columns*. Da ser dialogboksen vår slik ut:



Når vi klikker på *OK*, får vi følgende utskrift:

Blodtrykk * Behandling Crosstabulation

Count

		Behandling		
		Placebo	Aspirin	Total
Blodtrykk	Høyt blodtrykk	11	4	15
	Normalt blodtrykk	20	30	50
Total		31	34	65

Her er det viktig å merke seg følgende: SPSS presenterer alltid tabellen med laveste kode først. Det betyr at Normalt blodtrykk og Placebo legges i øverste venstre hjørne, og Høyt Blodtrykk og Aspirin i nederste høyre hjørne.

Vi kommer tilbake analyse av krysstabeller i kapittel 11.4.

8 Statistiske analyser via ordrefiler

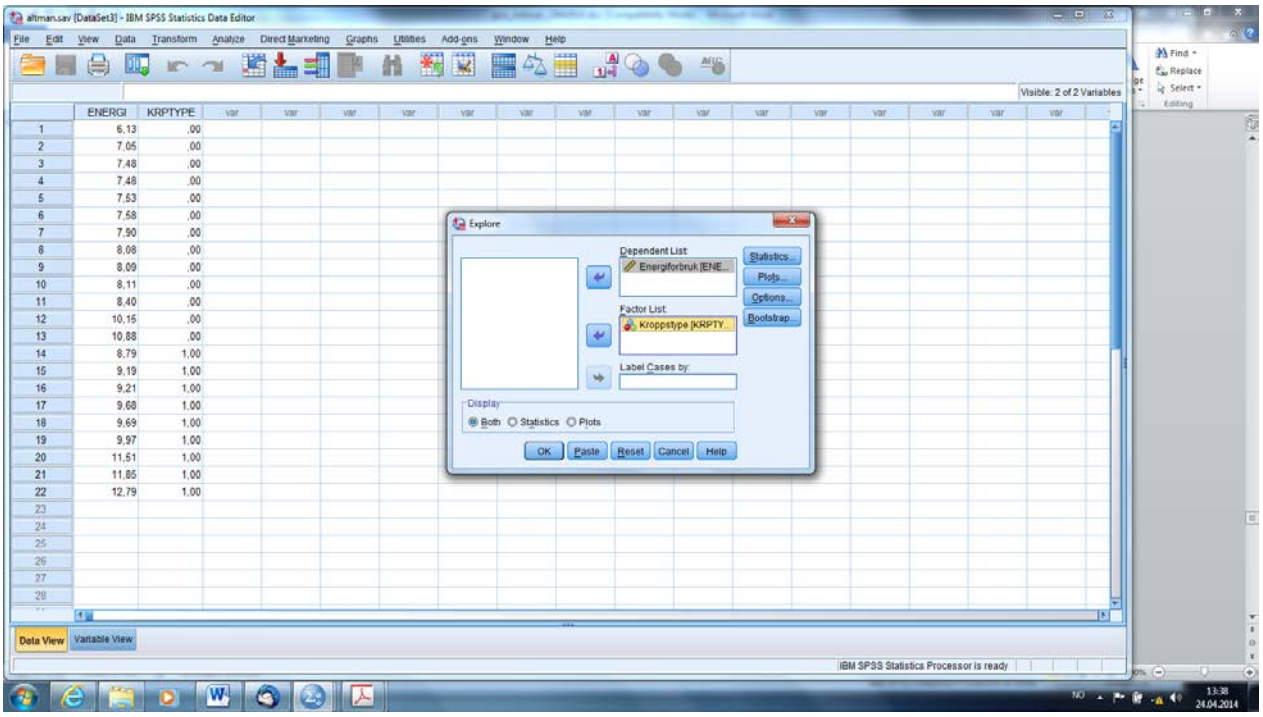
Læringsmål

Ordrefiler er en filer som inneholder en rekke SPSS-ordrer. En SPSS-ordre kan vi enten skrive selv i et ordrevindu eller la SPSS lage for oss ved å klikke på *Paste* når vi er inne i en dialogboks, og innholdet i dialogboksen blir da skrevet ned til en ordefil. En samling av slike ordrer som lagres på en fil til seinere bruk er en ordrefil. Den kan utføres samlet av SPSS.

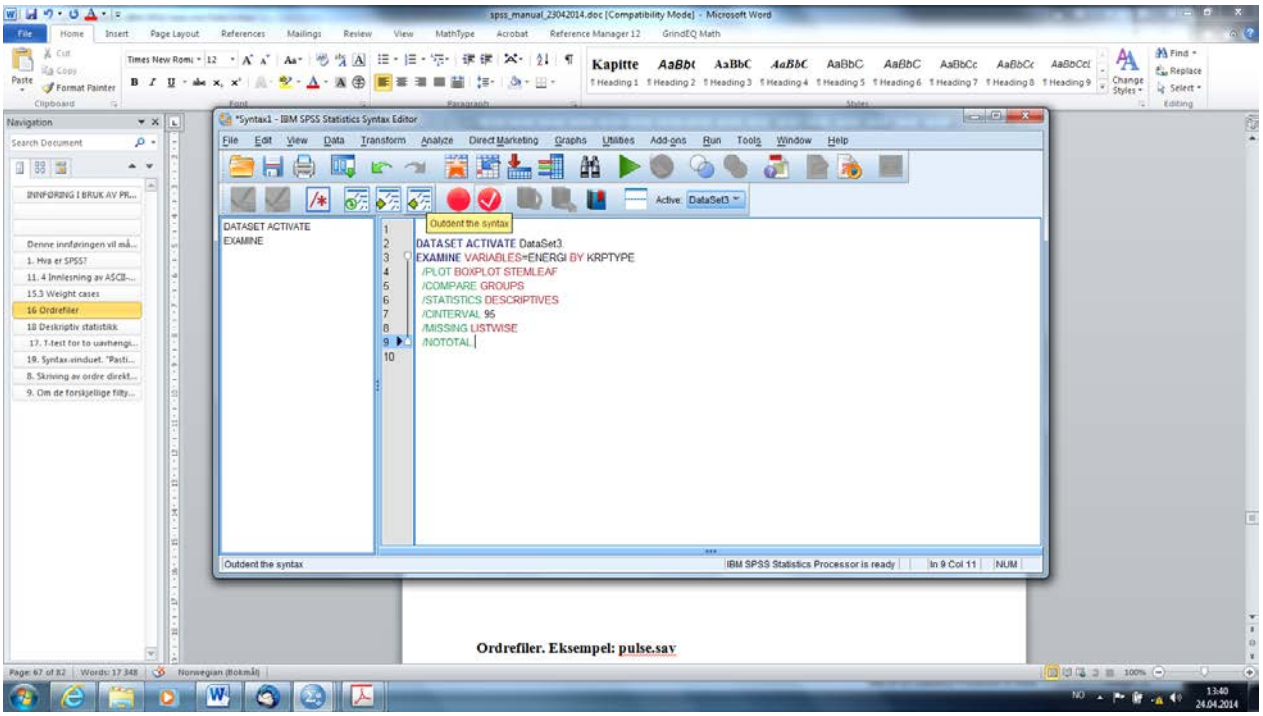
SPSS skriver også automatisk ut alle ordre som utføres fortløpende i utskriftsfilen. Men vi kan ikke kjøre analyser fra utskriftsfilen og vi kan heller ikke kopiere fra den. Det er derfor lurt å bruke ordrefiler. I dette kapittelet skal vi lære hvordan vi lager og kjører ordrefiler i SPSS.

8.1 Ordrefiler. Eksempel: altman.sav

Vi åpner datafilen **altman.sav**. Vi gjør nå en enkel analyse, som vi også har gjort tidligere. Vi går inn i *Analyze/Descriptive Statistics/Explore*. I dialogboksen trekker vi over ENERGI i *Dependent List* og KRPTYPE i *Factor List.* Da ser dialogboksen våre slik ut:



Men nå klikker vi på *Paste*. Da ender vi i et nytt vindu – ordrevinduet – som SPSS kaller *Syntax*. Det ser slik ut:



Her er det mange linjer vi ikke trenger å bekymre oss om, hverken nå eller senere. Hvis vi når går tilbake til dataarket, ser vi at dialogboksen med ordrene våre er borte. Ordrene ligger nå bare i ordrefilen. For å få kjøre ordrene må vi da være i ordrefilen. Vi markerer alle ordrene i ordrefilen. Så går vi opp til den øverste av de to knappelinjene og klikker på den grønne pilen. Da blir ordren utført og vi endrer i utskriftsfilen, der resultatene våre ligger. Vi går ikke inn på dem nå, men kommer tilbake til dem senere.

I SPSS skal i nå ha liggende tre typer filer. Vi har liggende én eller flere datafiler, én utskriftsfil og én ordrefil. Det kan være lurt å legge ordrefilen ned i katalogen våre. Vi gjør det ved å gå inn i ordrefilen. Da går vi inn i *File/Save as*. Da åpner det seg en dialogboks hvor SPSS foreslår å arkivere denne som en syntaxfil med ekstensjon **sps**. Det velger vi, og lar det stå. Men vi velger nytt filnavn, og foreslår **altman**. Vi velger også å legge denne i katalogen sammen med datafilene våre. Da vil filen **altman.sps** bli lagret i denne katalogen. Merk at vi nå har to filer **altman.sav** og **altman.sps**, som er henholdsvis datafilen og ordrefilen. Vi må ikke blande dem sammen!

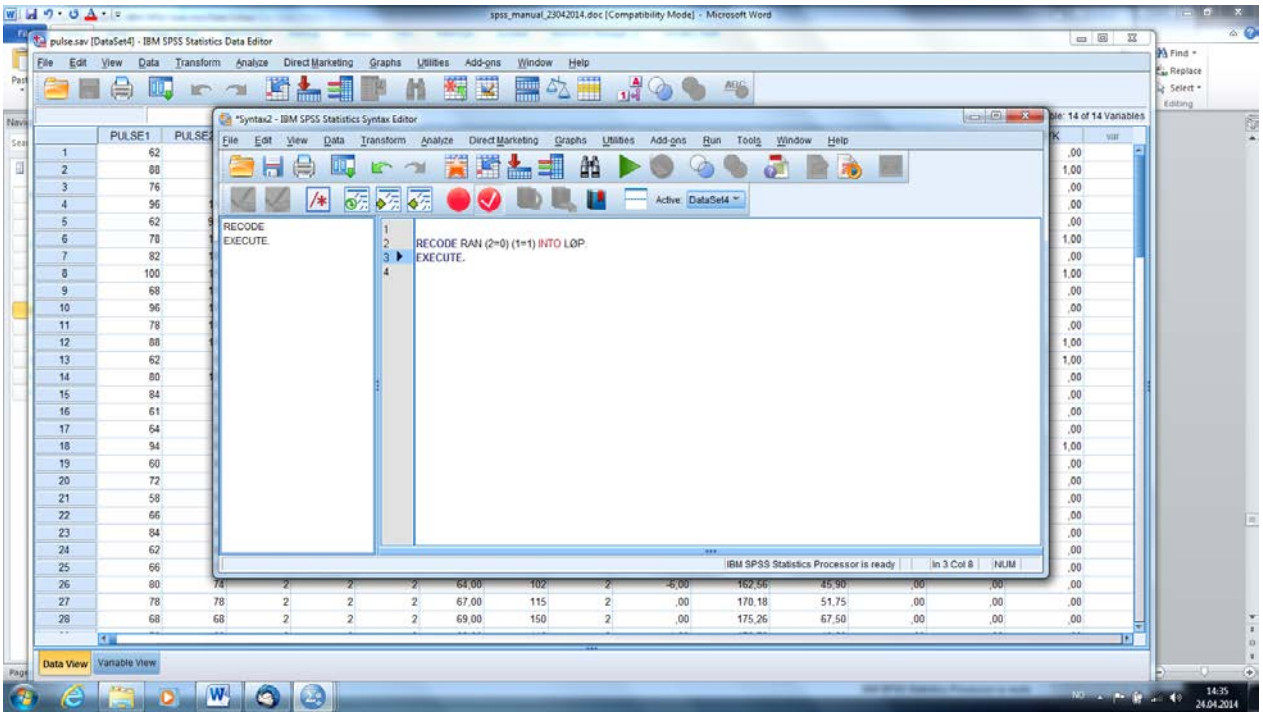
Når vi har lagret denne ordrefilen, stenger vi ned den ved å gå til den røde boksen i øvre hjørnet i ordrevinduet.

8.2 Ordrefiler. Eksempel: pulse.sav

Vi henter frem datfilen **pulse.sav**. Vi husker at vi gjorde en rekke omkodinger på den datafilen. La oss nå se om vi kan forenkle dem ved å bruke ordrefiler.

I kapittel 6 gjorde vi omkodinger av RAN til LØP, SEX til KJØNN og SMO til RØYK ved å bruke *Transform/Recode into Different Variables*.

La oss nå gjøre omkodingen av RAN på nytt. Vi går inn i *Transfrom/Recode into Different Variables*, og vi trekker RAN over i boksen i midten. I Output-vinduet som åpner seg skriver vi inn LØP, og vi skifter ved å klikke på *Change*. Da får vi opp en advarsel om vi har brukt dette navnet får. Vi klikker bare på *OK* i denne dialogboksen. Vi går så ned til *Old and New Values*, og da kommer vi inn i en ny dialogboks. I boksen med *Old Values* skriver vi inn 2 og i boksen med *New Values* skriver vi 0. For å få denne aktivisert må vi klikke på *Add*. Vi må også la koden 1 for RAN også være koden 1 for LØP. Det gjør vi enkelt ved å skrive 1 for *Old Values* og 1 for *New Values*. Til slutt klikker på *Add*. Da er vi klare til å klikke på *Continue*. Da ender vi i den første dialogboksen. Her klikker vi da på *Paste*. Da ender vi igjen i ordrevinduet, som nå ser slik ut:

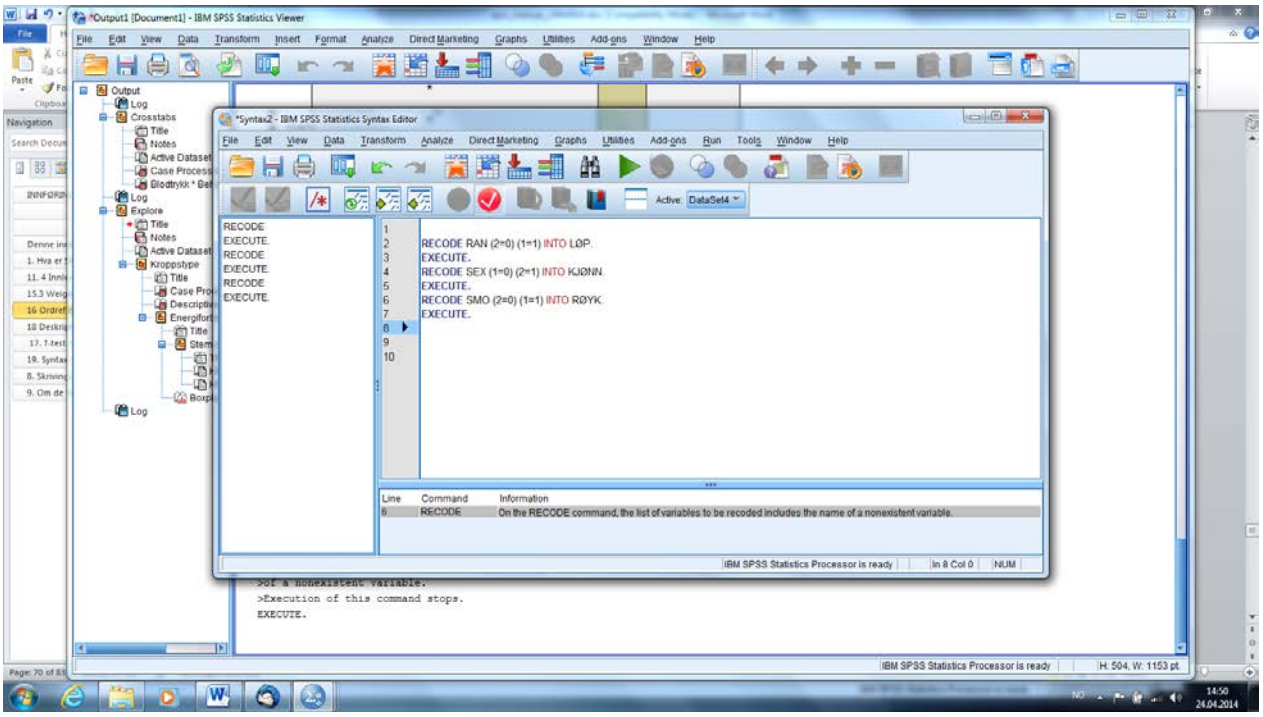


Merk at hvis vi ikke har stengt ned filen **altman.sps**, vil ordrene i den filen bli liggende over de nye ordrene.

Nå har vi mulighetene til å forenkle omkodingene på de andre variablene betydelig. Vi kan legge til følgende linjer i ordrefilen:

```
RECODE SEX (1=0) (2=1) INTO KJØNN.
EXECUTE.
RECODE SOM (2=0) (1=1) INTO RØYK.
EXECUTE.
```

Merk at hver ordre skal avsluttes med et punktum, og at hver ordre må etterfølges at et EXECUTE, for å bli utført. Da blir ordrefilen vår våre seende slik ut:



For å få utført disse omkodingene markerer vi alle linjene i ordrefilen, og går så opp til den grønne pilen på knappelinjen. Når vi klikker på den, får vi utført ordrene som er markert.

Når vi skal gjøre mange analyser, kan det være besparende å gjøre det via ordrefiler. Det kan ofte også være lurt å arkivere ordrefilene våre, slik at vi vet hvilke analyser som er gjort. Vi avslutter derfor dette kapittelet med å gå til *File/Save As* og velge filnavnet **pulse** på denne ordrefilen. Da blir **pulse.sps** lagret på den katalogen vi velger.

9 Deskriptiv analyse

Læringsmål

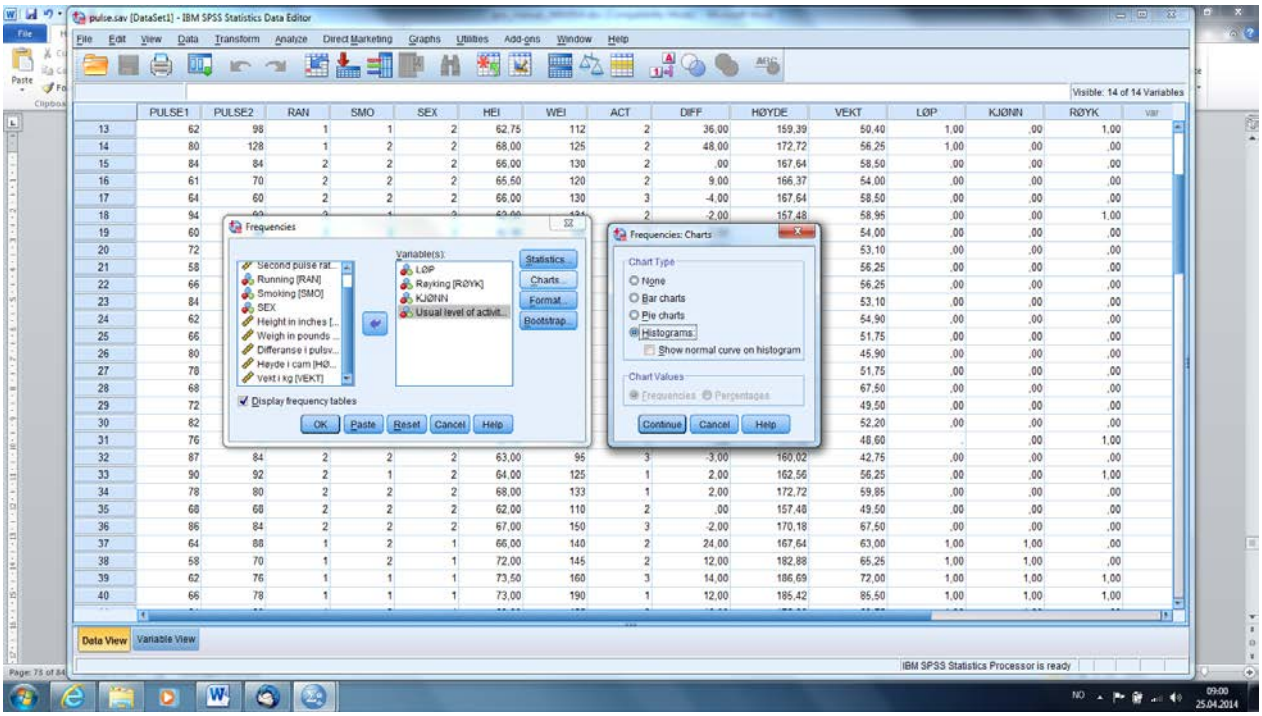
Vi starter gjennomgangen av de statistiske metodene med å vise til oversikten over datatyper og statistiske metoder i kapittel 3. Merk at analysemetodene er ulike avhengig av om vi har kategoriske eller kontinuerlige data.

Vi skiller mellom deskriptiv analyse, univariabel analyse og multivariabel analyse. Deskriptiv analyse ser på én og én variabel for seg. Vi starter alltid en statistisk analyse med en deskriptiv gjennomgang. De univariable metodene ser på sammenhengen mellom to variable. De multivariable metodene ser på sammenhengen mellom flere enn to variabler.

I dette kapittelet skal vi se på de deskriptive analysene. Dersom dataene er kategoriske finner vi kommandoene for en deskriptiv analyse under *Analyze/Descriptive Statistics/Frequencies*. Dersom dataene er kontinuerlige finner vi kommandoene under *Analyze/Descriptive Statistics/Descriptives.* Under begge disse kommandoene kan vi lage grafiske fremstillinger (diagrammer, plott og figurer). Kommandoen *Analyze/Descriptive Statistics/Explore* inneholder større muligheter til å analysere undergrupper i dataene våre, til å lage forskjellige plott og til å sjekke om data er normalfordelte.

9.1 Frequencies. Eksempel: pulse.sav

Frequencies brukes på kategoriske variable for å beskrive frekvensfordelingen til variablene. Vi går inn *Analyze/Descriptive Statistics/Frequencies.* Som vanlig åpner det seg en dialogboks. Der trekker over variablene LØP, KJØNN, RØYK og ACT. Vi ser også at det er en knapperekke til høyre for vinduet der variablene LØP, KJØNN, RØYK og ACT ligger. For de kategoriske variablene er det *Chart* som er av interesse. Vi klikker derfor på den. Da åpner det seg en ny dialogboks, der vi kan velge type grafisk fremstilling. Vi velger *Histogram*. Da ser dialogboksene våre slik ut:



Vi klikker på *Continue* her og *OK* på neste dialogboks. For LØP og ACT får vi følgende utskrift:

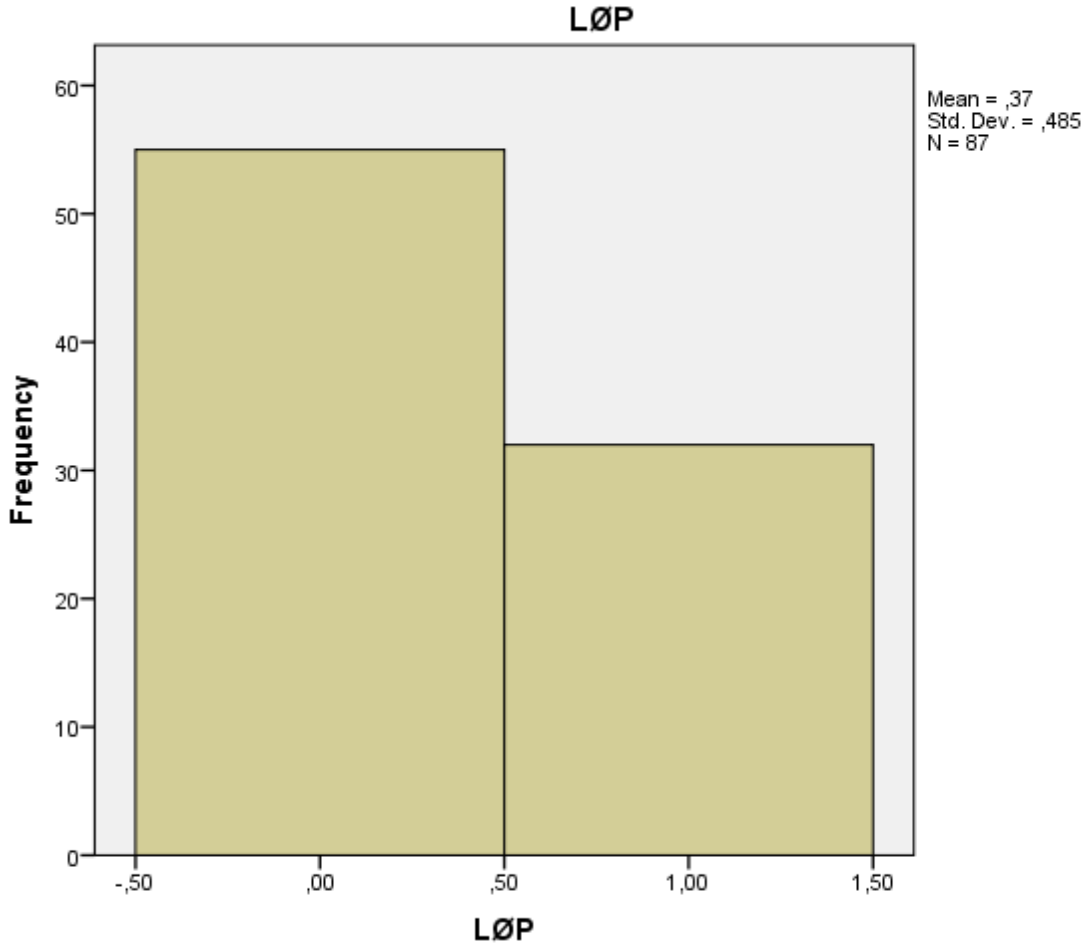
LØP

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Løp ikke	55	59,8	63,2	63,2
	Løp	32	34,8	36,8	100,0
	Total	87	94,6	100,0	
Missing	System	5	5,4		
Total		92	100,0		

Vi ser at undersøkelsen omfattet 92 individer. For 5 av disse ble det ikke registrert om personen løp, og for disse er det registrert *Missing*. Under Frequency og Valid Percent ser vi at det var 32 som løp (36.8%) og 55 som ikke løp (63.2%). Det er viktig at vi alltid bruker prosenfordelingen som er gitt i Valid Percent. Hvis vi hadde unnlatt å oppgi missing value, ville vi feilaktig funnet at det var 59.8% som ikke løp og 34.8% som løp, slik det står under Percent.

Vi ser også nytten av å ha *Value labels*. Da vet vi at kode 1 betyr at personen løp.

Nedenfor har vi historgrammet for LØP, som for en variabel med bare to kategorier er enkelt. Legg merke til at *Missing values* også er tatt ut av denne grafen.

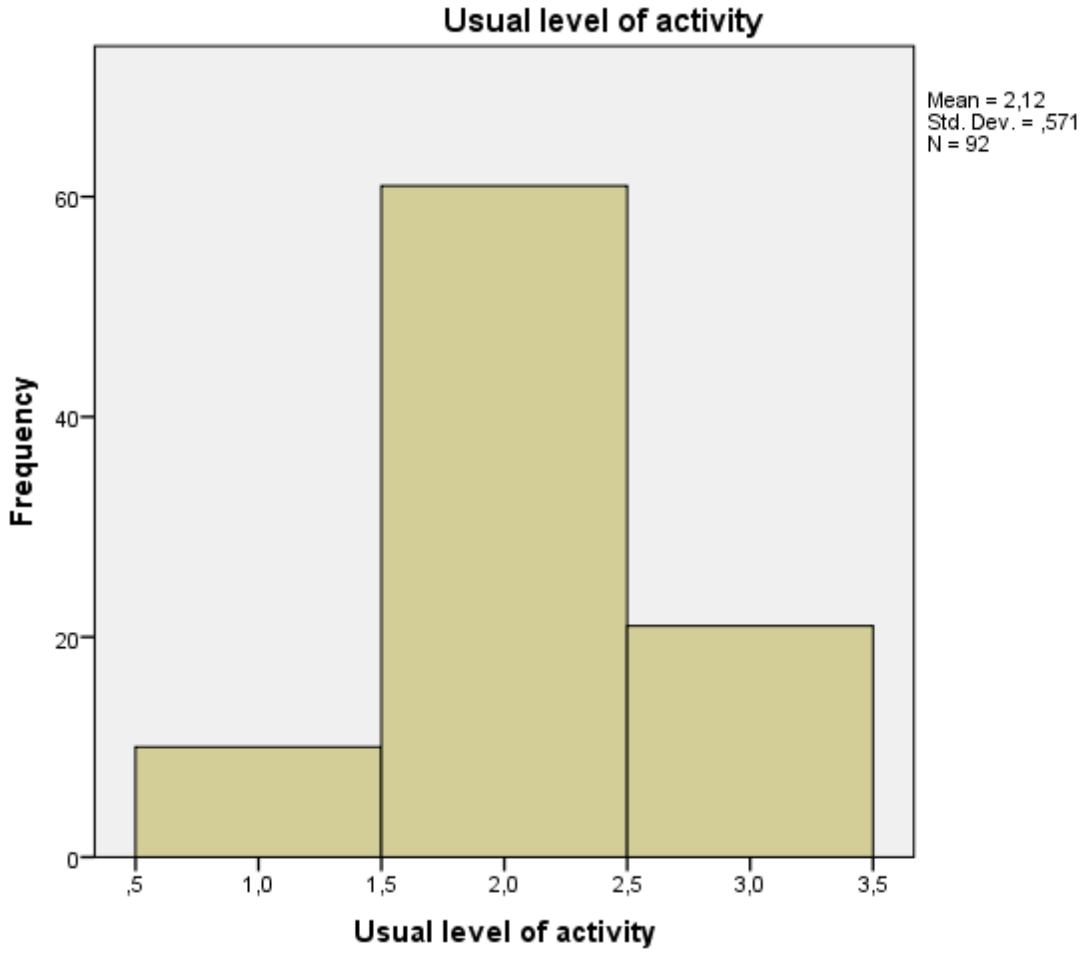


For ACT får vi følgende resultater:

Usual level of activity

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	slight	10	10,9	10,9	10,9
	moderate	61	66,3	66,3	77,2
	a lot	21	22,8	22,8	100,0
	Total	92	100,0	100,0	

Her ser vi at det er tre kategorier, og en litt mer interessant frekvensfordeling. Siden det ikke er *Missing values* for denne variabelen, blir fordelingen i Percent og i Valid Percent den samme. Grafen for ACT blir også litt mer interessant:



9.2 Frequencies. Eksempel: lowbwt.sav

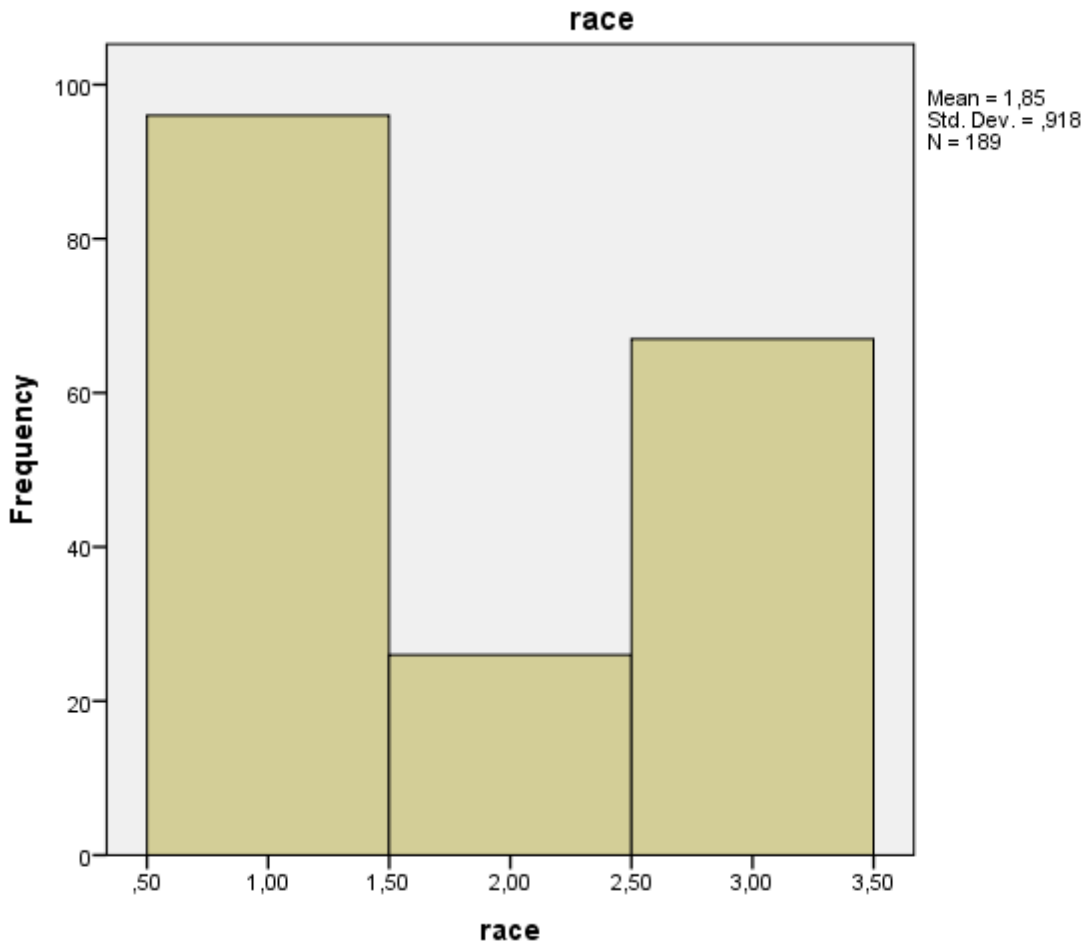
Vi leser inn datafilen **lowbwt.sav**. Vi går inn igjen i *Analyze/Descriptive Statistics/Frequencies*, og trekker over variablene LOW og RACE og under velger vi igjen *Histogram*. Da får vi følgende resultater:

low birth weight

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	bwt > 2500g	130	68,8	68,8	68,8
	bwt < 2500g	59	31,2	31,2	100,0
	Total	189	100,0	100,0	

race

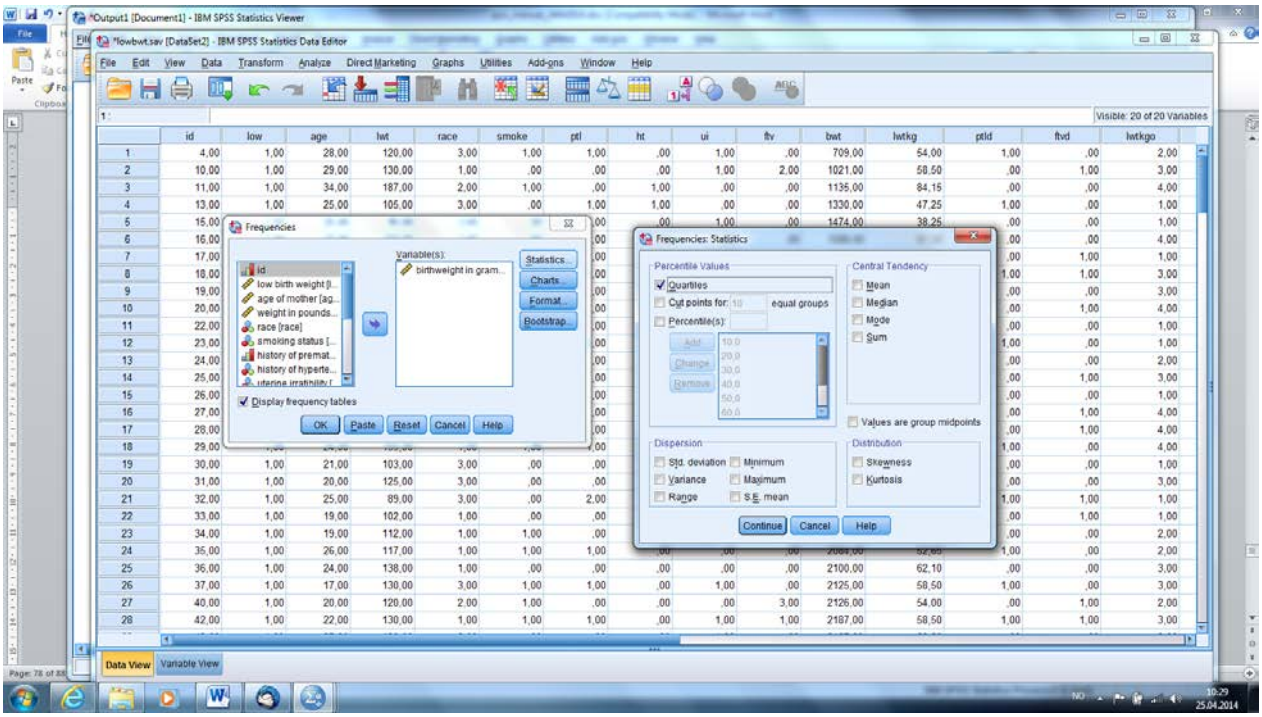
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	white	96	50,8	50,8	50,8
	black	26	13,8	13,8	64,6
	other	67	35,4	35,4	100,0
	Total	189	100,0	100,0	



Vi ser at resultatene er enkle å fortolke, enten ved selve frekvensoversikten eller ved histogrammet.

La oss nå se på en kommando som vi kan få utført gjennom *Analyze/Descriptive Statistics/Frequencies,* selv om vi nå snakker om en kontinuerlig variabel. Ofte er det interessant å finne kvartiler eller presentiler i en fordeling. Kvartilene er observasjoner som deler hele datamaterialet i fire like store deler. Nederste kvartil er den observasjonen som har 25% av observasjonene nedenfor seg, midterste kvartil har 50% av observasjonene nedenfor seg, og øverste kvartil har 75% nedenfor seg. Persentilene defineres i forhold til en gitt prosent. 10-persentilen har 10% av observasjonene nedenfor, 80-persentilen har 80% nedenfor seg.

Vi kan nå bruke *Analyze/Descriptive Statistics/Frequencies* til å finne kvartiler og persentiler for kontinuerlige variabler. Vi skal nå finne kvartiler og 10-90-persentiler for variabelen BWT. Vi trekker over over BWT i vinduet i dialogboksen og klikker på knappen *Statistics*. I dialogboksen som da åpner seg, klikker vi av på *Quartiles*. Da ser dialogboksene våre slik ut:



Og vi får følgende utskrift:

Statistics

birthweight in grams

N	Valid	189
	Missing	0
Percentiles	25	2412,0000
	50	2977,0000
	75	3481,0000

Vi ser at nedre kvartil er 2412 gram og øvre kvartil er 3481 gram.

For å få ut 10-90 persentilene kan vi klikke på *Cut points for 10 groups*, istedenfor Quartiles. Da får vi følgende utskrift:

Statistics

birthweight in grams

N	Valid	189
	Missing	0
Percentiles	10	1970,0000
	20	2325,0000
	25	2412,0000
	30	2495,0000
	40	2778,0000
	50	2977,0000
	60	3175,0000
	70	3374,0000
	75	3481,0000
	80	3629,0000
	90	3884,0000

Her ser vi at 10-persentilen er 1970 gram. Vi ser også at 50-persentilen er 2977 gram, som i tabellen over stemmer med at annen kvartil er 2977 gram.

9.3 Descriptives. Eksempel: pulse.sav

Denne kommandoen gir oss bl.a gjennomsnitt og standardavvik og er egnet for presentasjon av kontinuerlige variable. Vi henter frem igjen datafilen **pulse.sav.** Her skal vi nå se på de kontinuerlige variablene.

Vi klikker på *Analyze/Descriptive Statistics/Descriptives*. Når vi har kommet inn i *Descriptives* dialogboksen trekker vi over PULSE1, PULSE2, HØYDE og VEKT i vinduet og trykker *OK*. Dette gir denne utskriften:

Descriptive Statistics

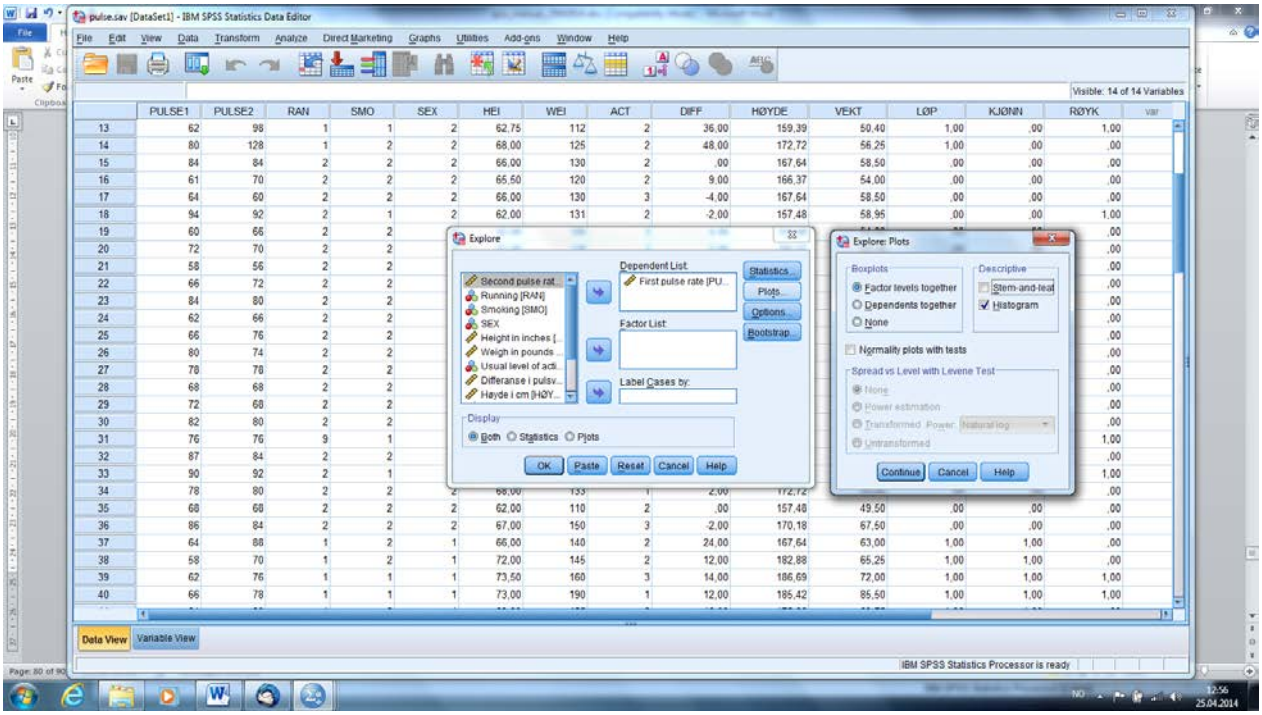
	N	Minimum	Maximum	Mean	Std. Deviation
First pulse rate	90	48	100	72,69	11,062
Second pulse rate	90	50	140	79,40	16,707
Høyde i cm	92	154,94	190,50	174,5422	9,29460
Vekt i kg	92	42,75	96,75	65,3185	10,68273
Valid N (listwise)	88				

Vi ser for eksempel at i gjennomsnitt, samlet de som løp og ikke løp, økte pulsen fra 72.7 til 79.4. Vi ser også at standardavviket er noe større etter løping, nemlig 16.7 mot 11.1, og maksimal pulseverdi er også vesentlig større ved annen måling.

9.4 Explore. Eksempel: pulse.sav

Kommandoen *Explore* gir oss mange muligheter til å se på hele fordelingen til en variabel, både i tall og grafisk fremstilling. *Explore* gir også mulighet til å se på sammenhengen mellom to variabler.

Vi skal se på variabelen PULSE1. Vi klikker på *Analyze/Descriptive Statistics/Explore* og kommer inn ny dialogboks. Denne dialogboksen har en avhengig variabel dvs. den variabelen vi vil forklare eller undersøke. Dette er variabelen som skal inn i *Dependent Variable(s).* For oss er det PULSE1. Vi flytter den over. Vi ser at det også her er en knapperekke til høyre. Der går vi inn i *Plots*. Da åpner det seg en dialogboks der vi tar bort haken på *Stem-and-leaf*, men setter inn en hake på *Histogram*. Da ser dialogboksen slik ut:



Vi klikker på *Continue* og *OK* og får da følgende resultat i utskriftsvinduet:

Descriptives

			Statistic	Std. Error
First pulse rate	Mean		72,69	1,166
	95% Confidence Interval for Mean	Lower Bound	70,37	
		Upper Bound	75,01	
	5% Trimmed Mean		72,43	
	Median		70,00	
	Variance		122,374	
	Std. Deviation		11,062	
	Minimum		48	
	Maximum		100	
	Range		52	
	Interquartile Range		16	
	Skewness		,441	,254
	Kurtosis		-,413	,503

Her er det mye informasjon. Det meste vil bli gjennomgått i statistikkundervisningen, men vi må gå gjennom noen viktige punkter her. Først får vi gjennomsnittet (Mean), som for PULSE1 er 72.69. Vi får også beregnet et 5% trimmet gjennomsnitt (5% Trimmed Mean) som er det gjennomsnittet vi får når vi tar ut de 5% største og de 5% minste observasjonene. Vi får også medianen, som er den observasjonen som deler datamaterialet i to like store deler. Median er da lik annen kvartil. Alle disse er mål for hvor observasjonene er sentrert.

Så har vi forskjellige mål for spredning. Vi får skrevet ut Std. Deviation som er standardavviket, og som er beregnet til 11.06. Variansen (Variance) er kvadratet av standavviket, og er 122.37. Interkvartil differansen (Interquartile Range) er avstanden mellom øvre og nedre kvartil. Differansen mellom største (Maximum) og minste (Minimum) observasjon er også et mål for spredning.

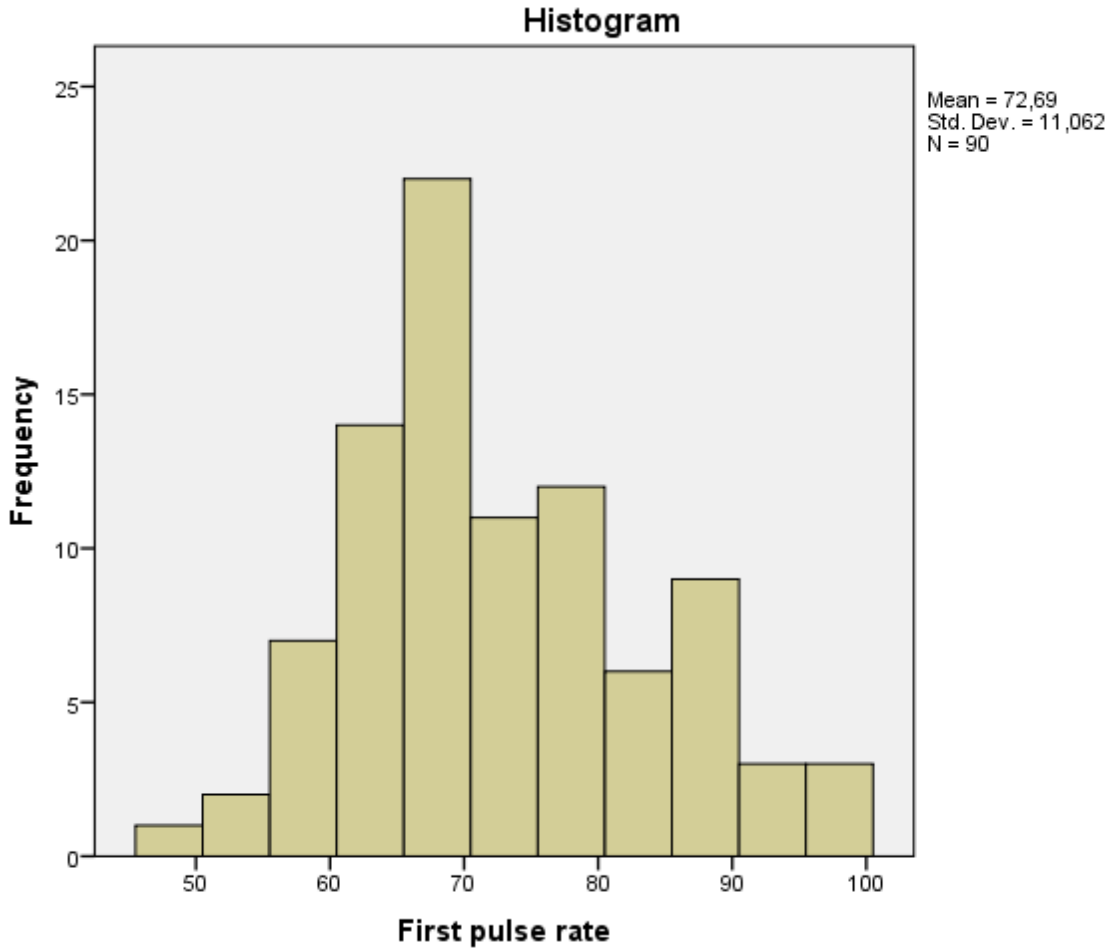
I tillegg til Std. Deviation får vi også skrevet ut Std. Error, som delt på kvadratroten av antall observasjoner, dvs. $11.06/\sqrt{92}=1.17$. Dette er spredningen til gjennomsnittet. Det er denne vi bruker til å beregne konfidensintervallet. Vi beregner nemlig konfidensintervallet som

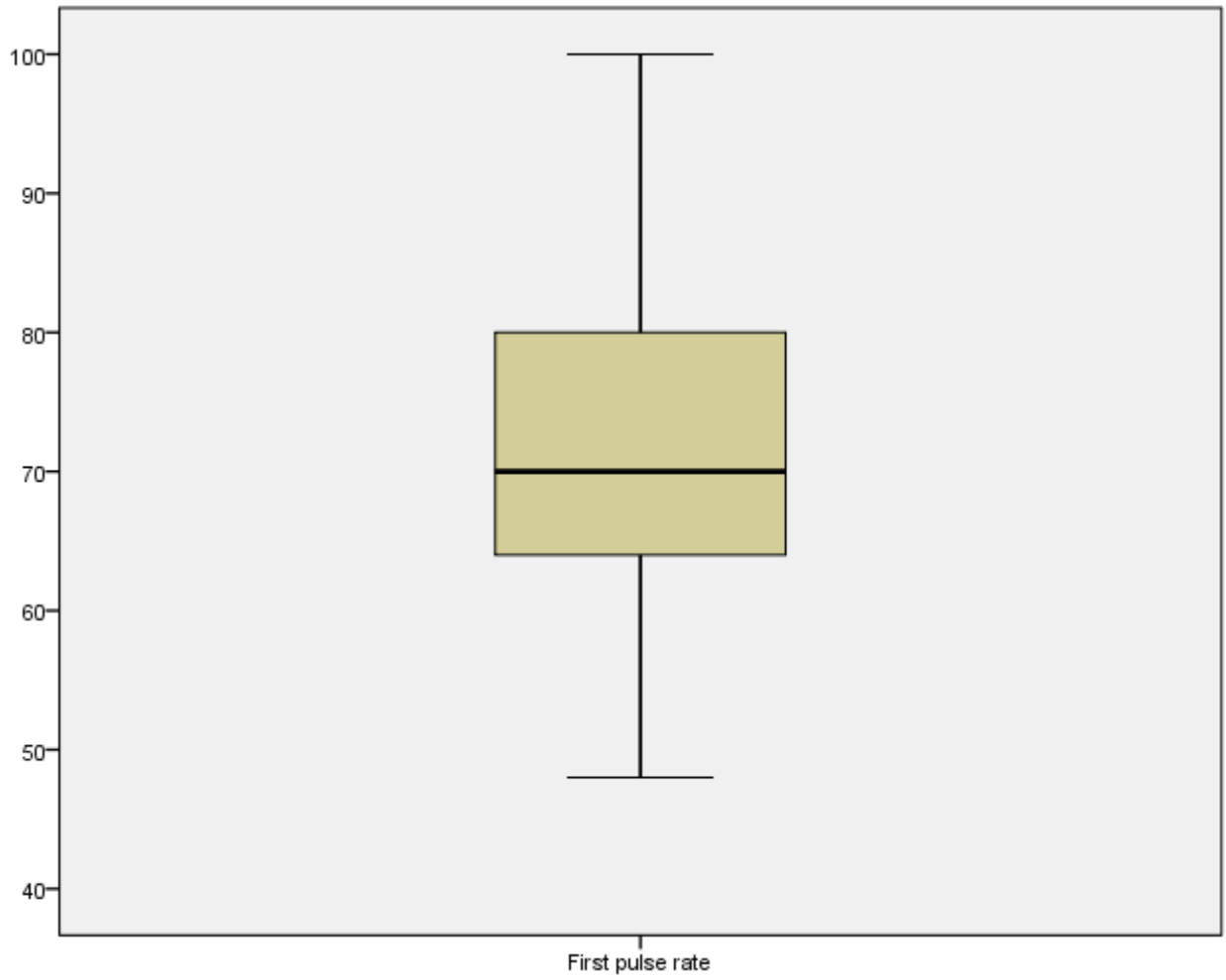
(Gjennomsnitt - 1.96 x Standardfeilen, Gjennomsnittet + 1.96 x Standardfeilen).

