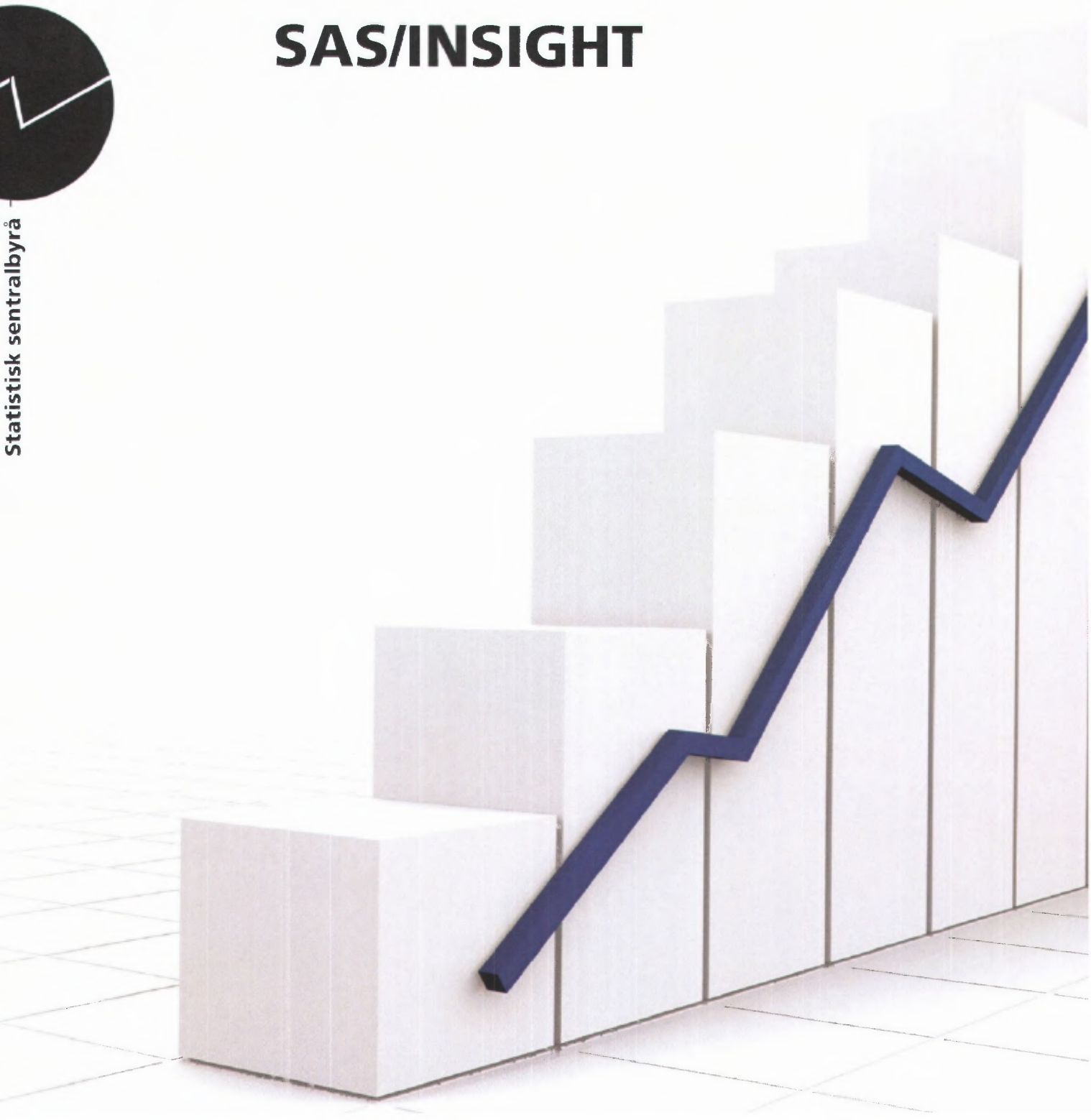




SAS/INSIGHT



Forord

SAS/INSIGHT kan være et nyttig verktøy for mange, fra erfarne SAS-brukere til de som unngår SAS og heller bruker regneark.