Når vi bruker denne formelen finner vi at konfidensintervallet er

(72.69 - 1.96 x 1.71,72.69 + 1.96 x 1.17) = (70.37, 75.01)

I tillegg til denne oversikten får vi også to grafer: et histogram og et boksplott.


Histogram
Mean = 72,69
Std. Dev. = 11,062
N = 90
Frequency
25
20
15
10
5
0
50
60
70
80
90
100
First pulse rate



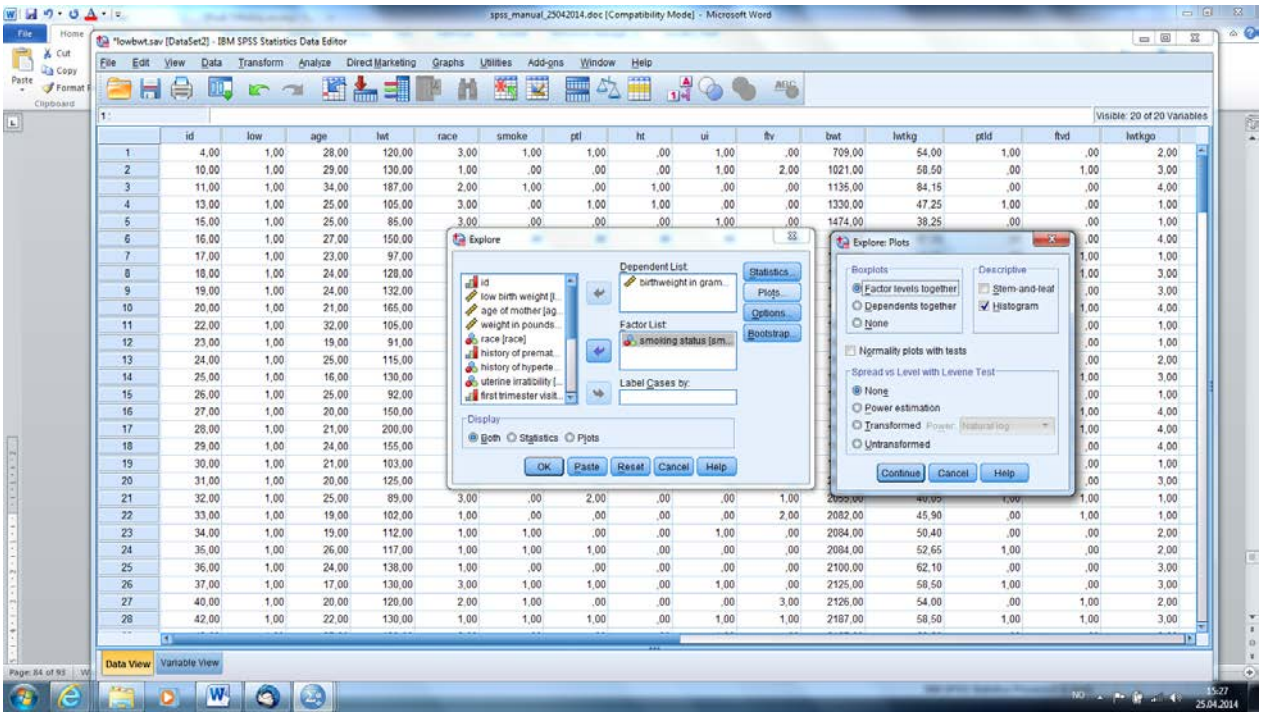
Histogrammet gir en enkel presentasjon av dataene i grupper, siden det gir en oversikt over antallet som faller de definerte gruppene. Histogrammet gir en fin oversikt over om fordelingen er symmetrisk om gjennomsnittet. Men det ligger mer informasjon i boksplottet. Det består av en boks, med en nedre kant som er nedre kvartil. I midten ligger det en strek som er medianen og den øvre kanten er øvre kvartil. Dersom medianen ligger midt i boksen, er det samme avstand mellom nedre kvartil og medianen, som mellom medianen og øvre kvartil. Da er fordelingen symmetrisk rundt medianen. Hvis også medianen er ganske lik gjennomsnittet, har vi en symmetri rundt gjennomsnittet. I vårt tilfelle er det en litt kortere avstand i nedre del av boksen, men det er ikke spesielt stort avvik fra symmetri.

Linjene som går ut boksen, oppover og nedover, går opp til høyeste verdi og ned til laveste verdi. Men hver linje strekker seg bare opp til 1.5 ganger boksens lengde. Observasjoner som er høyere eller lavere enn dette, kalles ekstremverdier (outliers) og plottes som enkeltstående observasjoner som sirkler.

9.5 Explore. Eksempel: lowbwt.sav

Vi går tilbake til filen **lowbwt.sav** for å vise en annen nyttig fremstilling av data som vi får frem via *Explore*. Vi er interessert i å få fremstilt forskjellen i fødselsvekt for mødre som røyker i forhold til mødre som ikke røyker. Vi går da inn i *Analyze/Descriptive*

Statistics/Explore. Nå trekker vi BWT over i *Dependent Variable(s)* vinduet og SMO over i *Factor List*. Med denne kommandoen får vi presentert variablene BWT etter kategoriene i SMO. Vi går også til *Plots* og klikker bort *Stem and Leaf* og klikker på *Histogram*. Da ser dialogboksene slik ut:



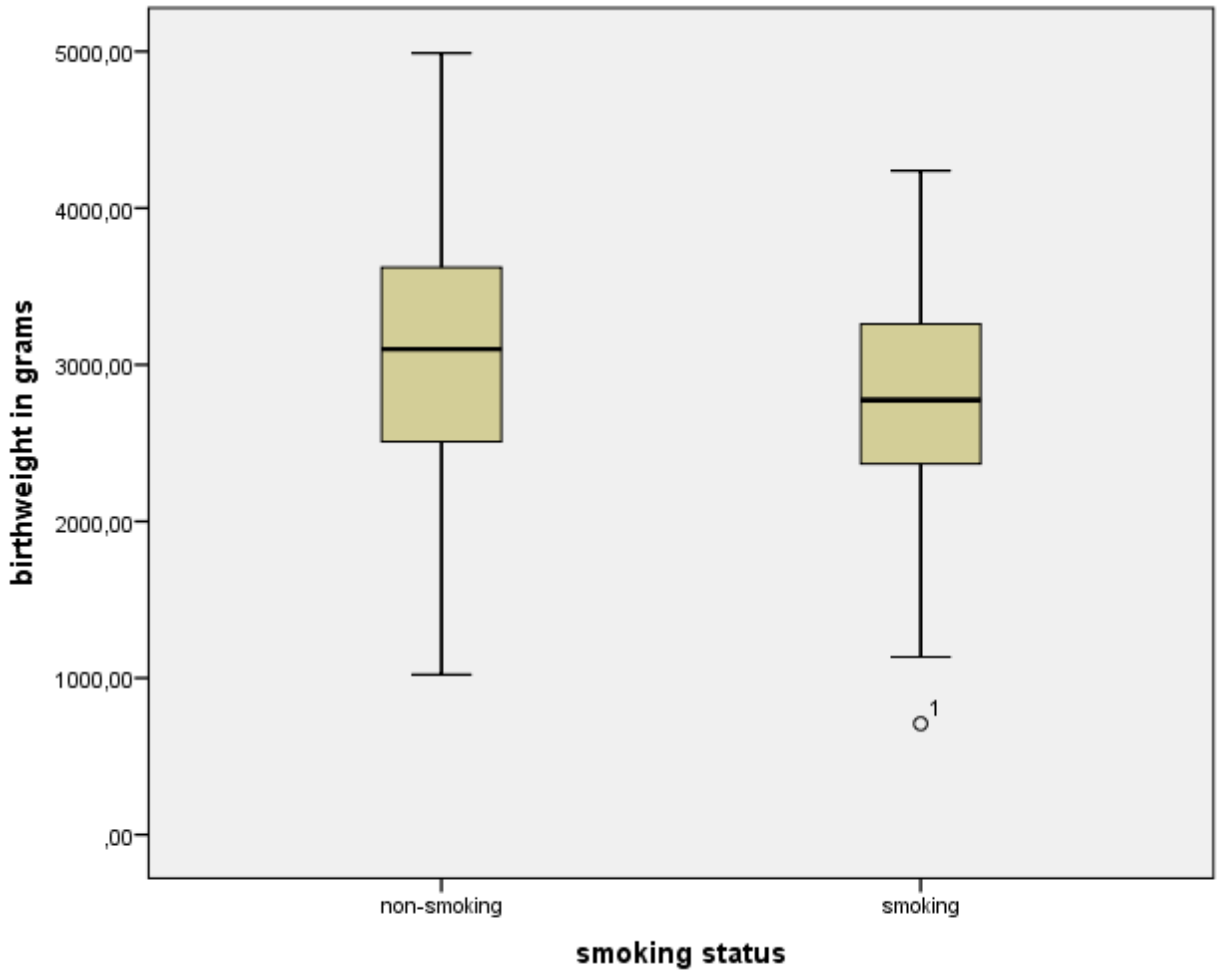
Når vi klikker på *Continue* og *OK*, får vi følgende resultater. Vi kommenterer ikke på histogrammene, siden de er så enkle.

Descriptives

	smoking status			Statistic	Std. Error
birthweight in grams	non-smoking	Mean		3054,9565	70,16250
		95% Confidence Interval for Mean	Lower Bound	2915,9651	
			Upper Bound	3193,9479	
		5% Trimmed Mean		3071,3333	
		Median		3100,0000	
		Variance		566119,323	
		Std. Deviation		752,40901	
		Minimum		1021,00	
		Maximum		4990,00	
		Range		3969,00	
		Interquartile Range		1134,00	
		Skewness		-,291	,226
		Kurtosis		-,227	,447
	smoking	Mean		2773,2432	76,73218
		95% Confidence Interval for Mean	Lower Bound	2620,3162	
			Upper Bound	2926,1703	
		5% Trimmed Mean		2786,0691	
		Median		2775,5000	
		Variance		435699,228	
		Std. Deviation		660,07517	
		Minimum		709,00	
		Maximum		4238,00	
		Range		3529,00	
		Interquartile Range		907,25	
		Skewness		-,296	,279
		Kurtosis		,420	,552

Vi ser at gjennomsnittlig fødselsvekt for røykende mødre er 2773 gram, mens den for ikke-røykende mødre er 3054 gram. Standardavvikene er henholdsvis 752 og 660 gram for ikke-røykende og røykende mødre. Standardfeilene er 70 og 77 gram.

Den samme bildet får vi i boksplottene nedenfor:



Vi ser at boksen for ikke-røykende mødre ligger over den for røykende mødre. Det gjenspeiler at fødselsvektene for ikke-røykende mødre ligger høyre enn for røykende mødre. For de røykende mødrene ser vi at det er én ekstremverdi (outlier), plottet som en sirkel, og med et nummer knyttet til seg. Dette er observasjonsnummeret. Hvis vi nå går tilbake til datavinduet, ser vi at den første observasjonen er en fødselsvekt på 709 gram (BWT = 709), og at den tilhører en røykende mor (SMOKE = 1).

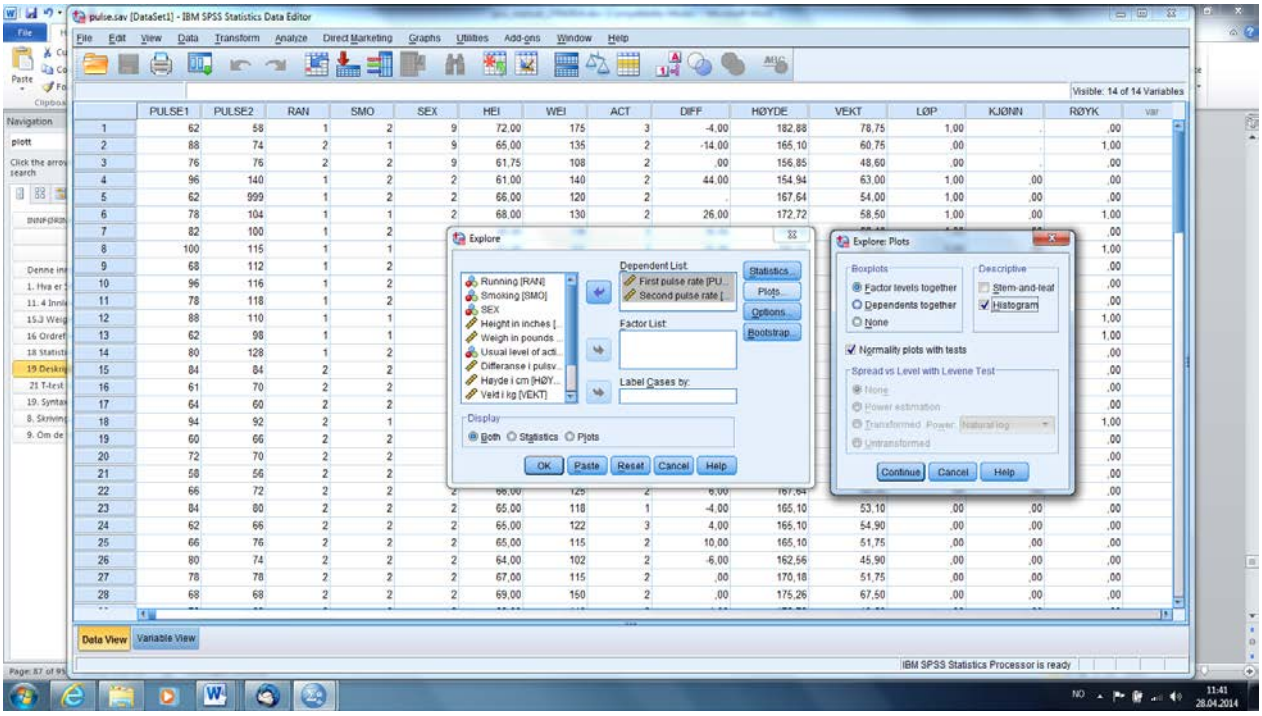
9.6 Sjekking av normalitet. Eksempel: pulse.sav

Antagelsen om normalitet er helt sentral i statistisk analyse. Mange av de metoden vi skal bruke, for eksempel t-tester og regresjon, bygger på antagelsen om at dataene er normalfordelte. Hvordan skal vi sjekke antagelsen om normalitet?

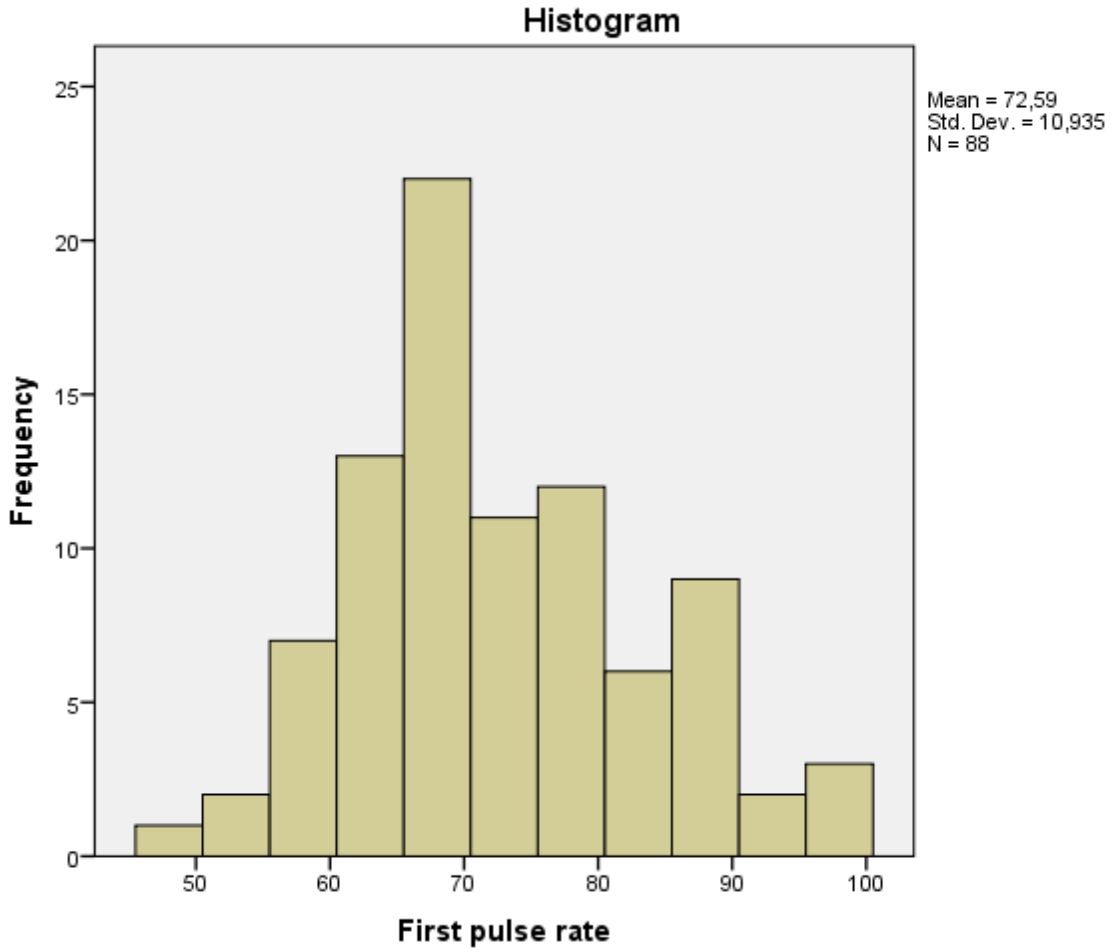
Det finnes noen tester for normalitet, men vi skal ikke bruke dem, men heller sjekke ved et såkalt normalfordelingsplott. Dette finner vi under *Analyze/Descriptive Statistics/Explore*.

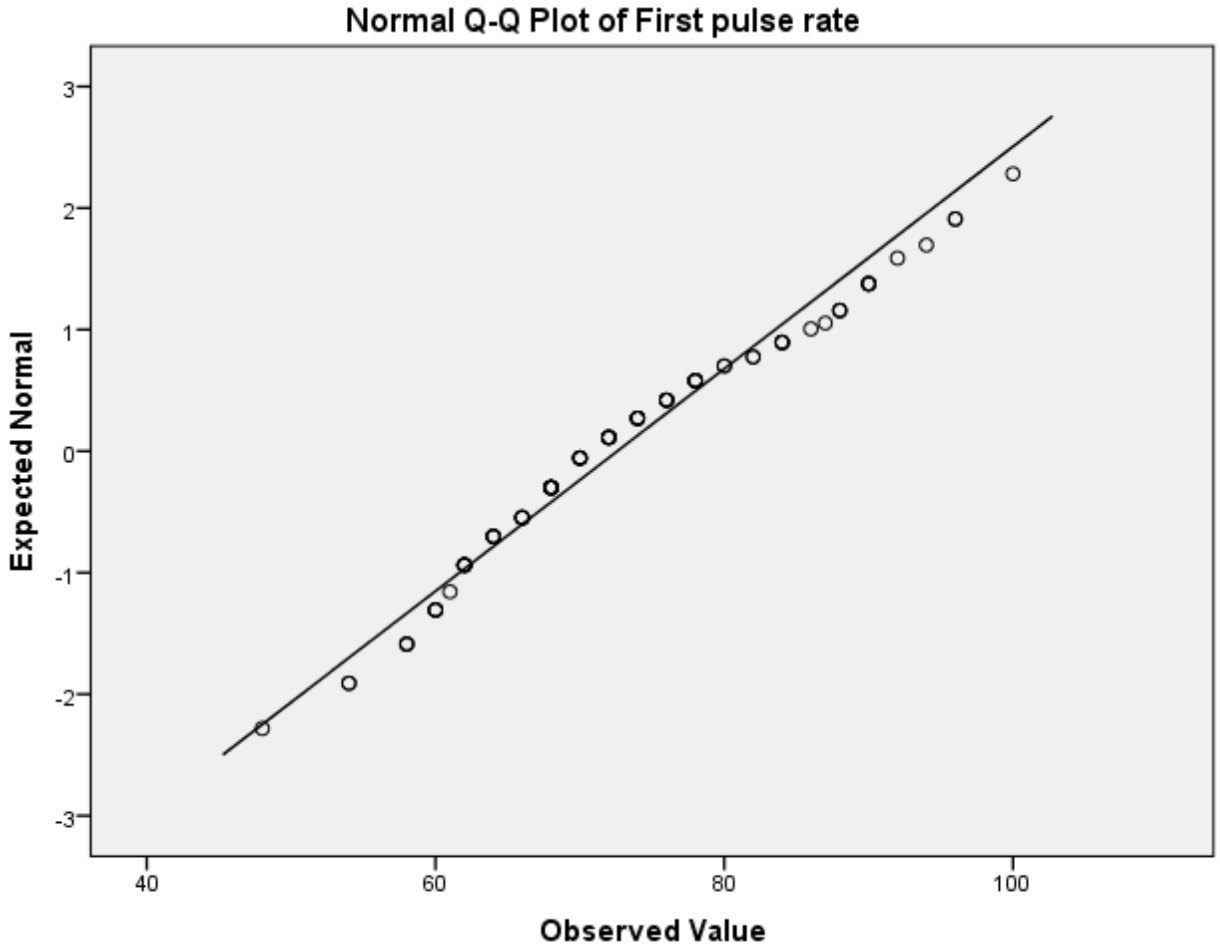
Vi henter frem datafilen **pulse.sav**. Vi skal her se på om variablene PULSE1 og PULSE2 kan betraktes som normalfordelte. Vi går da inn *Analyze/Descriptive Statistics/Explore* og trekker PULSE1 og PULSE2 over i vinduet *Dependent List*. Vi klikker så på *Plots* i knapperekken til

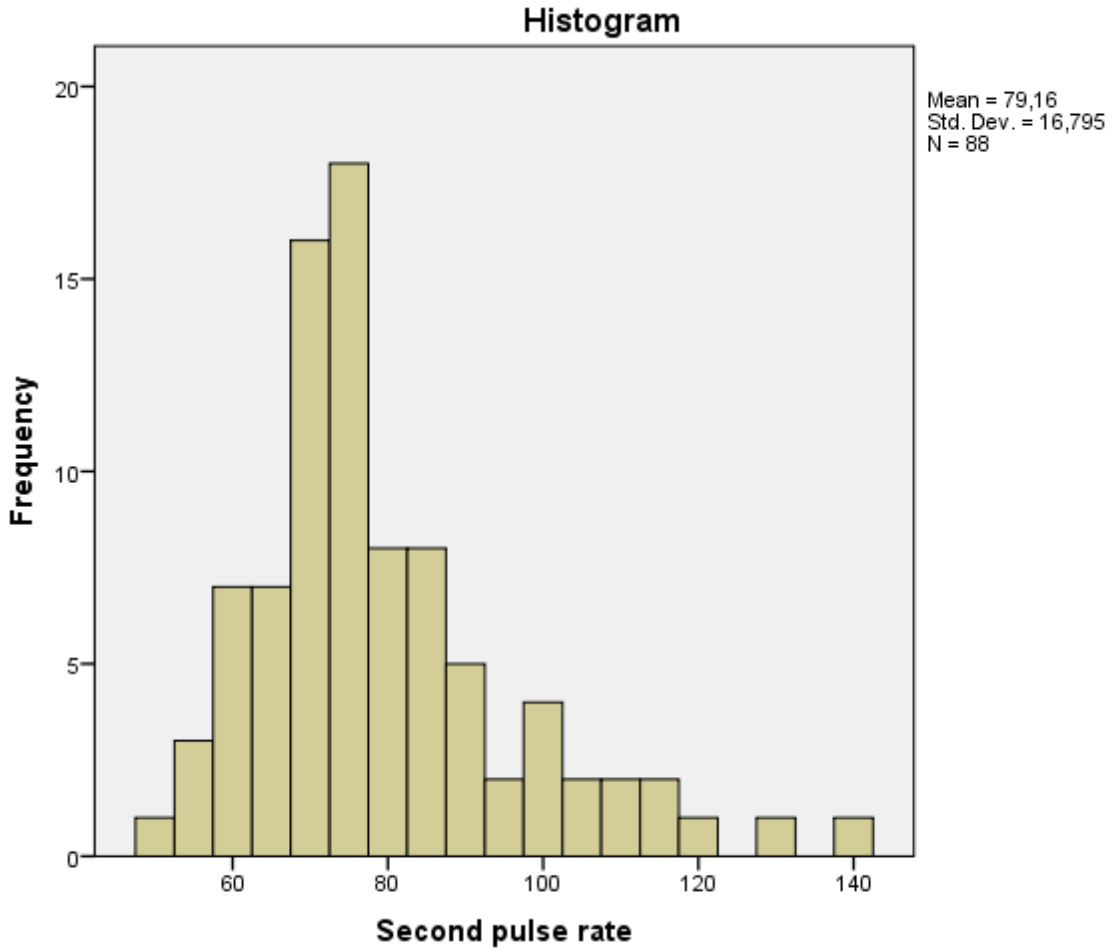
høyre. Midt i dialogboksen står det Normality plots with test. Der klikker vi av. Vi klikker også på at vi skal ha *Histogram* og ikke *Stem-leaf-plot*. Da ser dialogboksen slik ut:

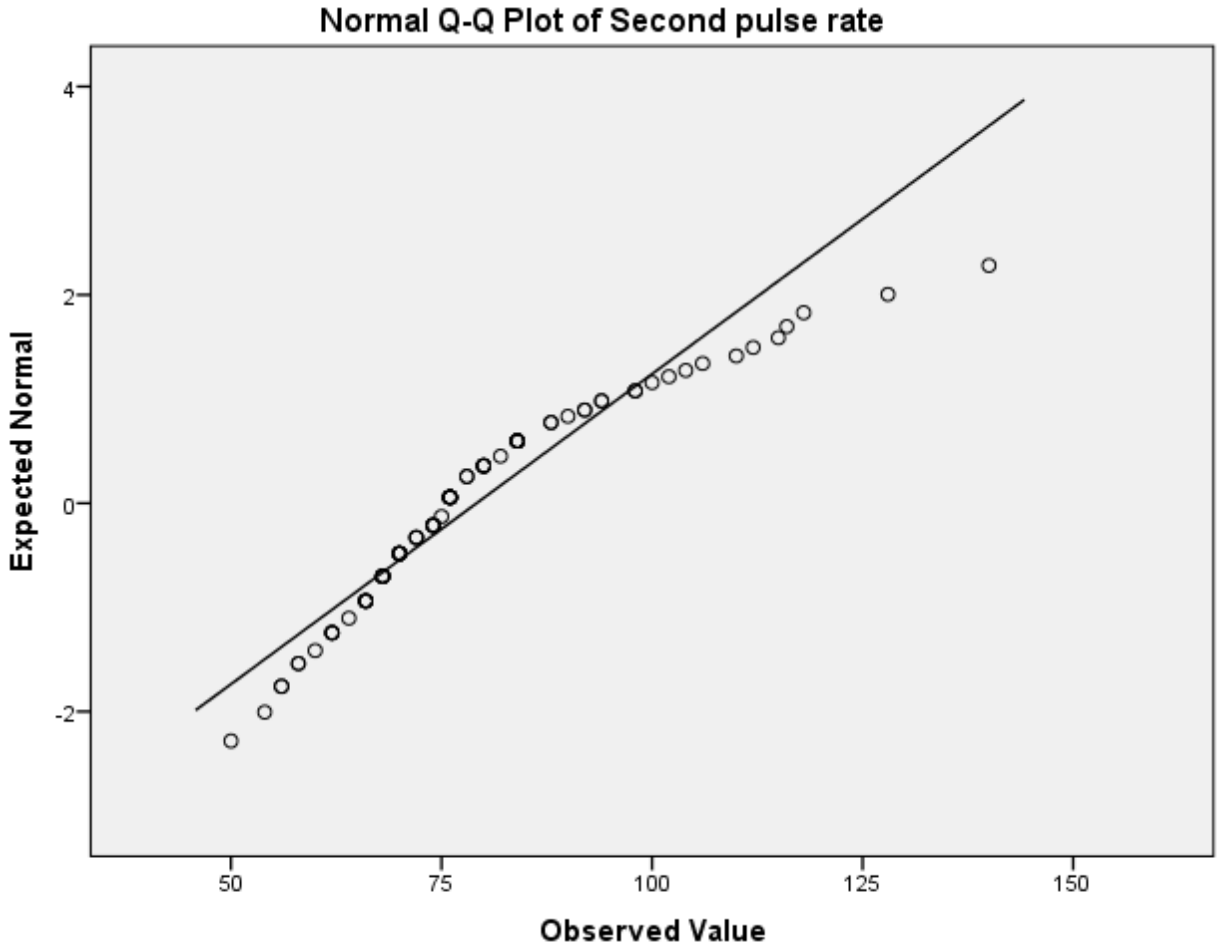


Vi klikker *Continue* og *OK*. Da får vi mye utskrift som vi kjenner igjen fra før. Her kopierer vi bare inn histogrammet og normalfordelingsplottet:


Histogram
Mean = 72,59
Std. Dev. = 10,935
N = 88
Frequency
25
20
15
10
5
0
50
60
70
80
90
100
First pulse rate


Normal Q-Q Plot of First pulse rate
Expected Normal
3
2
1
0
-1
-2
-3
40
60
80
100
Observed Value


Histogram
Mean = 79,16
Std. Dev. = 16,795
N = 88
Frequency
20
15
10
5
0
60
80
100
120
140
Second pulse rate



Vi ser at under normalfordelingsplottet ligger det et «detrended» normalfordelingsplottet. At det er «detrended» betyr at vi der ser på avstanden mellom observasjonene og linjen. Det er vanskeligere å tolke, så vi skal heller tolke normalfordelingsplottet.

I et normalfordelingsplott plotter vi alle observasjonenes verdi mot den tilsvarende verdien som observasjonen skulle ha vært, dersom vi hadde normalfordelte data. Dersom dataene er prefekt normalfordelt, skal de ligger på den rette linjen som også er tegnet i plottet. Men vi må forvente avvik, særlig når vi har lite data.

Først en liten kommentar til histogrammene. Vi kan bruke histogrammene til en vurdering av om data er normalfordelt, siden dette er en symmetrisk fordeling. Men det gir ikke en tilstrekkelig bakgrunn for å vurdere avvik fra normalfordelingen, særlig når det gjelder halene i fordelingen. Vi må derfor basere vurderingen på normalfordelingsplottet.

I normalfordelingsplottet for PULSE1 ser vi bare helt minimale avvik fra den rette linjen, og vi konkluderer med at PULSE1 er normalfordelt.

I plottet for PULSE2 ser vi noe større avvik. Vi ser for øvre del av fordelingen, dvs. i høyre hale, har vi to observasjoner som peker seg ut. For normalfordelingen skulle disse ha ligget på linjen, men det har altså for høye verdier, i forhold til normalfordelingen. Tilsvarende er det to verdier som som er for lave i forhold til normalfordelingen, i venstre hale. Men vi ser på disse

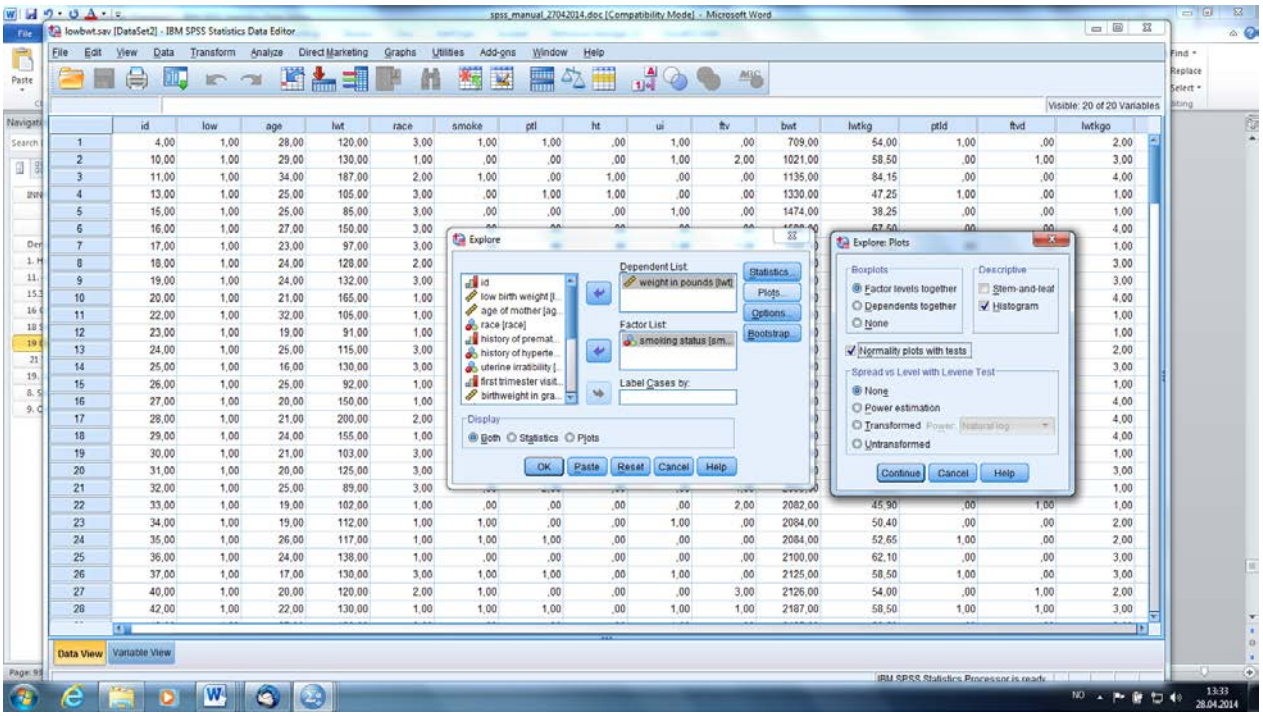
avvikene som mindre, og konkluderer med at vi kan gå videre med antagelsen om at PULS2E1 og PULSE2 er normalfordelte.

Generelt kan vi si at når vi har mange observasjoner som er ikke-normale i halene, må vi forlate antagelsen om normalfordelte data. Da kan vi ikke bruke statistiske metoder som bygger på denne antagelsen, og må bruke såkalte ikke-parametriske metoder, se kapittel 11.3.

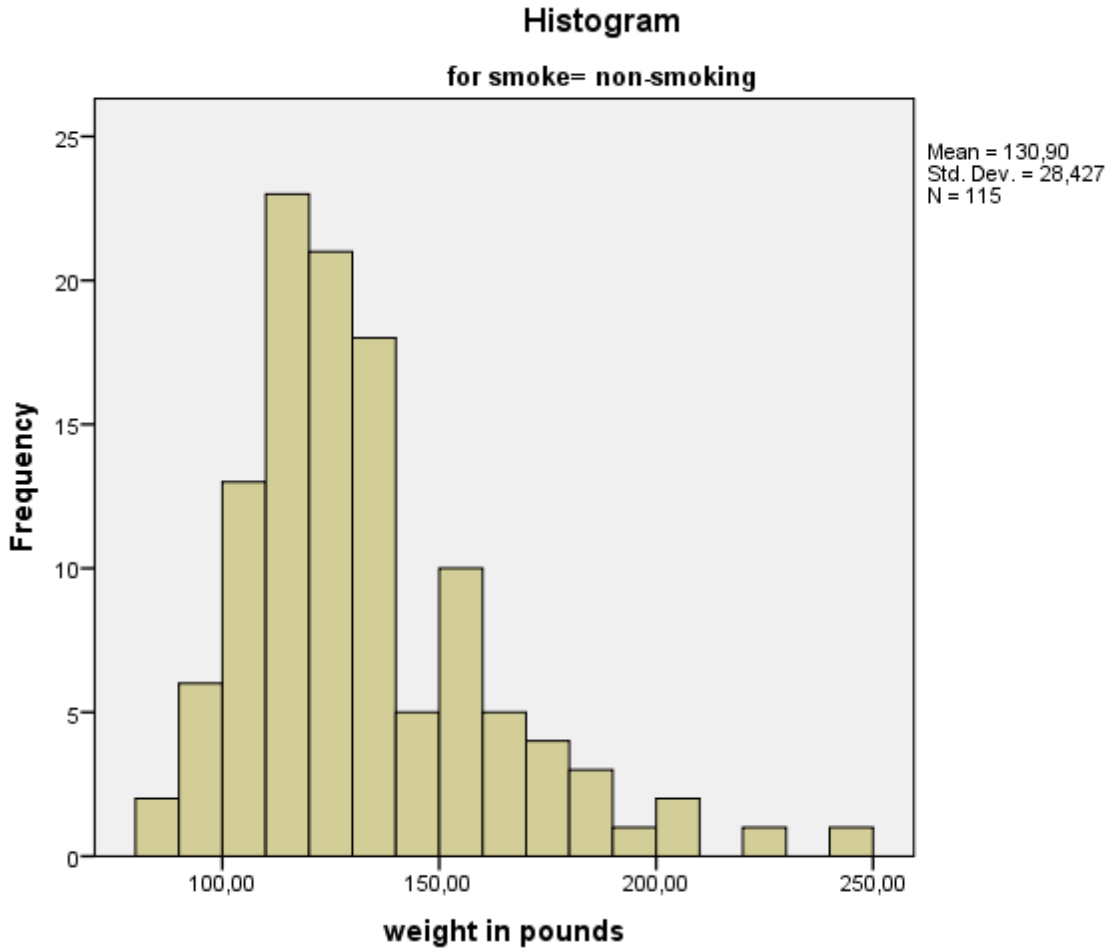
9.7 Sjekking av normalitet. Eksempel: lowbwt.sav

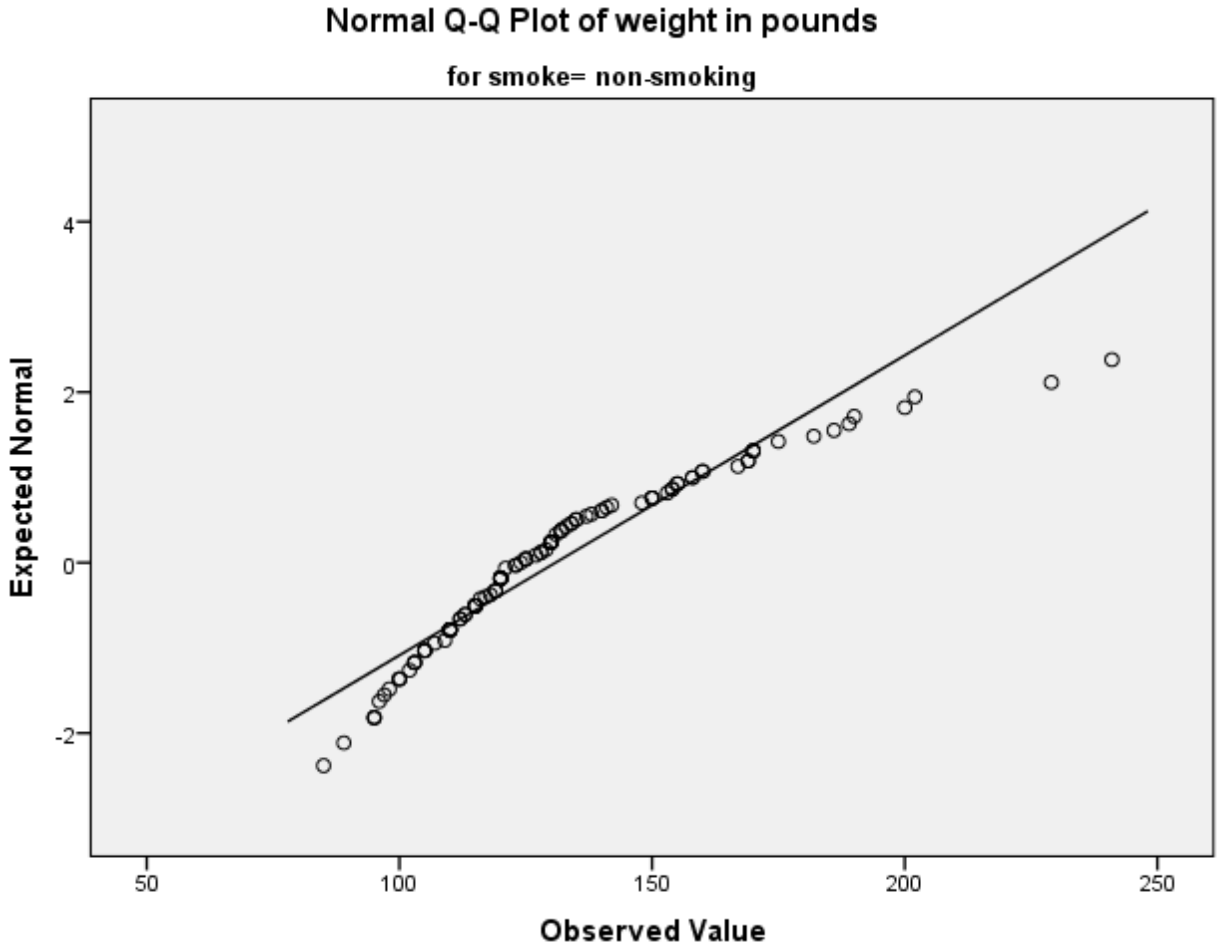
La oss nå hente frem dataene **lowbwt.sav**. Et formålene med denne studien er å undersøke om fødselvekten er ulik for røykende og ikke-røykende mødre. Da må vi undersøke om vi kan bruke statistiske metoder basert på normalfordelingen. Antagelsen er da at fødselsvekt (BWT) er normalfordelt for SMOKE = 1 og for SMOKE = 0.

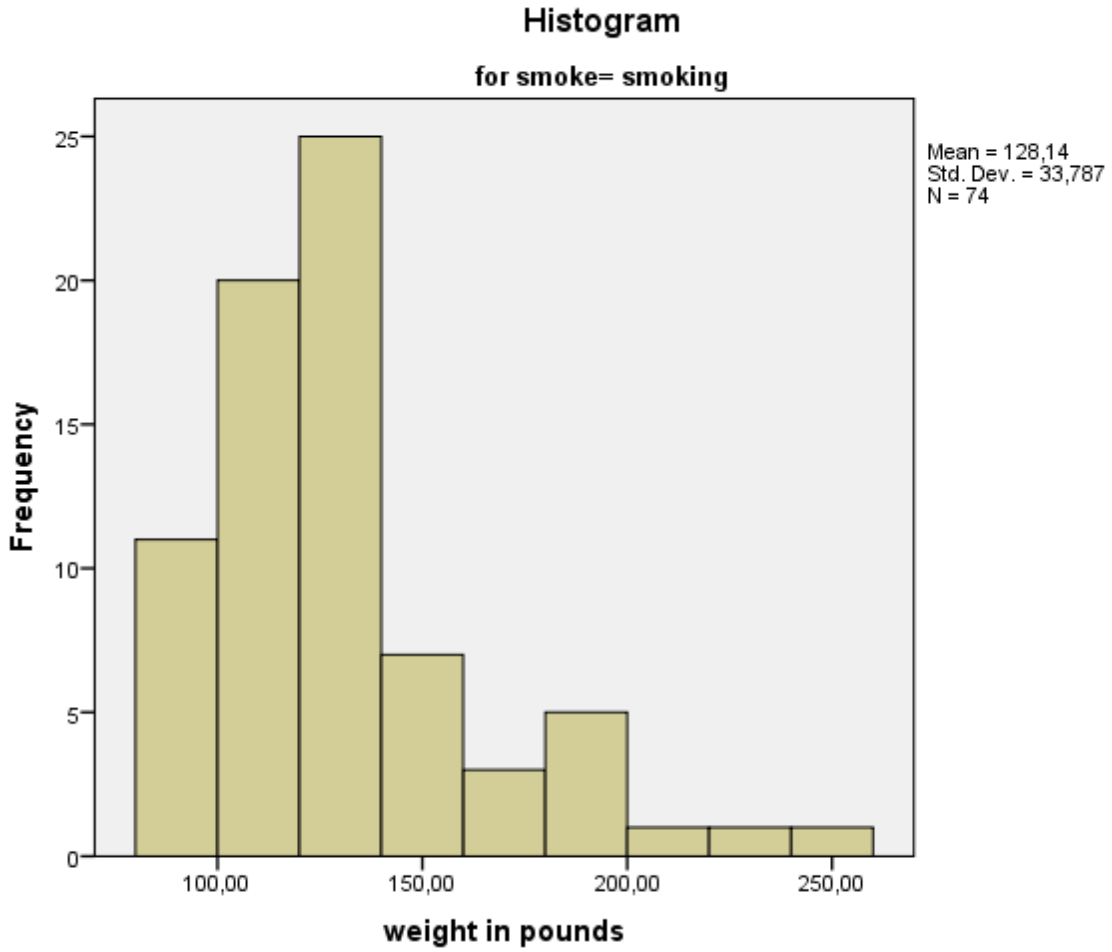
Vi går igjen inn i *Analyze/Descriptive Statistics/Explore* og trekker da nå over BWT i vinduet *Dependent List*. Deretter trekker vi SMOKE over i Factor List. Vi klikker så på *Plots* i knapperekken til høyre og klikker av på *Normality plots with test.* Til slutt klikker vi på at vi skal ha *Histogram* og ikke *Stem-leaf-plot*. Da ser dialogboksen slik ut:

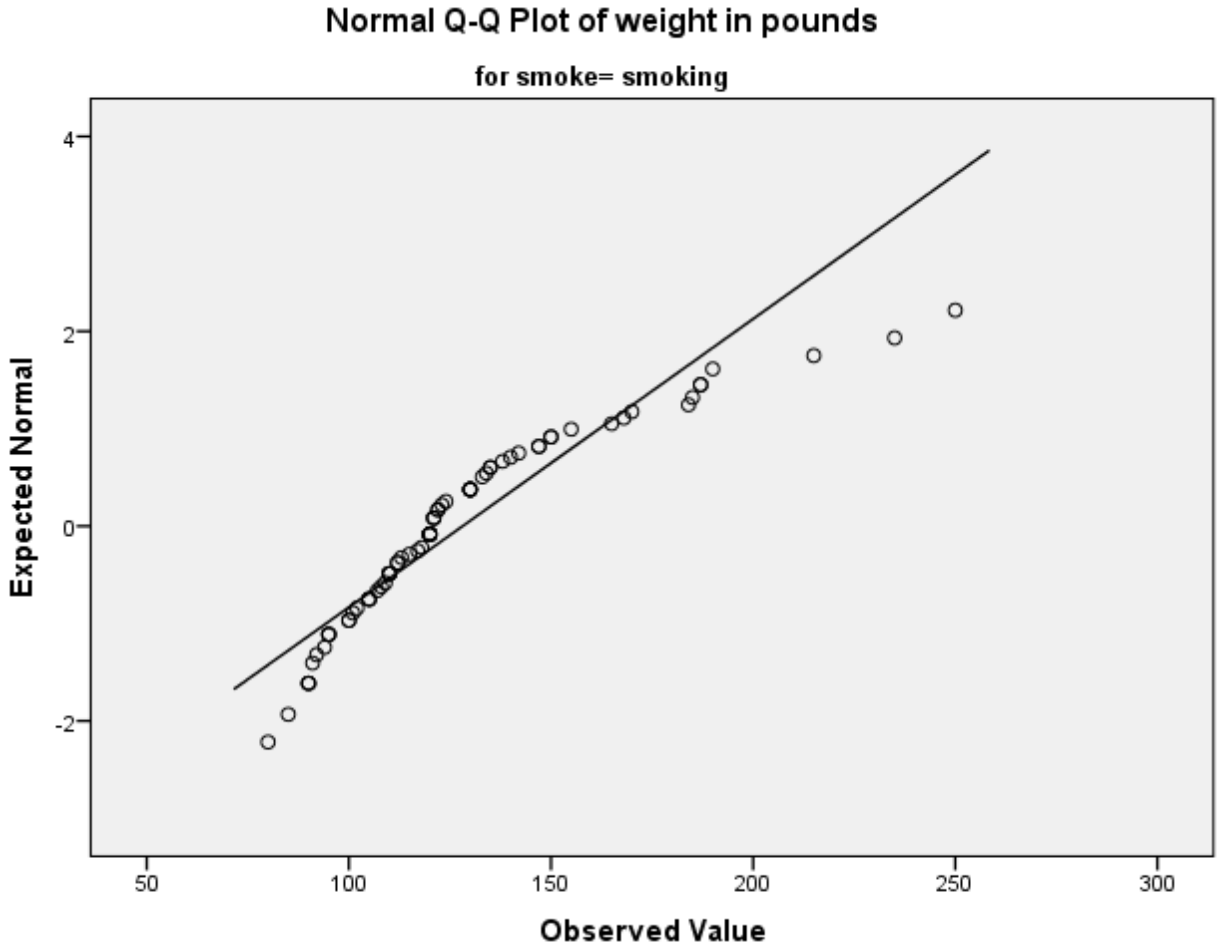


Vi klikker av på *Continue* og *OK* og får følgende i utskriftsvinduet:


Histogram
for smoke= non-smoking
Mean = 130,90
Std. Dev. = 28,427
N = 115
Frequency
25
20
15
10
5
0
100,00
150,00
200,00
250,00
weight in pounds


Normal Q-Q Plot of weight in pounds
for smoke= non-smoking
Expected Normal
4
2
0
-2
50
100
150
200
250
Observed Value


Histogram
for smoke= smoking
Mean = 128,14
Std. Dev. = 33,787
N = 74
Frequency
25
20
15
10
5
0
100,00
150,00
200,00
250,00
weight in pounds



Basert på histogrammene alene ser vi at vi kan mistro antagelsen om normalfordelte data i de to røyke-gruppene. Men normalfordelingsplottet viser igjen at vi bare har et par observasjoner som er avvikene. For både røykene og ikke-røykende mødre har vi to-tre observasjoner som er avvikene i høyre og venstre hale. I høyre hale er det observasjoner som ligger under linje, og derfor er for store i forhold til normalfordelingen. Men det samme mønsteret har vi også i venstre hale, der vi to-tre observasjoner som ligger under linje, og dermed også er for store.

Vi konkludere imidlertid med at det er bare små avvik, og vi går videre med antagelsen om at dataene i de to gruppene er normalfordelte.