SAS/INSIGHT er lett å tilegne seg samtidig som den gir mange muligheter innen analyse og grafikk.

Denne håndboken som er en videreføring av Interne dokumenter 2000/1, er ment å dekke flere behov:

- vise noen av mulighetene SAS/INSIGHT har for grafisk presentasjon og analyse av data,
- være en enkel oppslagsbok i SAS/INSIGHT for brukere i SSB,
- være en introduksjonsbok som kan brukes til kurs i SAS/INSIGHT ved Byråskolen.

SAS/INSIGHT brukes fra SAS på UNIX eller PC. Det er noen få forskjeller mellom brukermiljøene, det som har betydning er beskrevet.

Ytterligere dokumentasjon av SAS/ INSIGHT er tilgjengelig på Byrånettet - IT-info - SAS - SAS OnlineDoc, versjon 8.

Johan Heldal har gitt bidrag til håndboken, mens flere andre i seksjon 120 har kommet med tips og kommentarer.

Oslo, september 2007.

Anne Sofie Abrahamsen

Innhold

1. Innledning	4
1.1. Hva er SAS/INSIGHT	4
1.2. Bruk av SAS/INSIGHT	4
2. Komme i gang	5
2.1. Starte opp SAS/INSIGHT	5
2.2. Filstørrelser til SAS/INSIGHT?	5
3. Datavinduet.....	6
3.1. Informasjon i datavinduet.....	7
3.2. Endring av data i datavinduet	8
3.3. Markere variable/observasjoner	11
3.4. Transformasjon av variable	11
4. Histogram/stolpediagram	13
4.1. Nominelle variable	13
4.2. Menyer for redigering.....	14
4.3. Stolpehøyde etter andre variable	16
4.4. Markering	17
4.5. Intervallvariable.....	18
5. Boksplott/Mosaikkplott	19
5.1. Boksplott	19
5.2. Mosaikkplott.....	22
5.3. Tester for sammenligning av data fra flere grupper (kategorier)	23
6. X - Y - plott	25
6.1. Farger og symbol	26
6.2. Identifisering av punkter i plottet	26
6.3. Fjerne observasjoner fra plottet	27
6.4. Hente tilbake observasjoner i plottet	27
6.5. Transformasjon av variable	27
6.6. Punktstørrelse	28
6.7. Forstørrelse av plottet	28
6.8. Plottmatrise	28
7. Grafisk analyse av data	29
7.1. Flere grafer av samme type i et vindu	29
7.2. Analyse ved hjelp av stolpediagram	30
7.3. Fryse bildet	33
8. Analyse av en variabel	34
8.1. Fordeling	35
9. Multivariabel analyse.....	41
10. Modelltilpassing.....	45
10.1. Enkel lineær regresjon.....	45
10.2. Multippel regresjon/Variansanalyse.....	50
10.3. Logistisk regresjon	52
10.4. Modellkontroll.....	55

11.Linjeplott..... 56

12.Grafikk i flere dimensjoner 59

12.1. Roterende plott 59

12.2. Farger som en ekstra dimensjon 61

12.3. Konturplott 62

13.Lagring 63

13.1. Lagre data som et SAS-datasett 63

13.2. Lagre grafer og tabeller 64

13.3. Kjøre og lagre SAS/INSIGHT rutiner..... 66

13.4. SAS – kommandoer for automatisk SAS/Insight - analyse..... 66

14.Utskrift 68

14.1. Direkte utskrift. 68

14.2. Utskrift via Word 69

15.Tips 70

1. Innledning

1.1. Hva er SAS/INSIGHT

SAS/INSIGHT er et interaktivt vindu/meny-basert verktøy for å studere og analysere data. Det er en modul i SAS og baseres på basisversjonen av SAS. SAS/INSIGHT kan brukes på alle SAS - datasett. I SSB er SAS/INSIGHT tilgjengelig fra SAS på PC eller fra SAS på Unix.