10 Diagrammer og plott

Læringsmål

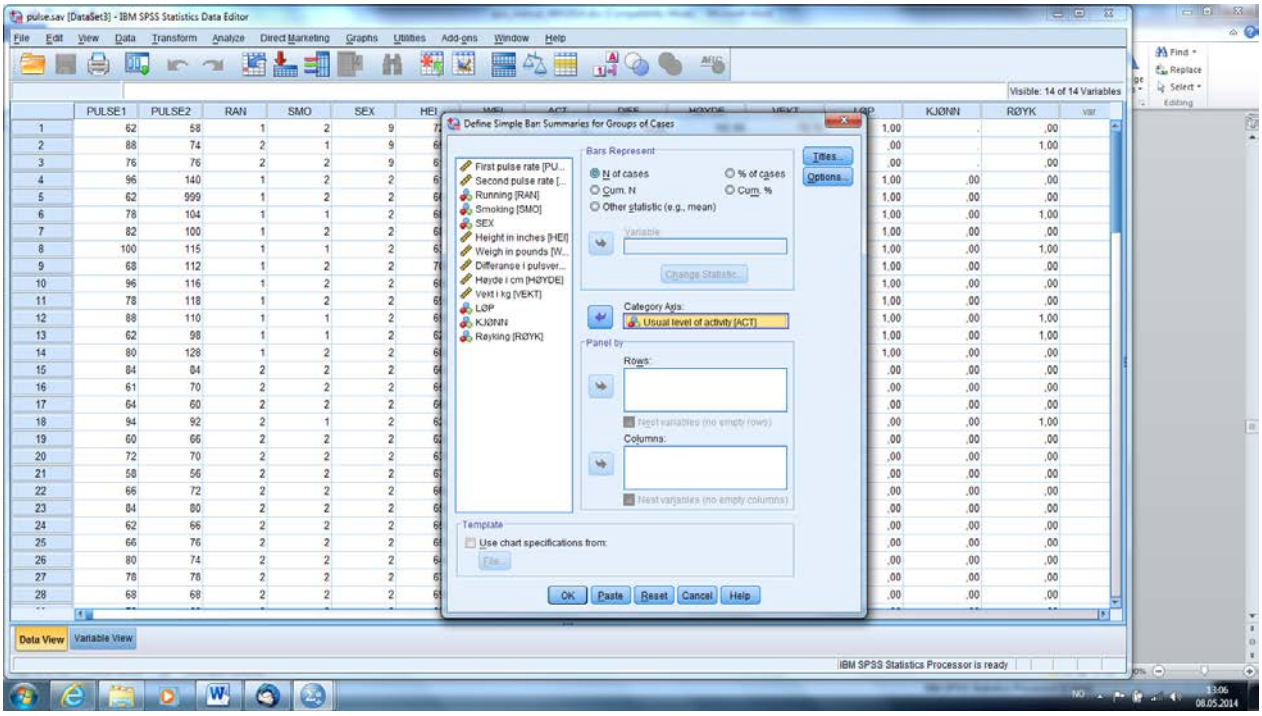
Man sier ofte at diagrammer og plott kan fortelle mer enn hundre analyser! Det er også delvis sant, men diagrammer og plott kan aldri erstatte en grundig statistisk analyse, som inneholder beregninger av effektestimater, konfidensintervall og p-verdier.

SPSS er ikke blant de beste programmene for å produsere diagrammer og plott. Men SPSS er tilstrekkelig for vårt formål, som er mer rettet mot statistisk analyse enn mot grafisk presentasjon.

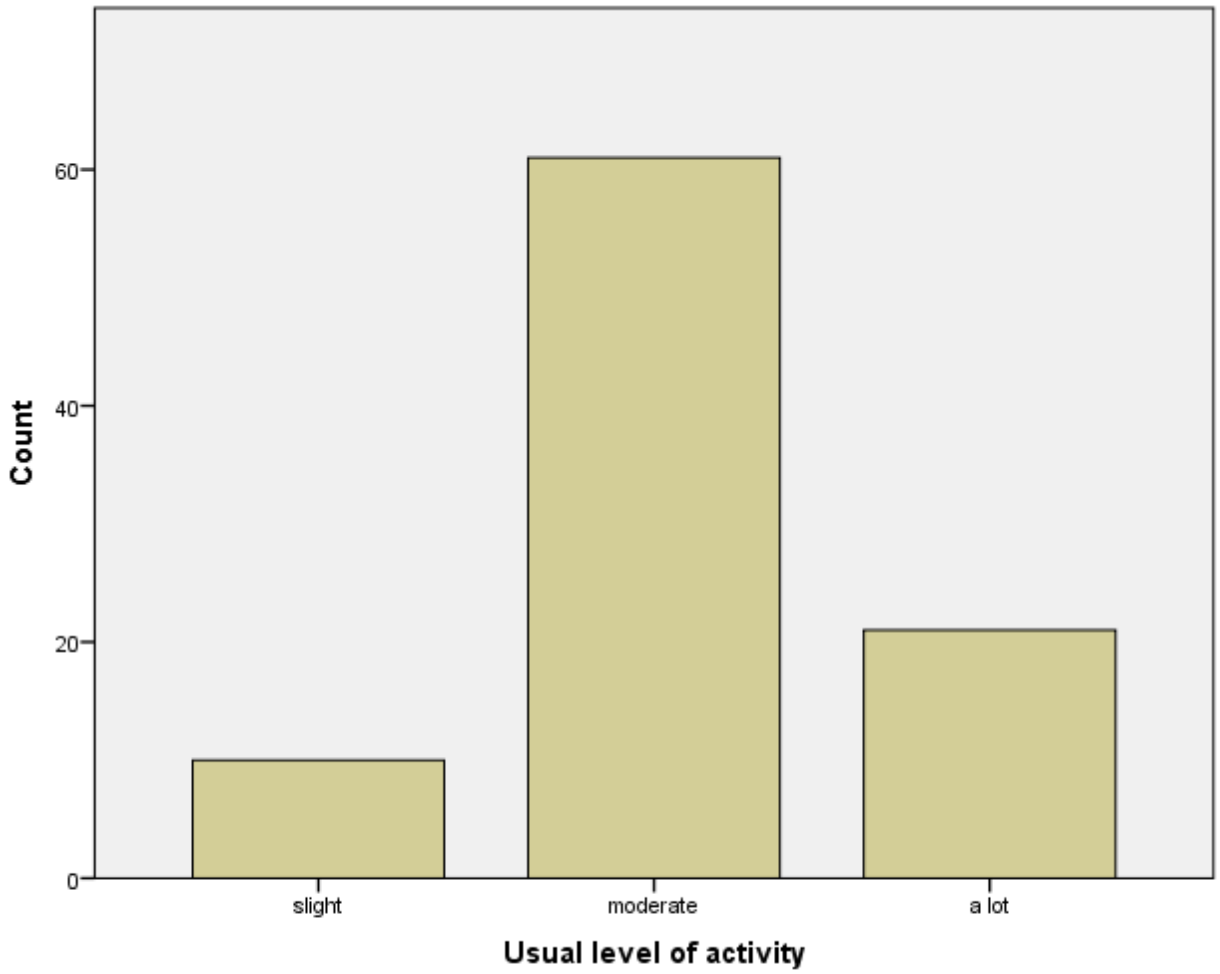
I dette kapittelet skal vi se på presentasjonen av stolpediagrammer, histogrammer, boksplott og spredningsplott.

10.1 Stolpediagrammer. Eksempel: pulse.sav

Et stolpediagram gir oss en frekvensfordeling for en variabel, fordelt i forhold til en annen variabel. La oss gå tilbake til datafilen **pulse.sav**. Vi ønsker første en grafisk fremstilling av variabelen ACT, og så en fremstilling av ACT fordelt på KJØNN. Da går vi inn i *Graphs/Legacy Dialogs/Bar*. Vi klikker på *Simple/Define*. I dialogboksen trekker vi Her trekker vi først over ACT i *Category Axis*. Da ser dilaogboksen slik ut:

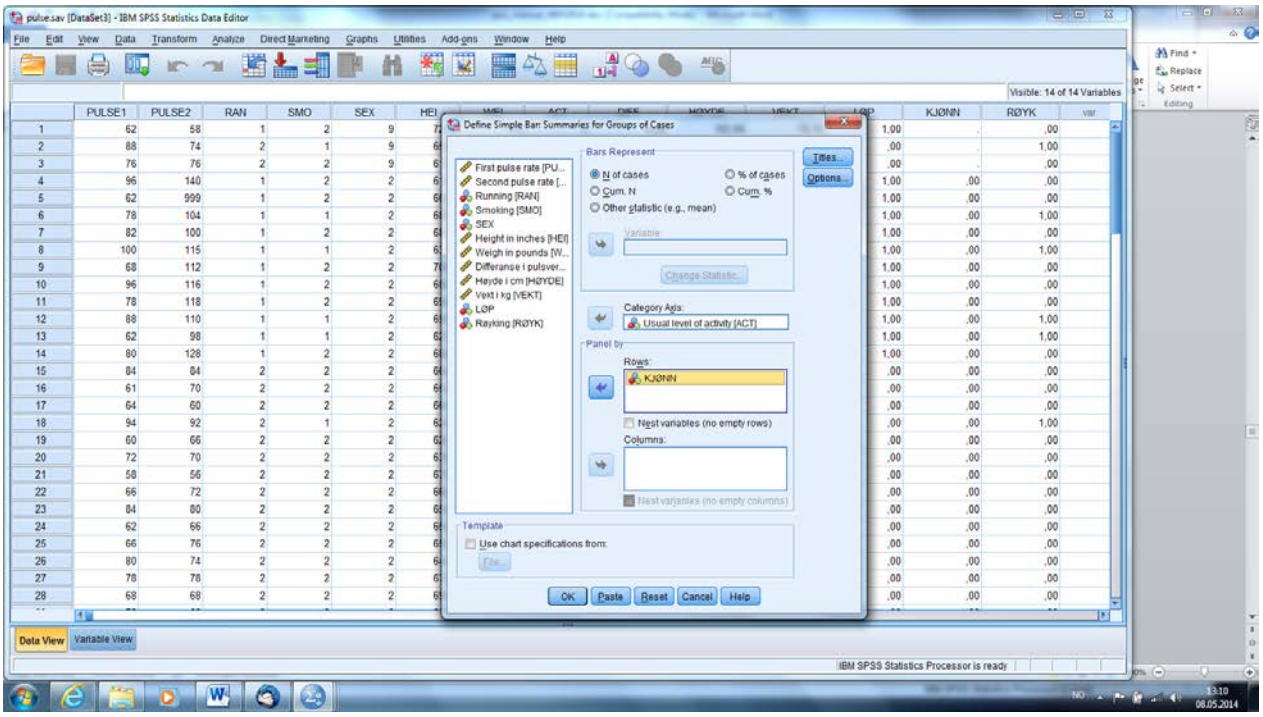


Vi klikker på *OK* og får følgende resultat:

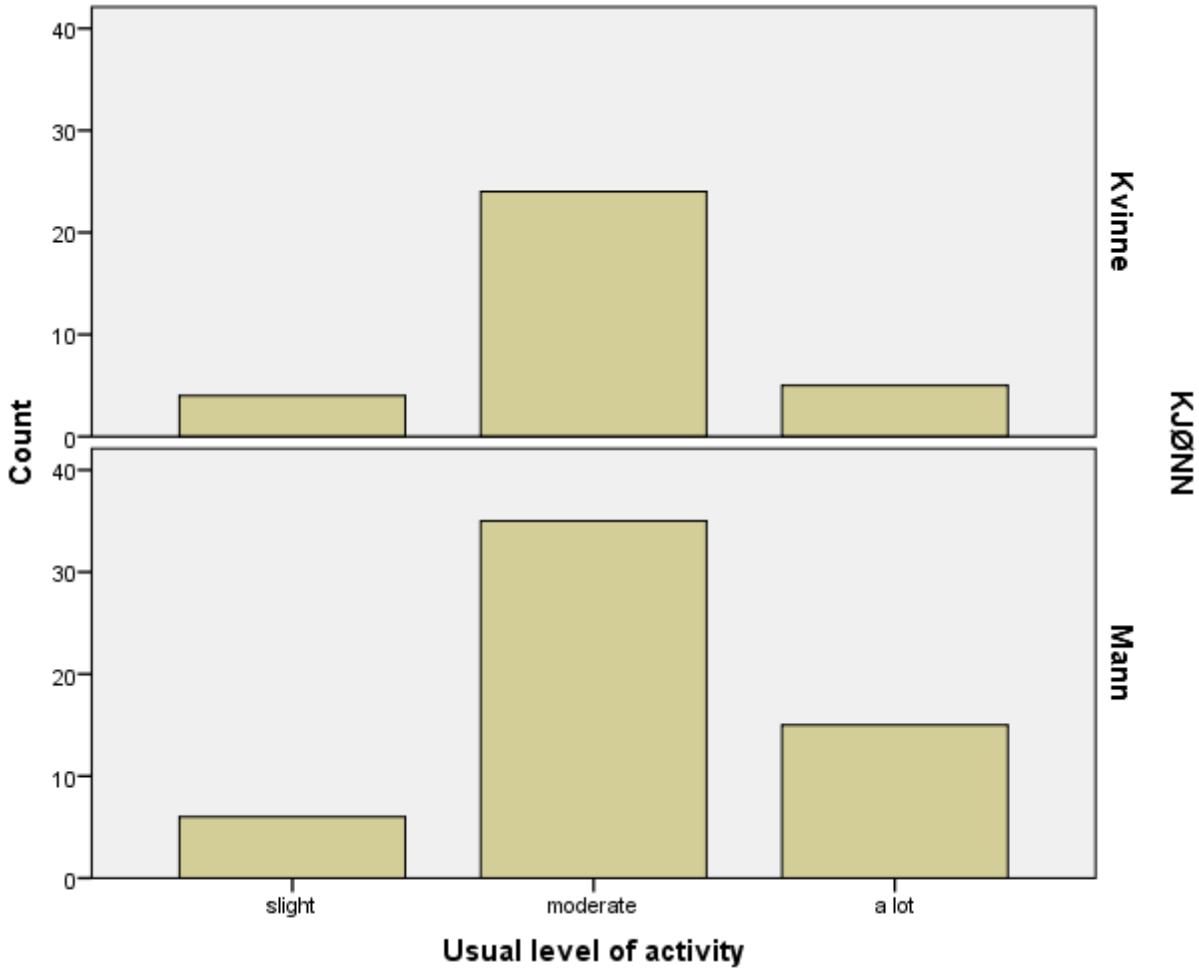


Vi ser en fin presentasjon av antall med de ulike nivåene av fysisk aktivitet.

Så går vi tilbake til *Graphs/Legacy Dialogs/Bar* med *Simple/Define*. Vi beholder ACT i *Category Axis*, men vi trekker over KJØNN i *Rows*. Da ser dilaogboksen slik ut:



Etter *OK* får vi følgende:

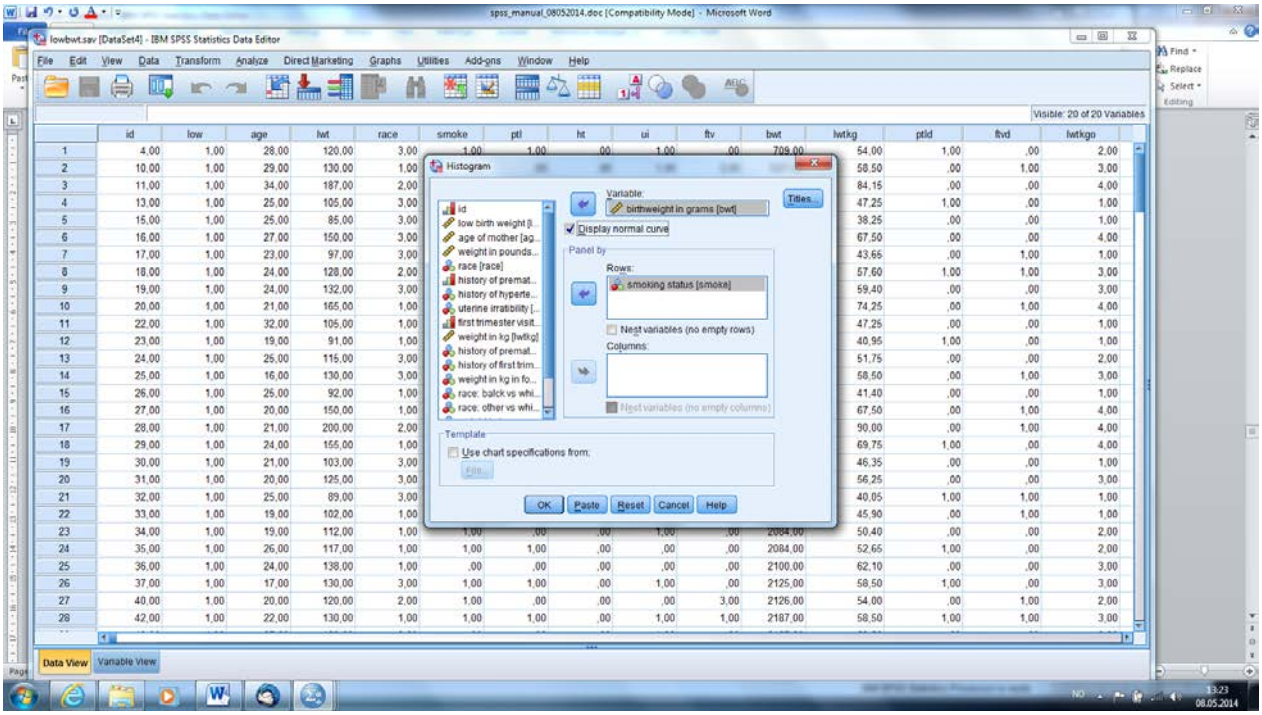


Vi ser en presentasjon som klart viser at menn tilsynelatende har en høyere grad av fysisk aktivitet enn menn. Men her må vi ikke la oss lure av at det er flere menn enn kvinner i denne studien.

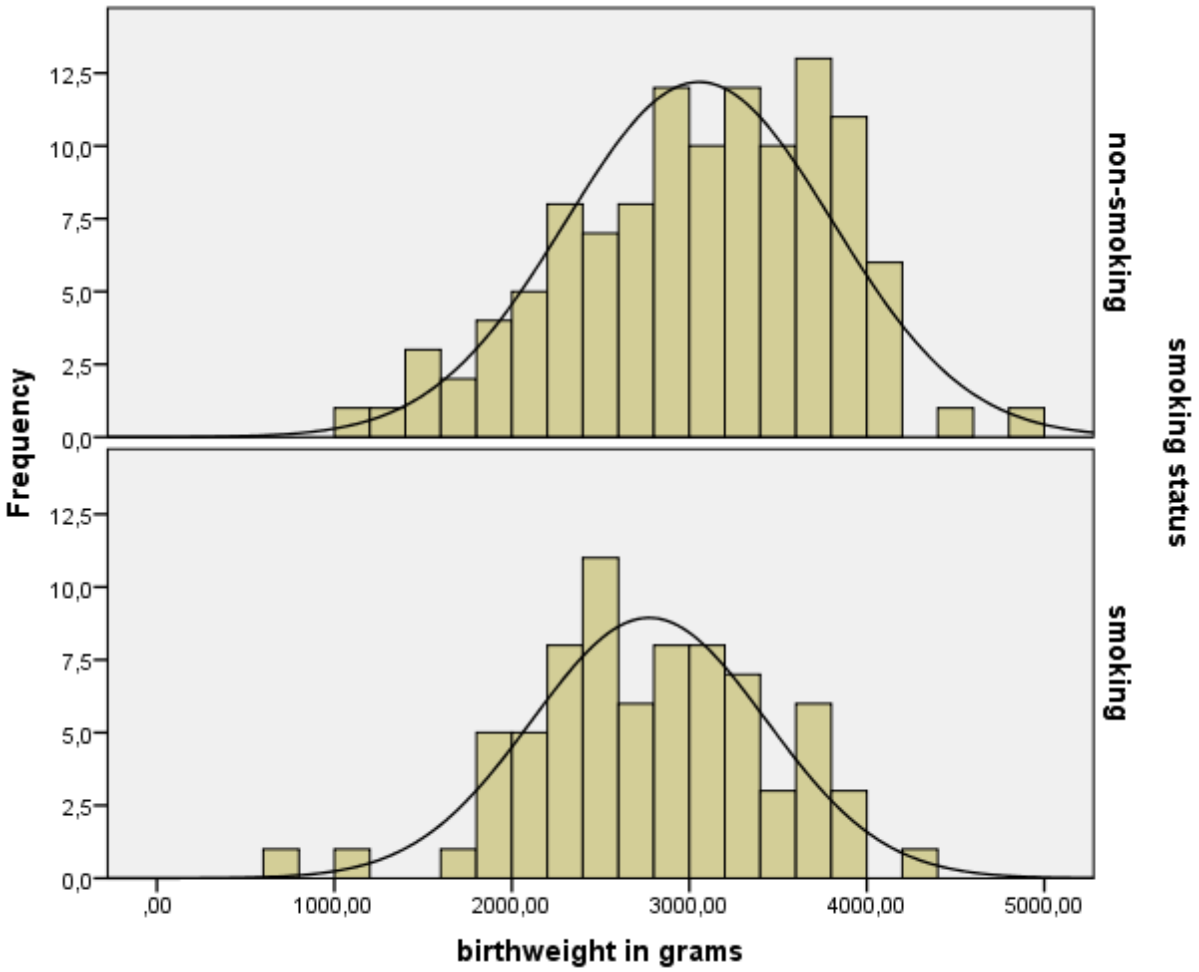
10.2 Histogrammer. Eksempel: lowbwt.sav

Vi bruker datafilen **lowbwt.sav**. Vi skal nå presentere fordelingen til en kontinuerlig variabel. Det kan grafisk gjøres ved et histogram. Ved å lage et histrogram via *Graphs* vi har noen flere muligheter enn nå vi gjør det via *Analyze/Descriptove Statistics*/*Explore*. Vi skal nå lage en presentasjon av variabelen BWT fordelt etter variabelen SMOKE.

Vi går da inn i *Graphs/Legacy Dialogs/Histogram.* Det åpner det seg en dialogboks, der vi trekker BWT over i *Variable* og SMOKE over i *Rows*. Vi klikker også på *Display normality curve*, for å få lagt normalfordelingen på histogrammet for å se om data ser normalfordelte ut. Da er dialogboksen slik ut:



Etter *OK*, har vi følgende:



Vi ser at dette er en fin presentasjon av de to fordelingene. De er lagt over hverandre, med samme skala, så vi ser med én gang at fordelingen av fødselvekter er forskjøvet til venstre for røykerne, mot lavere vekter for røykerne. Vi ser også at normalfordelingskurvene tyder på at dataene i begge grupper er normalfordelte.

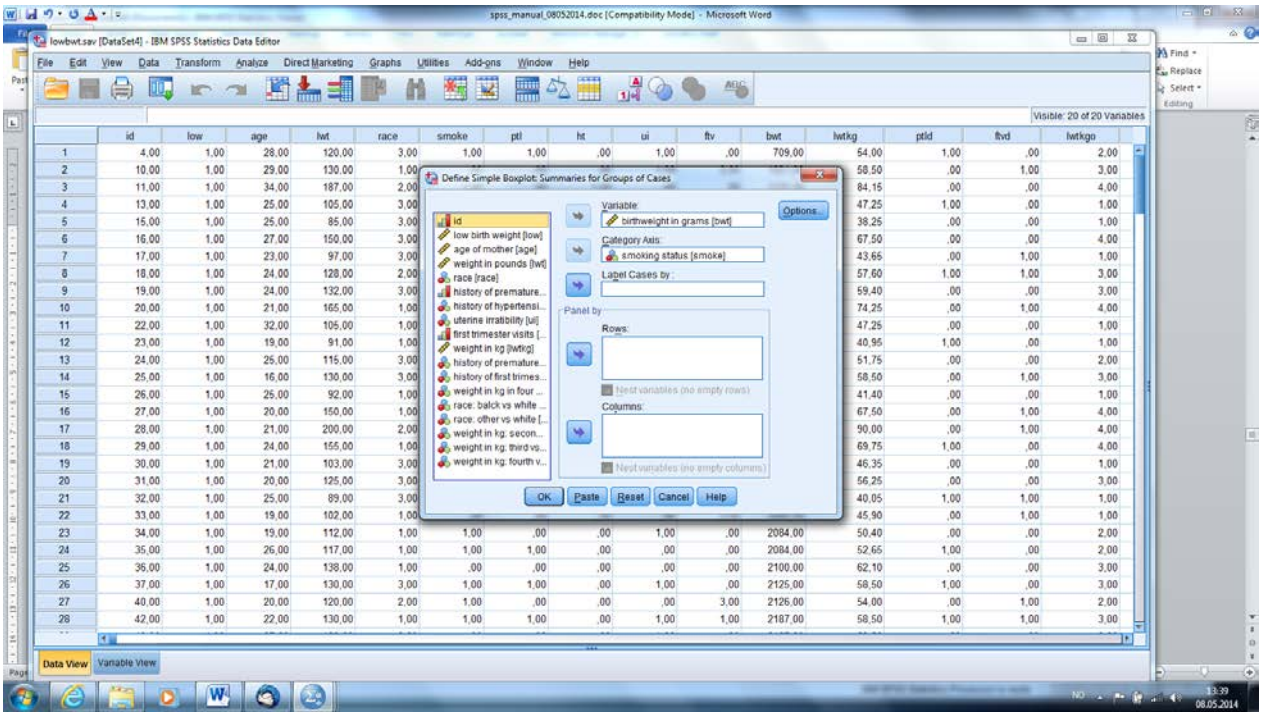
Vi har tidligere presentert disse histogrammene i *Analyze/Descriptive Statistics/Explore*, men da var histogrammene presentert for hver gruppe for seg. Denne presentasjonen er mye bedre.

Men *Analyze/Descriptive Statistics/Explore* gir oss mulighet for å lage et normalfordelingsplott, og det er viktig. Vi kan ikke basere våre vurderinger av om data er normalfordelte bare på historgrammer med overlagte normalfordelinger. Vi må bruke et normalfordelingsplott.

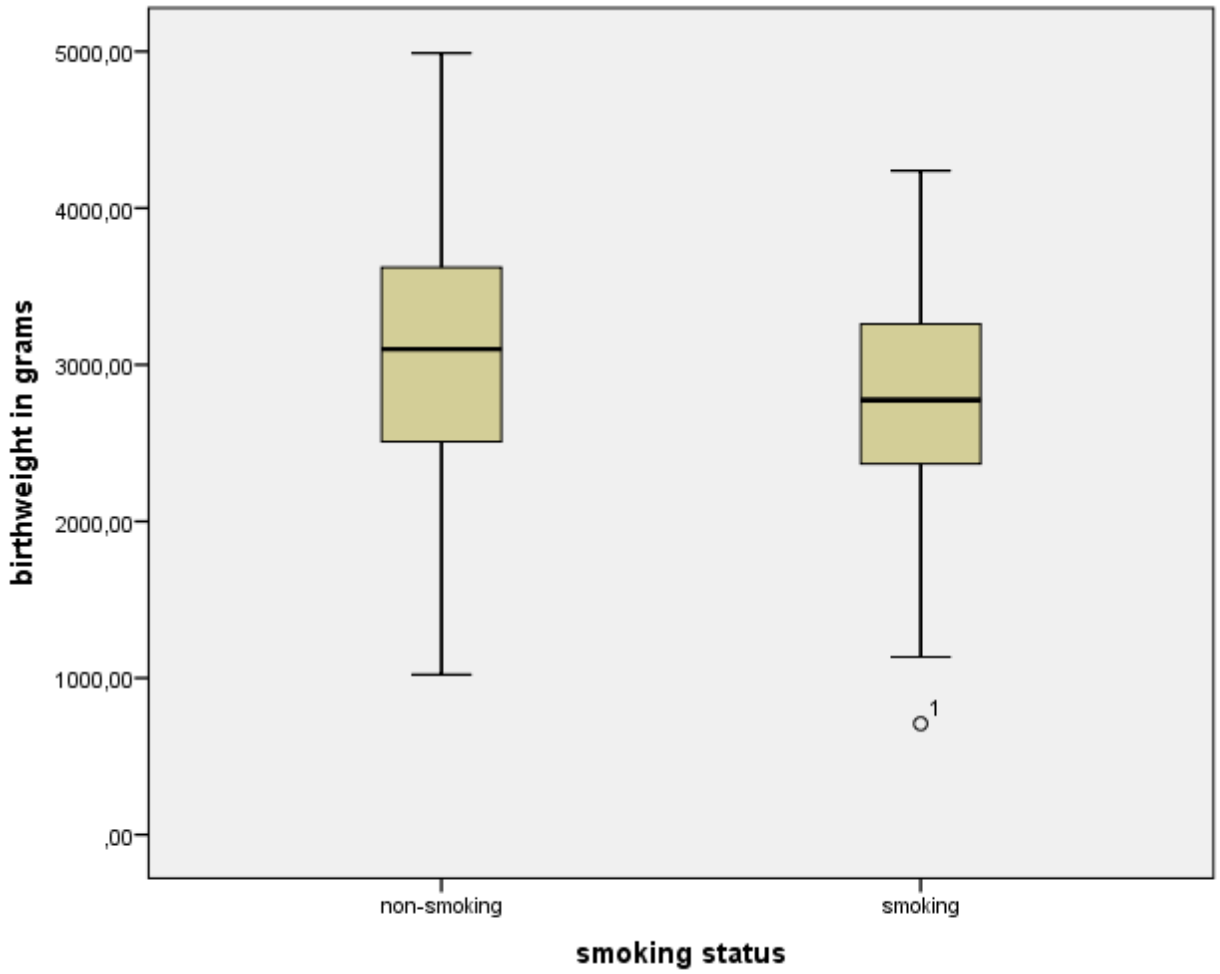
10.3 Boksplott. Eksempel: lowbwt.sav

Boksplott gir en fin presentasjon av hovedinntrykkene i dataene. Plottet gir oss informasjon om maksimums- og minimumsverdiene, om median, øvre og nedre kvartil, og det viser om det er skjeve fordelinger. Vi bruker fortsatt datafilen **lowbwt.sav**. Nå skal vi presentere sammenhengen mellom BWT og SMOKE via et boksplott.

Vi går da inn i *Graphs/Legacy Dialogs/Boxplott*. I den første dialogboksen klikker vi på Simple og Defien. Det åpner det seg en ny dialogboks, der vi trekker BWT over i *Variable* og SMOKE over i *Category Axis*. Da er dialogboksen slik:



Vi klikker på *OK* og det følgende plottet:

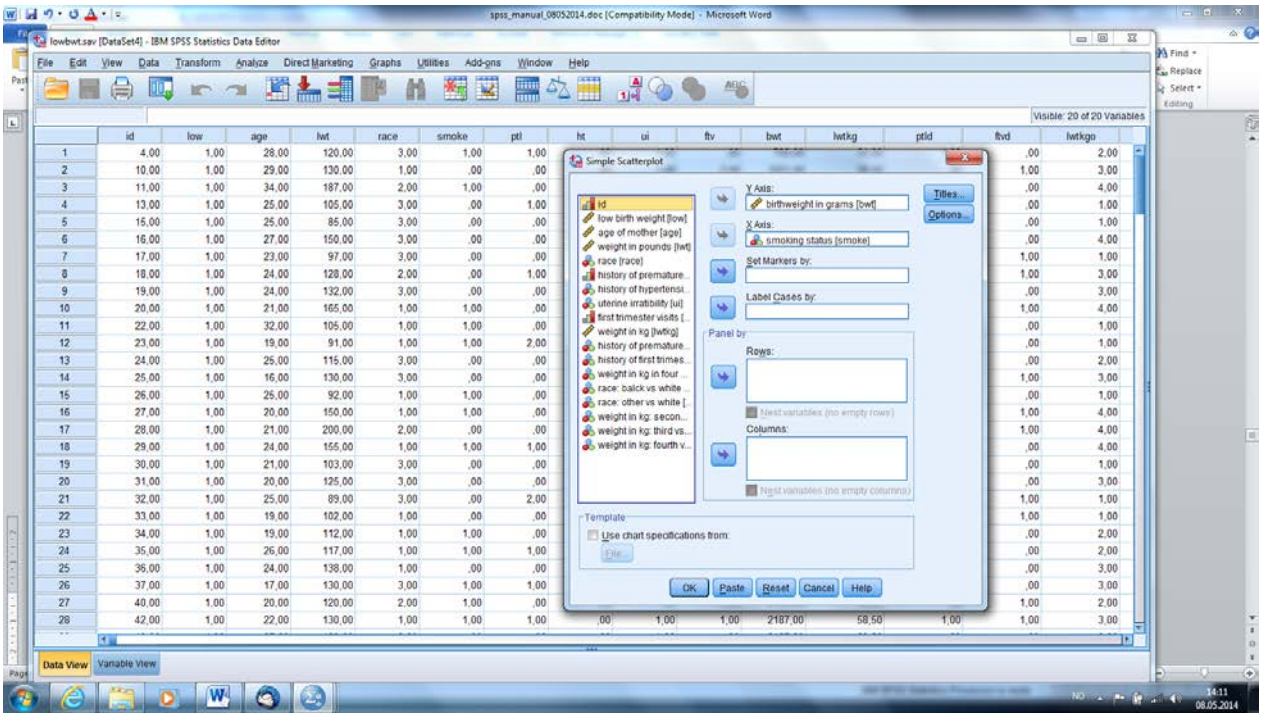


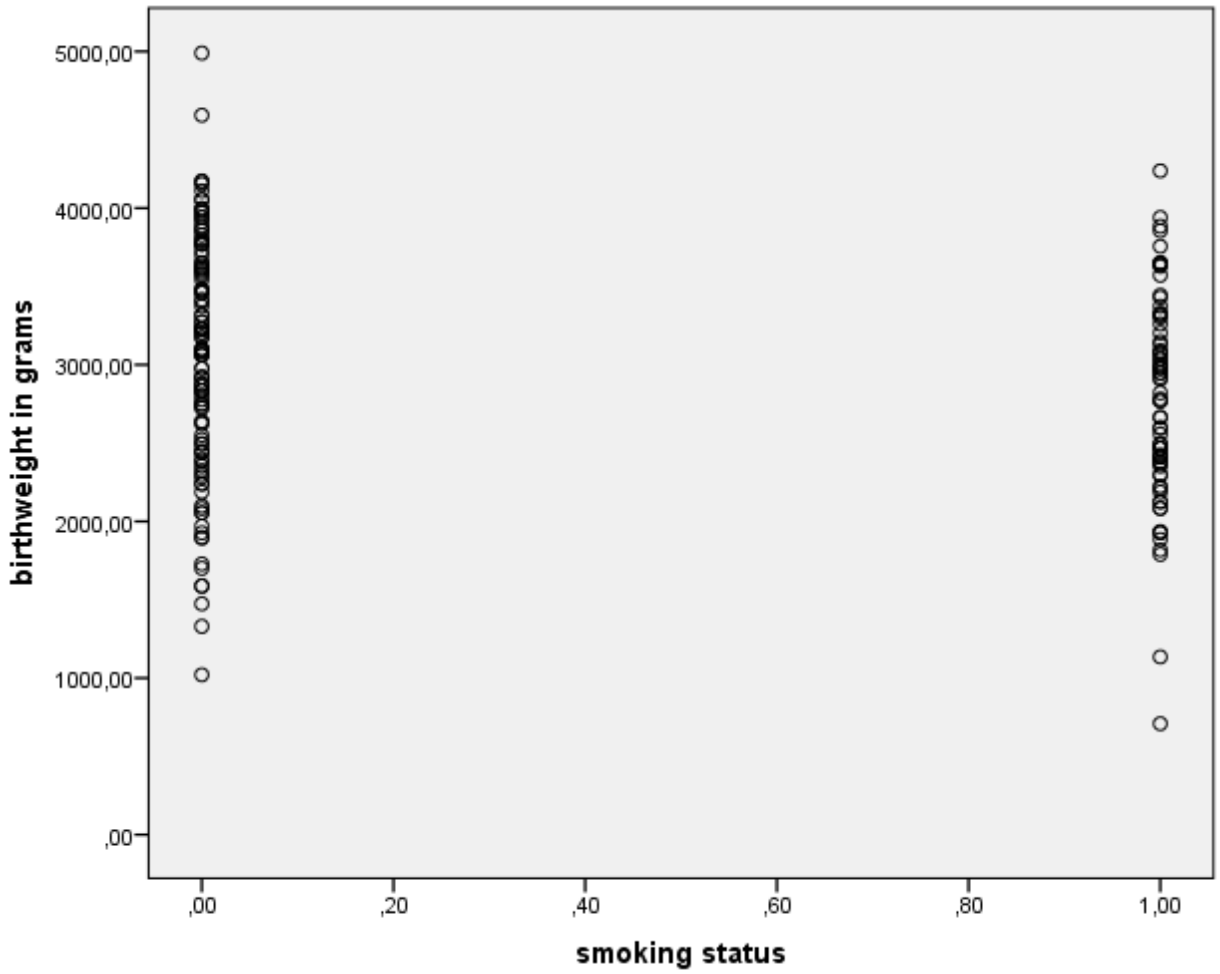
Igjen ser vi en fin presentasjon av boksplottet – og igjen bedre enn den vi får via *Analyze/Descriptive Statistics/Explore*. Siden boksplottene er plassert ved siden av hverandre kan vi lett sammenligne gruppene av mødre som røyker og ikke røyker. Vi ser at median fødselsvekt er høyere for ikke-røykerne enn for røykerne. Vi ser at fordelingene er ganske symmetriske, siden avstandene mellom øvre kvartil og median er omtrent den samme som mellom median og nedre kvartil. I tillegg er avstandene fra øvre kvartil til maksimumverdi omtrent lik avstanden fra nedre kvartil til minimum. Derfor vil median og gjennomsnitt være omtrent like store. Det er noe større variasjon blant røykerne enn blant ikke-røykerne. Én observasjon blant røykerne er å anse som ekstremverdi. Det er observasjon nummer 1, med en fødselsvekt på 709 gram.

10.4 Spredningsplott. Eksempel: lowbwt.sav

Spredningsplott er sannsynligvis den viktigste og mest vanlige måten å fremstille data på. Det er en fin måte å fremstille sammenhengen mellom to variabler på, ved å lage en to-dimensjonal presentasjon. De to variablene vi skal se på, kan enten være to kontinuerlige variabler eller én kontinuerlig og en kategorisk variabel. Vi skal bruke datasettet **lowbwt.sav** til å se på dette.

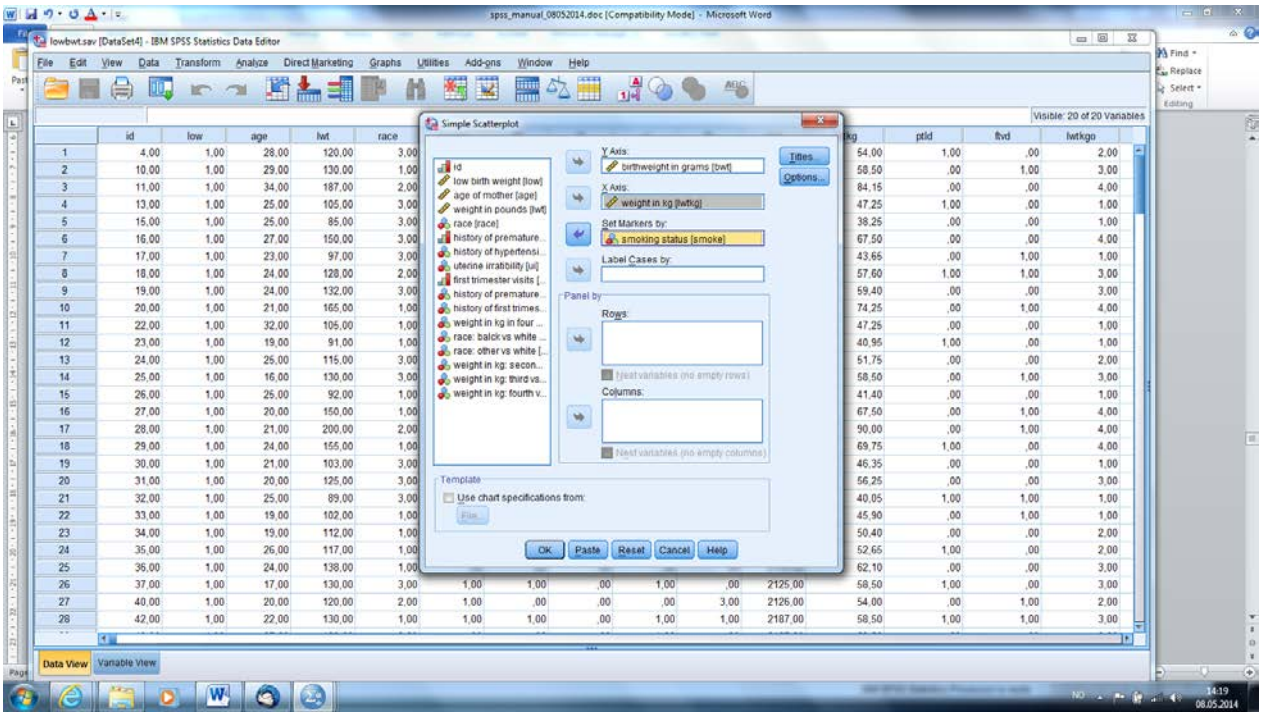
Først skal vi se på en fremstilling av sammenhengen mellom en kontinuerlig og en kategorisk variabel. Vi går til *Graphs/Legacy Dialogs/Scatter/Dot.* I den første dialogboksen klikker vi på *Simple Scatter* og *Define*. Det åpner det seg en ny dialogboks, der vi trekker BWT over i *Y axis* og SMOKE over i *X Axis*. Da er dialogboksen slik:



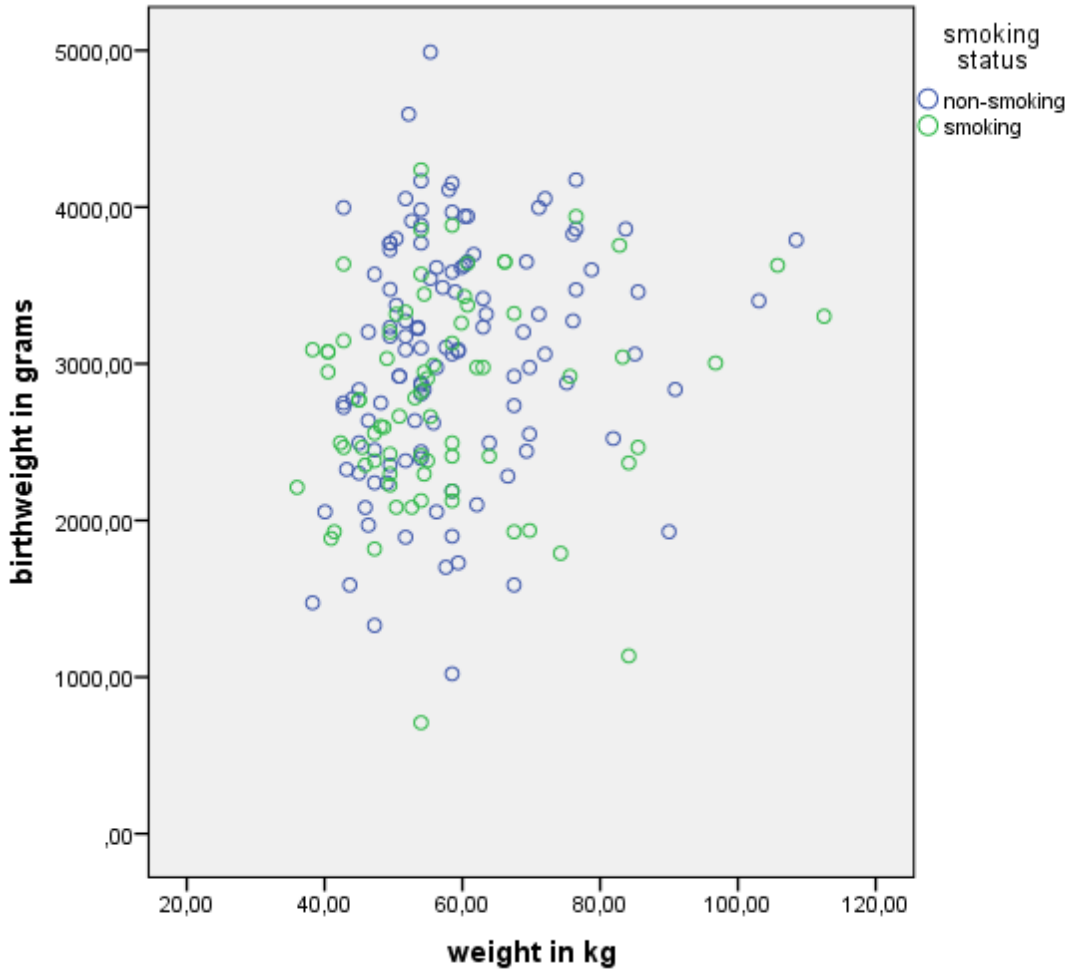


Her ser vi at det er to rekker med data, en for ikke-røykerne (SMOKE = 0) og en for røykerne (SMOKE = 1). Vi ser at det er større variasjon blant ikke-røykerne, men det er vanskelig å vurdere symmetri i fordelingen.

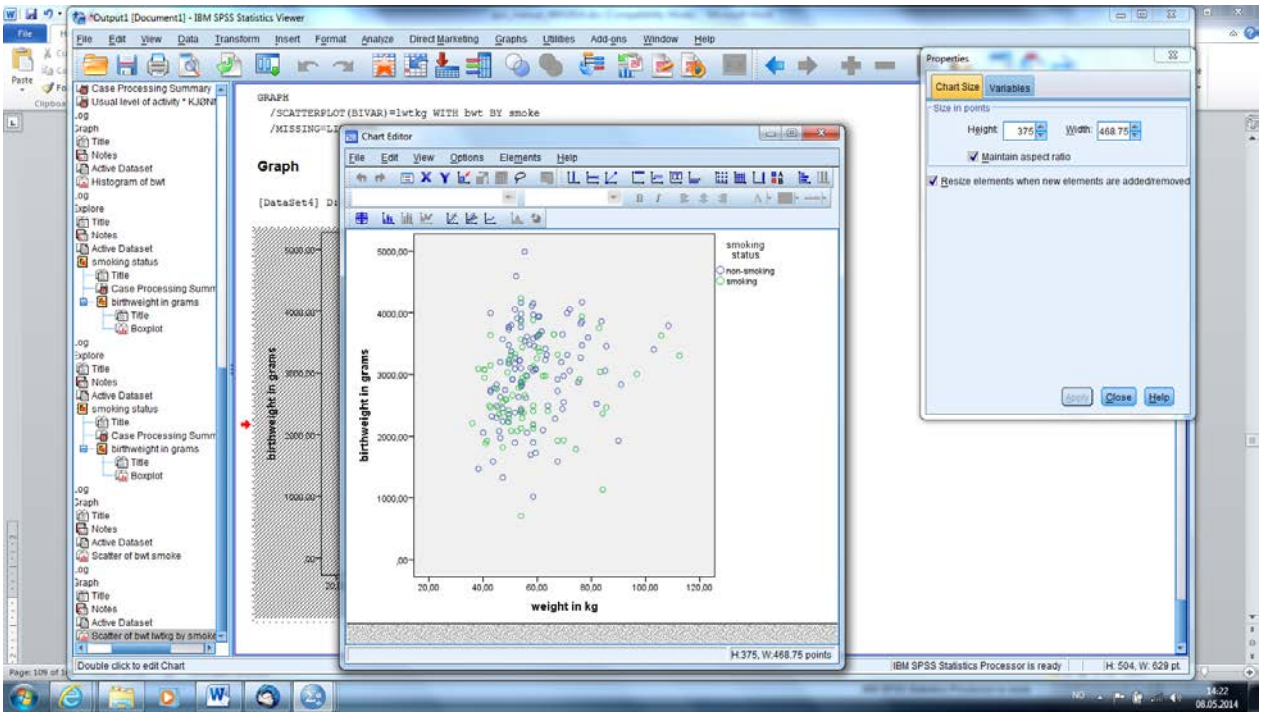
La nå til slutt se på sammenhengen mellom sammenhengen mellom barnets vekt (BWT) og mors vekt i kg (LWTKG) blant ikke-røykerne og røykerne. Vi går tilbake til *Graphs/Legacy Dialogs/Scatter/Dot.* I den første dialogboksen klikker vi igjen på *Simple Scatter* og *Define*. I neste dialogboks trekker vi over BWT i *Y axis* og LWTKG i *X Axis*. Til slutt trekker vi over SMOKE i *Set markers by*. Da er dialogboksen slik:



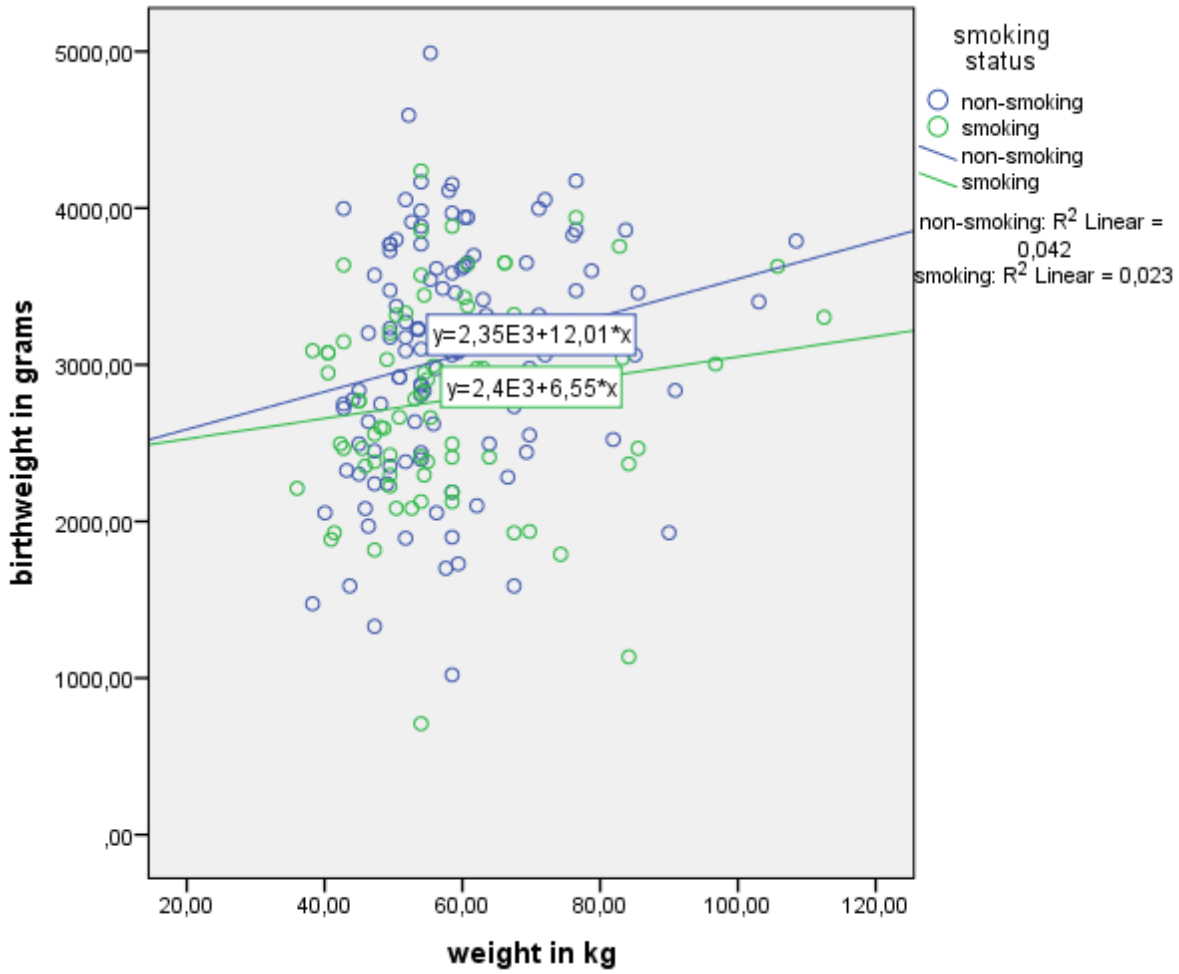
Etter *OK* får vi følgende plott:



Dette gir oss kanskje ikke så mye, siden sammenhengene virker litt rotete. Det kan være nyttig å få presentert sammenhengen mellom BWT og LWTKG på en enklere måte. Nå vi er i utskriftsvinduet, går vi til selve plottet og dobbeltklikker på det. Da kommer plottet opp i et nytt vindu. Da ser vinduene våre slik ut:



Nå går vi inn i grafikkvinduet *Chart Editor*. Her går vi til *Elements/Fit line at subgroups*. Den siste kommandoen gjør at vi får lagt inn to rette linjer for sammenhengen mellom BWT og LWTKG, én for ikke røykerne og én for røykerne. Hvis vi nå avslutter *Chart Editor* med å klikke på boksen med X øverst i høyre hjørne, kommer vi tilbake til utskriftfilen med det nye plottet. Det vi se slik ut:



Her ser vi de to linjene lagt inn. Vi har også fått lagt inn formelen for sammenhengen. For ikke-røykerne er sammenhengen:

BWT = 2350 + 12.0xLWTKG.

Merk at E3 betyr at vi flytter kommaet 3 plasser til høyre. Tilsvarende, E-3 betyr at vi flytter kommaet 3 plasser til venstre.

Vi har nå fått en fin presentasjon av sammenhengen mellom BWT og LWTKG.

11. Univariable statistiske metoder

Læringsmål

I dette kapittelet skal vi se på de mest sentrale metodene i statistisk analyse. En statistisk analyse vil alltid begynne med en deskriptiv analyse, gjerne sammen med diagrammer og plott for de viktige variablene.

Men statistisk analyse handler om å studere sammenhenger mellom to variabler eller mellom én variabel og flere andre variabler. Vi skiller mellom den avhengige variabelen og

forklaringsvariabelen. Den avhengige variabelen er den variabelen vi skal forklare. Forklaringsvariabelen er den variabelen vi forklarer den avhengige variabelen med. Når vi har bare én forklaringsvariabel, sier vi at analysen er univariabel. Dersom vi har flere enn én forklaringsvariabel, sier vi at analysen er multivariabel. Dette er tema for kapittel 12.

Målenivået på den avhengige variabelen og på forklaringsvariabelen avgjør hvilken analysemetode vi skal bruke. Dersom den avhengige variabelen er kontinuerlig og forklaringsvariabelen er kategorisk, med to kategorier, vil vi bruke t-tester eller ikke-parametriske metoder. Valget mellom de to metodene avhenger av hva slag fordeling den avhengige variabelen har. Dersom dataene er normalfordelte, vil vi bruke t-tester, ellers vil vi bruke ikke-parametriske metoder.

Dersom både den avhengige variabelen og forklaringsvariabelen er kategoriske, vil vi bruke krysstabeller for å analysere sammenhengen.

Dersom både den avhengige variabelen og forklaringsvariabelen er kontinuerlig, kan vi studere sammenhengen ved korrelasjon. Vi skal i kapittel 12 for denne situasjonen studere sammenhengen også ved regresjonsanalyse.

11.1 T-test for pardata

T-tester er de mest vanlige testene vi utfører. Det er to typer t-test: t-tester for pardata og t-tester for to uavhengige grupper. T-testene for pardata kalles også ofte for ett-utvalgs t-tester, siden vi bruker differansene mellom målingene i alle parene som utgangspunkt for vårt effektestimat og vår test.

T-testen som vi beregner for pardata er gitt som

t = Effektmålet/ Standardfeilen til effektmålet,

der effektmålet er gjennomsnittet på differansene, og standardfeilen er standardfeilen til differansen. På samme måte er konfidensintervallet tilnærmet gitt som

(Effektmålet – 1.96 x Standardfeilen til effektmålet, Effektmålet + 1.96 x Standardfeilen til effektmålet)

Konfidensintervallet er tilnærmet, siden vi egentlig bruker 97.5-persentilen i t-fordelingen, og ikke 1.96.

Før vi starter analysen må vi sikre oss at differansene er normalfordelte. Det er dette som «tillater» oss å bruke en t-test.

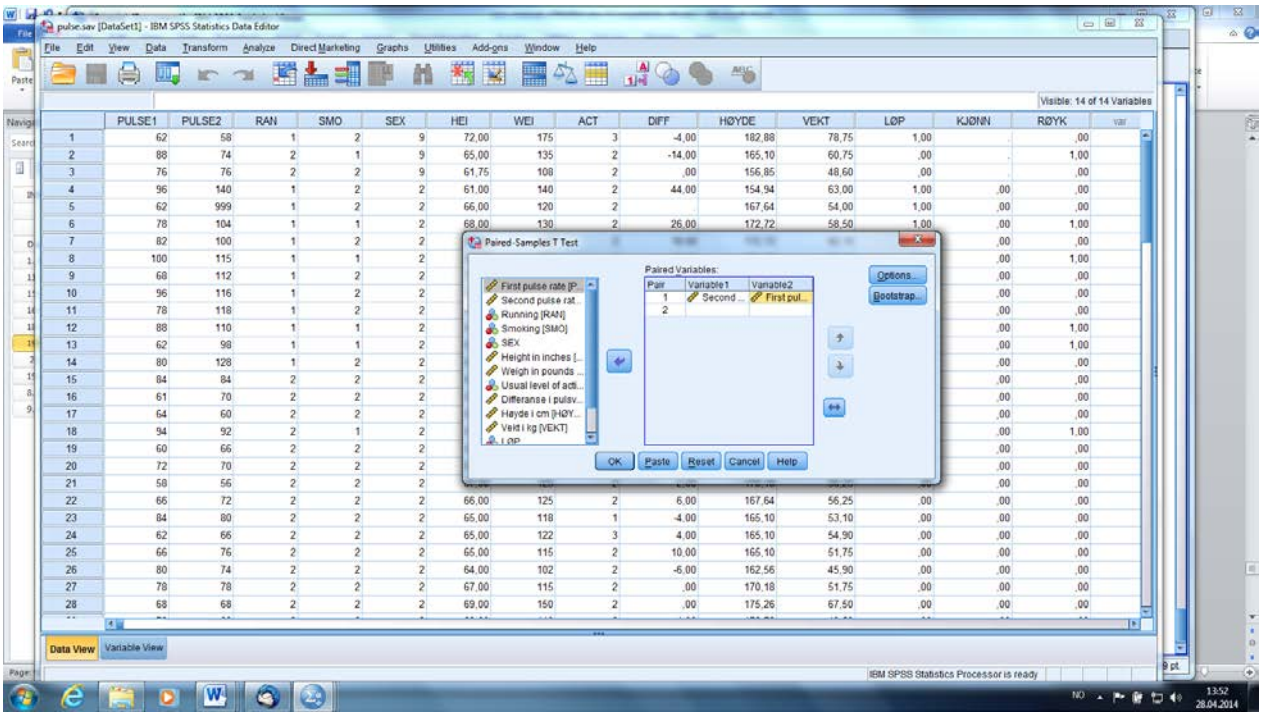
11.1.1 Eksempel: pulse.sav

Vi bruker igjen **pulse.sav** datasettet. Vi skal sammenligne pulsverdien før og etter løping. Dette ser vi på som en ett-utvalgs test for parede data. I SPSS kan vi enten bruke PULSE1 og PULSE2 verdiene i en paret test, eller vi kan bruke differansen DIFF i en ett-utvalgs test. Vi skal se på begge metodene, og starter med å analysere paret PULSE1, PULSE2.

Vi har allerede vurdert normalfordelingen til PULSE1 og PULSE2 og funnet ut at vi kan anta at dataene er normalfordelt. Da kan vi bruke en t-test for parede data.

Vi velger da *Analyze/Compare means/Paired Samples T-Tests*. Da kommer det opp en ny dialogboks, med overskriften *Paired Variabels*. Her flytter PULSE2 inn i *Variable 1* og PULSE1 inn i Variable 2. Grunnen til at vi flytter PULSE2 inn i *Variabel 2* og PULSE1 inn i *Variabel 1* er at SPSS alltid beregner differansen mellom *Variable 2* og *Variable 1*. Siden vi vet at PULSE2-verdiene er høyere enn PULSE1 verdiene får vi da en positiv differanse.

Da ser dialogboksen slik ut:



Vi klikker på *OK*, og får da denne utskriften:

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Second pulse rate	79,16	88	16,795	1,790
	First pulse rate	72,59	88	10,935	1,166

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Second pulse rate - First pulse rate	6,568	12,869	1,372	3,842	9,295	4,788	87	,000

Som vi viste tidligere steg pulsen fra 72.6 til 79.2. Effektmålet er differansen, nemlig det som er presentert under Mean i tabellen under. Effektmålet er 6.6. Vi ser at t = 4.8. Denne er beregnet som

t = Effektmålet/ Standardfeilen til effektmålet.

Siden gjennomsnittet er effektmålet og standardfeilen er Standard Error of the Mean (Std. Error Mean) finner vi at

t = 6.57/1.37 = 4.8.

Konfidensintervallet er tilnærmet gitt som

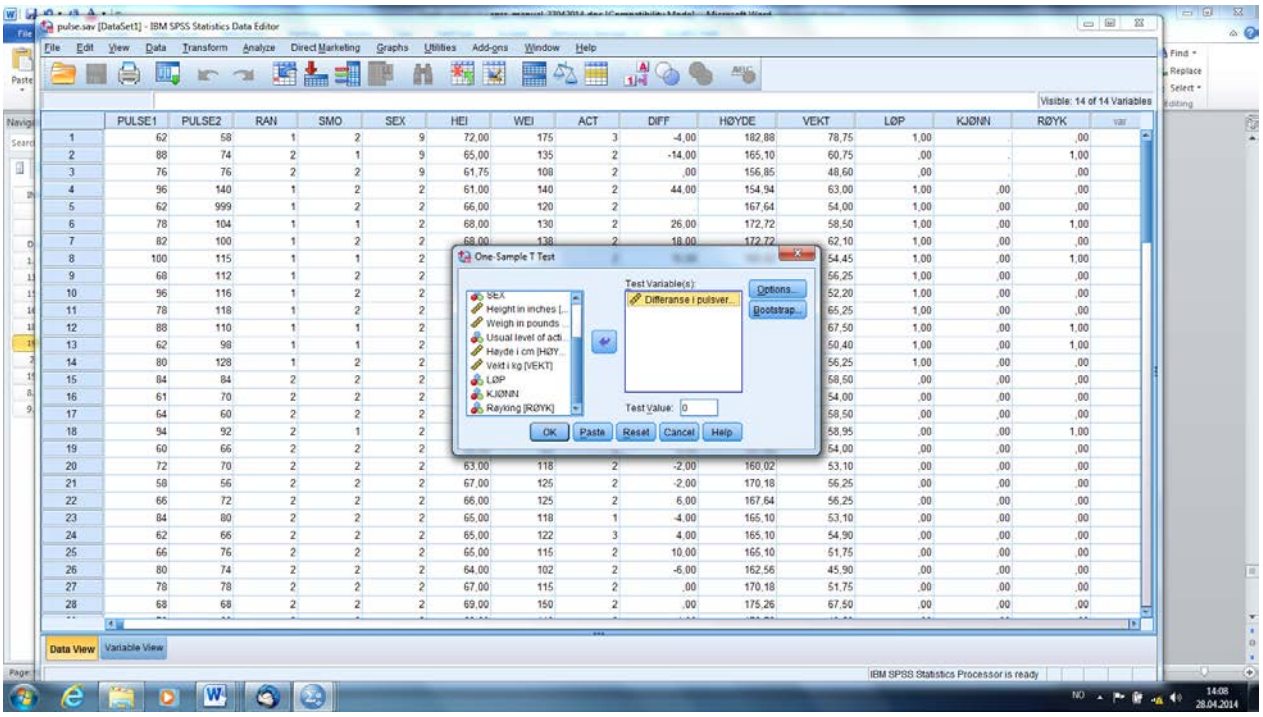
(Effektmålet – 1.96 x Standardfeilen til effektmålet, Effektmålet + 1.96 x Standardfeilen til effektmålet)

Dette intervallet er tilnærmet, siden vi egentlig bruker 97.5-persentilen i t-fordelingen, og ikke 1.96. Gjør vi det riktig, som SPSS gjør, får vi at et 95% konfidensintervallet er gitt som (3.8, 9.3). P-verdien regnes ut som sannsynligheten for å få den observerte eller enda større t-verdier. Denne beregner SPSS for oss, vi finner at p-verdien er $p < 0.001$. Vi kan altså si at det statistisk signifikant forskjell i de to pulsverdiene.

Merk at SPSS skriver ut tall med tre desimaler. Dersom tallet er < 0.0005, skriver SPSS 0.000. Dette har selvfølgelig ingen mening, spesielt for p-verdier, som er en sannsynlighet. Når SPSS skriver p = 0.000, skriver vi $p < 0.001$.

Vi finner altså at det er en klar statistisk forskjell mellom verdiene før og etter løping, selv om bare en del av forsøkspersonene løp (dem med LØP = 1).

La oss nå se at vi kan gjøre den samme analysen med en ett-utvalgs t-test, da for variabelen DIFF. Vi velger da *Analyze/Compare means/One-Sample T Test*. Igjen kommer det opp en ny dialogboks, nå med overskriften *One Sample T Test*. Her flytter vi DIFF over i Test Variable. Da får vi følgende dialogboks:



Når vi klikker på *OK* får vi følgende utskrift:

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Differanse i pulsverdier	88	6,5682	12,86874	1,37181

One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Differanse i pulsverdier	4,788	87	,000	6,56818	3,8416	9,2948

Vi ser at vi får akkurat den samme utskriften, og vi kan hente frem effektmålet (gjennomsnittet), konfidensintervallet og p-verdien like enkelt fra denne analysen som fra den parede analysen. Hva vi velger er smak og behag.

Legg også merke til at det meste av de resultatene som vi bruker i ett-utvalgs testen også fremkommer ved å gjøre en analyse ved *Analyze/Descriptive Statistics/Explore* og trekke over DIFF i *Dependent List*. Da får vi det følgende:

Descriptives

			Statistic	Std. Error
Differanse i pulsverdier	Mean		6,5682	1,37181
	95% Confidence Interval for Mean	Lower Bound	3,8416	
		Upper Bound	9,2948	
	5% Trimmed Mean		5,3687	
	Median		2,0000	
	Variance		165,604	
	Std. Deviation		12,86874	
	Minimum		-14,00	
	Maximum		48,00	
	Range		62,00	
	Interquartile Range		13,50	
	Skewness		1,558	,257
	Kurtosis		2,017	,508

Vi ser at gjennomsnittet, standardfeilen og konfidensintervallet er de samme som vi fikk i ett-utvalgs testen. Men i t-testen får vi også beregnet selve t-verdien og p-verdien for den statistiske testen.

11.2 T-test for to uavhengige utvalg

T-tester for to uavhengige utvalg er trolig den mest brukte testen i statistiske analyser. Den er vanlig i randomiserte kliniske studer, men også vanlig i observasjonelle studier, med to uavhengige grupper. Betingelsen for å bruke t-testen er at dataene i de to gruppene vi studerer er uavhengige og normalfordelte. Betingelsen om uavhengighet er vanligvis lett å avklare. Her kommer dataene fra ulike personer, i to grupper, og de er uavhengige. Men antagelsen om normalfordelingen må vi sjekke.

T-testen som vi beregner for to uavhengige utvalg er gitt som

t = Effektmålet/ Standardfeilen til effektmålet,

der effektmålet er differansen mellom gjennomsnittene i de to gruppene, og standardfeilen er standardfeilen til denne differansen. På samme måte er konfidensintervallet tilnærmet gitt som

(Effektmålet – 1.96 x Standardfeilen til effektmålet, Effektmålet + 1.96 x Standardfeilen til effektmålet)

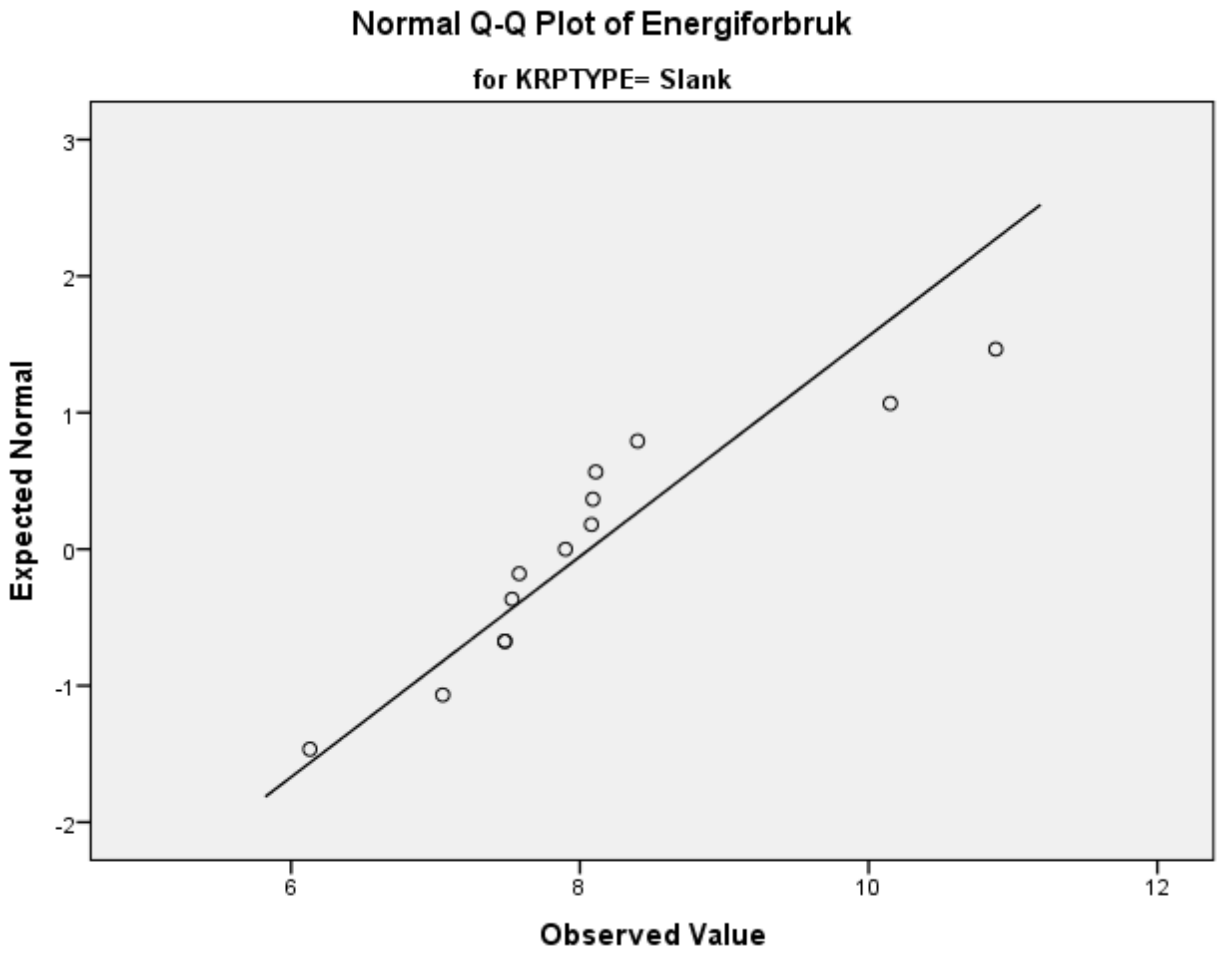
Igjen er det slik at konfidensintervallet er tilnærmet, siden vi egentlig bruker 97.5-persentilen i t-fordelingen, og ikke 1.96.

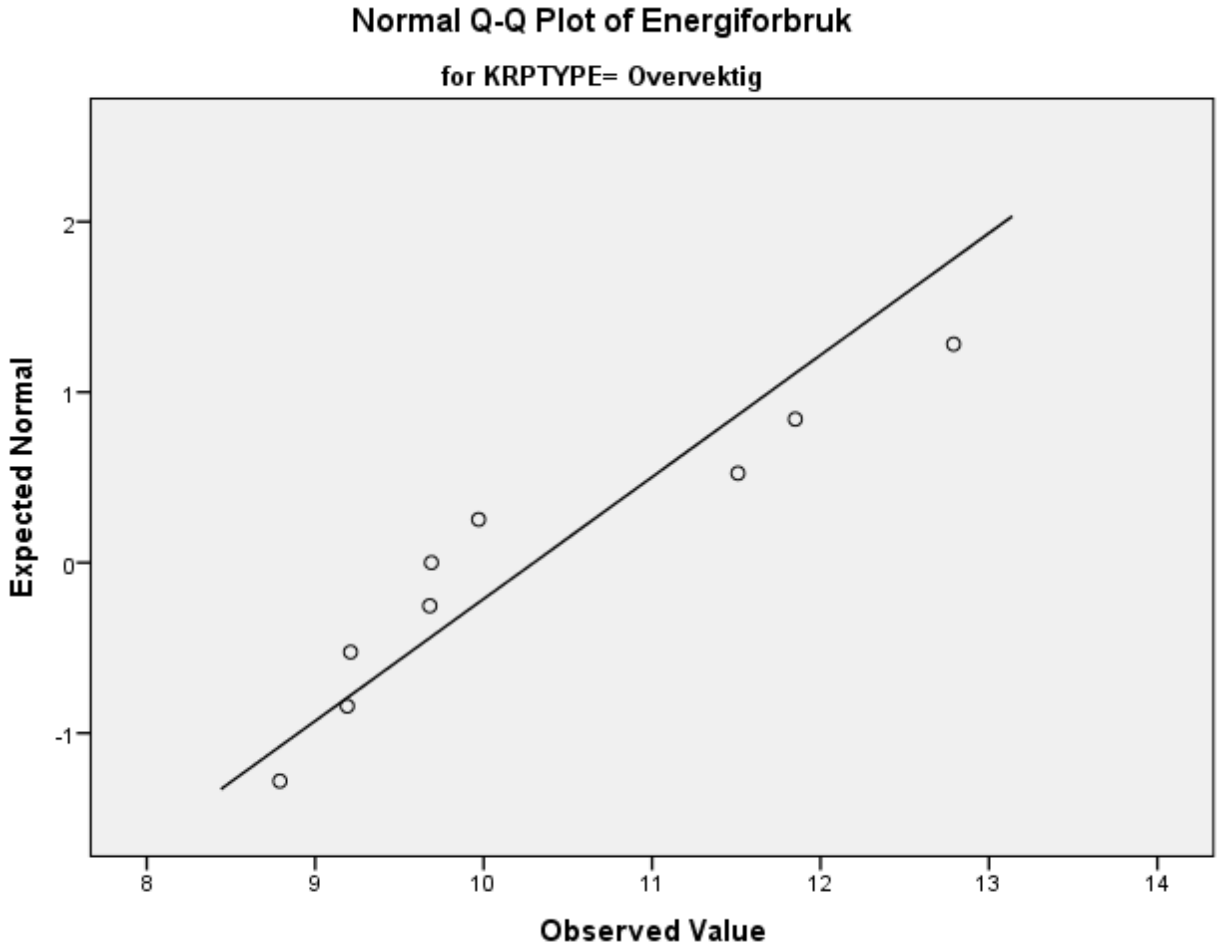
Merk at vi har to t-tester, en paret (ett-utvalgs) test og en to-utvalgs test (som av og til også en uparet test). Mange blander disse to sammen, men det er svært uheldig, og kan føre til gale konklusjoner. Merk også at data som er i par IKKE er uavhengige, siden de for eksempel er gjort på samme person. Men data i to utvalg, vil være uavhengige.

11.2.1 Eksempel: altman.sav

Vi skal nå se litt mer på eksemplet fra Altman som vi har brukt mange ganger. Disse dataene lagret vi som en SPSS data fil under navnet **altman.sav**. I kapittel 5.4 gjorde vi en liten statistisk analyse av disse dataene, der vi presenterte de to gruppene mht. energiforbruk. Vi gikk da inn i *Analyze/Descriptive Statistics/Explore* for å finne gjennomsnitt, standardavvik, median osv. for de to gruppene slanke og overvektige. Disse to gruppene er å betrakte som to uavhengige grupper, siden det er ulike personer som er målt i de to gruppene.

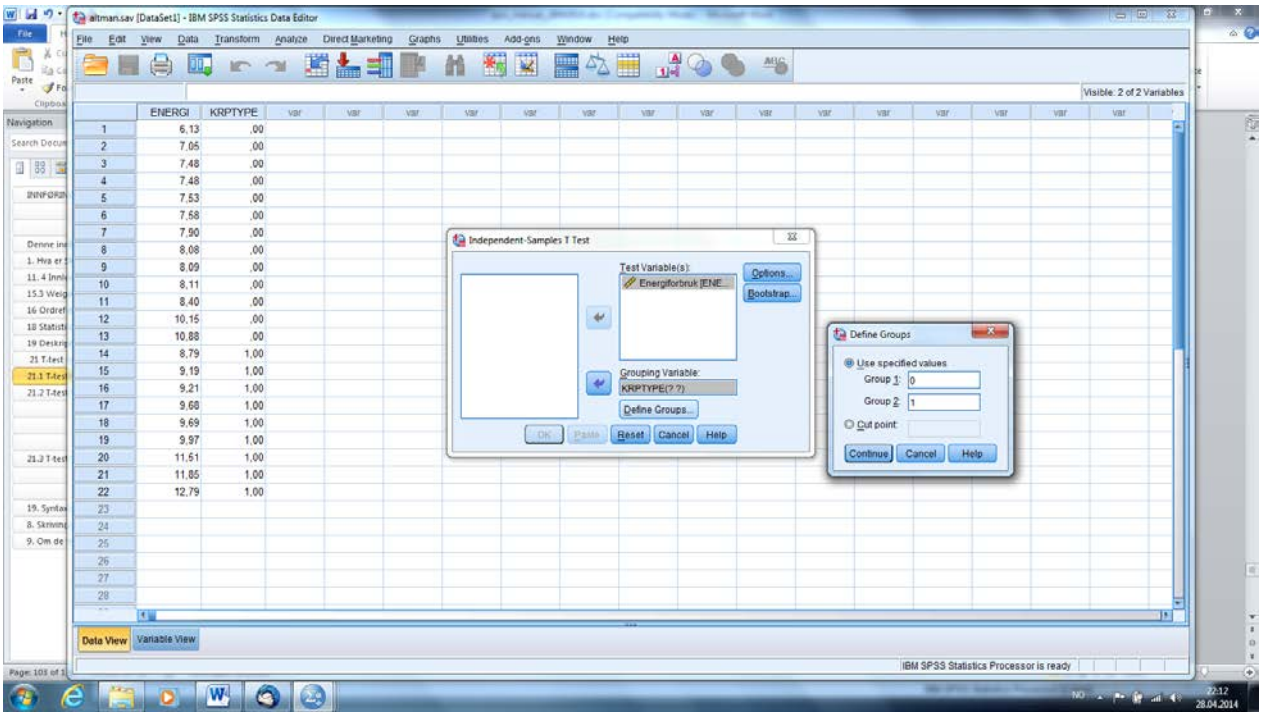
Når vi skal teste om det er forskjell i energiforbruk blant de slanke og overvektige kan vi bruke en t-test for to uavhengige utvalg, dersom dataene i de to gruppene er normalfordelte. Vi går inn i *Analyze/Descriptive Statistics/Explore* og trekker ENERGI over i *Dependent List* og KRPTYPE over i *Factor List*. Vi klikker på Plots, og merker av at vi skal ha *Normality Plots with Tests*. Da får vi følgende utskrift:





Her må vi være spesielt oppmerksom på at vi har lite data, og vi må da forvente at vi vil få avvike fra den rette linjen. I begge disse tilfellene er avvikene små, og vi antar at dataene er normalfordelte.

Da kan vi gå videre med en t-test for to uavhengige utvalg. Vi går da inn i *Analyze/Compare Means/ Independent Samples T Tests*. Her velger vi variabelen ENERGI som *Test Variable(s)* og variabelen KRPTYPE som *Grouping Variable*. Vi kan ennå ikke klikke på *OK*. Først må vi spesifisere hvilke verdier på vår grupperingsvariabel som skal sammenliknes, dvs. fortelle SPSS at verdiene er 0 og 1. Det gjør vi ved å klikke på *Define*. Da åpner det seg en ny dialogboks, og der skriver vi at 0 er gruppe 1 og 1 er gruppe 2. Med dette angir vi kodene for de to gruppene av variabelen KRPTYPE. Da ser dialogboksene våre slik ut:



Når dette er gjort, kan vi klikke på *Continue*. Da legges kodene 0 og 1 inn i variabelen KRPTYPE og vi kan klikke på *OK*, og få utført t-testen. Vi får da denne utskriften:

Group Statistics

	Kroppstype	N	Mean	Std. Deviation	Std. Error Mean
Energiforbruk	Slank	13	8,0662	1,23808	,34338
	Overvektig	9	10,2978	1,39787	,46596

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Energiforbruk	Equal variances assumed	1,002	,329	-3,946	20	,001	-2,23162	,56560	-3,41145	-1,05180
	Equal variances not assumed			-3,856	15,919	,001	-2,23162	,57882	-3,45917	-1,00408

Vi får først en sammenlikning av energiforbruket i de to gruppene. Så undersøker programmet om standardavviket i de to gruppene er like. Dette gjøres ved Levenes test til venstre i den andre tabellen. Når den har en P-verdi som er mindre enn 0.05 forkaster vi nullhypotesen om at gruppene har lik standardavviket. Da har vi altså grunnlag for å påstå at standardavvikene er ulike. I vårt tilfelle er p = 0.329 som klart gir grunnlag for å anta at gruppene har likt standardavvik. Da skal vi lese av resultatet av t-testen på første linje med Equal variances assumed.

Effektmålet er forskjellen i gjennomsnitt i de to gruppene. Det er gitt i Mean Differences, og er lik -2.2. Merk at SPSS alltid tar differansen mellom første og andre gruppe, altså de slanke minus de overvektige. Vi får da et negativt effektestimat.

Vi ser da at den to-sidige p-verdien er 0.001, følgelig er gruppene høysignifikant forskjellige med hensyn på energiforbruk. Et 95%-konfidensintervall for forskjellen står lengst til høyre, og det viser at intervallet er (-3.4, -1.1). Vi får altså presentert de tallene vi er interessert i: effektmålet, konfidensintervallet og p-verdien.

11.2.2 Eksempel: lowbwt.sav

Vi går inn i datafilen lowbwt.sav. Her er vi interessert i å undersøke om det er statistisk signifikant forskjell i fødelsvekter for røykende og ikke-røykende mødre. Vi har tidligere sett på en deskriptiv analyse av problemet, og også sett at dataene i de to gruppene kan antas å være normalfordelte. Da er vi klare til å gjøre en t-test for to uavhengige grupper.

Vi går da inn i *Analyze/Compare Means/ Independent Samples T Tests*. Her velger vi variabelen BWT som *Test Variable(s)* og SMOKE som *Grouping Variable*. Vi klikker så på *Define*. I dialogboksen skriver vi at 0 er gruppe 1 og 1 er gruppe 2, siden kodene for SMOKE er 0 og 1. Da kan vi klikke på *Continue*, og til slutt på *OK*, og få utført t-testen. Vi får da denne utskriften:

Group Statistics

	smoking status	N	Mean	Std. Deviation	Std. Error Mean
birthweight in grams	non-smoking	115	3054,9565	752,40901	70,16250
	smoking	74	2773,2432	660,07517	76,73218

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
birthweight in grams	Equal variances assumed	1,508	,221	2,634	187	,009	281,71328	106,96873	70,69274	492,73382
	Equal variances not assumed			2,709	170,001	,007	281,71328	103,97406	76,46677	486,95979

Av den øverste tabellen ser vi at forskjellen i gjennomsnittlig fødselsvekt er 281.7 gram (3054.9 – 2273.2). Vi ser også at standardavvikene er ganske like, 752 gram blant ikke-røykerne og 660 blant røykerne.

Levenes test gjør en statistisk test på om standardavvikene i de gruppene er forskjellige. Her er p = 0.221 og vi antar at gruppene har likt standardavvik. Vi finner da resultatet av t-testen på første linje med Equal variances assumed.

Der ser vi effektmålet vårt, som nettopp er forskjellen i gjennomsnittlig fødselsvekt, er 281.7 gram. Standardfeilen på differansen er 107.0. Formelen for å regne ut denne må vi finne i en lærebok i statistikk eller i statistikkforelesningene. Men nå har grunnlag for å regne ut t-

verdien som vi bruke som grunnlag for om vi kan påstå at det er forskjell i fødselsvektene. Vi har nemlig at t-verdien beregnet som

t = Effektmålet/ Standardfeilen til effektmålet,

og da har vi at

t = 281.7/107.0 = 2.63.

P-verdien er nå sannsynlighetene for å denne t-verdien og t-verdier som er enda større. SPSS regner ut denne p-verdien for oss, og vi har at p = 0.009. Siden p < 0.05, kan vi påstå at det er statistiske signifikante forskjeller i fødselsvekt i de to gruppene.

Konfidensintervallet er tilnærmet gitt som

(Effektmålet – 1.96 x Standardfeilen til gjennomsnittet, Effektmålet + 1.96 x Standardfeilen til gjennomsnittet)

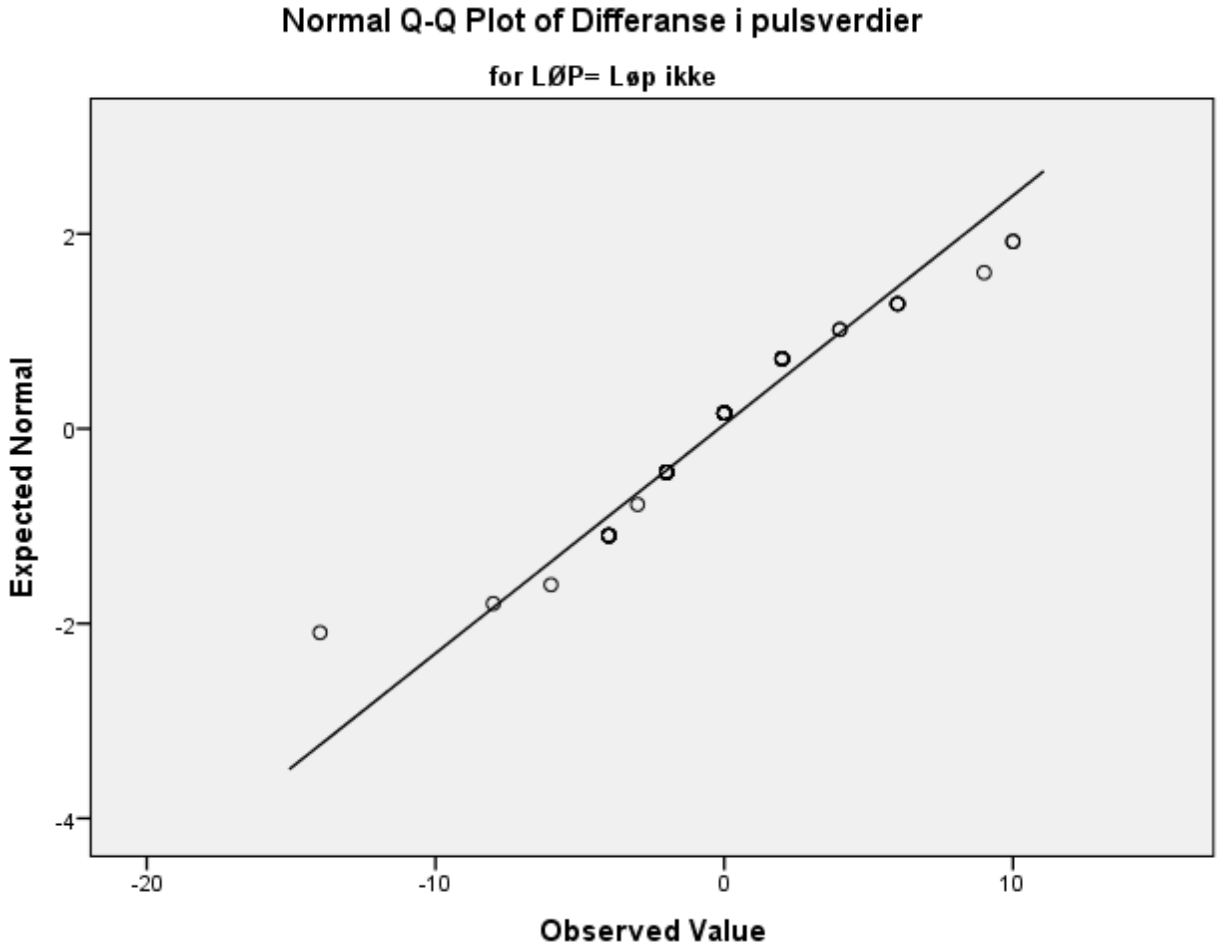
Igjen har vi at dette intervallet er tilnærmet, siden vi egentlig bruker 97.5-persentilen i t-fordelingen med 11 + 74 – 2 = 185 frihetsgrader, og ikke 1.96. Gjør vi det riktig, som SPSS gjør, får vi at et 95% konfidensintervallet er gitt som (70.7, 492.7).

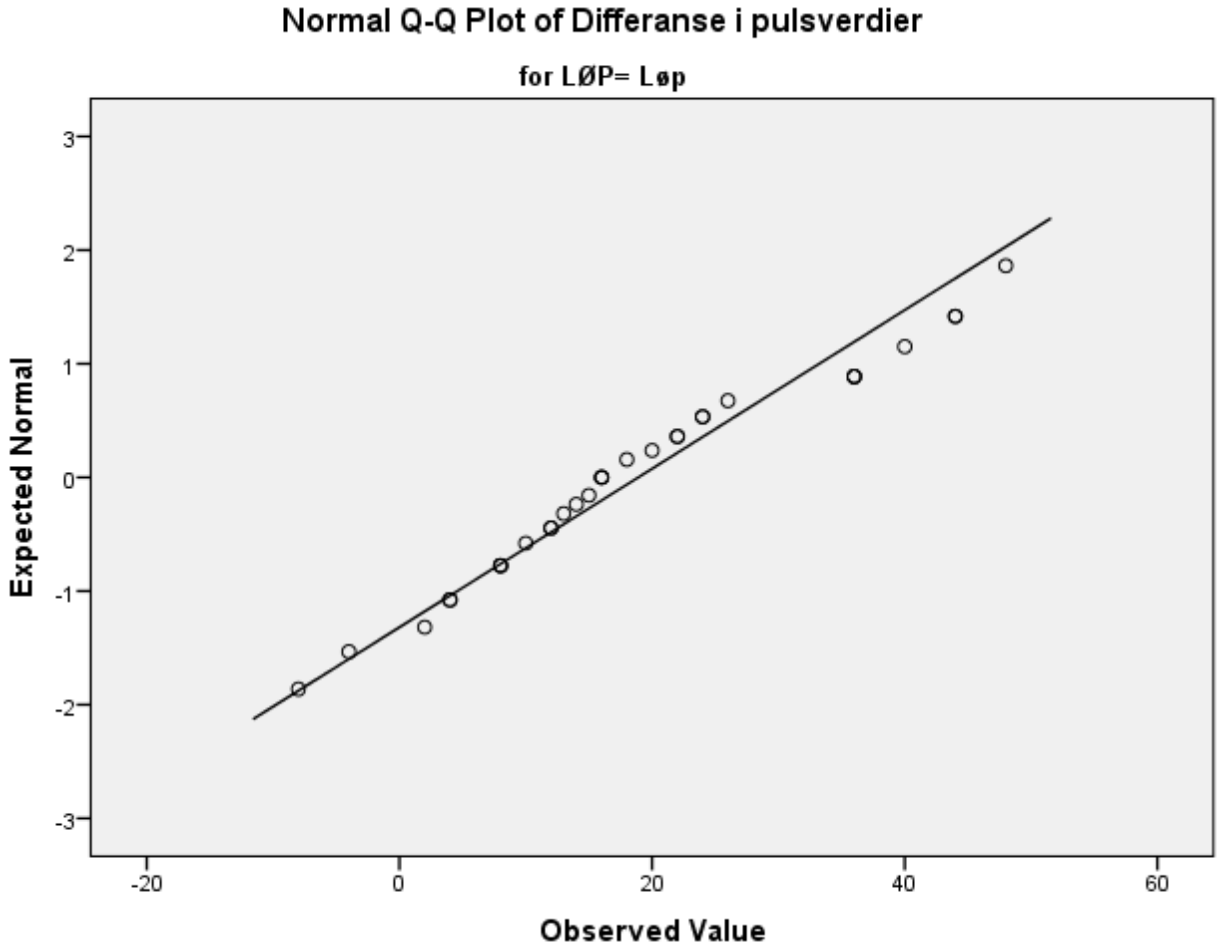
11.2.3 Eksempel: pulse.sav

I veldig mange studier har en blanding av parede og uparede data. Dette er tilfellet med **pulse.sav** dataene. Der observerer vi pulsverdiene før og etter en intervensjon (løping) og vi er interessert i effekten av intervensjonen. Dette vil også være situasjonen i kliniske forsøk, der vi måler utgangsverdien og verdien etter intervensjon. Intervensjonen består da av behandling og placebo, og studiepersonene er randomisert mht. om de kommer i behandlings- eller intervensjonsgruppen.

I slike situasjoner beregner vi differansen i måleverdiene før og etter intervensjon. Dette er da parede data. Så har vi to uavhengige grupper, behandlingsgruppen og placebogruppen. Det er forskjellen mellom disse to vi er interessert. Dette er uparede data, og vi tester forskjellen med en t-test for to uavhengige utvalg.

I **pulse.sav** har vi beregnet DIFF. Vi skal nå teste om det er ulikt gjennomsnitt av DIFF blant de som løp (LØP =1) og de som ikke løp (LØP = 0). Men først må vi undersøke om vi har grunnlag til å anta at dataene i de to gruppene er normalfordelte. Da går vi som vanlig inn i *Analyze/Descriptive Statistics/Explore* og trekker DIFF over i *Dependent List* og LØP over i *Factor List*. Vi går inn i *Plots* og klikker av på at vi vil ha *Normality plot with tests*. Da får vi følgende resultat:


Normal Q-Q Plot of Differanse i pulsverdier
for LØP= Løp ikke
Expected Normal
2
0
-2
-4
-20
-10
0
10
Observed Value



Her ser vi at det er gode grunner til å anta at dataene er normalfordelt i de to gruppene. Det er noen små avvik (to observasjoner) i den høyre halen for de som ikke løp. Men her må vi ta i betraktning at der lite data, og vi føler oss derfor trygge på at vi kan gå videre med en t-test for to utvalg.

Da går vi inn i *Analyze/Compare Means/ Independent Samples T Tests*, og velger DIFF som *Test Variable(s)* og LØP som *Grouping Variable*. I dialogboksen skriver vi igjen 0 og 1 for de to gruppene, siden kodene for LØP er 0 og 1. Etter å ha klikket på *Continue*, og til slutt på *OK*, får vi utført t-testen, med denne utskriften:

Group Statistics

	LØP	N	Mean	Std. Deviation	Std. Error Mean
Differanse i pulsverdier	Løp ikke	54	-,1852	4,25629	,57921
	Løp	31	18,9032	14,33028	2,57379

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Differanse i pulsverdier	Equal variances assumed	43,669	,000	-9,146	83	,000	-19,08841	2,08717	-23,23972	-14,93710
	Equal variances not assumed			-7,235	33,068	,000	-19,08841	2,63816	-24,45538	-13,72145

I tabellene over får vi nye resultater å forholde oss til. I øverste tabell ser vi at differansen i gjennomsnittlig pulsverdi er -19.09 (0.19 – 18.90). Men viktigere at at standardavviket er veldig forskjellig i de to gruppene. De som ikke løp har et standardavvik på 4.3, mens de som løp har et standaravvik på 14.3. Dette er ikke så rart, siden vi vet at løping både øker pulseverdien og øker spredningen.

Men de ulike standardavvikene fører også til at p-verdien på Levenes test i tabellen under er $p < 0.001$. Altså er det grunnlag til å påstå at standardavvikene er ulike i de to gruppene. Vi kan altså ikke lenger lese ut resultatene av den øverste linjen i t-testen. Vi må da gå til linjen under med Equal variances not assumed. Der finner vi at effektmålet er det samme, og det er naturlig siden det er gjennomsnittet som er vårt effektmål. Men vi ser at standardfeilen (Std. Error Difference) beregnes litt annerledes i denne situasjonen. Men uansett verdi av standardfeilen ser vi at t-verdien beregnes som

t = Effektmålet/ Standardfeilen til effektmålet,

som nå er

t = -19.09/2.09 = -9.15.

P-verdien er nå sannsynlighetene for å denne t-verdien og t-verdier som i dette tilfellet er enda mindre. SPSS regner ut denne p-verdien for oss, og vi har at $p < 0.001$, og vi kan påstå at det er statistiske signifikante forskjeller i pulsverdiene for de som løp og de som ikke løp.

Konfidensintervallet er igjen tilnærmet gitt som

(Effektmålet – 1.96 x Standardfeilen til gjennomsnittet, Effektmålet + 1.96 x Standardfeilen til gjennomsnittet)

Igjen har vi at dette intervallet er tilnærmet, men gjør vi det riktig, som SPSS gjør, får vi at et 95% konfidensintervallet er gitt som (-24.5, -13.7).

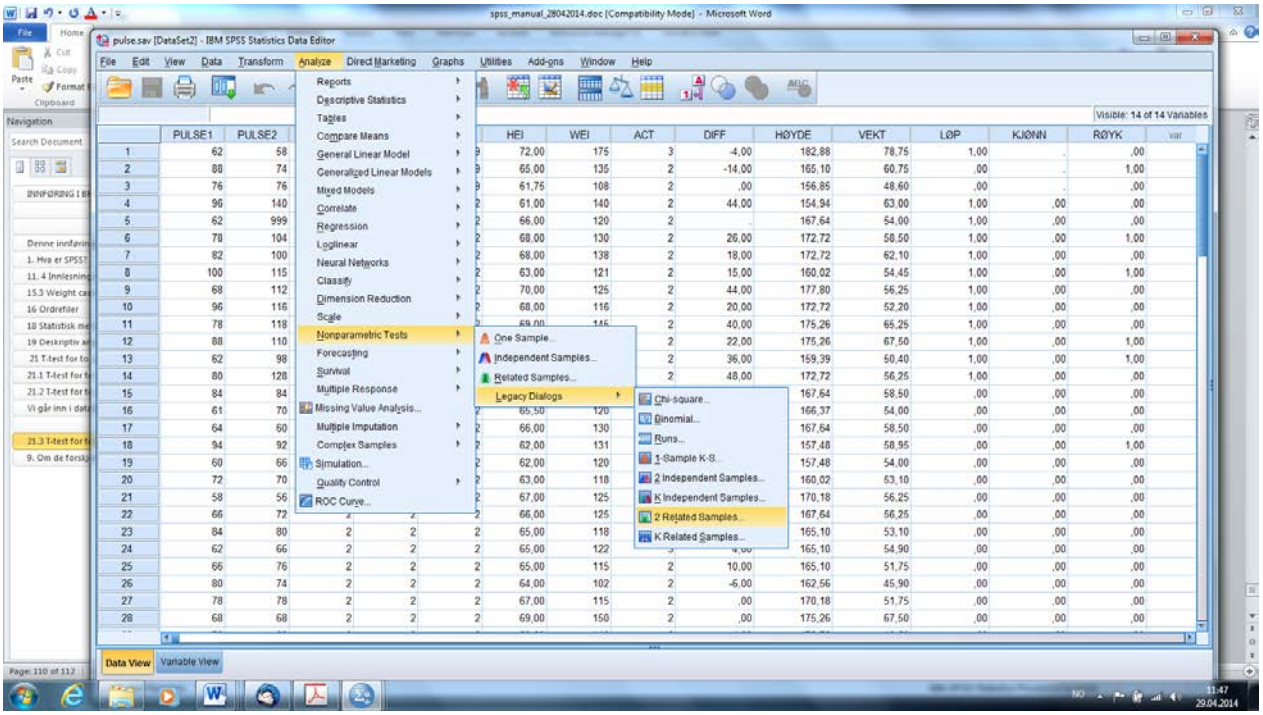
11.3 Ikke-parametriske metoder

Hvis dataene ikke er normalfordelt, må vi bruke ikke-parametriske metoder. Dette er metoder som er basert på rangene til observasjonene, og ikke selve verdiene. Vi har tidligere sett på persentiler, kvartiler og medianen. Alle disse observasjonene er rangbasert, siden vi først rangordner alle observasjonene våre og så finner vi persentiler, kvartiler og medianen ved å finne den riktige observasjonen blant de rangordnede. Nedenfor skal vi presentere metoder for

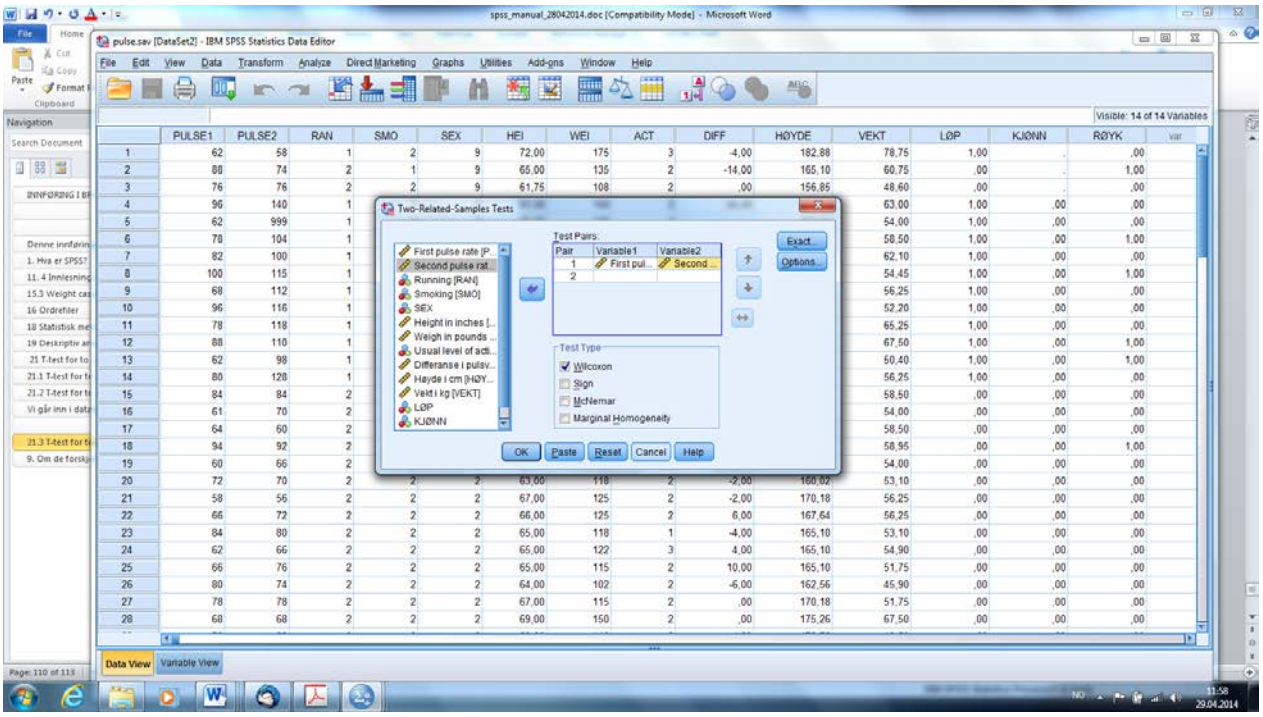
parede og uparede data, som da tilsvarer t-testene for parede data og for to uavhengige grupper (uparede data).