De viktigste rutineene er:

- Deskriptiv analyse - dvs. å beskrive egenskaper i et datasett
 - grafisk analyse - histogram, stolpediagram og ulike typer plott
 - markere og eventuelt utelate observasjoner i figurer og analyser
 - beregne gjennomsnitt, standardavvik, median, kvartiler og andre parametre
- Analysere fordelinger
 - tilpasse fordelingsfunksjoner (normalfordeling, lognormal, eksponensiell og Weibull fordeling, Kernel funksjoner)
 - teste hypoteser
 - grafisk Q-Q plott for fordelingsfunksjoner
- Analysere sammenhenger mellom variable
 - regresjonsanalyse - multipel regresjon, logistisk regresjon, variansanalyse etc.
 - kurvetilpasning - parametriske og ikkeparametriske
 - korrelasjonsmatriser og korrelasjonsanalyser
 - prinsippal komponent analyse

1.2. Bruk av SAS/INSIGHT

SAS/INSIGHT er et godt verktøy for å gjøre seg kjent med et datasett gjennom beskrivende grafikk og analyse. Man får lett oversikt over variable, verdiene de kan ha, fordeling av verdiene i datasettet og hvordan sammenhengen er mellom ulike variable. Utliggere kan identifiseres og effekten av utliggere kan analyseres ved å utelate dem fra analysen og eventuelt hente dem tilbake.

SAS/INSIGHT er svært lett å lære seg. Det er ikke nødvendig å kunne SAS (base), det er nok å ha et SAS-datasett. Grafer og analyser lages via menyer. Det er enkelt å redigere datasettet, sortere observasjoner, transformere variable og trekke ut deler av datasettet for videre analyse. Det er likevel en fordel å kunne litt SAS i tillegg, for eksempel når man ønsker en graf eller analyse som ikke er definert i menyene. Dette kan noen ganger løses ved å tilpasse datasettet.

SAS/INSIGHT har en mengde menyer. Standard rullegardinmenyer for «File», «Edit» og «Analyze» vises i alle vinduer. I tillegg er det spesialmenyer som kan hentes for et enkelt vindu og for enkelte funksjoner i vinduene. Flere menyvalg kan føre frem til samme analyse, og selv om mange menyvalg beskrives her, får vi ikke med alle. Man kan søke i menyene etter muligheter, begrensningene følger av manglende alternativer i menyene.

Grafer og beregninger som kan lages i SAS/INSIGHT kan også lages i basis - SAS, SAS/STAT, SAS/GRAPH eller andre SAS-moduler. Der får du imidlertid et helt annet brukergrensesnitt og ingen direkte koblinger mellom grafer og tabeller.

SAS/INSIGHT er et analyseverktøy. Resultatene i form av filer, tabeller eller grafikk kan lagres, redigeres og overføres til andre program som for eksempel Word.

2. Komme i gang

2.1. Starte opp SAS/INSIGHT

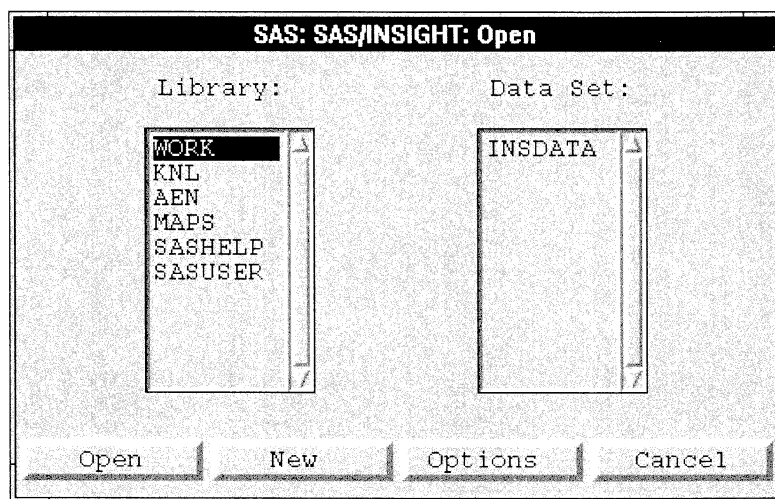
Start SAS på vanlig måte på PC eller på Unix.
Kjør dine libname - statements.

Det er to måter å starte SAS/INSIGHT på. Den vanlige måten (i) er via menyer i Program Editoren, den andre (ii) er ved å kjøre prosedyren "Proc insight" fra Program Editoren.