11.3.1 Wilcoxon test for pardata. Eksempel pulse.sav

Vi henter frem datafile **pulse.sav**. For å gjøre en ikke-parametrisk Wilcoxon test for pardata, går vi til *Analyze/Non-Parametric Tests/Legacy Dialogs* og velger i dialogboksen *2 Related Da ser Samples.*



I dialogboksen trekker vi over PULSE2 i Variable 1 og PULSE1 i Variable 2. Da ser dialogboksen våre slik ut:



Vi klikker på *OK* og får følgende resultat:

Ranks

		N	Mean Rank	Sum of Ranks
Second pulse rate - First pulse rate	Negative Ranks	25[a]	22,48	562,00
	Positive Ranks	46[b]	43,35	1994,00
	Ties	17[c]		
	Total	88		

a. Second pulse rate < First pulse rate

b. Second pulse rate > First pulse rate

c. Second pulse rate = First pulse rate

Test Statistics[a]

	Second pulse rate - First pulse rate
Z	-4,117[b]
Asymp. Sig. (2-tailed)	,000

a. Wilcoxon Signed Ranks Test

b. Based on negative ranks.

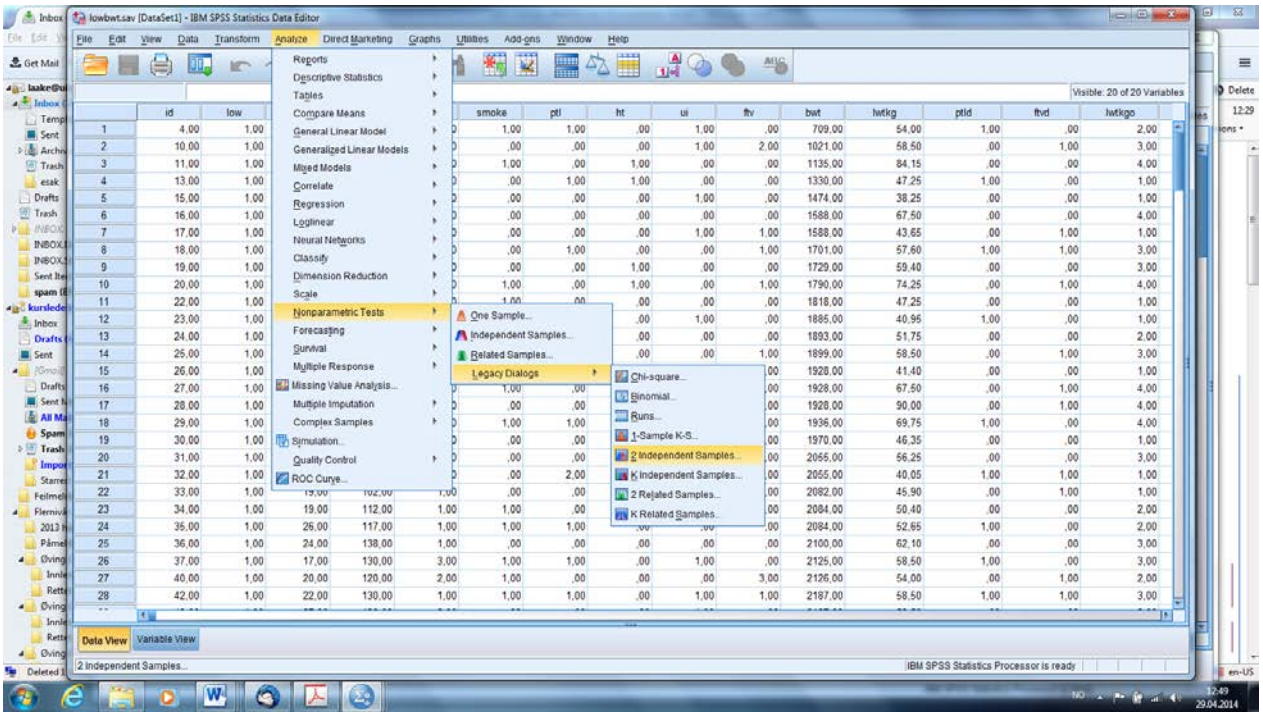
I den øverste tabellen får vi en oversikt over rangene til pulsverdiene. Vi ser at det er 17 ties. Dette er observasjoner som har samme verdi i PULSE1 som PULSE2. Disse observasjonene innholder ikke noen informasjon om endringer og de er tatt ut av analysen. Vi ser at det er 46 observasjoner som har PULSE2 verdier som er større enn PULSE1. I tabellen nedenfor ser vi

om dette er et signifikant resultat. Vi finner at $p < 0.001$, og vi konkluderer med at det er statistisk signifkant forskjell i PULSE1 og PULSE2 verdiene.

11.3.2. Wilcoxon-Mann-Whitney test for to uavhengige grupper. Eksempel lowbwt.sav

Den ikke-parametriske testen for to uavhengige grupper har mange navn, men den er mest kjent under navnet Wilxocon-Mann-Whitney testen. Den er analog til t-testen for normalfordelte data, men den er rangbasert.

Vi bruker dataene i **lowbwt.sav** og skal teste om det er forskjell i fødselsvekt for røykende og ikke-røykende mødre. Vi vet at det her er grunnlag for å kjøre en t-test, men vi vi nå vise resultatene for den tilsvarende ikke-parametriske testen. For å gjøre en Wilcoxon-Mann-Whitney test, går vi til *Analyze/Non-Parametric Tests/Legacy Dialogs* og velger i dialogboksen *2 Independent Samples.* Da ser dialogboksen slik ut:



Da kommer vi inn i en dialogboks som er lik med den vi har for t-testen for to uavhengige grupper. Vi trekker over BWT i *Test Variable List*, og SMOKE over i *Grouping Variable*. Så klikker vi på *Define* og bruker at de to gruppene for SMOKE er gitt ved kodene 0 og 1. Når vi klikker på *Continue* og *OK*, får vi følgende resultater:

Ranks

	smoking status	N	Mean Rank	Sum of Ranks
birthweight in grams	non-smoking	115	103,60	11913,50
	smoking	74	81,64	6041,50
	Total	189		

Test Statistics[a]

	birthweight in grams
Mann-Whitney U	3266,500
Wilcoxon W	6041,500
Z	-2,693
Asymp. Sig. (2-tailed)	,007

a. Grouping Variable: smoking status

Vi ser av den første tabellen at summen av rangene til ikke-røykende mødre er høyere enn for røykende mødre. Dette skyldes selvfølgelig at fødselsvekten for ikke-røykende mødre er gjennomgående høyere enn for røykende mødre, og da blir også rangene høyere.

I den andre tabellen finner vi at den to-sidige p-verdien (Asymp. Sig. (2-tailed)) er p =0.007. Konklusjonen er klar: Det er statistisk signifikant forskjell i fødselsvekt for ikke-røykende og røykende mødre.

11.4 Analyse av krysstabeller

Analyse av krysstabeller, eller tabellanalyse som det ofte kalles for enkelhets skyld, er like vanlig til analyse av kategoriske data som t-tester er for analyse av kontinuerlige data. Som navnet sier er en krysstabell en tabellering av data i celler som definert etter kategoriene for de kategoriske variablene.

Den mest vanlige krysstabellen er en 2x2 tabell, der vi har to variabler, begge med to kategorier. Vi hadde et slikt eksempel i kapittel 7.3, der vi så på sammenhengen mellom blodtrykk og behandling. BLODTRYKK hadde to kategorier, normalt blodtrykk og høyt blodtrykk, og BEHANDLING hadde også to kategorier, placebo og aspirin.

En generell krysstabell er av dimensjon rxc. I en slik tabell har vi r linjer (rows) og c kolonner (columns). Vi skal i 11.4.2 se på en 2x4 tabell. Det er en tabell med 2 linjer og 4 kolonner.

I kapittel 3.2 nevnte vi kort begrepene avhengig og uavhengig variabel. Dette blir viktige begreper når vi skal analysere krysstabeller. Vi gjentar kort: Den avhengige variabelen er den vi skal forklare, forklaringsvariabelen er den vi forklarer med. I eksempelet i kapittel 7.3 med datafilen **blodtrykk.sav** er det åpenbart at BLODTRYKK er den avhengige variabelen og BEHANDLING er forklaringsvariabelen.

Merk at vi alltid vil legge den avhengige variabelen i linjene og forklaringsvariabelen i kolonnene:

Blodtrykk * Behandling Crosstabulation

Count

		Behandling		
		Placebo	Aspirin	Total
Blodtrykk	Høyt blodtrykk	11	4	15
	Normalt blodtrykk	20	30	50
Total		31	34	65

Når vi nå skal prosentuere en tabell, skal vi alltid gjøre det ved å prosentuere etter forklaringsvariabelen. Dette skyldes at vi vil se på effekten som forklaringsvariabelen har på den avhengige variabelen, og det får vi uttrykt ved nettopp å prosentuere etter forklaringsvariabelen. Hvis vi gjør det (i kapittel 11.4.1 skal vi se hvordan), får vi følgende resultat:

Blodtrykk * Behandling Crosstabulation

			Behandling		
			Placebo	Aspirin	Total
Blodtrykk	Høyt blodtrykk	Count	11	4	15
		% within Behandling	35,5%	11,8%	23,1%
	Normalt blodtrykk	Count	20	30	50
		% within Behandling	64,5%	88,2%	76,9%
Total		Count	31	34	65
		% within Behandling	100,0%	100,0%	100,0%

Vi ser at andelen som oppnår normalt blodtrykk ved behandling med aspirin er 88.2%, mens den er 64.5% for placebo.

Det er viktig å merke seg at tabellen i SPSS er presentert annerledes enn i læreboken til Aalen og medforfattere, se side 130. Der presenteres tabellen også med den avhengige variabelen i linjene og forklaringsvariablene i kolonnene, men presentasjonen av kodene er snudd for begge variablene. Tabellen på side 130 ser slik ut:

	Eksponering		
Sykdom	Ja	Nei	
Ja	a	b	a+b
Nei	c	d	c+d
	a+c	b+d	n

Her ser vi at de syke (med kode 1) ligger i øverste linje og de eksponerte (med kode 1) ligger i venstre kolonne. Vi holder fast ved å bruke kodene 0 for de friske og de ikke-eksponerte, og 1 for de syke og eksponerte, men presentasjonen er annerledes enn slik SPSS gjør det.

Dersom vi skal lese inn en tabell som den over i SPSS og bruke *Weight cases*, slik vi gjorde i kapittel 7.3 blir tabellen med fire linjer, hver med tre kolonner:

1	1	a
1	0	b
0	1	c
0	0	d

Det er viktig å være klar over på hvilken måte tabellen er presentert på. Vi skal holde oss til presentasjonen til Aalen og medforfattere. Med denne notasjonen blir effektmålene som nedenfor:

Vi har tre effektmål som vi bruker til å måle effekten av forklaringsvariabelen på den avhengige variabelen. Det er

RD: Risiko differanse: a/(a+c) – b/(b+d)
RR: Relative risiko: [a/(a+c)]/[b/(b+d)]
OR: Odds ratio [a/c]/[b/d)]

Fortsatt med samme notasjon blir konfidensintervallene:

For RD:
RD – 1.96 x sqrt[1/(a+c) x a/(a+c) x c/(a+c) + 1/(b+d) x b/b+d) x d/(b+d)],
RD + 1.96sqrt[1/(a+c) x a/(a+c) x c/(a+c) + 1/(b+d) x b/b+d) x d/(b+d)]

For RR:
RR x exp[-1.96 x sqrt(1/a + 1/b – 1/(a+c) – 1/(b+d)],
RR x exp[1.96 x sqrt(1/a + 1/b – 1/(a+c) – 1/(b+d)].

For OR:
OR x exp[-1.96 x sqrt(1/a + 1/b + 1/c + 1/d],
RR x exp[1.96 x sqrt(1/a + 1/b + 1/c + 1/d].

Merk at null-verdien, altså verdien nå behandlet og ubehandlet gruppe har samme effekt er 0 for RD, og 1 for både RR og OR. Merk også at dersom RR > 1 har vi økt risiko, mend RR < 1 betyr redusert risiko. Tilsvarende betyr OR >1 økt odds og OR < 1 redusert odds.

Merk videre at vi bare definerer disse effektmålene for 2x2 tabeller. Dersom vi har større tabeller, for eksempel, en 2x4 tabell, må selektere (via *Select cases*) for å lage 2x2 tabeller der vi kan beregne effektmålene.

Merk til slutt at vi bruker begrepene risiko og sannsynlighet om hverandre. I statistikk bruker vi også tallene mellom 0 og 1 for sannsynligheter, og ikke 0% og 100%.

Når vi bruker standard presentasjon av tabellen for sammenhengen mellom blodtrykk og behandling, blir den.

	Behandling		
Blodtrykk	Aspirin	Placebo	
Normalt	30 (88.2%)	20 (64.5%)	50
Høyt	4 (11.8%)	11 (35.5%)	15
	34 (100%)	31 (100%)	65

RD er da differansen i sannsynligheten for å oppnå effekt med og uten behandling. I tilfellet over er da RD = 0.882 – 0.645 = 0.137. Det er altså en risikoreduksjon på 13.7 prosentpoeng. RR er forholdet mellom sannsynligheten for å oppnå effekt med behandling og uten behandling. I vår tilfelle er da RR = 0.882/0.645 = 1.37.

Odds ratioen er forholdet mellom oddsen for dem med behandling og oddsen for dem uten behandling. Odds er forholdet mellom sannsynligheten for å oppnå effekt dividert på sannsynligheten for ikke å oppnå effekt. I tabellen over er oddsen blant dem som får behandling med aspirin lik 0.882/0.118 = 7.47. Oddsen blant dem som får placebo er 0.645/0.355 = 1.82. Da er odds ratioen OR = 7.47/1.82 = 4.13.

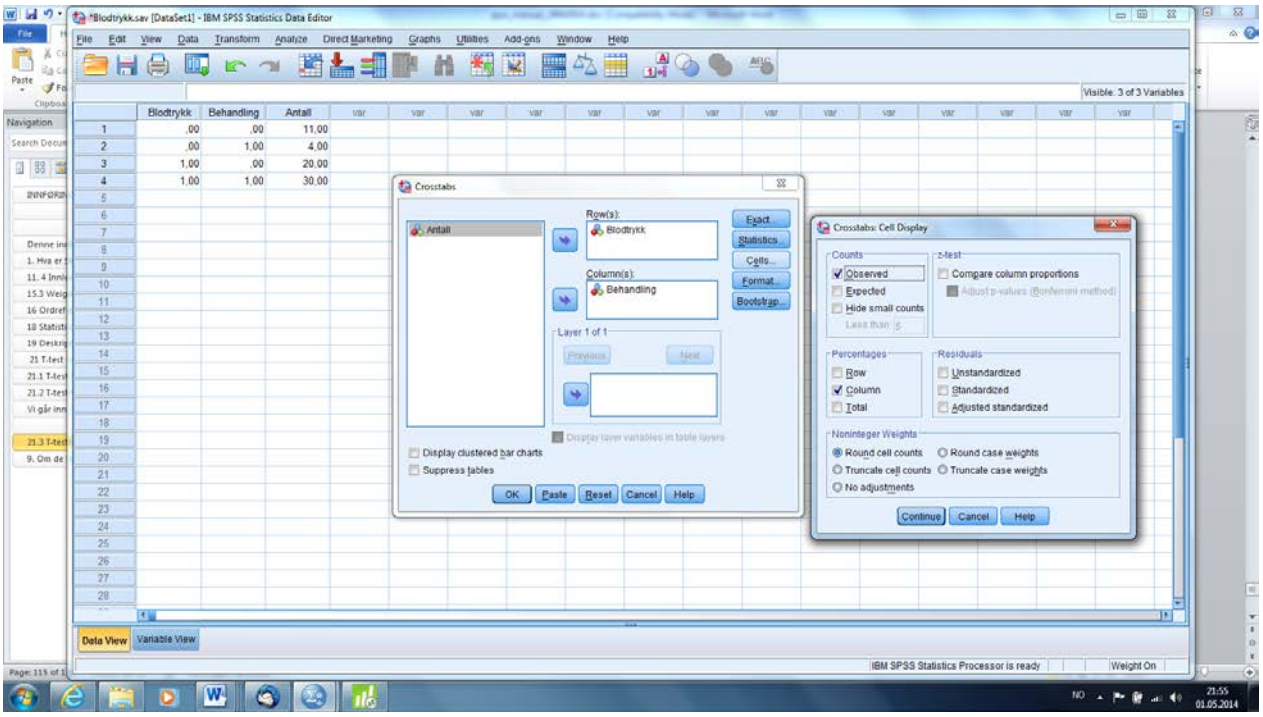
Enklere er det å bruke regelen om at odds ratioen er lik kryssproduktet i tabellen. Kryssproduktet er gitt ved å multiplisere celleantallet fra øvre venstre hjørne til nedre høyre hjørne og dele dette på det vi får ved å multiplisere celleantallet i øvre høyre hjørne og dele på antallet i cellen i nedre venstre hjørne. Dette gir OR = 30 x 11/20 x 4 = 4.13.

Testen for å undersøke om det er sammenheng mellom de to variablene i krysstabellen, altså mellom avhengig og forklaringsvariabel kalles kji-kvadrat testen. Selve teststørrelsen er kji-kvadratfordelt, og p-verdien for denne testen må regnes ut ved å bruke en tabell over kji-kvadrat fordelingen. Betingelsen for at vi kan bruke kji-kvadratfordelingen er at det er rimelig antall observasjoner i hver celle i tabellen. Under tabellen skriver SPSS ut hvor mange celler som har expected count less than 5. Dette antallet skal være 0. Dersom SPSS viser at antallet er større enn 0, må vi bruke en såkalt Fisher test. Den kommer vi tilbake til i kapittel 11.4.2.

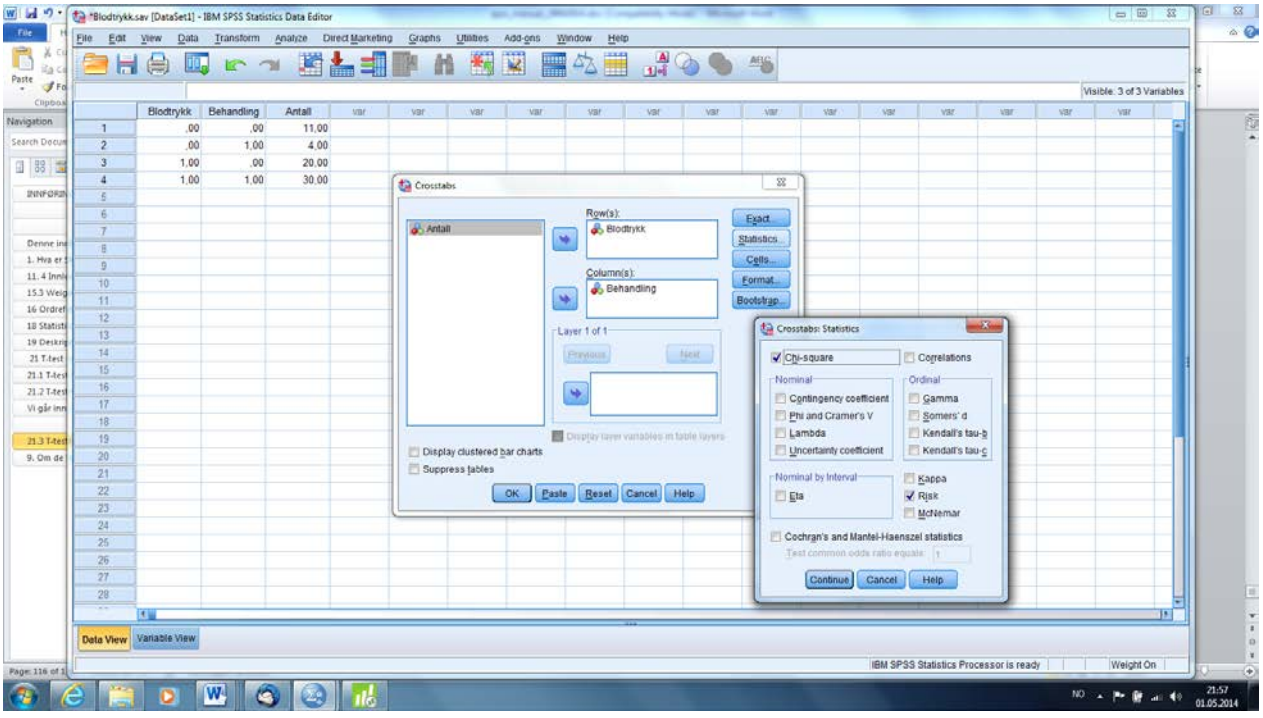
Vi skal i neste kapittel se hvordan SPSS regner ut RR, OR og p-verdien for kji-kvadratfordelingen. Dessverre regner ikke SPSS ut RD, men den er heldigvis lett å regne ut for hånd.

11.4.1 Eksempel: blodtrykk.sav

Vi fortsetter leser inn datafilen **blodtrykk.sav**. Husk å vekte opp via Data/Weight cases før vi analyserer på filen. Vi går da inn *Analyze/Descriptive Statistics/Crosstabs*. Siden BLODTRYKK er den avhengige variabelen trekker vi den over i *Row(s)* og BEHANDLING over i *Column(s)*. Så går vi inn i *Cells*. Der klikker vi på *Column* under *Percentages*. Dette gjør vi for å få prosentuert tabellen etter kolonnene, siden forklaringsvariabelen nettopp ligger i kolonnene. Da ser dialogboksen vår slik ut:



Vi klikker på *Continue*. Men vi er ennå ikke ferdig. Vi går så inn i *Statistics*. Her klikker vi av på *Chi-square* øverst til venstre (som gjør at vi får ut kji-kvadrattesten), og på *Risk* nede til høyre. Da ser dialogboksen slik ut:



Da er vi klare til å klikke på *Continue* og *OK*. Da får vi følgende resultat:

Blodtrykk * Behandling Crosstabulation

			Behandling		
			Placebo	Aspirin	Total
Blodtrykk	Høyt blodtrykk	Count	11	4	15
		% within Behandling	35,5%	11,8%	23,1%
	Normalt blodtrykk	Count	20	30	50
		% within Behandling	64,5%	88,2%	76,9%
Total		Count	31	34	65
		% within Behandling	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5,139[a]	1	,023		
Continuity Correction[b]	3,890	1	,049		
Likelihood Ratio	5,272	1	,022		
Fisher's Exact Test				,038	,024
Linear-by-Linear Association	5,060	1	,024		
N of Valid Cases	65				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 7,15.

b. Computed only for a 2x2 table

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Blodtrykk (Høyt blodtrykk / Normalt blodtrykk)	**4,125**	**1,151**	**14,786**
For cohort Behandling = Placebo	1,833	1,161	2,894
For cohort Behandling = Aspirin	,444	,186	1,060
N of Valid Cases	65		

Resultatene over viser først tabellen som er prosentuert etter kolonnene. Det betyr at vi kan lese ut andelene med normalt blodtrykk etter behandling direkte. Det er disse tallene vi bruker til å beregne effektmålene våre.

Deretter følger en tabell med tester. I denne analysen skal vi bruke Pearsons kji-kvadrattest, som står i første linje i tabellen. Teststørrelsen har en verdi på 5.14 og p-verdien for testen (altså sannsynligheten for å få en så høy og høyere verdi på teststørrelsen) er p = 0.023. Under tabellen ser vi at vi ingen celler som har færre enn 5 observasjoner. Det er den opplysningen som gir oss lov til å bruke Pearsons kji-kvadrat test. Siden p-verdien er < 0.05, konkluderer vi med at det er statistisk signifikant forskjell i effekt på placebo og behandling med aspirin.

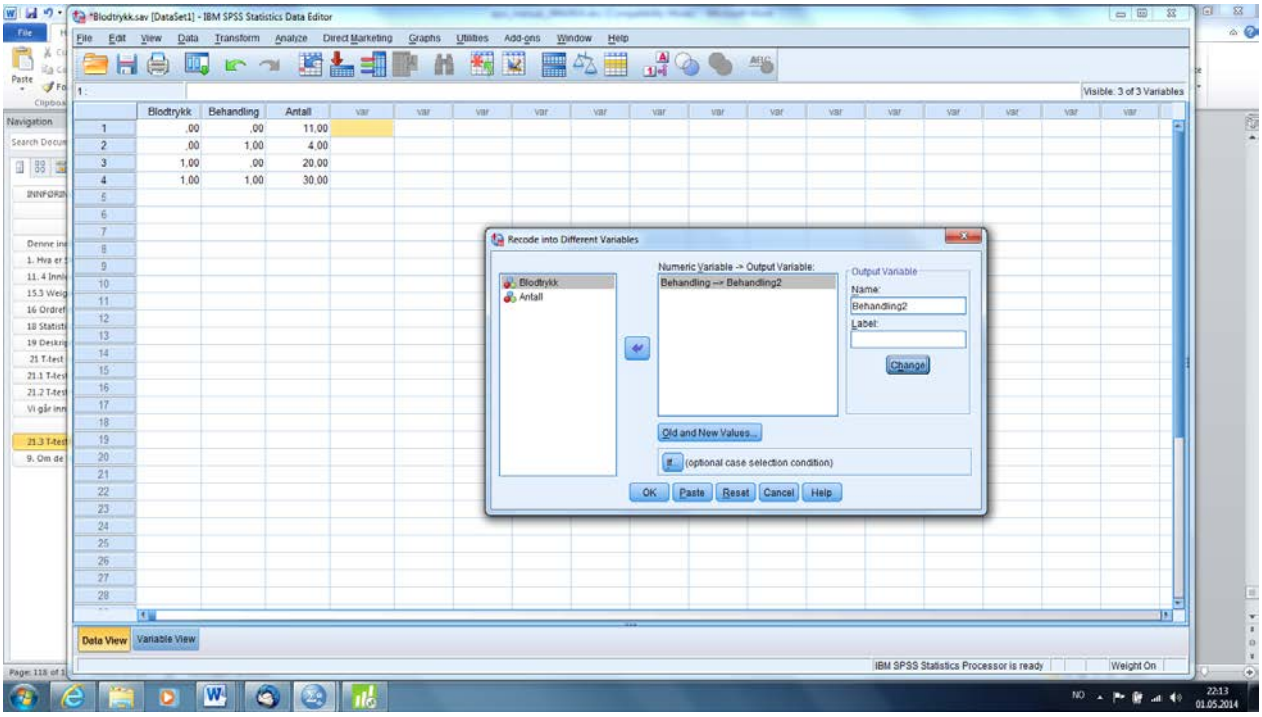
Til slutt følger en tabell med odds ratio for sammenhengen. Vi finner odds ratioen i øverste linje, som her er markert i **uthevet skrift**. Vi finner at OR = 4.13. Dette betyr at oddsen er 4.1 ganger så stor for å få et normalt blodtrykk etter behandling med aspirin, som med placebo. Sagt på en annen måte: Oddsen for normalt blodtrykk ved aspirinbehandling øker med 313%.

Vi finner et konfidensintervall for OR på (1.15, 14.79). Siden null-verdien (som er 1 for OR) ligger utenfor konfidensintervallet, vet vi også at p-verdien for testen av sammenheng er nødt til å være < 0.05. Det stemmer!

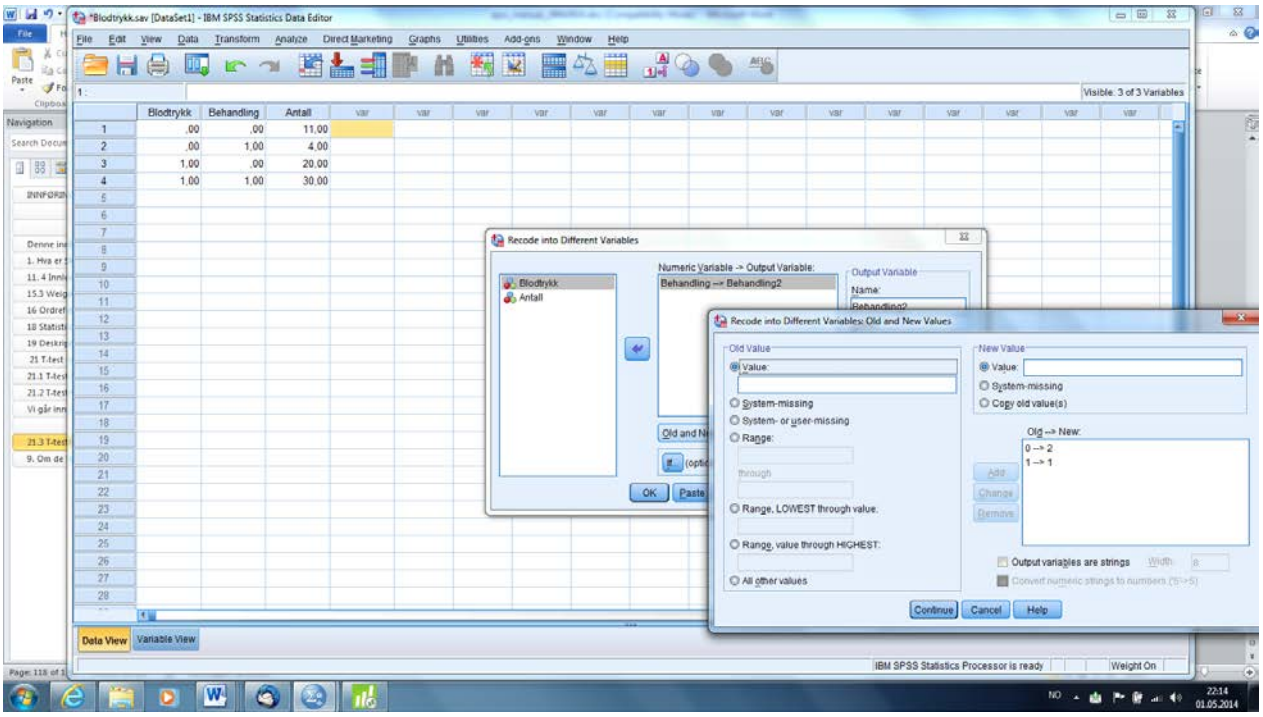
For å regne ut RR må vi dessverre gjøre to «grep». Disse er ulogiske, men vi må bare lære oss dem. Grunnen er at SPSS ikke direkte gir utskriften av RR, men vi må foreta en omkoding og en ny presentasjon av tabellen.

Først må vi altså omkode forklaringsvariabelen. I dette eksempelet må vi omkode BEHANDLING. Vi skal omkode slik at verdien 0 omkodes til 2, mens 1 forblir 1. Vi lager oss en ny variabel som vi kaller BEHANDLING2.

Dette gjør vi ved å gå inn i *Transform/Recode into Different variable*. Vi trekker BEHANDLING inn i vinduet i midten og skriver BEHANDLING2 som *Output Variable.* Da klikker på *Change.* Da ser dialogboksen slik ut:



Så klikker vi på *Old* and *New Values*, og skriver at 0 skal bli til 2 og 1 skal forbli 1. Da ser dialogboksen slik ut:



Vi klikker på *Continue* og *OK*. Da blir BEHANDLING2 lagt til datafilen vår.

Det neste «grepet» vi må gjøre er å lage en ny tabellanalyse. Men denne gangen må vi legge BEHANDLING2 i linjene og BLODTRYKK i kolonnene, til tross for at BEHANDLING er forklaringsvariabelen vår. Vi passer på at *Chi-square* og *Risk* fortsatt er krysset av under *Statistics*.

I den utskriften vi da får, er det bare tabellen for Risk Estimate vi er interessert i. Den ser lik ut:

Risk Estimate

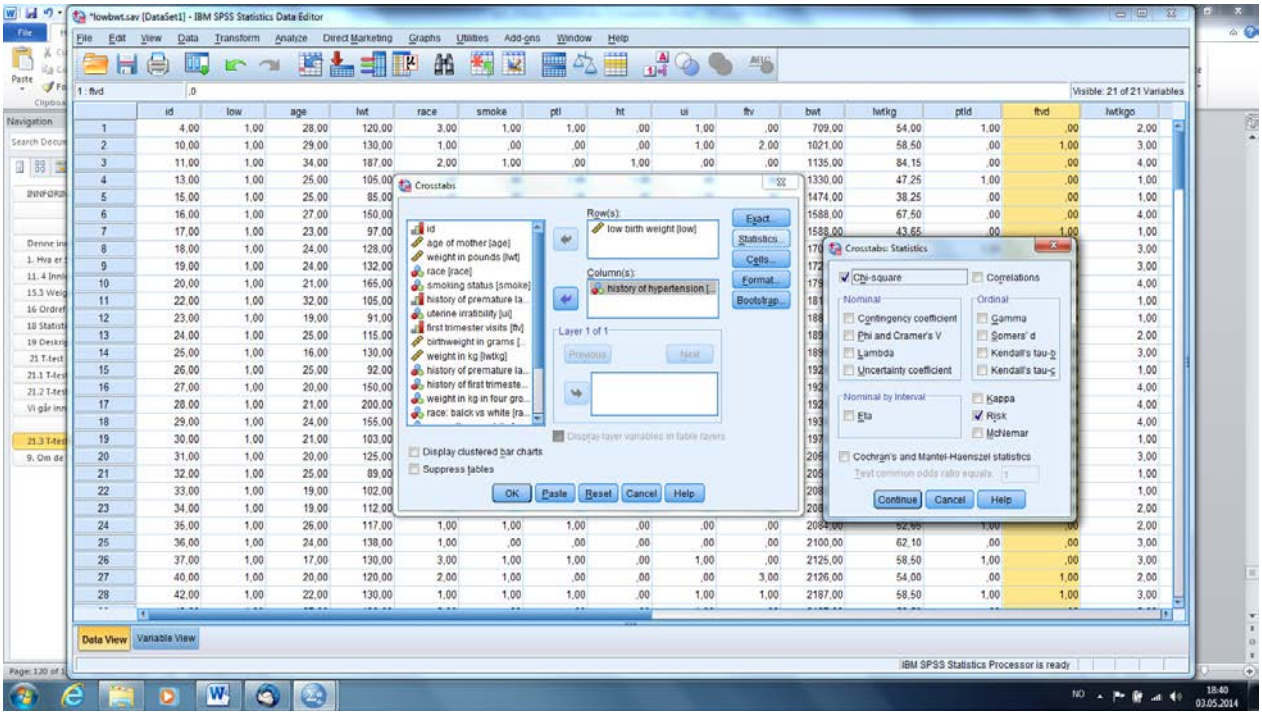
	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Behadling2 (1,00 / 2,00)	,242	,068	,869
For cohort Blodtrykk = Høyt blodtrykk	,332	,118	,934
For cohort Blodtrykk = Normalt blodtrykk	**1,368**	**1,025**	**1,825**
N of Valid Cases	65		

I denne tabellen er det bare den tredje linjen, som vi her har uthevet, som er av interesse. Vi ser at RR = 1.37 som er det vi fikk ved å regne for hånd over. Men det som er viktigere er at vi får beregnet konfidensintervallet for RR. Det er nemlig like (1.03, 1.83). Som for OR er null-verdien for RR også lik 1. Vi ser at konfidensintervallet for RR ligger over 1, som igjen passer med at p-verdien for kji-kvadrat testen er mindre enn 0.05.

11.4.2 Eksempel: lowbwt.sav

Vi går tilbake til datafilen **lowbwt.sav**. Vi starter med å gjøre en analyse av sammenhengen mellom lav fødselsvekt på barnet (LOW) og om mor er hypertensive (HT). Vi husker at LOW er en kategorisk variabel som angir om fødselsevekten er over 2500 gram (LOW = 0) eller under 2500 gram (LOW = 0) og HT er en kategorisk variabel med to kategorier (HT = 0 for de normotensive og HT = 1 for de hypertensive).

Vi vet at LOW er den avhengige variabelen. Da lager vi en tabell mellom LOW og HT ved å gå inn *Analyze/Descriptive Statistics/Crosstabs*. LOW er den avhengige variabelen, så vi trekker vi den over i *Row(s).* HT er forklaringsvariabelel, og den trekker vi over i *Column(s).* Så går vi inn i *Cells*. Der klikker vi på *Column* under *Percentages*, siden vi vil ha prosentuert tabellen etter forklaringsvariabelen. Vi klikker på *Continue*. Vi går så inn i *Statistics*. Her klikker vi av på *Chi-square* øverst til venstre, og på *Risk* nede til høyre. Da ser dialogboksen slik ut:



Vi klikker på *Continue* og *OK* og får følgende:

low birth weight * history of hypertension Crosstabulation

			history of hypertension		
			no	yes	Total
low birth weight	bwt > 2500g	Count	125	5	130
		% within history of hypertension	70,6%	41,7%	68,8%
	bwt < 2500g	Count	52	7	59
		% within history of hypertension	29,4%	58,3%	31,2%
Total		Count	177	12	189
		% within history of hypertension	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	4,388[a]	1	,036		
Continuity Correction[b]	3,143	1	,076		
Likelihood Ratio	4,022	1	,045		
Fisher's Exact Test				**,052**	**,042**
Linear-by-Linear Association	4,365	1	,037		
N of Valid Cases	189				

a. 1 cells (25,0%) have expected count less than 5. The minimum expected count is 3,75.

b. Computed only for a 2x2 table

Risk Estimate

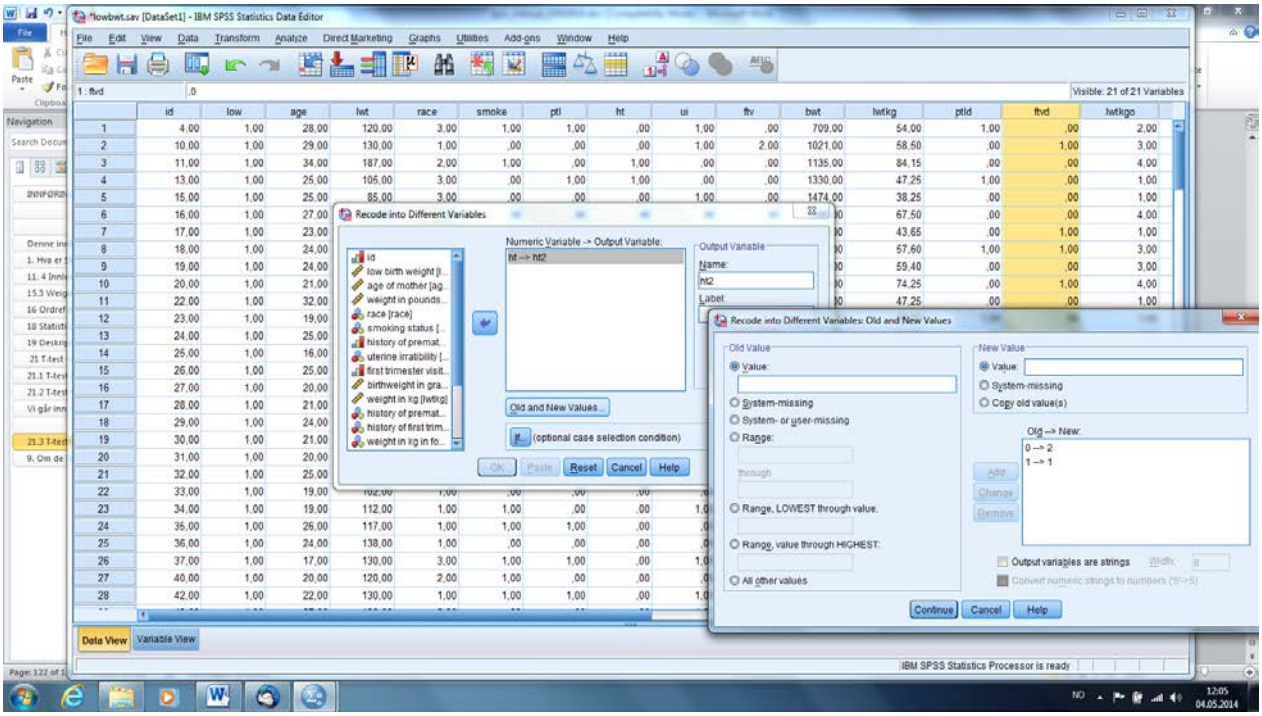
	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for low birth weight (bwt > 2500g / bwt < 2500g)	**3,365**	**1,021**	**11,088**
For cohort history of hypertension = no	1,091	,987	1,205
For cohort history of hypertension = yes	,324	,107	,979
N of Valid Cases	189		

Vi ser at det blant de hypertensive er det 58.3% av mødrene som føder et barn under 2500 gram, mens blant de normotensive er det 29.4%. Det kan se ut som om det er stor forskjell i sannsynlighetene. Spørsmålet er om det er statistisk signifikant forskjell! Det ser vi i neste tabell.

Men under tabellen med Chi-Square Tests er vi at det står at 1 celle har forventet at antall mindre enn 5. Det betyr at vi ikke bør bruke kji-kvadrattesten i denne tabellen. Da bør vi heller bruker Fishers eksakte test. Ved å lese av i kolonnen for tosidig p-verdi finner vi at p =

0.052. Merk at denne er en del høyre enn Pearsons p-verdi. Det vil skje ganske ofte at p-verdien er høyere når vi bruke Fishers eksakte p-verdi for tabeller med få observasjoner i cellene. Altså må vi konkludere med at det ikke er en statistisk signifikant sammenheng mellom lav fødselsvekt og hypertensjon. Mer her må vi tolke resultatet i lys av at vi har lite data. Hadde vi hadde flere kvinne som var hypertensive, hadde vi nok funnet en statistisk signifikant sammenheng.

La oss nå finne RR i denne tabellen. Da må vi gjøre våre to grep: Først må vi omkode forklaringsvariabelen HT, slik at 0 blir omkodet til 2. Det gjør vi via *Data/Transform Into Different Variable*. Vi lager oss HT2, og der omkoder vi 0 til 2 og lar 2 fortsatt være 2. Da ser dialogboksen slike ut:



Vi klikker på *Continue* og *OK*. Deretter må vi lage en tabell med HT2 i linjene og LOW i konlonnene. Husk å gå via *Statistics* og merk av på *Risk*. Når vi har gjort denne analysen fpr vi følgende svar. Vi gjengir bare resultatet for Risk Estimate:

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for ht2 (1,00 / 2,00)	,297	,090	,979
For cohort low birth weight = bwt > 2500g	,590	,300	1,160
For cohort low birth weight = bwt < 2500g	**1,986**	**1,169**	**3,373**
N of Valid Cases	189		

Vi leser av i tredje linje og ser at RR = 1.99. Nå vi går tilbake til tabellen, ser vi sannsynligheten for å få et lite barn når mor er hypertensiv er 58.3%, mens den er 29.4 når mor er normotensiv. Med hoderegning ser vi at RR er omtrent lik 2, som stemmer med vi ser over. Men det som er viktigere at SPSS gir oss konfidensintervallet for RR, som er ganske vanskelig å regne ut for hånd.

Merk også at konfidensintervallet ligger over 1. Dette skulle bety at p-verdien også skulle være mindre enn 0.05. Men det gjelder bare når vi bruker Pearsons kji-kvadrattest. I dette tilelle, med Fishers eksakte test, vil ikke dette nødvendigvis være tilfelle.

Siden vi her finner at det er én celle med forventet antall mindre enn 5, kan det være naturlig å finne ut hvilken celle dette er. Det kan vi få SPSS til å finne ut. Vi går inn *Analyze/Descriptive Statistics/Crosstabs*, med LOW i *Rows* og HT i *Columns*. Vi går inn i *Cells*. Der klikker vi som vanlig av på *Columns*, men nå klikker vi også av på *Expected*. Ved å gå ut av dialogboksene med *Continue* og *OK*, får vi følgende resultat for tabellen:

low birth weight * history of hypertension Crosstabulation

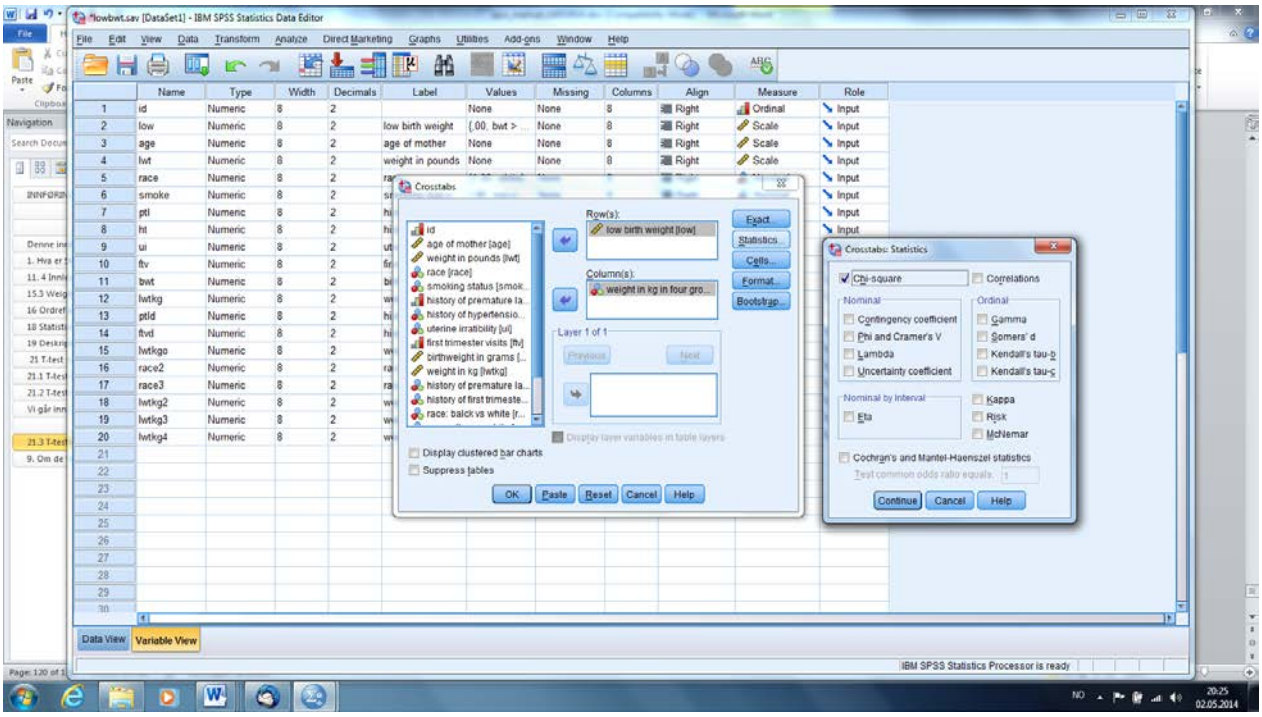
			history of hypertension		
			no	yes	Total
low birth weight	bwt > 2500g	Count	125	5	130
		Expected Count	121,7	8,3	130,0
		% within history of hypertension	70,6%	41,7%	68,8%
	bwt < 2500g	Count	52	7	59
		Expected Count	55,3	3,7	59,0
		% within history of hypertension	29,4%	58,3%	31,2%
Total		Count	177	12	189
		Expected Count	177,0	12,0	189,0
		% within history of hypertension	100,0%	100,0%	100,0%

Vi ser at det er cellen nederst til høyre som har færre enn 5. Vi kunne kanskje ha trodd at det var cellen over, med 5, som hadde færre enn 5. Men det er slik at det er produktet av antallet i marginalene i linjenen og kolonnene som bestemmer forventet antall. Siden marginalen i nederste linje bare er 59, og vesentlig mindre enn linjen over, som er 130, er det nederste celle om har færrest forventet antall.

Så skal vi vise analysen av en 2x4 tabell. Vi lager da en tabell mellom LOW og LWTKGO. Vi husker at LOW er en kategorisk variabel som angir om fødselsevekten er over 2500 gram (LOW = 0) eller under 2500 gram (LOW = 0). LWTKGO er en kategorisk variabel som angir vekten til mor i fire kategorier, nemlig om vekten ligger i første kvartil, annen kvartil, tredje kvartil eller i fjerde kvartil. Vi lager da en krysstabell mellom LOW og LWTKGO.

Vi går da inn *Analyze/Descriptive Statistics/Crosstabs*. LOW er den avhengige variabelen, så vi trekker vi den over i *Row(s).* LWTKGO er forklaringsvariabelel og den trekker vi over i *Column(s).* Så går vi inn i *Cells*. Der klikker vi på *Column* under *Percentages*, siden vi vil ha prosentuert tabellen etter forklaringsvariabelen. Vi klikker på *Continue*. Vi går så inn i

Statistics. Her klikker vi av på *Chi-square* øverst til venstre. Men siden vi nå lager en 2x4 tabell kan ikke få beregnet *Risk*. Vi klikker derfor ikke på den. Da ser dialogboksen slik ut:



Da er vi klare til å klikke på *Continue* og *OK*. Da får vi følgende resultat:

low birth weight * weight in kg in four groups Crosstabulation

			weight in kg in four groups				
			First quartile	Second quartile	Third quartile	Fourth quartile	Total
low birth weight	bwt > 2500g	Count	28	33	34	35	130
		% within weight in kg in four groups	52,8%	76,7%	73,9%	74,5%	68,8%
	bwt < 2500g	Count	25	10	12	12	59
		% within weight in kg in four groups	47,2%	23,3%	26,1%	25,5%	31,2%
Total		Count	53	43	46	47	189
		% within weight in kg in four groups	100,0%	100,0%	100,0%	100,0%	100,0%

Chi-Square Tests

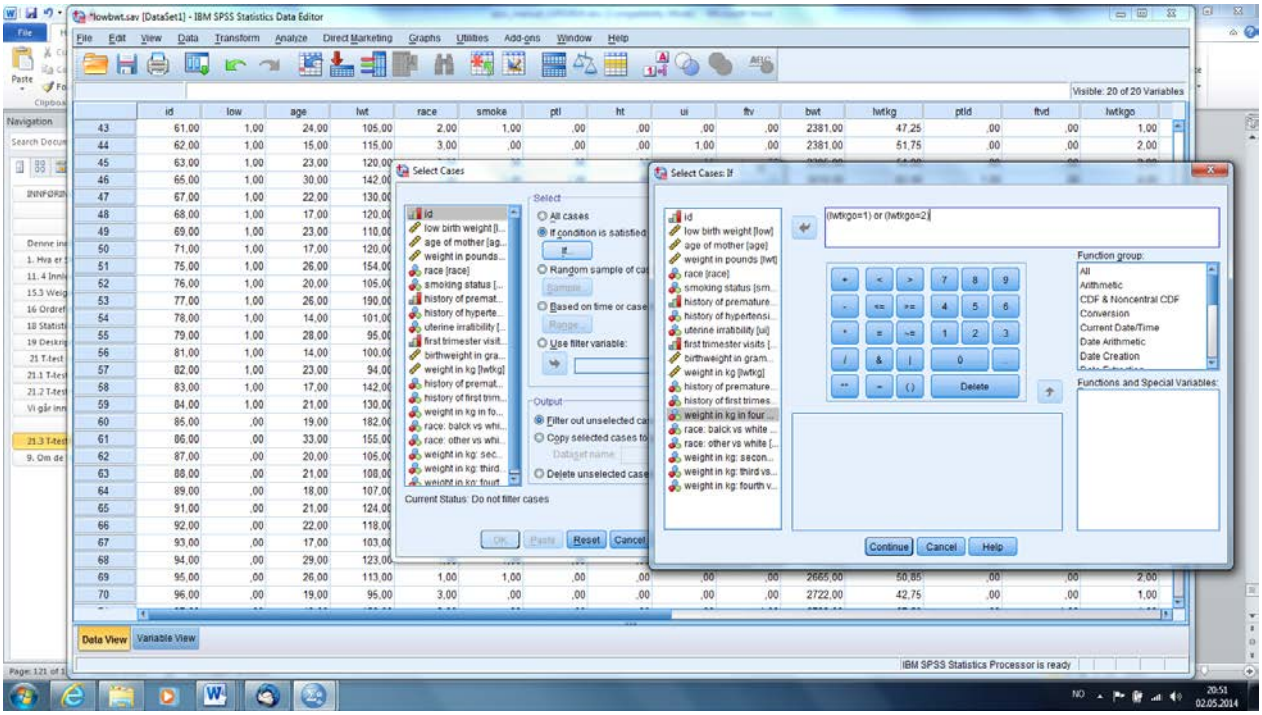
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	**8,822[a]**	**3**	**,032**
Likelihood Ratio	8,520	3	,036
Linear-by-Linear Association	4,898	1	,027
N of Valid Cases	189		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 13,42.

Vi ser at sannsynligheten for å få et lite barn (altså et barn under 2500 gram) er 47.2% dersom mors vekt er i første kvartil. Men når mors vekt er i annen, tredje eller fjerde kvartil, er sannsynligheten omtrent den samme og ca. 25%.

I tabellen under ser vi at Pearsons kji-kvadrattest er 8.82 og p-verdien er p = 0.032. Vi har altså en statistisk signifikant sammenheng mellom mors vekt i kvartiler og barnets fødselsvekt. Merk at det under tabellen står at det er 0 celler med forventet antall mindre enn 5. Vi kan altså bruke Pearsons test.

Som nevnt tidligere kan vi ikke få beregnet OR eller RR for en 2x4 tabell. Vi skal nå beregne OR i 3 2x2 tabeller. Grunnen til at vi får 3 tabeller er at vi holder første kvartil fast og sammenligner de 3 andre kvartilene opp mot første kvartil La oss starte med å sammenligne første og annen kvartil. Da må vi selektere ut kvinner med vekt i første og annen kvartil. Det gjør vi ved å gå inn i *Data/Select Cases*. Der merker vi at på *If condition is satisfied* og klikker så på *If*. Da åpner det seg en ny dialogboks. Nå må vi spesifisere seleksjonen vår. Her skal vi nå plukke ut dem som har verdier LWTKGO = 1 eller verdier LWTKG = 2. Vi skriver da inn (lwtkg0=1) or (lwtkgo=2) i vinduet. Legg merke til at vi må bruke parenteser mellom de to verdiene. Da ser dialogboksen vår slik ut:



Da kan vi klikke på *Continue* og *OK*. Da er seleksjonen gjort.

Nå må vi lage tabellen. Vi går da tilbake til *Analyze/Descriptive Statistics/Crosstabs* med LOW som den avhengige variabelen, og LWTKGO som forklaringsvariabelel. Så går vi inn i *Statistics*. Vi passer på at *Chi-square* er klikket av øverst til venstre. Men siden vi nå lager en 2x2 tabell kan vi nå få beregnet *Risk*. Vi klikker derfor på den. Ved å gå via *Continue* og *OK*, får vi følgende resultat:

low birth weight * weight in kg in four groups Crosstabulation

			weight in kg in four groups		
			First quartile	Second quartile	Total
low birth weight	bwt > 2500g	Count	28	33	61
		% within weight in kg in four groups	52,8%	76,7%	63,5%
	bwt < 2500g	Count	25	10	35
		% within weight in kg in four groups	47,2%	23,3%	36,5%
Total		Count	53	43	96
		% within weight in kg in four groups	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	5,860[a]	1	,015		
Continuity Correction[b]	4,874	1	,027		
Likelihood Ratio	6,008	1	,014		
Fisher's Exact Test				,019	,013
Linear-by-Linear Association	5,799	1	,016		
N of Valid Cases	96				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 15,68.

b. Computed only for a 2x2 table

Risk Estimate

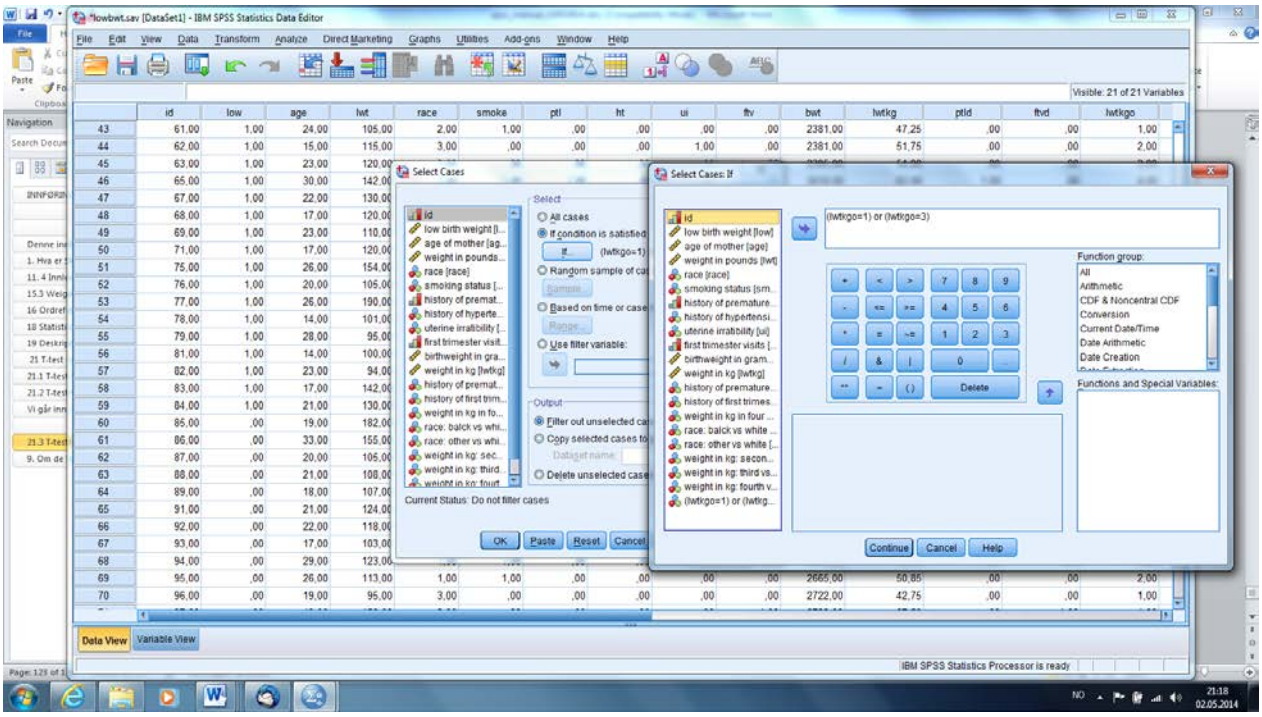
	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for low birth weight (bwt > 2500g / bwt < 2500g)	**,339**	**,139**	**,826**
For cohort weight in kg in four groups = First quartile	,643	,456	,906
For cohort weight in kg in four groups = Second quartile	1,893	1,068	3,357
N of Valid Cases	96		

For det første ser vi at vi har fått riktig gruppe selektert, med kvinner med fødselsvekt i første eller annen kvartil. Deretter ser vi at det er en statistisk signifikant sammenheng mellom lav fødselsvekt og om mors vekt er i første og annen kvartil.

Den siste tabellen viser det viktigste resultatet. Den viser at oddsen for å føde et lite barn reduseres til 0.339 når mors vekt går fra første til annen kvartil. Det er kanskje enklere å forklare dette ved at oddsen reduseres med 66% for å få et lite barn når mors vekt går fra første til annen kvartil.

La oss nå gjøre det sammen for neste seleksjon. Da beholder kvinner med vekt i første kvartil som referanse, men sammenlinger nå dem med kvinner med vekt i tredje kvartil. Merk at vi først må velge *All cases* og klikke på *OK*, før vi går løs på neste seleksjon.

Da går vi gjennom seleksjonen som over, men endrer fra LWTKG0 = 2 til LWTKG = 3. Merk at vi først må velge All cases og klikke på *OK*, får vi går løs på neste seleksjon. Da vil dialogboksen for seleksjonen bli som under:



Vi klikker på *Continue* og *OK*. Så lager vi tabellen slik vi også gjorde over. Vi behøver ikke gjøre noen endringer i *Analyze/Descriptive Statistics/Crosstabs*. Når vi gjør denne analysen får vi følgende resultat. Vi gjengir her bare resultatet for OR:

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for low birth weight (bwt > 2500g / bwt < 2500g)	**,395**	**,169**	**,926**
For cohort weight in kg in four groups = First quartile	,668	,469	,952
For cohort weight in kg in four groups = Third quartile	1,691	1,008	2,836
N of Valid Cases	99		

Vi finner at p-verdien ved å bruke Pearsons kji-kvadrattest er p = 0.031. Altså er det statistisk signifkant forskjell i sannsynlighetene for å få et lite barn når mor er i tredje kvartil i forhold til i første kvartil.

I tabellen over ser vi at OR = 0.395. Altså reduseres oddsen for å få et lite barn med 60% når mors vekt er i tredje kvartil i forhold til i første kvartil.

Hvis vi gjør det samme også for sammenligningen mellom mors vekt i første kvartil og i fjerde kvartil, finner vi følgende resultat for OR:

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for low birth weight (bwt > 2500g / bwt < 2500g)	**,384**	**,164**	**,897**
For cohort weight in kg in four groups = First quartile	,658	,461	,938
For cohort weight in kg in four groups = Fourth quartile	1,713	1,024	2,866
N of Valid Cases	100		

Her ser vi at OR = 0.384, altså en reduksjon av oddsen for å føde et lite barn på 62% når kvinnen er i fjerde kvartil mht. vekt i forhold til i første kvartil.

Konklusjonen her er at mors vekt har en statistisk signifikant betydning for om barnet får en fødselsvekt under 2500 gram. Men vi ser at oddsen for å føde et lite barn synker med ca. 60%, uansett om kvinnen er i annen, tredje eller fjerde kvartil mht. vekt.

Vi starter med å gjøre en analyse av sammenhengen mellom lav fødselsvekt på barnet (LOW) og om mor er hypertensive (HT).

Til slutt skal vi lage en 4x3 tabell. Vi skal se på sammenhengen mellom LWTKGO og RACE. Vi går da inn i *Analyze/Descriptive Statistics/Crosstabs* og trekker LWTKGO over i *Row(s*)

og RACE over i *Columns*. Vi går inn i *Cells* og merker av på at vi skal ha tabellen prosentuert eller kolonne (*Columns*). Til slutt går vi inn i *Statistics*, og merker av på *Pearson's Chi-Square*. Vi kan ikke merke av på *Risk*, siden dette ikke er en 2x2 tabell.

Da får vi følgende resultat:

weight in kg in four groups * race Crosstabulation

			race			
			white	black	other	Total
weight in kg in four groups	First quartile	Count	23	3	27	53
		% within race	24,0%	11,5%	40,3%	28,0%
	Second quartile	Count	19	7	17	43
		% within race	19,8%	26,9%	25,4%	22,8%
	Third quartile	Count	28	5	13	46
		% within race	29,2%	19,2%	19,4%	24,3%
	Fourth quartile	Count	26	11	10	47
		% within race	27,1%	42,3%	14,9%	24,9%
Total		Count	96	26	67	189
		% within race	100,0%	100,0%	100,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	15,358[a]	6	,018
Likelihood Ratio	15,586	6	,016
Linear-by-Linear Association	6,807	1	,009
N of Valid Cases	189		

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 5,92.

Vi ser at selve tabellen at det er forskjell i fordelingen for hvite, svarte og andre. For de hvite er det omtrent like mange i hver kvarti. For de svarte er det flest i den øverste kvartilen, dvs. med tunge mødre, mens det for andre er flest i den laveste kvartilen.

Vi ser at disse forskjellene er statistisk signifikante, siden p = 0.018. Vi kan bruke Pearsons kji-kvadrattest, siden forventet antall er større en 5 i alle cellene.

11.5 Korrelasjon

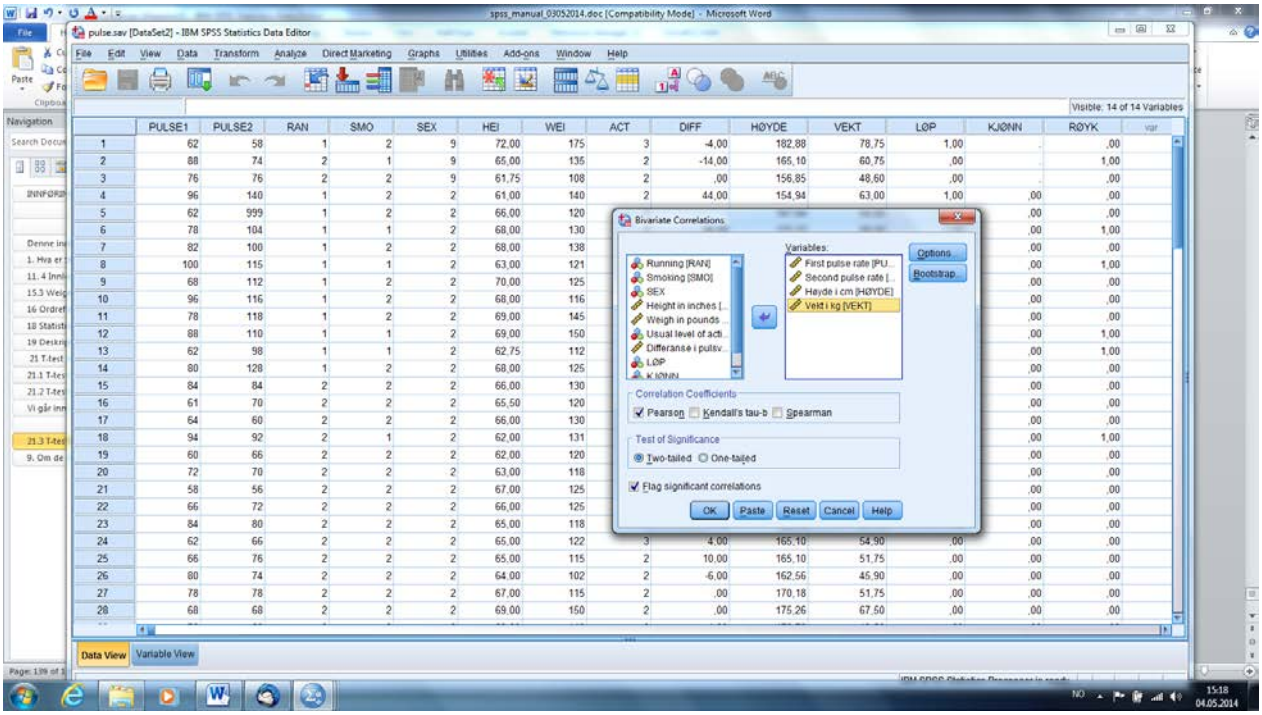
Korrelasjon er et symmetrisk mål for sammenheng mellom to kontinuerlige variabler. Den utrykker lineær sammenheng mellom variablene. Vi skal her se på Pearsons korrelasjonskoeffisient, som vi skriver som *r*. Men korrelasjonskoeffisienten spiller også en viktig rolle i regresjonsanalyse. For det første er det enkel algebraisk sammenheng mellom regresjonskoeffisienten og korrelasjonskoeffisienten, slik at når korrelasjonskoeffisienten er

null, er også regresjonskoeffisienten lik null. Eller sagt på en annen måte: Hvis det ikke er korrelasjon er det heller ikke regresjon.

I en regresjonsanalyse ser vi sammenhengen mellom en avhengig variabel og en forklaringsvariabel. Fortolkningen er derfor ikke lenger symmetrisk som i korrelasjon, men asymmetrisk, siden vi snakker om effekten av forklaringsvariabelen på den avhengige variabelen. Spørsmålet er da hvor mye forklaringsvariabelen forklarer av variasjonen i den avhengige variabelen. Korrelasjonskoeffisienten gir svar på det. Det er nemlig slik at kvadratet av korrelasjonskoeffisienten, r^2, har en fortolkning som forklart varians i regresjonsanalysen. Vi har at r^2 er tall som ligger mellom 0 og 1. Jo nærmere tallet er 1, jo mer av variasjonen i den avhengige variabelen blir forklart av forklaringsvariabelen. Vi kommer tilbake til det også i kapittelet om regresjon i kapittel 12.2.

11.5.1 Eksempel: pulse.sav

Vi vil se på korrelasjonen (Pearsons) mellom PULSE1, PULSE2 og HØYDE og VEKT. Vi går til *Analyze/Correlate/Bivariate* og legger de fire variablene over i *Variables* boksen. Da ser dialogboksen vår slik ut:



Vi klikker på *OK*.

Dette gir følgende resultatet i utskriftsvinduet:

Correlations

		First pulse rate	Second pulse rate	Høyde i cm	Vekt i kg
First pulse rate	Pearson Correlation	1	,643**	-,234*	-,223*
	Sig. (2-tailed)		,000	,027	,034
	N	90	88	90	90
Second pulse rate	Pearson Correlation	,643**	1	-,147	-,159
	Sig. (2-tailed)	,000		,167	,135
	N	88	90	90	90
Høyde i cm	Pearson Correlation	-,234*	-,147	1	,785**
	Sig. (2-tailed)	,027	,167		,000
	N	90	90	92	92
Vekt i kg	Pearson Correlation	-,223*	-,159	,785**	1
	Sig. (2-tailed)	,034	,135	,000	
	N	90	90	92	92

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

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

Dette kalles en korrelasjonsmatrise og viser korrelasjonen mellom alle de valgte variablene. Korrelasjonen mellom PULSE1 og PULSE ligger i første linje, annen kolonne. Vi ser at korrelasjonen er 0.643. Det er altså en forholdsvis høy korrelasjon mellom pulsverdien. Hvis går ned til tredje linje, fjerde kolonne, ser vi at korrelasjonen mellom høyde og vekt er 0.785. Det er (som forventet) en høy korrelasjon mellom høyde og vekt.

I andre linje oppgis antall observasjoner. Vi ser at det varierer fra korrelasjon til korrelasjon. Det skyldes at *Missing values* er tatt ut fra de enkelte analysene. Siden antall missing varierer fra variabel til variabel, vil også antall som inngår i beregningene av korrelasjonene, variere.

I siste linje får vi p-verdien testen for om for at korrelasjonen er statistisk forskjellig fra 0. Vi ser at for både PULSE1 og PULSE2 og for HØYDE og VEKT er disse to korrelasjonskoeffisientene signifikant forskjellige fra 0, begge med $p < 0.001$.

11.5.2 Eksempel: lowbwt.sav

Vi går tilbake til datafilen lowbwt.sav. Nå er vi interessert i korrelasjonen mellom BWT og LWTKG. Da går vi inn i *Analyze/Correlate/Bivariate* og legger de to variablene over i *Variables* boksen, og klikker på *OK*. Da får vi følgende resultat:

Correlations

		birthweight in grams	weight in kg
birthweight in grams	Pearson Correlation	1	,186*
	Sig. (2-tailed)		,010
	N	189	189
weight in kg	Pearson Correlation	,186*	1
	Sig. (2-tailed)	,010	
	N	189	189

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

Vi ser at korrelasjonskoeffisienten er 0.186. Vi ser også at p-verdien for null hypotesen om at korrelasjonen er lik 0, er p = 0.010. Altså er det en ikke så veldig korrelasjon mellom BWT og LWTKG, men den er klart statistisk signifikant.

Vi har nå at r^2 = 0.186x0.186 = 0.035. Hvis vi nå tenker på BWT som en avhengig variabel som skal forklares av LWTKG, har vi altså at bare 3.5% av variasjonen i BWT er forklart av LWTKG.

12 Multivariable statistiske metoder

Læringsmål

I forrige kapittel så vi på de univariable statistiske metodene. Dette er metoder der vi forklarer den avhengige variabelen med bare én forklaringsvariabel, I dette kapittelet skal vi se på hvordan vi kan utvide de univariable metodene til multivariable metoder, altså metoder med flere enn én forklaringsvariabel.

Igjen er det slik at målenivået på den avhengige variabelen og på forklaringsvariabelen avgjør hvilken analysemetode vi skal bruke. Dersom den avhengige variabelen er kontinuerlig og forklaringsvariablene er kategoriske, bruker vi variansanalyse.

Dersom den avhengige variabelen er kontinuerlig og forklaringsvariablene er enten kontinuerlige eller kategoriske, skal vi bruke lineær regresjonsanalyse. Til slutt skal vi se på overlevelsesanalyse. Her registrerer vi tiden til en begivenhet, som sykdom eller død. Analysen krever en spesiell behandling, siden vi ikke registrerer tiden til begivenheten vi studerer, for eksempel ved at studien avsluttes før personen vi studerer er blitt syk eller er død. Dette er en form for *Missing values*, som vi i denne sammenhengen kaller sensurering.

12.1 Variansanalyse – ANOVA (ANalysis Of VAriance)

Variansanalysen er både enveis og flerveis. Enveis variansanalyse er analyse med bare én forklaringsvariabel. Forklaringsvariabelen har da mer enn to kategorier. Enveis variansanalyse er da en direkte videreføring av t-tester for to uavhengige utvalg, siden dette er situasjonen

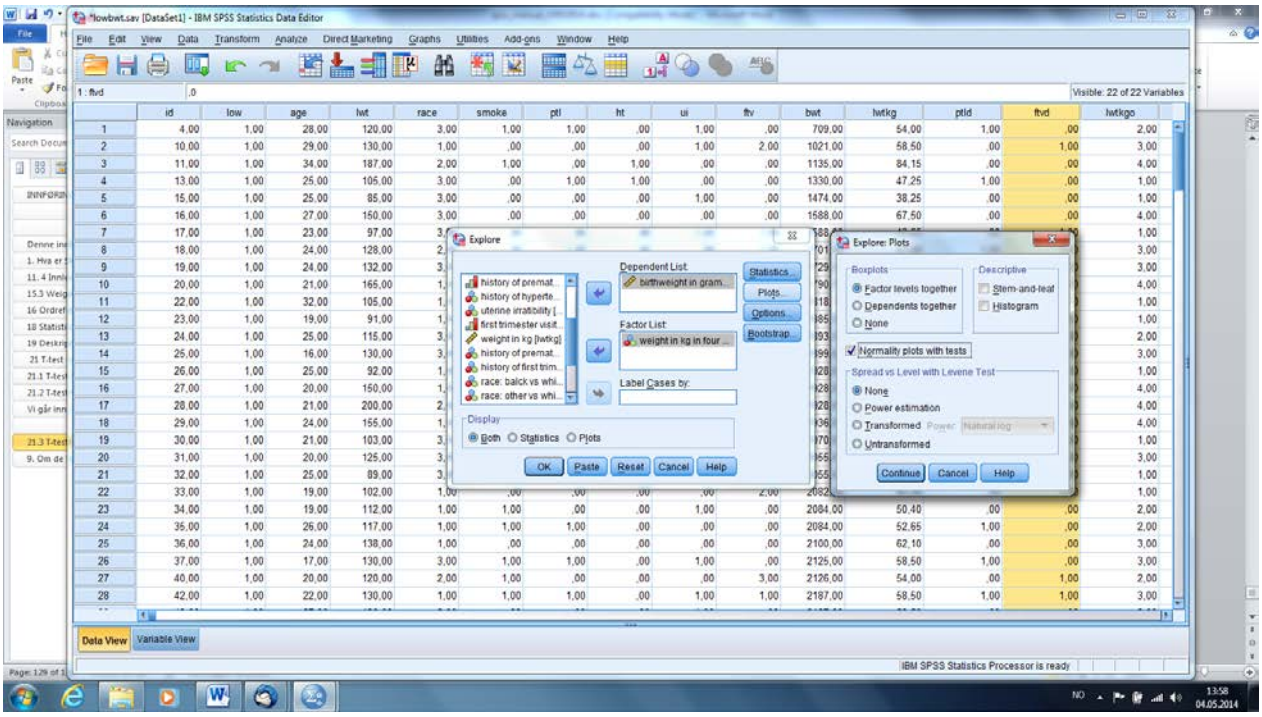
med én forklaringsvariabel som har to kategorier. Dersom vi har tre eller flere grupper som vi skal sammenligne gruppene, må vi bruke enveisvariansanalyse.

I all multivariabel analyse har vi behov for å innføre flere enn to forklaringsvariabler. Dette skyldes at det er flere forklaringsvariabler som påvirker den avhengige variabelen, og vi ønsker å studere effekten av den enkelte variabelen når vi kontrollerer for effekten av de andre forklaringsvariablene.

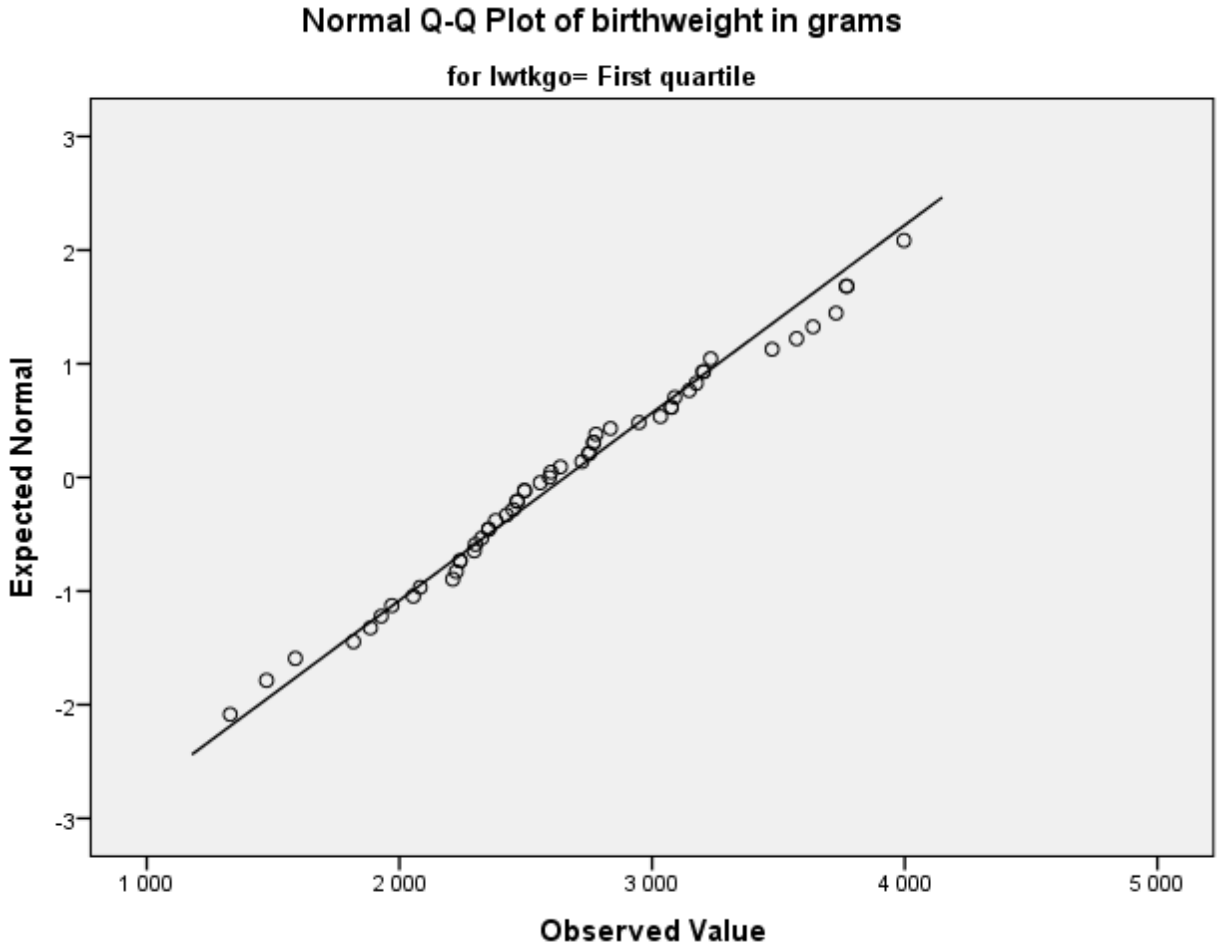
Vi skal nedenfor først se på enveis variansanalyse. Deretter skal vi se på flerveis variansanalyse.

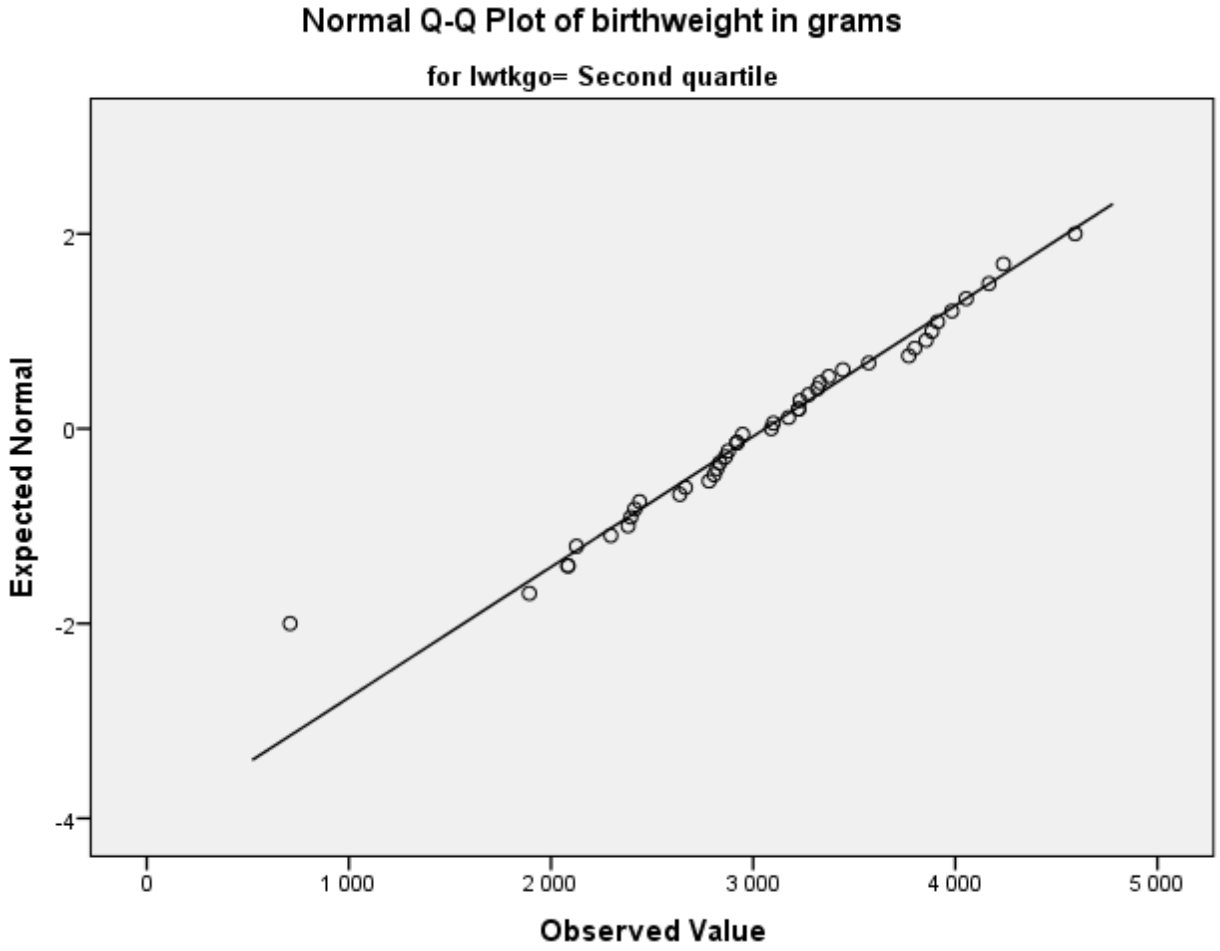
12.1.1 Enveis variansanalyse. Eksempel: lowbwt.sav

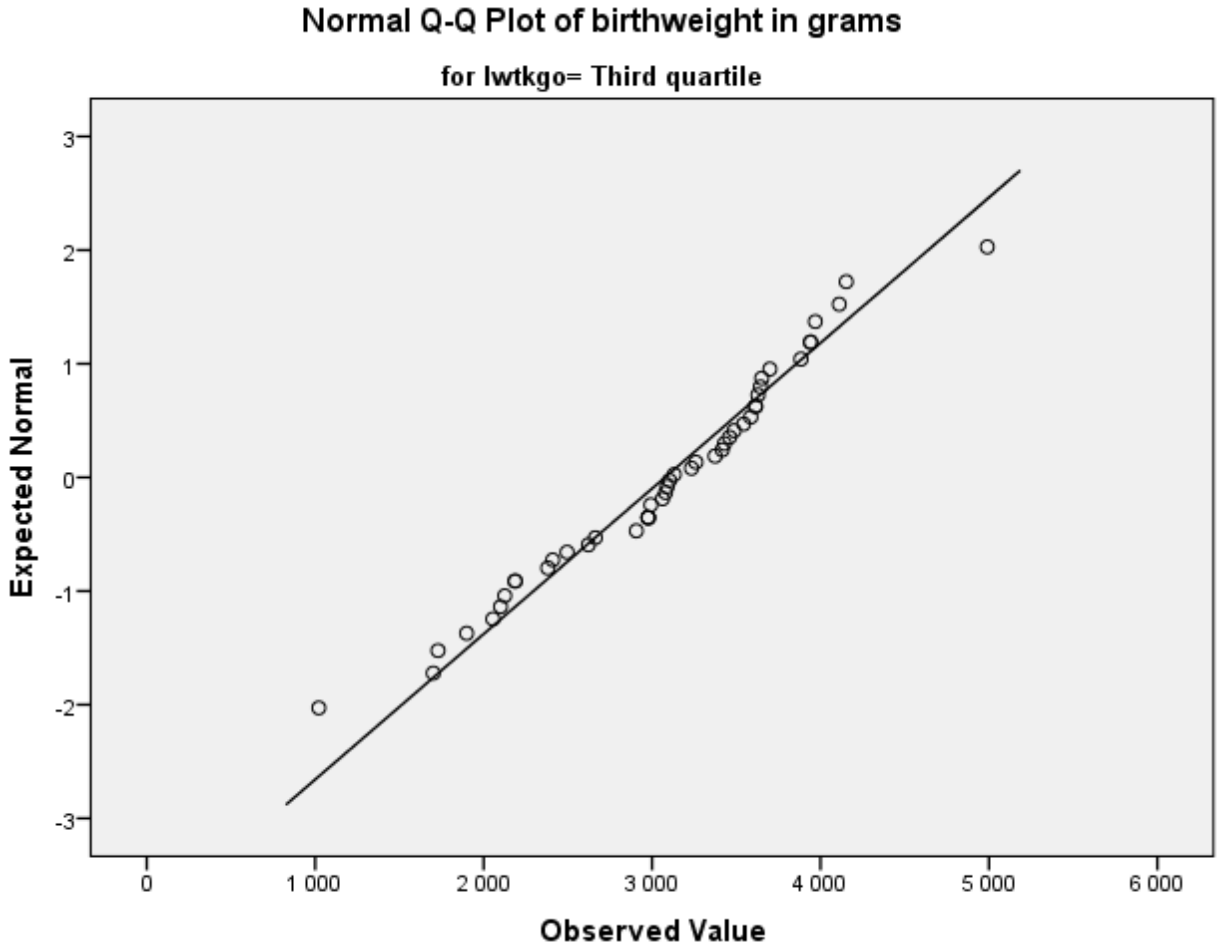
Vi skal nå sammenligne de de fire gruppene i LWTKGO med hensyn til fødselsvekt (BWT). Siden ANOVA baserer seg på antagelsen om normalfordeling i de gruppene vi skal sammenligne, må vi først lage normalfordelingsplott for BWT i de fire gruppene til LWTKGO. Dette gjør vi via *Analyze/Descriptive Statistics/Explore*. Vi trekker BWT over i *Dependent List* og LWTKGO over i *Factor List*. Vi går så inn i *Plots*, der vi klikker av på *Normality plots with tests*. Da ser dialogboksen vår slik ut:

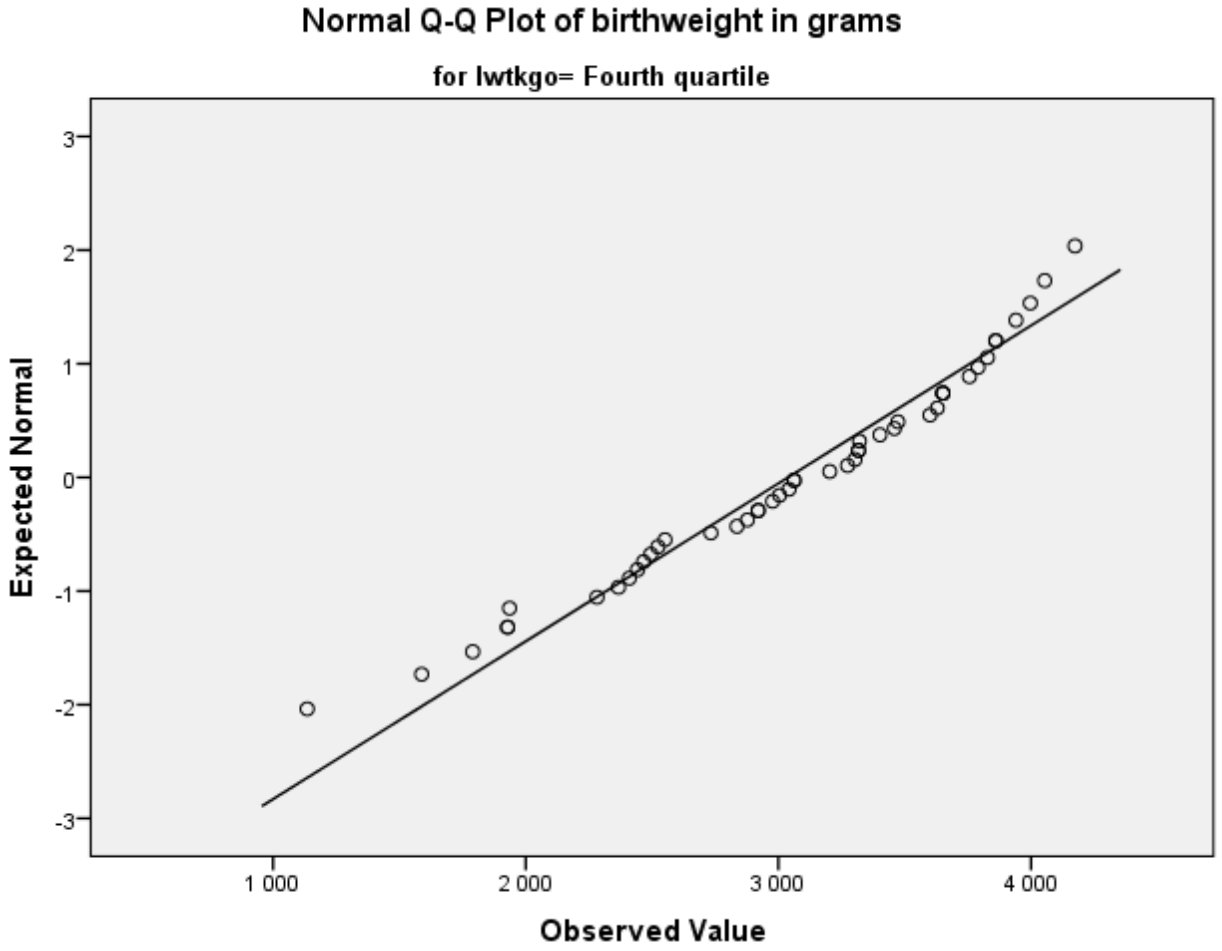


Ved å klikke på *Continue* og *OK*, får vi følgende resultat. Her gjengir vi bare selve normalfordelingsplottene.


Normal Q-Q Plot of birthweight in grams
for lwtkgo= First quartile
Expected Normal
3
2
1
0
-1
-2
-3
1 000
2 000
3 000
4 000
5 000
Observed Value

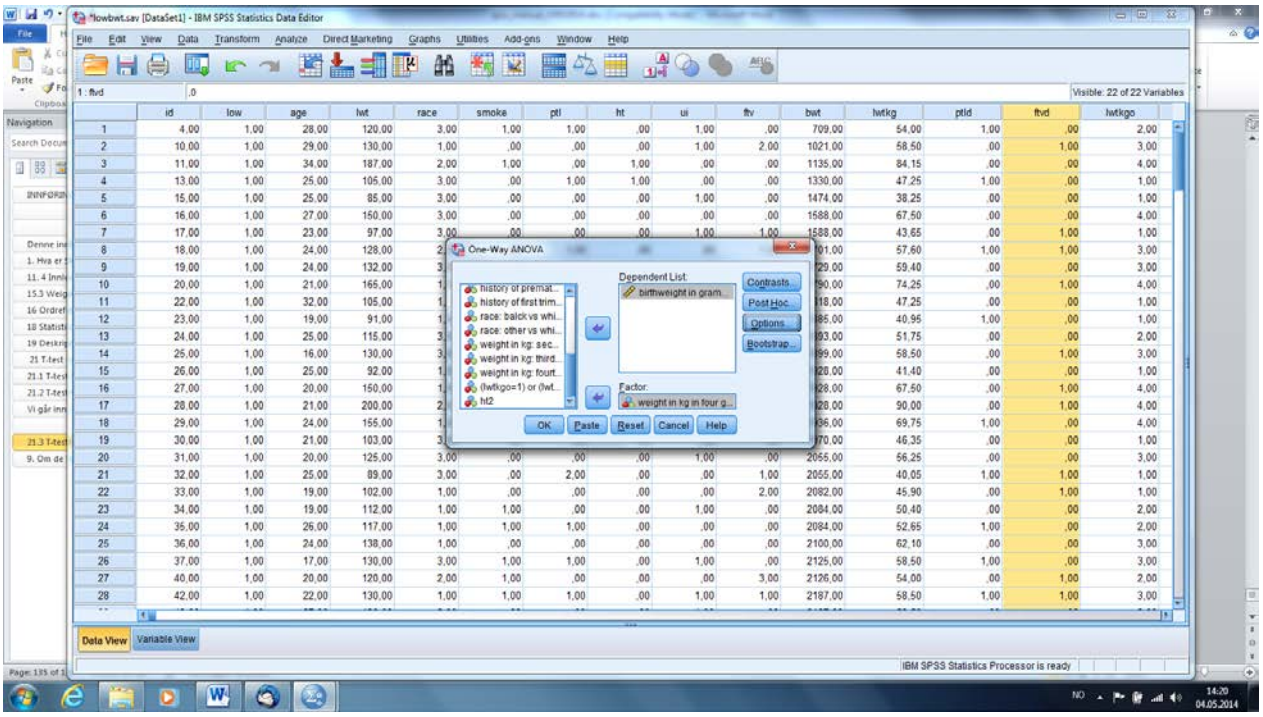

Normal Q-Q Plot of birthweight in grams
for lwtkgo= Second quartile
Expected Normal
2
0
-2
-4
0
1 000
2 000
3 000
4 000
5 000
Observed Value


Normal Q-Q Plot of birthweight in grams
for lwtkgo= Third quartile
Expected Normal
3
2
1
0
-1
-2
-3
0
1 000
2 000
3 000
4 000
5 000
6 000
Observed Value



Vi ser at helhetsbildet er at dataene i alle gruppene er normalfordelte. Det er noen få avvik i venstre hale i gruppene som er lagd av andre, tredje og fjerde kvartil for LWTKG. Vi går derfor videre med antagelsen om at dataene er normalfordelte.

Da skal vi kjøre en variansanalyse for BWT med hensyn til LWTKGO. Vi går inn i *Analyze/Compare Means/One-Way ANOVA*. Der trekker vi BWT over i vinduet med *Dependent List* og LWTKGO over i *Factor*. Da ser dialogboksen vår slik ut:



Vi klikker på OK og får det følgende:

ANOVA

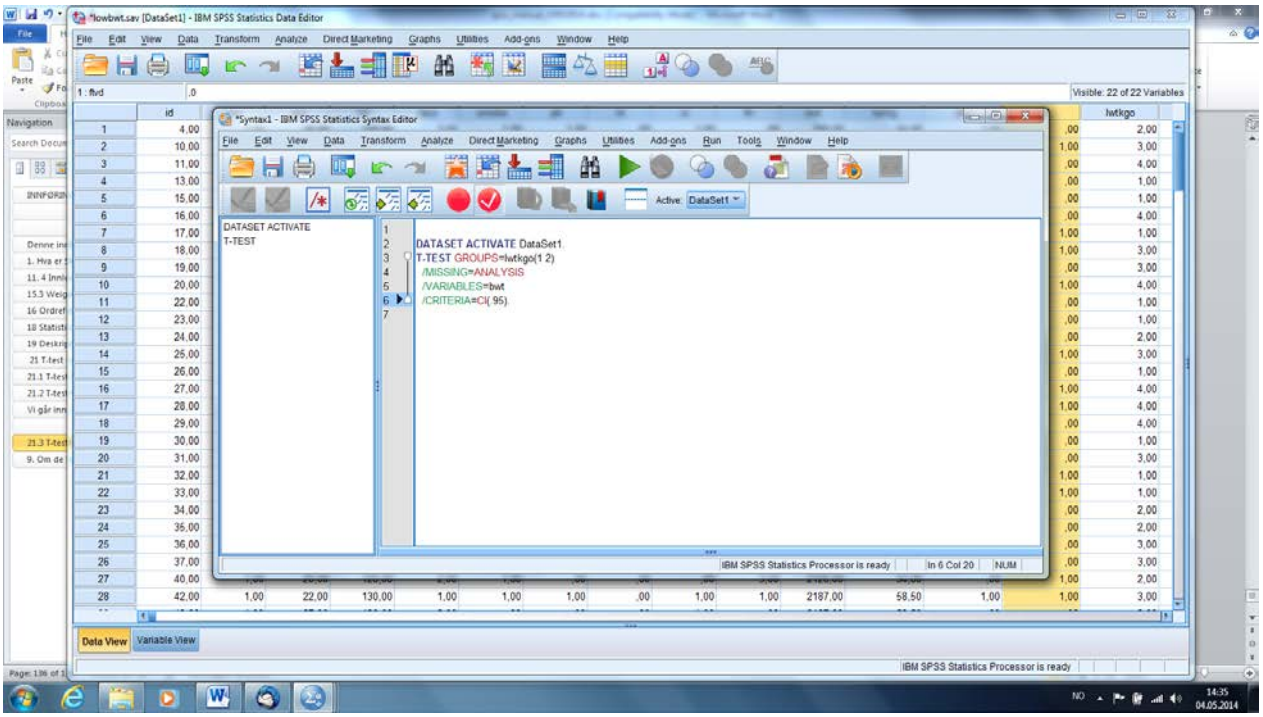
birthweight in grams

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	6190272,335	3	2063424,112	4,073	,008
Within Groups	93726780.31	185	506631,245		
Total	99917052.65	188			

Dette gir oss en oversikt over hvordan variasjonen fordeler seg mellom og innen gruppene. Vi ser at variasjonen mellom grupper er stor i forhold til den innen grupper. P-verdien i kolonnen lengst til høyre, sier oss om variasjonen mellom grupper er statistisk signifikant. Siden p = 0.008, som er mye mindre enn 0.05, er det klart at det er signifikant forskjell mellom gruppene.

Spørsmålet er nå om mellom hvilke grupper det er forskjell. Da må vi gjøre separate t-tester fo å avgjøre dette. Vi trenger egentlig bare å gjøre tre separate t-tester. Hvis vi velger laveste kategori som våre referansekategori, som her betyr at vi velger første kvartil som referanse, trenger vi bare å gjøre tre sammenligninger: Annen kvartil mot første kvartil, tredje kvartil mot første kvartil og fjerde kvartil mot første kvartil. Vi skal se hvordan vi gjør dette. Siden det er flere t-tester som er ganske like, velger vi å bruke en ordrefil få å gjøre dette enkelt for oss selv.

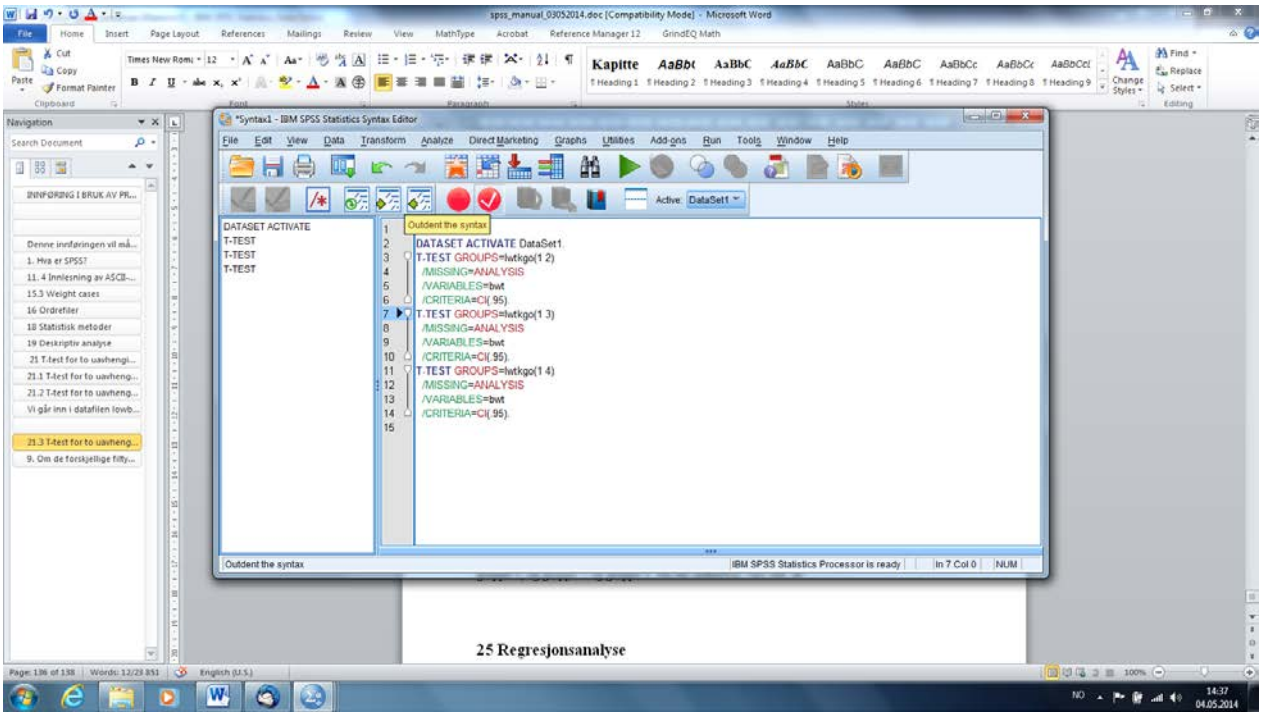
Vi starter da med en t-test mellom første og annen kvartil. Da går vi inn i *Analyze/Compare means/Independett-Samples T-test*. Her trekker vi over BWT i *Test Variable(s)* og LWTKGO i *Grouping Variable*. Da åpner det seg en *Define Groups* knapp som vi klikker på. Her skriver vi inn 1 og 2 i de vinduene. Da går vi tilbake til den forrige dialogboksen ved å klikke på OK. Men nå klikker vi på *Paste*. Da åpner det seg et ordrevindu som ser slik ut:



Nå kan vi gjøre alle analysene samtidig ved å utvide ordrefilen vår. Vi legger til

```
T-TEST GROUPS=lwtkgo(1 3)
  /MISSING=ANALYSIS
  /VARIABLES=bwt
  /CRITERIA=CI(.95).
T-TEST GROUPS=lwtkgo(1 4)
  /MISSING=ANALYSIS
  /VARIABLES=bwt
  /CRITERIA=CI(.95).
```

bare ved å endre første linje i hver analyse, slik at vi også får sammenlignet gruppe 1 og gruppe 3, og gruppe 1 og gruppe 4. Da ser ordrefilen våre slik ut:



Da markerer vi alle linjene i denne filen og klikker på den grønne pilen i knappelinjen på toppen av filen. Da får vi følgende resultat:

Group Statistics

	weight in kg in four groups	N	Mean	Std. Deviation	Std. Error Mean
birthweight in grams	First quartile	53	2655,5472	605,75822	83,20729
	Second quartile	43	3058,3721	745,63533	113,70832

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
birthweight in grams	Equal variances assumed	1,171	,282	-2,921	94	,004	-402,82492	137,89407	-676,61682	-129,03302
	Equal variances not assumed			-2,859	80,402	,005	-402,82492	140,90080	-683,20493	-122,44492

Group Statistics

	weight in kg in four groups	N	Mean	Std. Deviation	Std. Error Mean
birthweight in grams	First quartile	53	2655,5472	605,75822	83,20729
	Third quartile	46	3076,1304	781,08550	115,16477

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
birthweight in grams	Equal variances assumed	2,340	,129	-3,013	97	,003	-420,58326	139,57451	-697,60003	-143,56650
	Equal variances not assumed			-2,960	84,352	,004	-420,58326	142,07878	-703,10526	-138,06127

Group Statistics

	weight in kg in four groups	N	Mean	Std. Deviation	Std. Error Mean
birthweight in grams	First quartile	53	2655,5472	605,75822	83,20729
	Fourth quartile	47	3037,9574	719,91337	105,01016

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
birthweight in grams	Equal variances assumed	1,682	,198	-2,884	98	,005	-382,41028	132,59848	-645,54765	-119,27291
	Equal variances not assumed			-2,854	90,379	,005	-382,41028	133,97980	-648,56932	-116,25124

Vi ser at tabellene er ganske like. Det er en vektøkning på barna på omtrent 400 gram når mors vekt øker fra første kvartil til enten annen, tredje eller fjerde kvartil. Vi ser at spredningene i alle gruppene er omtrent like store, og Levenes test har p-verdier som alle er lang over 0.05. Altså kan vi anta at variansene er like store, og sammenligningene mellom fødselsvektene kan vi lese fra første linje i utskriften for Independent Samples Test. Vi finner at alle p-verdiene er klart mindre enn 0.05.

Merk at alle differansene i fødselsvekt er negative, siden vi tar differansen mellom fødselsvekt i første kvartil og i de andre kvartilene. Fødselsvektene i annen, tredje og fjerde kvartil er høyere enn i første kvartil. Når vi skal presentere resultatene, kan der være naturlig å snu om fra negative til positive tall. Når vi for eksempel skal presentere tallene for sammenligningen mellom første og annen kvartil, har vi at forkjellen i fødselsvekt er 382 gram, med et konfidensintervall på (119, 646) og en p-verdi på 0.005.

Dersom vi vil sammenligne fødselvekten mellom annen, tredje og fjerde kvartil kan vi bare endre verdien i ordrefilen vi nå har lagd. Vi finner at det ikke er statistisk signifikant forskjell i fødselsvekt mellom noen av gruppene. Dersom vi kjører en analyse mellom annen og fjerde kvartil finner vi følgende:

Group Statistics

	weight in kg in four groups	N	Mean	Std. Deviation	Std. Error Mean
birthweight in grams	Second quartile	43	3058,3721	745,63533	113,70832
	Fourth quartile	47	3037,9574	719,91337	105,01016

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
birthweight in grams	Equal variances assumed	,009	,924	,132	88	,895	20,41465	154,53562	-286,69242	327,52172
	Equal variances not assumed			,132	86,646	,895	20,41465	154,77957	-287,24422	328,07351

Vi ser at forskjellen i fødselsvekt i annen og fjerde kvartil er 20 gram. Det er gode grunner til å lese resultatene for sammenligningen fra øverste linje i utskriften for t-testen. Da finner vi at konfidensintervallet er (-287, 327). Husk at vi her IKKE behøver snu konfidensintervallet, siden vi har en positiv forskjell i differansen mellom gruppene.

Vi gjør deretter en vararіansanalyse for sammenhengen mellom BWT og RACE. Pass på at det fortsatt er datafilen **lowbwt.sav** som er vår aktive datafil. Først må vi sjekke at BWT-dataene er normalfordelte for de tre kategoriene av RACE. Vi viser ikke normalfordelingsplottene her, men konkluderer bare med at det er gode grunner til å anta at dataene er normalfordelte. Vi setter bare opp selve ordrene, som kopieres inn i en ordrefil, og som kjøres ved å markeres og så klikke på den grønne pilen. Merk at vi her i t-testen bare sammenligner black (RACE = 2) mot white (RACE = 1) og other (RACE = 3) mot white (RACE = 1). Ordrene som vi skal kjøre er:

```
ONEWAY bwt BY race
  /MISSING ANALYSIS.
T-TEST GROUPS=race(1 2)
  /MISSING=ANALYSIS
  /VARIABLES=bwt
  /CRITERIA=CI(.95).
T-TEST GROUPS=race(1 3)
  /MISSING=ANALYSIS
  /VARIABLES=bwt
  /CRITERIA=CI(.95).
```

Da blir resultatene:

ANOVA

birthweight in grams

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	5070607,632	2	2535303,816	4,972	,008
Within Groups	94846445.01	186	509927,124		
Total	99917052.65	188			

Group Statistics

	race	N	Mean	Std. Deviation	Std. Error Mean
birthweight in grams	white	96	3103,7396	727,72424	74,27304
	black	26	2719,6923	638,68388	125,25621

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
birthweight in grams	Equal variances assumed	,822	,367	2,446	120	,016	384,04728	156,99086	73,21629	694,87826
	Equal variances not assumed			2,637	44,232	,011	384,04728	145,62144	90,61007	677,48448

Group Statistics

	race	N	Mean	Std. Deviation	Std. Error Mean
birthweight in grams	white	96	3103,7396	727,72424	74,27304
	other	67	2804,0149	721,30115	88,12096

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
birthweight in grams	Equal variances assumed	,002	,963	2,597	161	,010	299,72466	115,42969	71,77317	527,67614
	Equal variances not assumed			2,601	142,958	,010	299,72466	115,24664	71,91695	527,53237

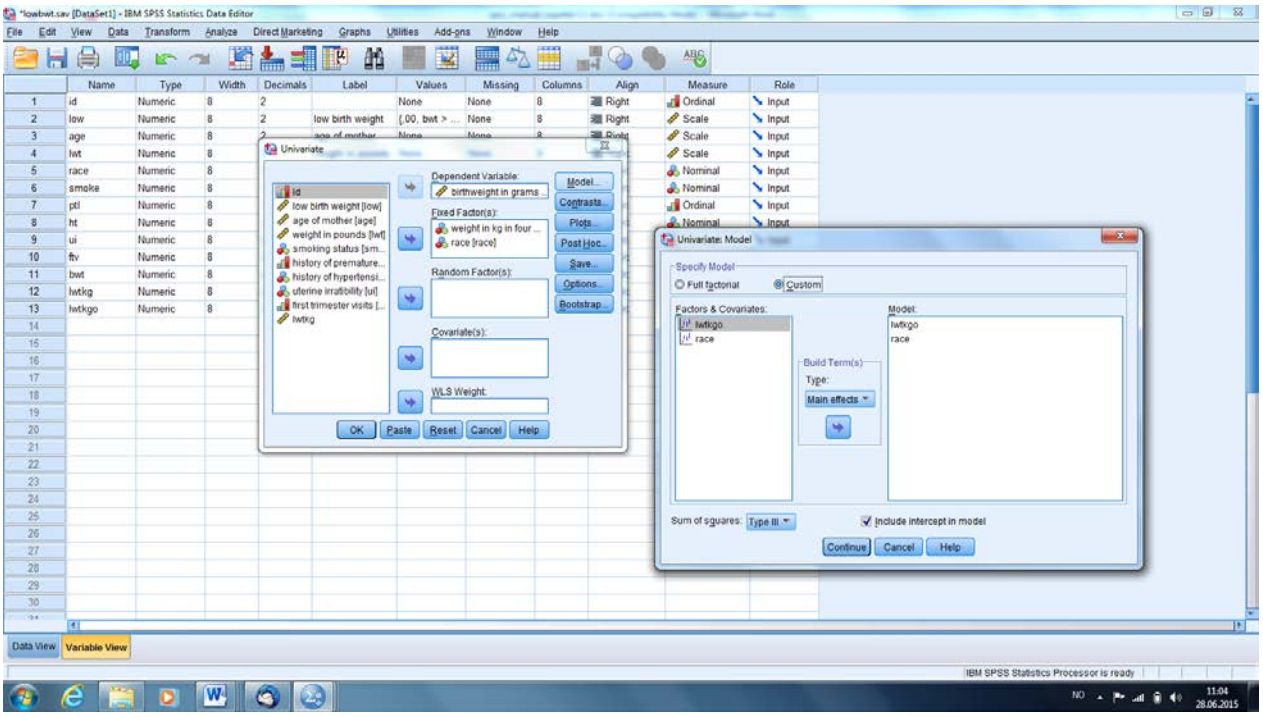
Av den første tabellen ser vi at det er en signifikant forskjell mellom de tre kategoriene av RACE med hensyn til fødselsvekt BWT. Dette ser vi av p-verdien, som her er p = 0.008. Da er spørsmålet om det er forskjell mellom white og black, og mellom white og other. I den føste av t-testene ser vi differansen i fødselsvekt er 384 gram. Et 95% konfidensintervall er (73.2, 694.9) og p-verdien er 0.016. Merk at 0 ligger utenfor konfidensintervallet, slik at vi bare ved å se på kan si at p-verdien må være < 0.05. Siden Levenes test viser en p-verdi på 0.367, kan vi lese av resultatene i øverste linje.

Tilsvarende finner vi at forskjellen i fødselvekt mellom white og black er på 299 gram, med et 95% konfidensintervall på (71.8, 527.7). P-verdien er p = 0.010. Altså statistisk signifikant forskjell også i fødselsvekt mellom barna til white og other mødre.

12.1.2 Flerveis variansanalyse

I forrige eksempel (enveis variansanalyse) så vi at variablene LWTKGO og RACE begge hadde statistisk signifikant effekt på BWT (begge p-verdier 0.008). Vi er nå interessert i om effekten av LWTKGO endrer seg når vi kontrollerer for effekten som RACE har på BWT. For å få dette til må vi analysere begge variablene samtidig i en varianslyse. Dette er det vi kaller en flerveis variansanalyse.

Da går vi til *Analyze/Generalized Linear Model/Univariate*. Der trekker vi BWT over i *Dependent Variable* og RACE og LWTKGO over i *Fixed Factor(s).* Vi går så inn i *Model*. Der går vi til i midten og endrer til *Main effects.* Vi trekker over de to variablene RACE og LWTKGO. Da blir dialogboksene seende slik ut:



På vanlig måte klikker vi på *Continue* og *OK*. Da får vi følgende resultat:

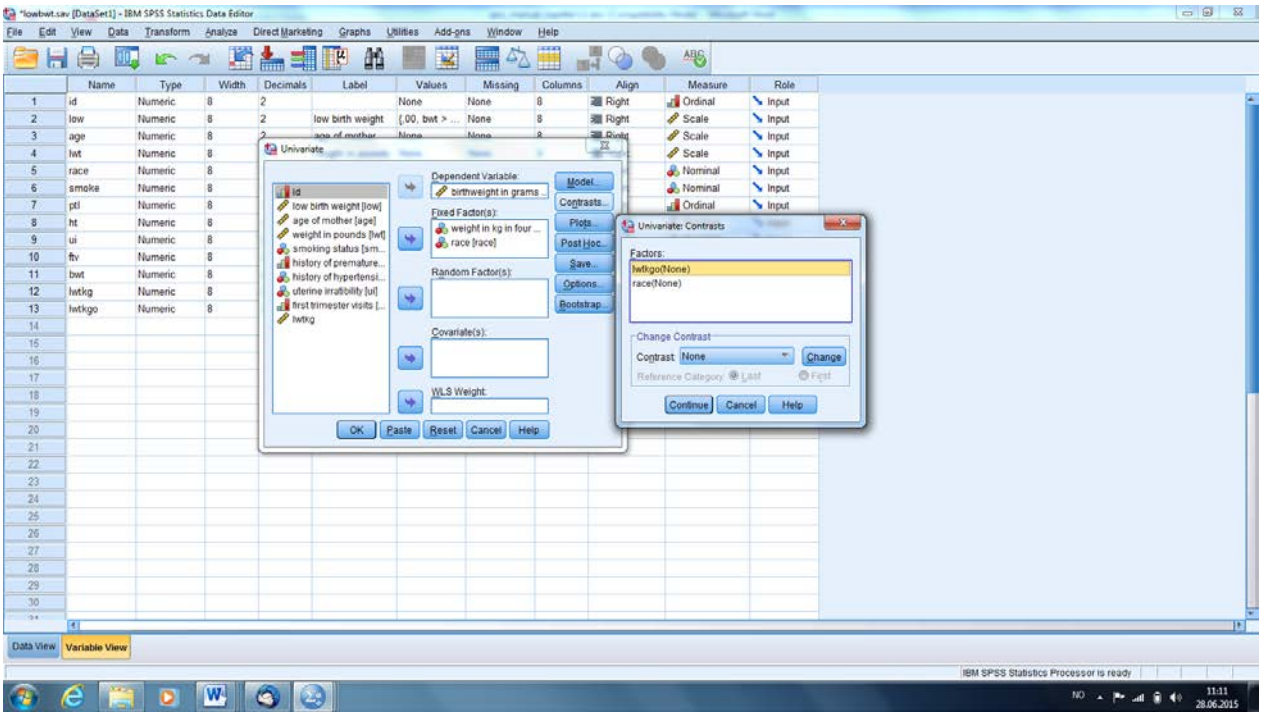
Tests of Between-Subjects Effects

Dependent Variable: birthweight in grams

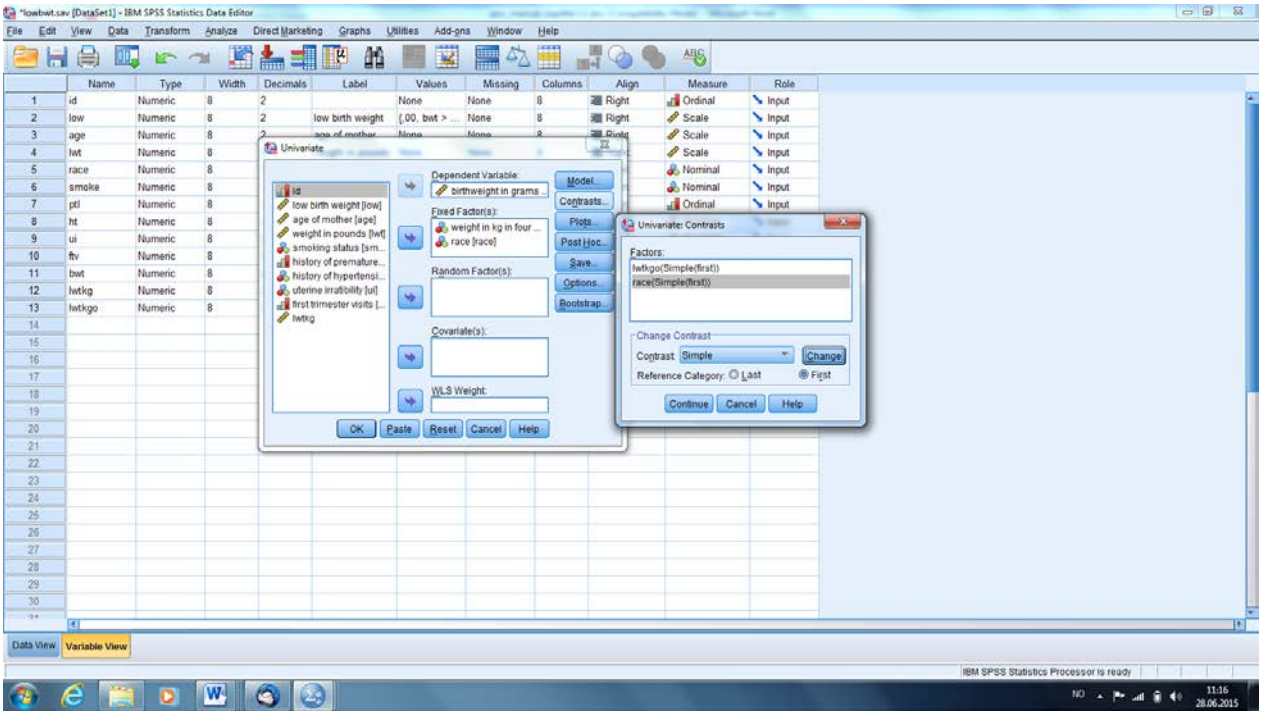
Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	10907160.0[a]	5	2181431,996	4,485	,001
Intercept	1161811490	1	1161811490	2388,628	,000
lwtkgo	5836552,347	3	1945517,449	4,000	,009
race	4716887,645	2	2358443,822	4,849	,009
Error	89009892.67	183	486392,856		
Total	1738735950	189			
Corrected Total	99917052.65	188			

a. R Squared = ,109 (Adjusted R Squared = ,085)

Da ser vi at både RACE og LWTKGO er statistisk signifikante, med p-verdier på 0.009. Da er vi interessert i å se hvor effektene ligger. Da går vi tilbake til *Analyze/Generalized Linear Model/Univariate*. Her går vi til *Contrasts*. Da åpner det seg nye dialogbokser som ser slik ut:



Markeringen står på variabelen LWTKGO. Der det står *Contrast(s): None* skifter vi til *Contrast(s): Simple*. Så kikker vi på *First* og til slutt på *Change*. Vi gjør så det samme med RACE. Da ser dialogboksene slik ut:



Når vi klikker på *Continue* og *OK*, får vi følgende resultat:

Contrast Results (K Matrix)

weight in kg in four groups Simple Contrast[a]			Dependent Variable: birthweight in grams
Level 2 vs. Level 1	Contrast Estimate		421,831
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		421,831
	Std. Error		144,272
	Sig.		,004
	95% Confidence Interval for Difference	Lower Bound	137,181
		Upper Bound	706,482
Level 3 vs. Level 1	Contrast Estimate		389,247
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		389,247
	Std. Error		142,712
	Sig.		,007
	95% Confidence Interval for Difference	Lower Bound	107,674
		Upper Bound	670,820
Level 4 vs. Level 1	Contrast Estimate		388,937
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		388,937
	Std. Error		144,640
	Sig.		,008
	95% Confidence Interval for Difference	Lower Bound	103,560
		Upper Bound	674,314

a. Reference category = 1

Contrast Results (K Matrix)

race Simple Contrast[a]			Dependent Variable: birthweight in grams
Level 2 vs. Level 1	Contrast Estimate		-434,668
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-434,668
	Std. Error		156,019
	Sig.		,006
	95% Confidence Interval for Difference	Lower Bound	-742,495
		Upper Bound	-126,841
Level 3 vs. Level 1	Contrast Estimate		-237,977
	Hypothesized Value		0
	Difference (Estimate - Hypothesized)		-237,977
	Std. Error		113,556
	Sig.		,037
	95% Confidence Interval for Difference	Lower Bound	-462,025
		Upper Bound	-13,930

a. Reference category = 1

Første tabell viser resultatene for LWTKGO. Siden vi har valgt *Simple* og *First* i *Contrast*, sammenlignes alle gruppe mot den første gruppen. Da ser vi at fødselsvekten for dem som har vekt i annen kvartil er 421 gram høyere enn for dem som som er i første kvartil. Tilsvarende for dem som er i høyeste kvartil har en fødselsvekt som er 388 gram høyere enn dem som er i første kvartil. Alle effektene er statistisk signifikante.

Andre tabell er for RACE. Her er kategorien White første kategori, som vi da sammenligner med ved å bruke *Simple*. Da viser resultatene at forskjellen i fødselsvekt for mor som er Black i forhold til mor som er White er en nedgang på 435 gram. Tilsvarende er nedgangen for Other 238 gram. Begge effektene er statistisk signifikante.

12.2 Lineær regresjonsanalyse

Regresjonsanalyse er en svært viktig metode i all statistisk analyse. I dette kapittelet skal vi se på lineær regresjon, som er analysemetoden vi bruker når vi har en kontinuerlig avhengig variabel. I all statistisk analyse er vi interessert i å studere sammenhengen mellom et sett med forklaringsvariabler og den avhengige variabelen. I enkel lineær regresjon ser vi på sammenhengen mellom den avhengige variabelen og én forklaringsvariabel. Men vanligvis har vi flere forklaringsvariabler som bidrar til forklaringen av den avhengige variabelen. Vi må da innføre alle de forklaringsvariablene vi er interessert, for å se om den effekten som vi hadde for én forklaringsvariabel beholdes når vi tar inn flere forklaringsvariabler. Vi sier at vi kontrollerer for effekten av de andre forklaringsvariablene.

La oss bruke datasettet **lowbwt.sav** som eksempel. Vi antar at det er en sammenheng mellom barnets vekt (BWT) og mors vekt (LWTKG). Den studerer vi en i enkel lineær regresjon, med

BWT som avhengig variabel og LWTKG som (eneste) forklaringsvariabel. Men vi vet at om mor røyker (SMOKE = 1) vil påvirke barnet vekt (BWT) og mors vekt (LWTKG). Vil da mors vekt fortsatt ha effekt på barnets vekt når vi tar bort, eller kontrollerer for som vi sier, effekten av røyking. Dette får vi ut ved å analysere dette via en multippel regresjonsanalyse med BWT som avhengig variabel og LWTKG og SMOKE som forklaringsvariabler.

Vanligvis tenker vi på lineær regresjonsanalyse som en metode for å studere sammenhengen mellom en kontinuerlig avhengig variabel og en kontinuerlig forklaringsvariabel. Det er ufravikelig at den avhengige variabelen skal være kontinuerlig, men forklaringsvariabelen behøver ikke være det. Vi kan bruke en kategorisk variabel – med to kategorier – direkte i en regresjonsanalyse. Dette betyr at vi kan bruke variablene SMOKE, PTLD og FTVD direkte i regresjonsanalysen.

Men kategoriske variabler med flere enn to kategorier kan vi ikke bruke som de er. Her må vi omkode dem til dummy-variabler, som hver er kategoriske med to kategorier. Antallet dummy-variabler vi må lage, er lik antallet kategorier i variabelen minus 1. Vi kan altså ikke bruke RACE direkte, men vi må lage to dummy-variabler ut av RACE. Det var det vi gjorde i kapittel 14.5, da vi lagde RACE2 og RACE3. Når vi skal bruke RACE i en regresjonsanalyse, må vi altså bruke både RACE2 og RACE3 som våre forklaringsvariabler.

Gangen i en regresjonsanalyse vil vanligvis være som følger:

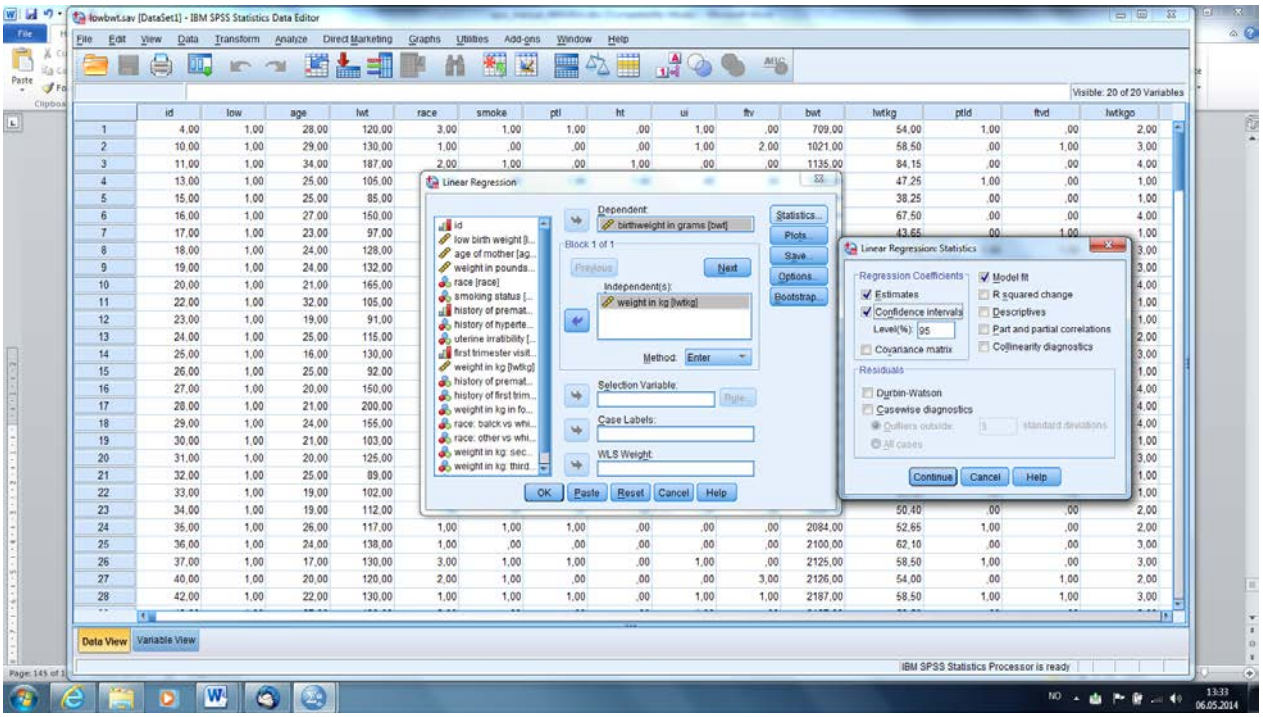
1. Kjør enkel lineær regresjon med hver enkelt forklaringsvariabel. Dersom p-verdien er < 0.20 tar vi den med videre som en kandidat for en multippel regresjon. Ta også med variable som er av biologisk eller medisinsk interesse, selv om de har en p-verdi > 0.20.
2. Kjør multippel regresjon med alle variablene som er inkludert på trinn 1.
3. I den multiple regresjonsmodellen tar vi ut den forklaringsvariabelen som har høyest p-verdi, og kjør en multippel regresjon uten den.
4. Gjenta trinn 3 inntil alle forklaringsvariablene er statistisk signifikant, med $p < 0.05$. Behold forklaringsvariable som er av biologisk eller medisinsk interesse

12.2.1 Enkel lineær regresjon. Eksempel: lowbwt.sav

Vi skal i dette eksempelet analysere BWT som avhengig variabel og med LWTKG, AGE, SMOKE, HT, RACE, PTLD og FTVD som forklaringsvariabler. SMOKE, HT, PTLD og FTVD er kategoriske variabler med to kategorier, og de trenger vi ikke gjøre noe med. RACE er en kategorisk variabel, med tre kategorier. Da må vi lage to dummy-variabler, RACE2 og RACE3, som vi har gjort tidligere.

Da er vi klare til å gjøre trinn 1 i metoden som vi beskrev i innledningen til dette kapittelet. Vi gjør da enkle regresjonsanalyser med hver enkelt av variablene LWTKG, AGE, SMOKE, HT, RACE, PTLD og FTVD.

Vi starter med en regresjon av BWT på LWTKG. Vi går da inn i *Analyze/Regression/Linear*. Vi trekker over BWT i *Dependent* og LWTKG over i *Independent*. Så gå vi til *Statistics*. Der klikker vi av på *Confidence Interval Level*, og lar det stå 95 der, siden vi ønsker et 95% konfidensintervall. Da ser dialogboksen vår slik ut:



Da får vi følgende resultat:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,186[a]	,035	,029	718,24270

a. Predictors: (Constant), weight in kg

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2369,672	228,431		10,374	,000	1919,040	2820,304
	weight in kg	9,843	3,807	,186	2,586	,010	2,333	17,352

a. Dependent Variable: birthweight in grams

I den første tabellen ser vi at r = 0.186 og vi ser at r^2 = 0.035. Disse tallene kjenner vi igjen fra korrelasjonsanalysen som vi gjorde i kapittel 11.5.

I neste tabell finner vi resultatene fra selve regresjonsanalysen. I kolonnen under B finner vi selve regresjonskoeffisienten. For LWTKG er den 9.84, med en standardfeil på 3.81. At effekten av LWTKG er 9.84 betyr at for hver kilo økning av mors vekt gir en vektøkning på barnet på 9.84 gram.

Siden teststørrelsen er lik effektmålet delt på standardfeilen, har vi at t = 9.84/3.81 = 2.59. Den to-sidige p-verdien finner vi under Sig. Den er p = 0.010. Den kjenner vi også igjen fra p-verdien på korrelasjonskoeffisienten. P-verdien for testen på om regresjonskoeffisienten er 0, er den samme som for om korrelasjonskoeffisienten er lik 0, siden de to testene er identiske.

Til slutt har vi konfidensintervallet for regresjonskoeffisienten. Som vi har fra tidligere er dette tilnæremet gitt som

(Effektmålet – 1.96xStandardfeilen, Effektmålet + 1.96xStandardfeilen)

Vi ser at dette er tilnærmet likt det som SPSS gir som sitt konfidensintervall, nemlig (1919.0, 2820.3). Vi ser at konfidensintervallet i sin helhet ligger over 0, som betyr at p-verdien er nødt til å være < 0.05. Vi oppsummerer altså med at effekten av mors vekt er 9.84 gram per kilo vektøkning, KI = (1919.0, 2820.3), p = 0.010.

Da gjør vi en helt tilsvarende analyse for sammenhengen mellom BWT og AGE. Vi går da tilbake til *Analyze/Regression/Linear*. Vi beholder BWT i *Dependent* men trekker LWTKG tilbake fra *Independent*. Istedenfor trekker vi AGE over i *Independent*. Ved å klikke på *OK* får vi følgende resultat.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,090[a]	,008	,003	728,01145

a. Predictors: (Constant), age of mother

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2657,333	238,804		11,128	,000	2186,236	3128,429
	age of mother	12,364	10,021	,090	1,234	,219	-7,404	32,132

a. Dependent Variable: birthweight in grams

Vi ser av første tabell at $r^2 = 0.008$. Det betyr at 0.8% av variasjonen i fødelsvekt er forklart av mors alder. Det er ikk mye! Dette stemmer også bra med resultatet i neste tabell. Der ser vi at fødselsvekten øker med 12 gram for hver år mor blir eldre, KI = (-7.4, 32.1) og p = 0.090. Siden p = 0.090, som er > 0.05, er sammenhengen mellom fødselsvekt og mors alder ikke statistisk signifikant. Dette får vi også bekreftet ved å se på konfidensintervallet som dekker verdien 0. Men p-verdien er lavere enn det kravet som vi satte i punkt 1 i innledningen til dette kapittelet.

Men mors alder er en viktig biologisk og medisinsk forklaringsvariabel så vi velger å ta denne med videre i analysen, uansett om p-verdien hadde vært > 0.20.

Så går vi videre variabelen SMOKE. Dette er en to-kategorisk variabel, men den kan tas direkte inn i regresjonsanalysen, og regresjonskoeffisienten kan tolkes som effekten av SMOKE når den går fra 0 (ikke-røyker)) til 1 (røyker).

Vi gjør da som over og går til *Analyze/Regression/Linear*. Vi trekker AGE tilbake fra *Independent* og trekker over SMOKE. Ved å klikke på *OK* får vi følgende resultat:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,189[a]	,036	,031	717,77898

a. Predictors: (Constant), smoking status

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3054,957	66,933		45,642	,000	2922,915	3186,998
	smoking status	-281,713	106,969	-,189	-2,634	,009	-492,734	-70,693

a. Dependent Variable: birthweight in grams

Her ser vi SMOKE forklarer bare 3.6% av variansen i BWT. Det er også lite, men av tabellen nedenfor ser vi at effekten av SMOKE på BWT er statistisk signifikant, p = 0.009. Vi finner at regresjonskoeffisienten er -281.7, 95% KI = (-493, -71). Snur vi på fortolkningen av dette har at fødselsvekten reduseres med 281.7 gram når mor er røyker, i forhold til å være ikke-røyker. Et 95% konfidensintervall på reduksjonen i vekt er (71, 493).

Disse resultatene kjenner vi igjen fra kapittelet om t-tester. Da vi gjorde en t-test for uavhengige utvalg i analysen av BWT og SMOKE hadde vi følgende resultat (som er kopiert fra kapittel 11.2.2):

Group Statistics

	smoking status	N	Mean	Std. Deviation	Std. Error Mean
birthweight in grams	non-smoking	115	3054,9565	752,40901	70,16250
	smoking	74	2773,2432	660,07517	76,73218

Independent Samples Test

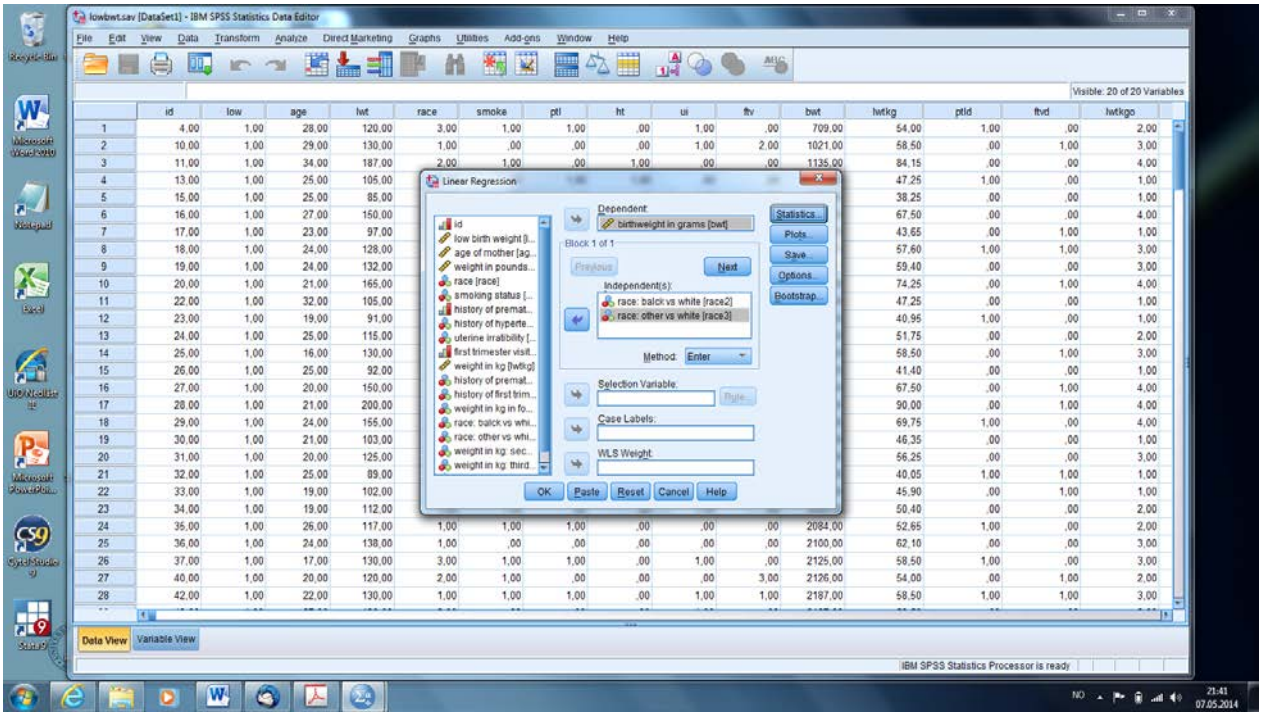
		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
birthweight in grams	Equal variances assumed	1,508	,221	2,634	187	,009	281,71328	106,96873	70,69274	492,73382
	Equal variances not assumed			2,709	170,001	,007	281,71328	103,97406	76,46677	486,95979

Her ser vi effektmålet, forskjellen i gjennomsnittlig fødselsvekt, er 281.7 gram. Vi har et 95% konfidensintervallet som er (70.7, 492.7). I tabellen ser vi at p-verdien er p = 0.009. Disse resultatene er identiske med dem vi fikk i regresjonsanalysen, bortsett fra at fortegnene er snudd.

Dette er et generelt prinsipp: Det å gjøre en regresjonsanalyse med en kategorisk forklaringsvariabel med to kategorier er det samme som å gjøre en t-test for to uavhengige utvalg.

La oss så går videre med RACE. Her må vi huske at RACE er en kategorisk variabel med tre kategorier. Da må vi bruke to dummy-variabler for å studere effekten av to kategorier i forhold til den tredje. Vi har valgt å lage RACE2 og RACE3 slik at kategorier White er referansekategorien og variabelen RACE2 måler effekten av Black i forhold til White og RACE3 måler effekten av Other i forhold til White. Når vi bruker dummy-variabler til å studere effekten av kategoriske variabler, er det viktig at vi alltid tar med alle dummy-variablene i analysen. Vi må altså ta med både RACE2 og RACE3 for å finne den samlede effekten av RACE.

Vi går da inn i *Analyze/Regression/Linear*. Vi trekker SMOKE tilbake fra *Independent* og trekker over RACE2 og RACE3 i *Independent*. Da ser dialogboksen vår slik ut:



Ved å klikke på *OK* får vi følgende resultat:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,225[a]	,051	,041	714,09182

a. Predictors: (Constant), race: other vs white, race: balck vs white

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3103,740	72,882		42,586	,000	2959,959	3247,521
	race: balck vs white	-384,047	157,874	-,182	-2,433	,016	-695,502	-72,593
	race: other vs white	-299,725	113,678	-,197	-2,637	,009	-523,988	-75,462

a. Dependent Variable: birthweight in grams

Av den første tabellen ser vi at r = 0.051, dvs. at 5.1% av variasjonen er forklart av variabelen RACE (altså samlet av RACE2 og RACE3). Av den neste tabellen ser vi at mødre som er Black føder barn som er 384.1 gram lettere enn White mødre, og Other mødre føder barn som er 300.0 gram lettere enn White mødre. Vi ser at p-verdiene er henholdsvis 0.016 og 0.009, som betyr at begge disse effekten er statistisk signifikante. Vi ser at konfidensintervallet for Black i forhold til White er (72.6, 695.5) og (75.5, 524.0) for Other i forhold til White. Merk at i her har snudd fortegnet for å få en enklere presentasjon av effektene.

I kapittel 12.1.1 brukte vi ANOVA til å sammenligne effektene mellom gruppene white, black og other. Hvis vi nå går tilbake til resultatene for sammenligningen mellom black og white hadde vi der:

Group Statistics

	race	N	Mean	Std. Deviation	Std. Error Mean
birthweight in grams	white	96	3103,7396	727,72424	74,27304
	black	26	2719,6923	638,68388	125,25621

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
birthweight in grams	Equal variances assumed	,822	,367	2,446	120	,016	384,04728	156,99086	73,21629	694,87826
	Equal variances not assumed			2,637	44,232	,011	384,04728	145,62144	90,61007	677,48448

Vi ser at selve effekten er den samme, nemlig 384.0 gram. Men p-verdien og konfidensintervallet er litt forskjellige. Grunnen til det er at i regresjonsanalysen regnes standardfeilen ut på en litt annen måte enn i t-testene. Det blir derfor en liten forskjell i resultatene for p-verdiene og konfidensintervallene.

Men som for t-testene har vi at et generelt prinsipp: Det å gjøre en regresjonsanalyse med en kategorisk forklaringsvariabel med flere kategorier er det samme som å gjøre en variansanalyse.

Da går vi videre med HT. Da finner vi følgende resultat.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,155[a]	,024	,019	722,16567

a. Predictors: (Constant), history of premature labor

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2989,418	56,535		52,877	,000	2877,890	3100,947
	history of premature labor	-228,651	106,760	-,155	-2,142	,034	-439,260	-18,041

a. Dependent Variable: birthweight in grams

HT forklarer også bare en liten del av variansen i BWT, 2.4%. Men effekten av HT på BWT er statistisk signifikant, p = 0.034. Regresjonskoeffisienten er -228.7, og et 95% KI = (-439, -18). Fødselsvekten reduseres med 228.7 gram når mor er hypertensive, i forhold til å være normotensive, og et 95% konfidensintervall på reduksjonen i vekt er (18, 439).

Vi må fortsette med PTLD. For den variabelen finner vi følgende resultat:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,218[a]	,048	,043	713,35439

a. Predictors: (Constant), history of premature labor

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3013,572	56,573		53,269	,000	2901,970	3125,175
	history of premature labor	-434,172	141,996	-,218	-3,058	,003	-714,293	-154,052

a. Dependent Variable: birthweight in grams

Igjen ser vi PTLD forklarer bare en liten del av variasjonen i BWT, men effekten er klart statistisk signifikant (p < 0.001). Når PTLD = 1 får vi en redusert fødselsvekt på barnet på 434.2 gram, med et 95% konfidensintervall (154.1, 714.3).

Til slutt analyserer vi FTVD. Da får vi på lignende måte:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,116[a]	,013	,008	726,05715

a. Predictors: (Constant), history of first trimester visits

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2865,270	72,606		39,463	,000	2722,038	3008,502
	history of first trimester visits	168,584	105,805	,116	1,593	,113	-40,141	377,309

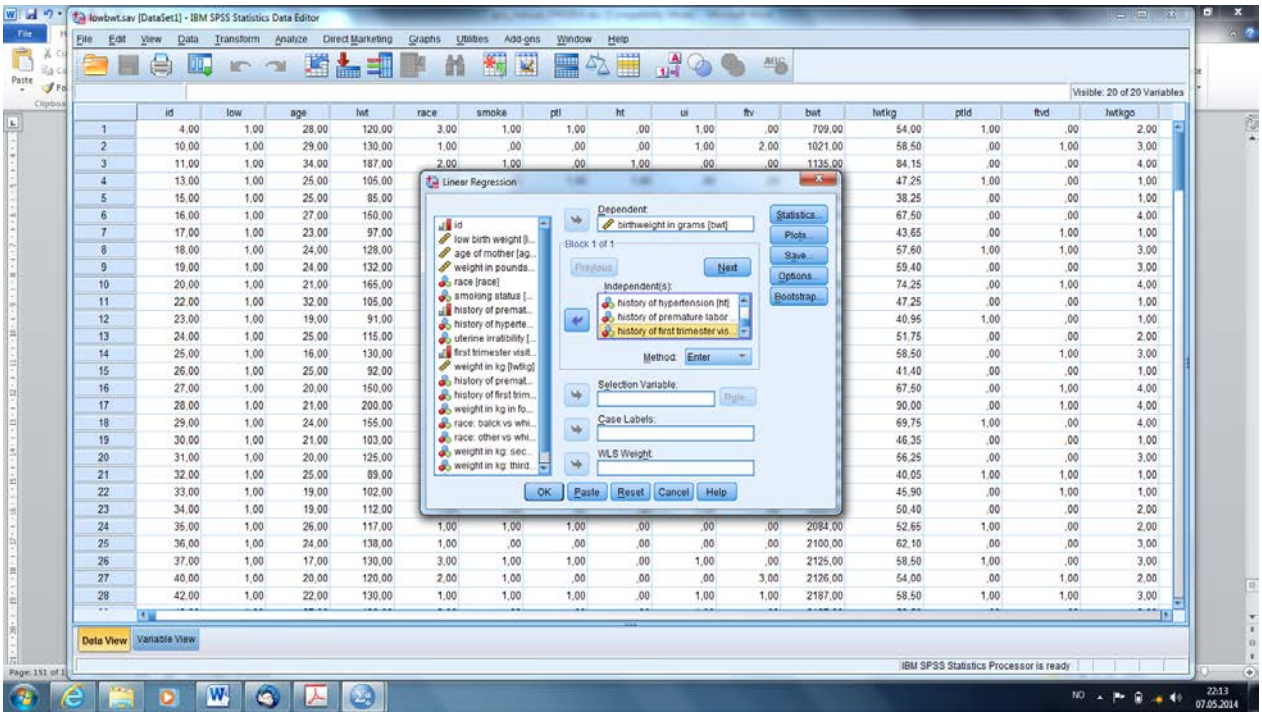
a. Dependent Variable: birthweight in grams

FTVD forklarer også bare en liten del av variasjonen i BWT (0.8%). Denne effekten er ikke statistisk signifikant (p = 0.113). Når FTVD = 1 får vi en økt fødselsvekt på barnet på 168.6 gram, med et 95% konfidensintervall på (-40.1, 377.3).

12.2.2 Multippel regresjon. Eksempel: lowbwt.sav

Nå har vi gjort alle de enkle lineære analysene som vi må for å kunne vurdere effekten i forhold til den avhengige variabelen BWT. Ikke alle effektene var statistisk signifikante, men alle p-verdiene var < 0.20. I henhold til punkt 2 i oversikten i innledningen til dette kapittelet tar med alle variablene inn i en multippel regresjonsanalyse.

Da går vi igjen til *Analyze/Regression/Linear*. Vi trekker nå over LWTKG, AGE, SMOKE, RACE2, RACE3, HT, PTLD og FTVD over i *Independent*. Da ser dialogboksen vår slik ut:



Når vi klikker på *OK*, får vi følgende resultat.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,444[a]	,197	,162	667,48850

a. Predictors: (Constant), history of first trimester visits, race: balck vs white, history of premature labor, history of hypertension, smoking status, age of mother, weight in kg, race: other vs white

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2737,172	317,356		8,625	,000	2110,955	3363,389
	weight in kg	10,297	3,921	,194	2,626	,009	2,560	18,035
	age of mother	-,923	9,963	-,007	-,093	,926	-20,583	18,737
	smoking status	-330,360	111,588	-,222	-2,961	,003	-550,549	-110,171
	race: balck vs white	-468,422	154,396	-,222	-3,034	,003	-773,080	-163,763
	race: other vs white	-343,334	120,485	-,226	-2,850	,005	-581,079	-105,590
	history of hypertension	-506,507	206,339	-,170	-2,455	,015	-913,661	-99,352
	history of premature labor	-289,537	138,519	-,146	-2,090	,038	-562,867	-16,207
	history of first trimester visits	44,639	103,884	,031	,430	,668	-160,348	249,625

a. Dependent Variable: birthweight in grams

Vi ser at alle disse variablene samlet sett forklarer 19.7% av variasjonen i BWT. Nå begynner det å bli bra!

Når vi går til neste tabell ser vi AGE og FTVD ikke er statistisk signifikante. AGE har en p-verdi på 0.926 og FTVD på 0.668. Etter regelen i punkt 3 skal vi nå ta ut begge disse variablene. Men vi har bestemt oss for at AGE er så viktig at vi lar den være med i den videre analysen. Men vi tar nå ut FTVD av listen i *Independent*. Når vi klikker på *OK* får vi følgende resultat:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,443[a]	,197	,165	665,98337

a. Predictors: (Constant), history of premature labor, race: balck vs white, history of hypertension, age of mother, smoking status, weight in kg, race: other vs white

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	2746,193	315,947		8,692	,000	2122,780	3369,606
	weight in kg	10,269	3,912	,194	2,625	,009	2,551	17,988
	age of mother	-,015	9,715	,000	-,002	,999	-19,183	19,153
	smoking status	-339,431	109,326	-,228	-3,105	,002	-555,148	-123,715
	race: balck vs white	-470,871	153,943	-,223	-3,059	,003	-774,624	-167,118
	race: other vs white	-353,756	117,753	-,233	-3,004	,003	-586,101	-121,412
	history of hypertension	-512,103	205,463	-,172	-2,492	,014	-917,514	-106,692
	history of premature labor	-286,585	138,037	-,144	-2,076	,039	-558,953	-14,217

a. Dependent Variable: birthweight in grams

Vi ser at r^2 = 0 0197 som betyr at alle disse forklaringsvariableen samlet sett forklarer 19.7% av variasjonen i BWT. Ikke dårlig!

Vi ser at alle variablene (med unntak av AGE som vi likevel vil ha med) er statistisk signifikante. Effekten av hver forklaringsvariabel må nå tolkes som effekten av den forklaringsvariabelen når vi kontrollerer for (dvs. tar bort) effekten av de andre forklaringsvariablen. Altså er effekten av røyking (SMOKE) at fødselsvekten reduseres med 339.4 gram (p = 0.002, 95% KI = (123.7, 555)) når vi kontrollerer for effekten av alle de andre forklaringsvariablene. Tilsvarende er effekten av mors vekt at en vektøknig på 1 kg for mor, betyr at barnets fødselsvekt øker med 10.3 gram (p = 0.009, 95% KI = (2.6, 18.0). Slik kan vi forklare alle variablene som nå har vist seg å være viktige for forklaringen av fødselsvekt.

12.3 Logistisk regresjonsanalyse

I avsnitt 12.2 så vi på lineær regresjon, som er analysemetoden vi bruker når vi har en kontinuerlig avhengig variabel, og vi er interessert i å studere sammenhengen mellom et sett med forklaringsvariabler og den avhengige variabelen. Når vi har flere forklaringsvariabler som bidrar til forklaringen av den avhengige variabelen, må vi innføre alle de forklaringsvariablene vi er interessert, for å se om den effekten som vi hadde for én forklaringsvariabel beholdes når vi tar inn flere forklaringsvariabler. Vi sier at vi kontrollerer for effekten av de andre forklaringsvariablene.

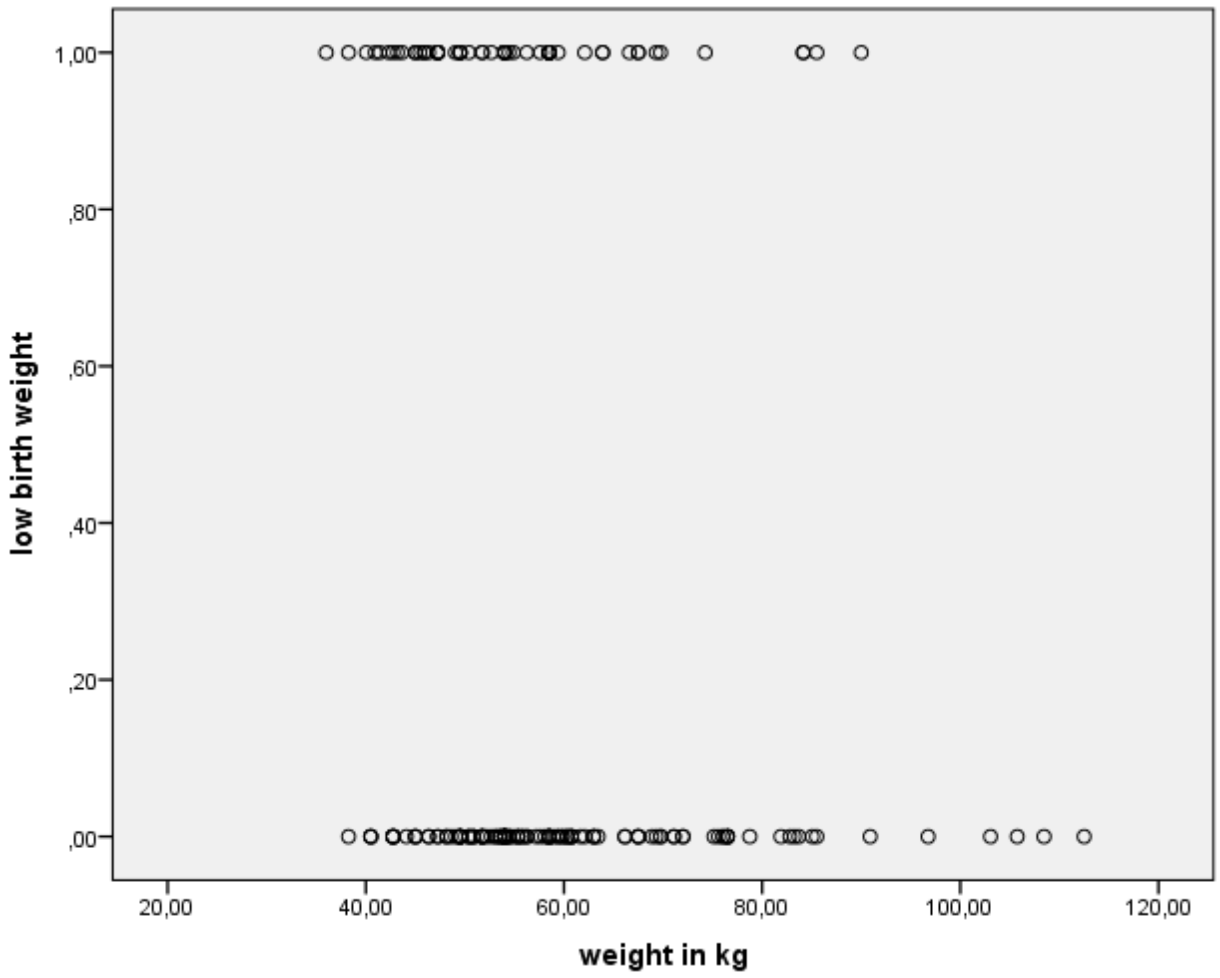
I lineær regresjonsanalyse er det ufravikelig at den avhengige variabelen skal være kontinuerlig, men forklaringsvariabelen behøver ikke være det. Vi kan for eksempel bruke en kategorisk variabel – med to kategorier – direkte i en regresjonsanalyse. Men kategoriske variabler med flere enn to kategorier kan vi ikke bruke som de er. I regresjonsanalysen må vi omkode dem til dummy-variabler, som hver er kategoriske med to kategorier. Antallet dummy-variabler vi må lage, er lik antallet kategorier i variabelen minus 1, se avsnitt 12.2. I SPSS sitt program for logistisk regresjon, behøver vi ikke lage dummy-variabler. Programmet gjør det for oss, forutsatt at vi markerer at vi har en kategorisk variabel som skal gjøres om til dummy-variabler

I logistisk regresjon ser vi på sammenhengen mellom en binær avhengig variabel og forklaringsvariabler som enten er kontinuerlige eller kategoriske. Den avhengige variabelen skal alltid ha to kategorier, 1 og 0.

La oss igjen bruke datasettet **lowbwt.sav** som eksempel. I kapittel 12.2, lineær regresjon, så vi først på sammenhengen barnets vekt (BWT) og mors vekt (LWTKG) i en enkel lineær regresjon. Deretter så vi på om variablene AGE, SMOKE, HT, RACE, PTLD og FTVD hadde noen effekt på BWT. Til slutt så vi på om hvilke av disse variablene som vil påvirke barnets vekt (BWT). Dette gjør vi via en multippel regresjonsanalyse med BWT som avhengig variabel og de andre variablene som mulige forklaringsvariabler.

Når den avhengige variabelen er binær, må vi analysere sammenhengen mellom denne variabelen og én forklaringsvariabel eller et sett av forklaringsvariabler med logistisk regresjon. Grunner til at vi ikke kan analysere en binær variabel med lineær regresjon, er så enkel som at sammenhengen ikke er lineær.

I datasettet **lowbwt.sav** er variabelen LOW binær. Dersom vi plotter sammenhengen mellom LOW og LWTKG finner vi



Vi ser at det ikke kan være en lineær sammenheng mellom lav fødselsvekt og mors vekt. Derimot viser det seg at en logistisk sammenheng passer fint. Siden den avhengige variabelen

er binær, er det sannsynligheten for at den avhengige variabelen har verdien 1, vi er interessert i. Den logistiske regresjonsmodellen kan da skrives

P(LOW = 1) = exp(a + b LWTKG)/([1 + exp(a + b LWTKG)].

Her betyr exp(x) eksponentialfunksjonen av x.

Denne formelen kan også skrives som

ln{P(LOW = 1)/[1 - P(LOW = 1)] = a + b LWTKG,

eller som

P(LOW = 1)/[1 - P(LOW = 1)] = exp(a + b LWTKG)

Venstre side av den første at disse to formlene over kjenner vi igjen som logaritmen til oddsen for at barnet skal ha lav fødselsvekt. Høyre side viser en lineær sammenheng mors vekt.

Når vi nå skal fortolke verdien av regresjonskoeffisienten b, gjør vi dette vi oddsforholdet. Dersom vi endrer LWTKG med 1, vil da exp(b) være endringen i odds for å føde et lite barn (altså for at LOW = 1).

Dersom vi nå formulerer den enkle logistiske regresjonsmodellen med y som avhengig variabel og x som forklaringsvariabel, vil formelen være

ln{P(y = 1)/[1 - P(y = 1)] = a + b x

og fortolkningen av regresjonskoeffisienten b, vil være at dersom vi endrer forklaringsvariabelen med 1 enhet, vil oddsen for at y = 1 endre seg med exp(b) enheter.

Tilsvarende som for enkel og multippel lineær regresjon kan vi finne effekten av én forklaringsvariabel, kontrollert for andre, ved å gjøre en multippel logistisk regresjonsanalyse. Hvis vi generelt har p forklaringsvariabler, vil effekten av én forklaringsvariabel, kontrollert for de p – 1 andre, uttrykkes ved at oddsen for at y = 1 endrer seg med exp(b) enheter, når vi endrer forklaringsvariabelen med 1 enhet.

Gangen i en regresjonsanalyse vil vanligvis være som i lineær regresjon:

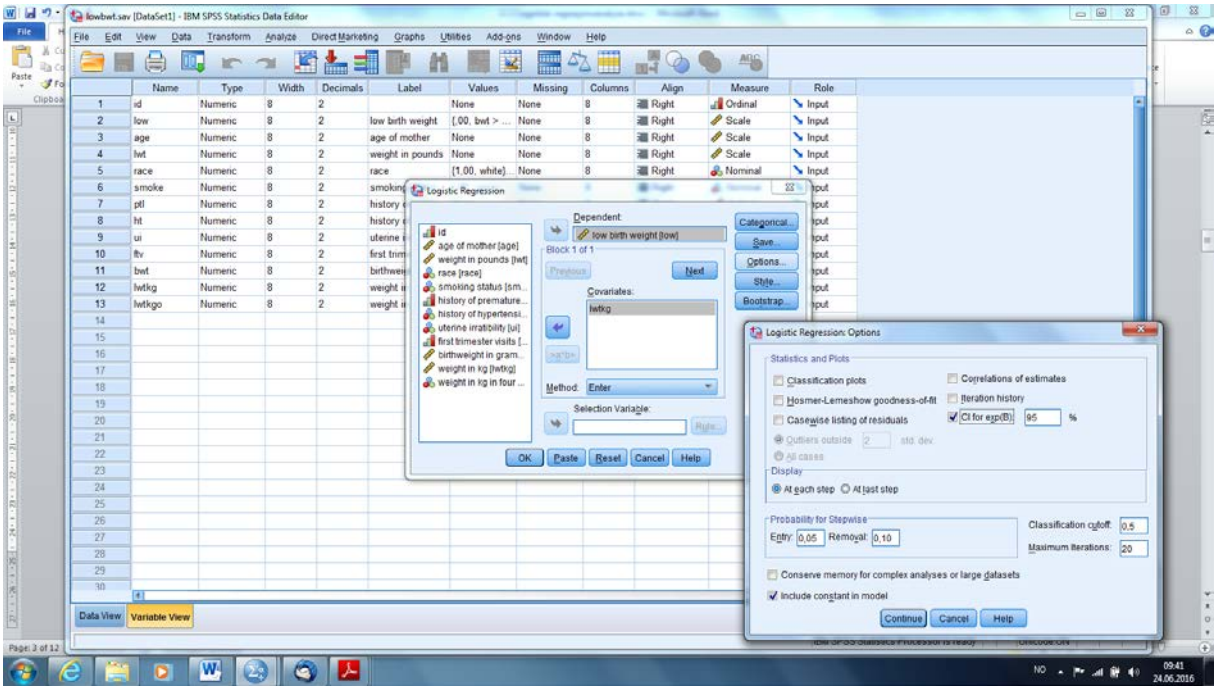
5. Kjør enkel logistisk regresjon med hver enkelt forklaringsvariabel. Dersom p-verdien er < 0.20 tar vi den med videre som en kandidat for en multippel regresjon. Ta også med variable som er av biologisk eller medisinsk interesse, selv om de har en p-verdi > 0.20.
6. Kjør multippel logistisk regresjon med alle variablene som er inkludert på trinn 1.
7. I den multiple regresjonsmodellen tar vi ut den forklaringsvariabelen som har høyest p-verdi, og kjør en multippel regresjon uten den.
8. Gjenta trinn 3 inntil alle forklaringsvariablene er statistisk signifikant, med p < 0.05. Behold forklaringsvariabler som er av biologisk eller medisinsk interesse

12.3.1 Enkel logistisk regresjon. Eksempel: lowbwt.sav

Vi skal i dette eksempelet analysere LOW som avhengig variabel og – som i lineær regresjon – skal vi bruke LWTKG, AGE, SMOKE, HT, RACE, PTLD og FTVD som forklaringsvariabler. SMOKE, HT, PTLD og FTVD er kategoriske variabler med to kategorier, og de trenger vi ikke gjøre noe med. RACE er en kategorisk variabel, med tre kategorier. Programmet for logistisk regresjon er enklere enn programmet for lineær regresjon, så nå trenger vi ikke lage dummy-variabler for RACE.

Da er vi klare til å gjøre trinn 1 i metoden som vi beskrev i innledningen til dette kapittelet. Vi gjør da enkle regresjonsanalyser med hver enkelt av variablene LWTKG, AGE, SMOKE, HT, RACE, PTLD og FTVD.

Vi starter med en enkel logistisk regresjon av LOW på LWTKG. Vi går da inn i *Analyze/Regression/Binary Logistisk*. Vi trekker over LOW i *Dependent* og LWTKG over i *Covariates*. Så gå vi til *Options*. Der klikker vi av på *CI for exp(B)*, og lar det stå 95 der, siden vi ønsker et 95% konfidensintervall. Da ser dialogboksen vår slik ut:



Ved å trykke på *Continue* og *OK*, får vi følgende resultat (som i denne sammenhengen er det viktigste for oss):

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	lwtkg	-,031	,014	5,192	1	,023	,969	,944	,996
	Constant	,998	,785	1,616	1	,204	2,714		

a. Variable(s) entered on step 1: lwtkg.

I tabellen finner vi resultatene fra selve regresjonsanalysen. I kolonnen for B finner vi selve regresjonskoeffisienten for den logistiske regresjonsanalysen. Men den verdien vi skal fortolke finner vi under Exp(B). Denne verdien angir hvor mye oddsen for å få et lite barn (dvs. LOW = 1) endrer seg når vi endrer LWTKG med 1 kg. Vi ser at oddsen reduseres med 3% (1 – 0.969), når mors vekt øker med 1 kg. Konfidensintervallet for endringen i odds er (0.944, 0.996)

Den to-sidige p-verdien finner vi under Sig. Den er p = 0.023. Vi oppsummerer altså med at effekten av mors vekt er redusert odds på 3% for å føde et lite lite barn for hver kilos vektøkning, KI = (0.944, 0.996), p = 0.023.

Da gjør vi en helt tilsvarende analyse for sammenhengen mellom LOW og AGE. Vi går da tilbake til *Analyze/Regression/Binary Logistic*. Vi beholder LOW i *Dependent* men trekker LWTKG tilbake fra *Covariates.* Istedenfor trekker vi AGE over i *Covariates*. Ved å klikke på *OK* får vi følgende resultat.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	age	-,051	,032	2,635	1	,105	,950	,893	1,011
	Constant	,385	,732	,276	1	,599	1,469		

a. Variable(s) entered on step 1: age.

Vi ser av tabellen over at for hvert år mor blir eldre, reduseres oddsen for å føde et lite barn med 5%. Siden p = 0.105, som er > 0.05, er sammenhengen mellom fødselsvekt og mors alder ikke statistisk signifikant. Dette får vi også bekreftet ved å se på konfidensintervallet som dekker verdien 1. Husk at for oddsforholdet er «nullverdien» lik 1. Men p-verdien er lavere enn det kravet som vi satte over. Men mors alder er en viktig biologisk og medisinsk forklaringsvariabel så vi velger å ta denne med videre i analysen, selv om p-verdien hadde vært > 0.20.

Så går vi videre med variabelen SMOKE. Dette er en to-kategorisk variabel. Som for lineær regresjon, kan den taes direkte inn i regresjonsanalysen, og regresjonskoeffisienten kan tolkes som effekten SMOKE når den går fra 0 (ikke-røyker) til 1 (røyker).

Vi gjør da som over og går til *Analyze/Regression/Binary Logistic*. Vi trekker AGE tilbake fra *Covariates* og trekker over SMOKE. Ved å klikke på *OK* får vi følgende resultat:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	smoke	,704	,320	4,852	1	,028	2,022	1,081	3,783
	Constant	-1,087	,215	25,627	1	,000	,337		

a. Variable(s) entered on step 1: smoke.

Av tabellen ser vi at effekten av SMOKE på LOW er statistisk signifikant, p = 0.028. Vi oddsen for å få et lite barn mer enn dobles (2.022) når mor røyker. Et 95% konfidensintervall for økningen i odds er (1.08, 3.78).

Denne analysen kan vi også gjøre ved en analyse av krysstabellen mellom LOW og SMOKE, se avsnitt 11.4.2. Da finner vi følgende krysstabell og oddsforhold:

low birth weight * smoking status Crosstabulation

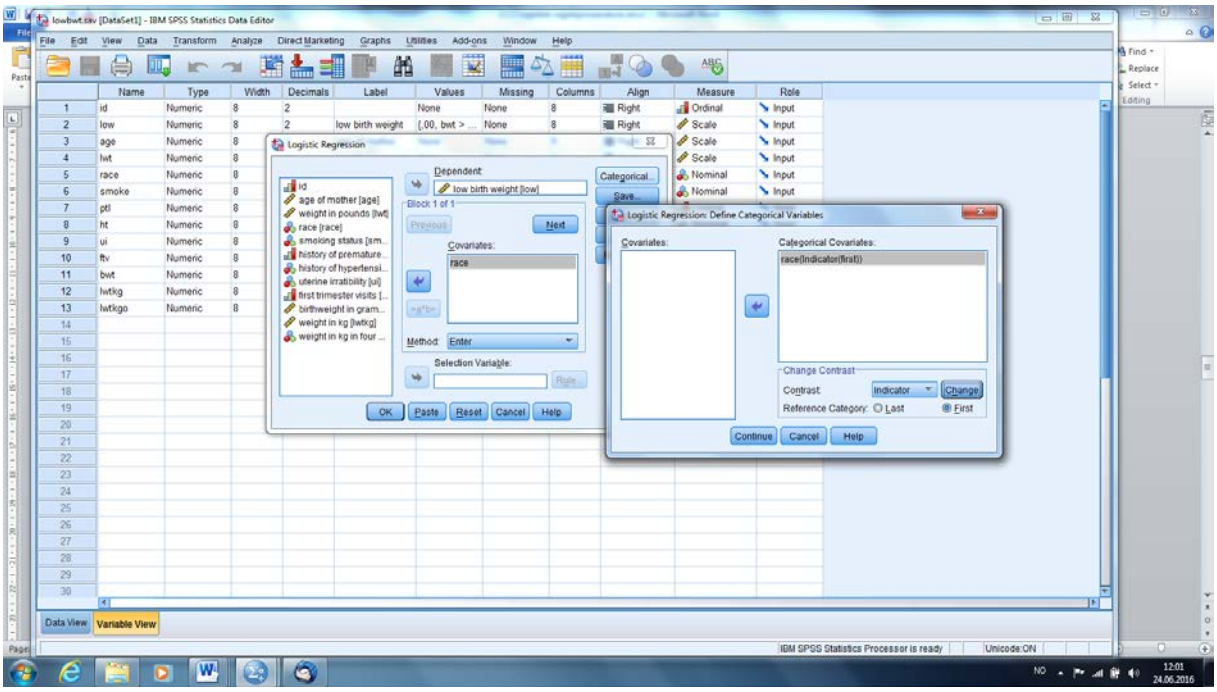
			smoking status		Total
			non-smoking	smoking	
low birth weight	bwt > 2500g	Count	86	44	130
		% within smoking status	74,8%	59,5%	68,8%
	bwt < 2500g	Count	29	30	59
		% within smoking status	25,2%	40,5%	31,2%
Total		Count	115	74	189
		% within smoking status	100,0%	100,0%	100,0%

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for low birth weight (bwt > 2500g / bwt < 2500g)	2,022	1,081	3,783
For cohort smoking status = non-smoking	1,346	1,010	1,794
For cohort smoking status = smoking	,666	,470	,942
N of Valid Cases	189		

Her ser vi oddsforholdet er 2.022, med et konfidensintervall på (1.08, 3.78). Dette er det samme som vi fant ved logistisk regresjon. Som for lineær regresjon, der vi hadde at en regresjonsanalyse med en kategorisk forklaringsvariabel med to kategorier er det samme som å gjøre en t-test for to uavhengige utvalg, finner vi her at det er det samme å lage en krysstabell og beregne OR derfra, som å gjøre en logistisk regresjonsanalyse.

La oss så gå videre med RACE. Her må vi huske at RACE er en kategorisk variabel med tre kategorier. Vi går da inn i *Analyze/Regression/Binary Logistic*. Vi trekker SMOKE tilbake fra *Covariates* og trekker over RACE i *Covariates*. Nå må vi markere at RACE er en kategorisk variabel. Det gjør vi ved klikke av på *Categorical*. Her trekker vi RACE over i *Categorical Covariates*. Til slutt må vi merke av på at vi ønsker å ha første kategori som referansekategori. Dette er i samsvar med slik vi har definert dummy-variablene. Da vi trykke på *Change*. Da ser dialogboksen vår slik ut:



Ved å klikke på *Continue* og *OK* får vi følgende resultat:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	race			4,922	2	,085			
	race(1)	,845	,463	3,323	1	,068	2,328	,939	5,772
	race(2)	,636	,348	3,345	1	,067	1,889	,955	3,736
	Constant	-1,155	,239	23,330	1	,000	,315		

a. Variable(s) entered on step 1: race.

Aller først ser vi at p-verdien for RACE, samlet sett, er 0.085. Variabelen er ikke signifikant, men vi tar den med videre i analysen. Deretter ser vi at oddsforholdet for de to variablene RACE(1) = Black og RACE(2) = Other, vurdert i forhold til referansen RACE = White, er omtrent like store, og ca. 2.0.

Da går vi videre med HT. Da finner vi følgende resultat.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	ht	1,214	,608	3,979	1	,046	3,365	1,021	11,088
	Constant	-,877	,165	28,249	1	,000	,416		

a. Variable(s) entered on step 1: ht.

Effekten av HT på BWT er statistisk signifikant, p = 0.046. Det å være hypertensive (HT =1) reduserer oddsen for å få det lite barn, til omtrent det halve.

Vi må fortsette med PTLD. For den variabelen finner vi følgende resultat:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	ptld	1,463	,414	12,455	1	,000	4,317	1,916	9,726
	Constant	-1,057	,181	34,003	1	,000	,347		

a. Variable(s) entered on step 1: ptld.

Vi ser at effekten av PTLD er klart statistisk signifikant (p < 0.001). Når PTLD = 1 får vi økt odds på 146% for lav fødselsvekt på barnet.

Til slutt analyserer vi FTVD. Da får vi på lignende måte:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	ftvd	-,479	,319	2,247	1	,134	,620	,331	1,159
	Constant	-,575	,208	7,627	1	,006	,563		

a. Variable(s) entered on step 1: ftvd.

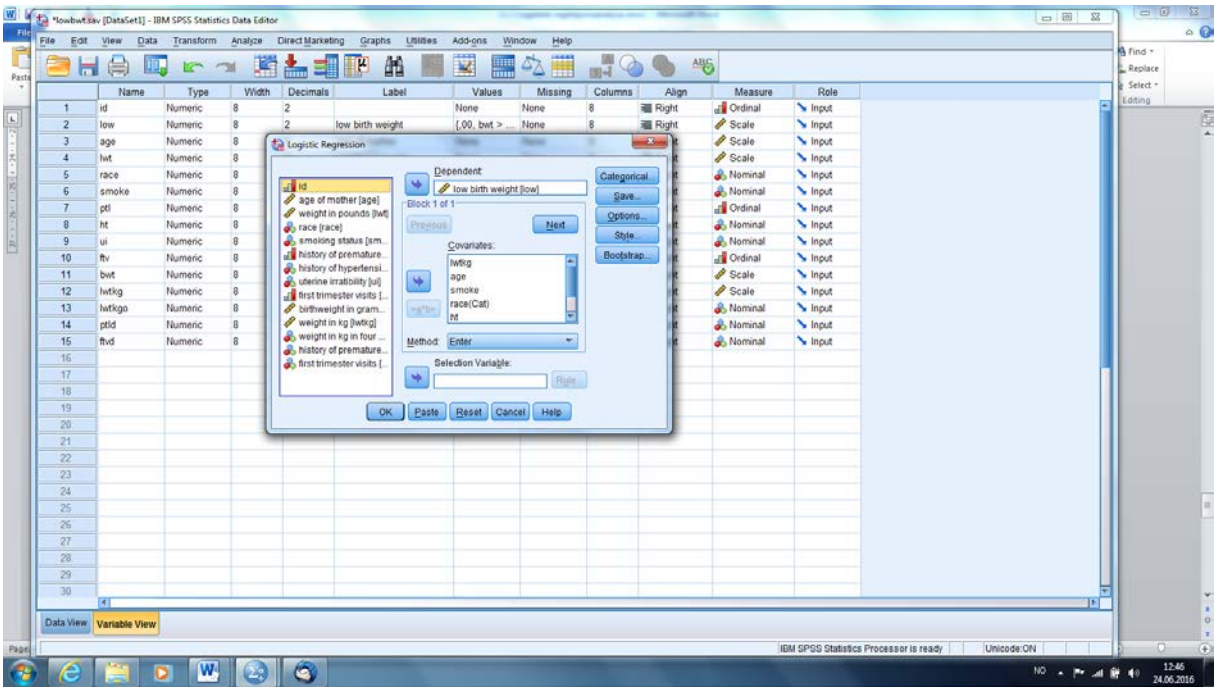
Effekten av FTVD er ikke statistisk signifikant (p = 0.134). Når FTVD = 1 får vi en klart odds for lav fødselsvekt på barnet, nemlig en reduksjon på 38%.

12.3.2 Multippel regresjon. Eksempel: lowbwt.sav

Nå har vi gjort alle de enkle logistiske analysene som vi må for å kunne vurdere effekten i forhold til den avhengige variabelen BWT. Ikke alle effektene var statistisk signifikante, men alle p-verdiene var < 0.20. I henhold til punkt 2 i oversikten i innledningen til dette kapittelet tar med alle variablene inn i en multippel logistisk regresjonsanalyse.

Da går vi igjen til *Analyze/Regression/Binary Logistic*. Vi trekker nå over LWTKG, AGE, SMOKE, RACE, HT, PTLD og FTVD over i *Covariates*. Merk at RACE nå er tatt inn som en kategorisk variabel, som vises ved RACE(CAT).

Da ser dialogboksen vår slik ut:



Når vi klikker på *OK*, får vi følgende resultat.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	lwtkg	-,034	,016	4,778	1	,029	,966	,937	,996
	age	-,040	,038	1,082	1	,298	,961	,892	1,036
	smoke	,818	,417	3,854	1	,050	2,266	1,001	5,130
	race			5,616	2	,060			
	race(1)	1,158	,533	4,715	1	,030	3,184	1,119	9,055
	race(2)	,775	,454	2,910	1	,088	2,170	,891	5,283
	ht	1,723	,705	5,979	1	,014	5,604	1,408	22,304
	ptld	1,349	,460	8,594	1	,003	3,852	1,564	9,491
	ftvd	-,157	,372	,177	1	,674	,855	,413	1,772
	Constant	,959	1,211	,628	1	,428	2,610		

a. Variable(s) entered on step 1: lwtkg, age, smoke, race, ht, ptld, ftvd.

Når vi går til neste tabell ser vi AGE, RACE og FTVD ikke er statistisk signifikante. AGE har en p-verdi på 0.298, RACE en p-verdi på 0.060 og FTVD på 0.764. Etter regelen i punkt 3 skal vi nå ta ut begge disse variablene. Men vi har bestemt oss for at AGE er så viktig at vi lar

den være med i den videre analysen. Men vi tar nå ut RACE og FTVD av listen i *Covariates*. Når vi klikker på *OK* får vi følgende resultat:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	lwtkg	-,033	,015	4,865	1	,027	,967	,939	,996
	age	-,056	,036	2,441	1	,118	,946	,882	1,014
	smoke	,497	,347	2,054	1	,152	1,644	,833	3,243
	ht	1,783	,703	6,423	1	,011	5,946	1,498	23,606
	ptld	1,418	,449	9,987	1	,002	4,130	1,714	9,954
	Constant	1,797	1,091	2,711	1	,100	6,030		

a. Variable(s) entered on step 1: lwtkg, age, smoke, ht, ptld.

Vi ser at SMOKE nå er ikke-signifikant (p = 0.118). Men røyking er en viktig forklaringsvariabel og vi beholder den inne i modellen. Da er alle alle variablene (med unntak også av AGE som vi også vil ha med) statistisk signifikante.

Effekten av hver forklaringsvariabel må nå tolkes som effekten av den forklaringsvariabelen når vi kontrollerer for (dvs. tar bort) effekten av de andre forklaringsvariablen. Altså er effekten av røyking (SMOKE) at oddsen for å føde et lite barn øker med en faktor på 1.644 (p = 0.118, 95% KI = (0.833, 3.243)) når vi kontrollerer for effekten av alle de andre forklaringsvariablene.

Tilsvarende er effekten av mors vekt at en vektøknig på 1 kg for mor, betyr at oddsen for at mor skal få et barn med lav fødselsvekt redusert med ca. 3% (p = 0.027, 95% KI = (0.939, 0.996). Slik kan vi forklare alle variablene som nå har vist seg å være viktige for forklaringen av fødselsvekt.

12.3 Overlevelsesanalyse

Overlevelsesanalyse spiller en viktig rolle i medisinsk forskning. Vi registrerer tiden til en begivenhet, som sykdom eller død. Siden ikke alle individene som vi undersøker vil oppleve den begivenheten vi studerer, for eksempel vil ikke alle individene være døde før vi avslutter studien, tiden før begivenheten inntreffer, har vi vanligvis *Missing values*. Dette kaller vi i overlevelsesanalysen for sensurering. Vi må da alltid ha med en variabel i datasettet som viser om vi har observert tiden til begivenheten eller oppfølgingstiden til begivenheten faktisk er missing.

I overlevelsesanalyse vil vi være interessert i plot overlevelseskurven, som vi kaller Kaplan Meier plott. Vi vil være interessert i om overlevelseskurvene er forskjellige mellom to eller flere grupper, som vi skal teste ved en såkalt log-rank test. Til slutt vil vi være interessert i tid for overlevelse i forhold til flere forklaringsvariabler. Dette er en variant av regresjonsanalyse, som vi kaller Cox-regresjon.

I dette kapittelet skal vi bruke datasettet CVD Risk fra nettsiden http://www.umass.edu/statdata/statdata/stat-survival.html. Oppfølgingstiden til død etter innleggelse på sykehus er registrert for 65 pasienter. Oppfølgingstiden er registrert i variabelen TIME. Det er 32 pasienter som døde av kardiovaskulær sykdom (CVD) og de andre 33 er sensurerte observasjoner. Noen av disse er sensurerte fordi de er døde av andre årsaker, mens for andre har vi ikke registrert tid til død. Informasjon om sensurering ligger i variabelen EV_TYP. Kodene for EV_TYP er EV_TYP = 1 for Death of CVD, EV_TYP = 2 for Death of other Causes, EV_TYP = 3 for Censored. Vi har også data om alder (AGE), og kjønn (GENDER, GENDER = 0 for male, GENDER = 1 for female). Denne datafilen ligger på nettsiden for kurset under navnet **cvdrisk.sav**. Beskrivelsen av filen på hjemmesiden er som følger:

CVD Risk Data (cvdrisk.dat)

SIZE:
65 Observations, 6 Variables

REFERENCE:
Hosmer, D.W. and Lemeshow, S. and May, S. (2008): Applied Survival Analysis: Regression Modeling of Time to Event Data: Second Edition, John Wiley and Sons Inc., New York, NY

DESCRITPTIVE ABSTRACT:
Description of the variables in the CVD Risk competing risk example in Section 9.6 of reference text.

DISCLAIMER:
This data is also available at the following Wiley's FTP site:
ftp//ftp.wiley.com/public/sci_tech_med/survival

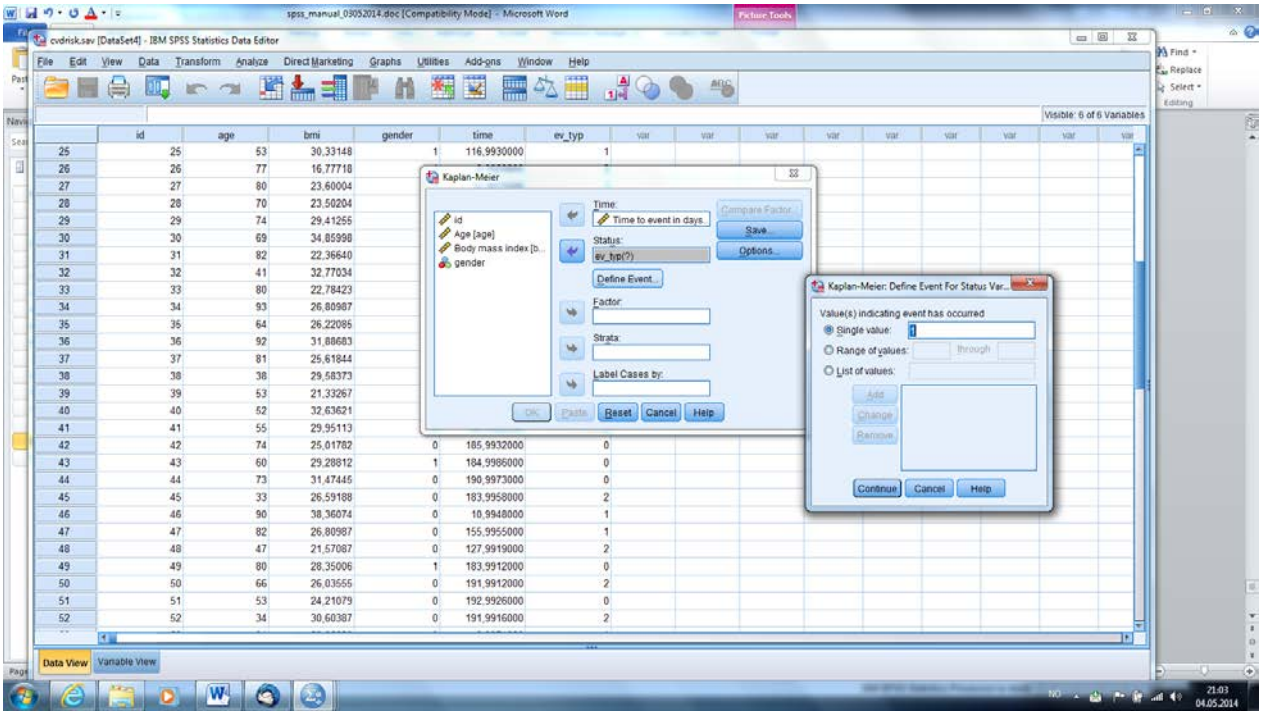
LIST OF VARIABLES:

Variable	Name	Description	Codes/Values
1	id	Study ID	1 - 65
2	age	Age	Years
3	bmi	Body Mass Index	kg/m**2
4	gender	Gender	0 = Male, 1 = Female
5	time	Follow Up Time	Days
6	ev_typ	Event Type	1 = CVD 2 = Other Cause 0 = Censor

12.3.1 Overlevelsesanalyse. Eksempel: cvdrisk.sav

Vi henter frem datafilen **cvdrisk.sav.** Først er vi interessert i å lage et overlevelsesplott for død av CVD. Da går vi inn i *Analyze/Survival/Kaplan-Meier*. Her trekker vi over TIME i vinduet med *Time*, og EV_TYP over i *Status*. Da åpner det seg et vindu med *Define Event*.

Her må vi angi hva som er begivenheten vår. Vi skriver derfor inn 1 i vinduet Single event, slik at dialogboksen ser slik ut:

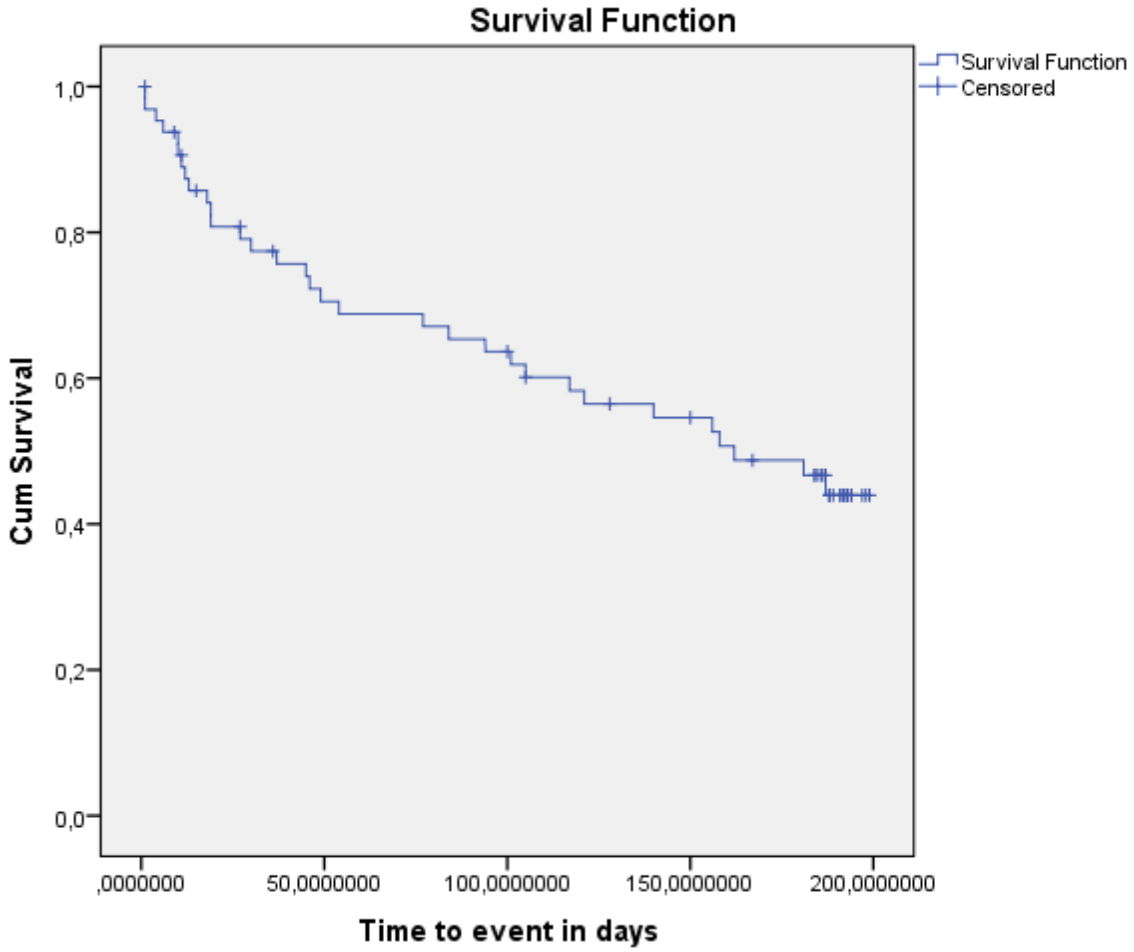


Da klikker vi på *Continue* og *OK*. Da får vi følgende resultater.

Means and Medians for Survival Time

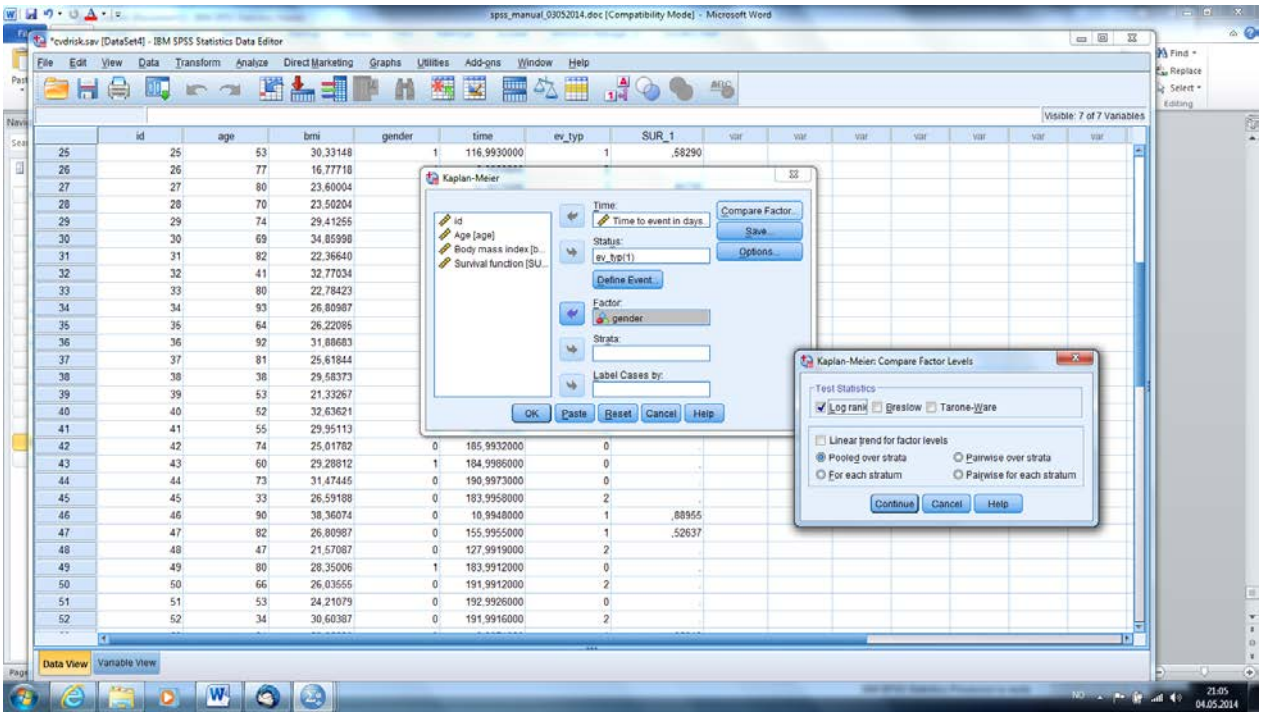
Mean[a]				Median			
		95% Confidence Interval				95% Confidence Interval	
Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
127,511	10,098	107,719	147,302	161,991	35,033	93,327	230,655

a. Estimation is limited to the largest survival time if it is censored.



I den første tabellen får vi oversikt over gjennomsnittlig og median tid for overlevelse. Vi ser at gjennomsnittet er 127

La oss nå de på sammenligningen av overlevelse for de to kjønnene. Da skal vi lage en log-rank test for å gjøre sammenligningen. Vi lager også et Kaplan-Meier plott for de to gruppene, for å få en visuell fremstilling av overlevelsen. Dette får vi frem ved å gå til *Analyze/Survival/Kaplan-Meier*, og trekke TIME over i *Time* og GENDER over i *Factor*. Vi beholder EV_TYP = 1 som koden for *Event*, under *Status*. Etter det går vi til *Compare Factors*. Der klikker vi at på *Log-rank* øverst til venstre. Så klikker vi på *Continue*. Til slutt går vi til *Options*. Her klikker vi av på *Survival* under *Plots*. Da ser dialogboksen vår slik ut:



Etter *Continue* og *OK* får vi følgende utskrift:

Means and Medians for Survival Time

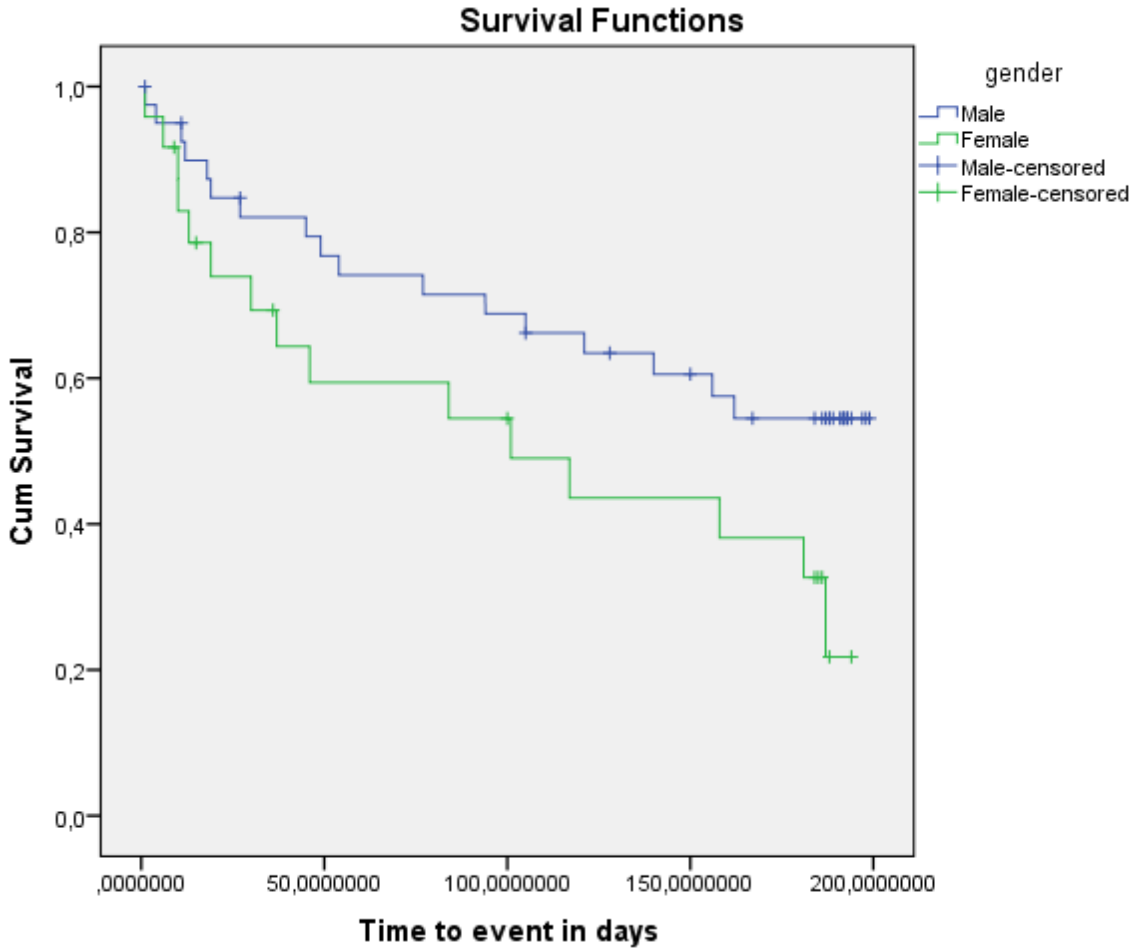
gender	Mean[a]				Median			
			95% Confidence Interval				95% Confidence Interval	
	Estimate	Std. Error	Lower Bound	Upper Bound	Estimate	Std. Error	Lower Bound	Upper Bound
Male	139,057	12,151	115,241	162,874	.	.	.	.
Female	105,246	16,675	72,563	137,928	100,991	49,197	4,565	197,416
Overall	127,511	10,098	107,719	147,302	161,991	35,033	93,327	230,655

a. Estimation is limited to the largest survival time if it is censored.

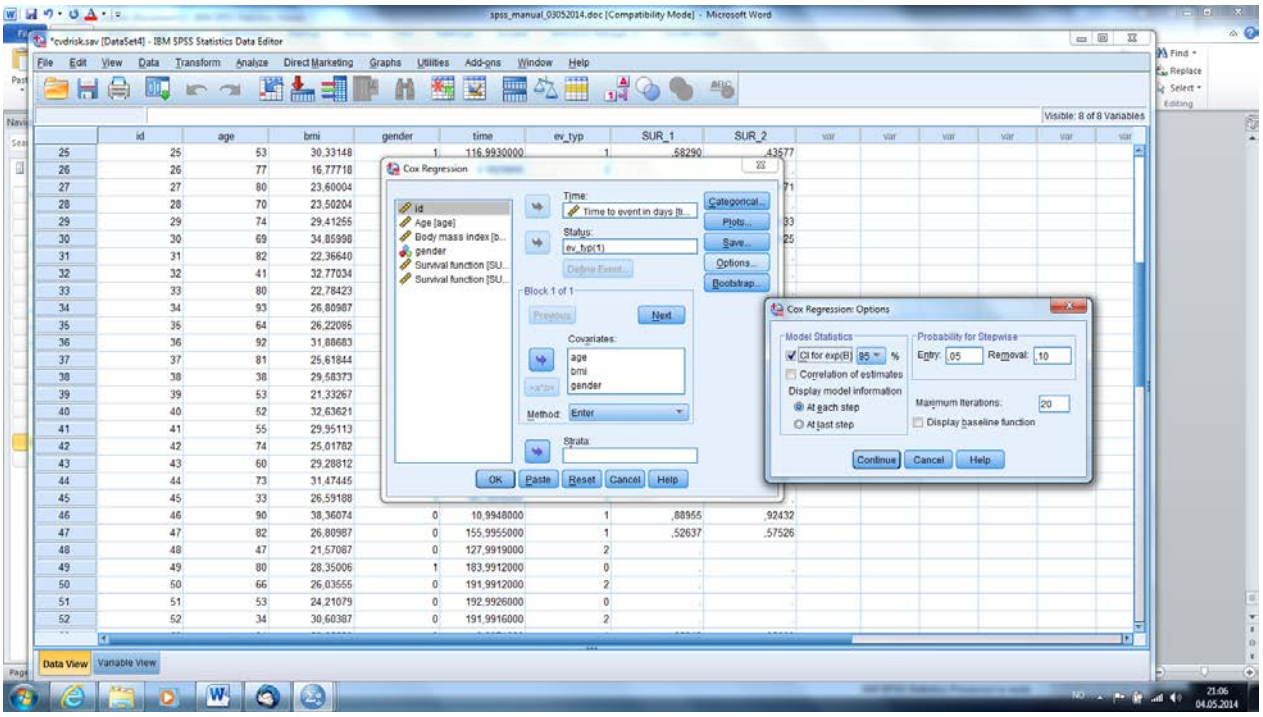
Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	3,824	1	,051

Test of equality of survival distributions for the different levels of gender.



Til slutt skal vi kjøre en Cox regresjon med variablene AGE, BMI og GENDER. Da går vi inn i *Analyze/Survival/Cox Regression*. Som før legger vi over TIME i *Time* og CENSORING i *Status*. I *Define Event* bruker vi fortsatt 1 som Event. Vi trekker over AGE, BMI og GENDER i *Covariates*. Før vi gjør analysen må vi klikke på *Options*. Der klikker vi at på CI for exp(B), som betyr at vi skal ha presentert konfidensintervallet for hazard ratio. Det er nemlig hazard ratio som er effektmålet i en Cox regresjon, og da vet vi at vi må få presentert selve hazard ratioen, konfidensintervallet for den og p-verdien. Da ser dialogboksen vår slik ut:



Vi klikker på Continue og OK, og får resultatet:

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)	95,0% CI for Exp(B)	
							Lower	Upper
age	,059	,013	21,542	1	,000	1,061	1,035	1,088
bmi	,094	,037	6,469	1	,011	1,099	1,022	1,181
gender	,307	,380	,654	1	,419	1,360	,646	2,863

Vi ser at AGE er klart statistisk signfikant ($p < 0.001$), og BMI er det også ($p = 0.011$). Vi ser at hazard ratioen for alder er 1.06, mens den er 1.10 for BMI. Dette betyr at risikoen for å død av CVD øker med 6 prosent for hvert dag som går og med 10% for hver enhet i økt BMI.

Men GENDER har ingen signifikant betydning for risikoen for død av CVD. Dette er kanskje litt rart, siden vi fant en mye klarere, men likevel ikke signifikant effekt, i logrank testen. Dette skyldes at når vi kontrollerer for AGE og BMI, blir effekten av GENDER veldig redusert, og så mye redusert at den er langt fra å være statistisk signifikant.

13 Oppsummering om de forskjellige filtyper i SPSS

Læringsmål

Ved bruk av SPSS kommer vi i kontakt med en rekke forskjellige filer som vi gir en oversikt over her. De forskjellige filtypene må ikke blandes sammen. For at vi selv skal vite hva slags filtype vi har foran oss, lønner det seg derfor å være konsekvente med hvilken ekstensjon (dvs. de tre tegnene etter punktum i filnavnet) filene får. Her følger vi internasjonal standard og de regler SPSS setter opp. SPSS foreslår selv som regel riktig filekstensjon. I oversikten under står filekstensjonen i fete typer.

13.1 Oversikt over filtypene i SPSS

RÅDATAFILER Dette er filer som bare inneholder tall. De er i et format i samsvar med

ASCII-filer en amerikansk standard og kalles derfor også ASCII-filer (American

*****.DAT** Standard Code for Information Interchange). Vi leser disse filene i SPSS ved å klikke på *File/Read Text Data*. Avhengig av hvordan filen er bygget opp leses den i fritt eller fast format. ASCII-filtypen sikrer muligheten for å utveksle data mellom de aller fleste dataprogrammer. I kapittel 4 gikk vi gjennom hvordan vi går fra en ASCII-fil til en SPSS-fil.

SPSS DATAFILER Dette er filer som *bare* kan leses av SPSS. De inneholder både data,

******.SAV** variabel navn, variabel label og value label. De leses inn i SPSS gjennom *File/Open*, se kapittel 4. Vi har gitt en rekke eksempler på slike filer, for eksempel **lowbwt.sav** og **altman.sav**.

SPSS ORDREFILER Dette er filer som inneholder SPSS-ordrer. En SPSS-ordre kan

******.SPS** vi enten skrive selv i et ordrevindu eller la SPSS lage for oss ved å *Paste* innholdet i en dialogboks. En samling av slike ordrer som lagres på en fil til seinere bruk er en ordre-fil. Den kan utføres samlet av SPSS. Se kapittel 8.

SPSS lager automatisk alle ordrene som utføres fortløpende. Disse skrives ut på resultatfilen vår. Her finner vi syntaxen for alt vi bruker SPSS til. Filen kan være nyttig å gå inn på når vi lurer på hva vi har gjort, eller hvis vi vil gjenta analysene fra før. Men problemet med denne utskriftsfilen er at vi ikke kan kopiere fra denne og inn i en ordrefil.

RESULTATFILER	Resultatene fra de analyser vi får SPSS til å gjennomføre, skrives ut i
******.SPV**	Resultatvinduet. Hvis vi lagrer dette på disk, får vi en listefil. Den kan redigeres og overføres til andre tekstbehandlere for videre bearbeiding, se kapittel 5.