(i) SAS/INSIGHT startes fra «Program Editor» - vinduet via menyen:
Solutions → Analysis → Interactive Data Analysis.

Det åpnes et vindu hvor tilgjengelige libraries og SAS-datasett listes:

Figur 2.1. Åpning av SAS/INSIGHT



Her velger man et «Library» og datasettene listes under «Data Set». Når du velger det datasettet du skal arbeide med og klikker på «Open», kommer du direkte inn i datavinduet til SAS/INSIGHT. Alternativt kan man velge «New» og få opp et tomt datavindu. Der kan dataene skrives rett inn. Fordelen ved denne vegen inn i SAS/INSIGHT er at man kan bruke SAS til andre program/analyser samtidig som man er inne i SAS/INSIGHT.

(ii) SAS/INSIGHT kan startes ved prosedyren "Proc insight" i program editoren. Dette er mest hensiktsmessig hvis det er en SAS/INSIGHT-analyse som skal gjentas mange ganger. Analysen kan både lages og kjøres ved "proc insight" (Se kapittel 13.3 og 13.4). Vær oppmerksom på at du må gå ut av SAS/INSIGHT før andre program kan kjøres videre.

2.2. Filstørrelser til SAS/INSIGHT?

SAS/INSIGHT er en modul i SAS og baseres på basisvesjonen av SAS. Det interaktive grafiske brukergrensesnittet gjør at SAS/INSIGHT krever store dataressurser. Det begrenser størrelsen på filer som kan behandles i SAS/INSIGHT. Vi kan ikke gi absolutte grenser for filstørrelse, det avhenger av maskinkapasitet (memory). Når datamaskinene får økt kapasitet, øker også størrelsen på filer som kan brukes i SAS/INSIGHT. Selv om en fil kan lastes inn i SAS/INSIGHT, er det ikke sikkert at man kan gjøre alle analyser som ønskes. Det går greit å prøve seg frem, du får beskjed om filen er for stor for innlasting eller du kastes rett ut av SAS. Hvis det datasettet du har tenkt å arbeide med er for stort, kan

du kanskje selekttere ut mindre datasett, for eksempel ved å analysere en gruppe(næring, alder, oa.) om gangen. Man bør også legge tilside variable som ikke er nødvendige i analysen, før dataene hentes inn i SAS/INSIGHT.

3. Datavinduet

Muligheter i datavinduet:

- Studere og justere datasettet
 - endre verdier
 - endre og lage nye variable
 - finne og markere observasjoner
- Utgangspunkt for analyser

Figur 3.1. Datavinduet

SAS: WORK.INSDATA									
File Edit Analyze Tables Graphs Curves Vars Help									
	13	Nom	Nom	Int	Nom	Int	Int	Int	
287		BOK2	FYLKE	ANTSYSS	STRATUM	OMSETN	TILGANG	EFFEKT	
■	1	DB	12	131	1	55850	35153	38528	
■	2	DB	06	106	1	27970	19760	13350	
■	3	DB	11	89	1	25555	14411	12341	
■	4	DB	05	147	1	37148	21324	19832	
■	5	DB	10	134	1	38926	20500	20500	
■	6	DB	11	19	1	5567	6100	1800	
■	7	DB	18	140	1	51880	23611	37009	
■	8	DB	15	43	1	17833	6500	7400	
■	9	DB	01	50	1	17822	18500	17900	
■	10	DB	01	95	1	32995	23699	28201	
■	11	DB	12	23	1	11613	0	0	
■	12	DB	07	219	1	99564	63791	63791	

Hele SAS-datasett vises i Datavinduet. Observasjonene listes under hverandre med observasjonsnummer helt til venstre i tabellen. Navn på alle variable listes mot høyre med variabelverdiene i kolonner under.

Datasettet vi bruker som eksempel i mesteparten av dette heftet er plukket ut fra et utvalg til ordrestatistikken. Variabelnavnene står for:
BOK2 = 2-bokstav næring,
STRATUM = størrelsesgruppe (1 = totaltelling, dvs min 100 sysselsatte eller min 10% av omsetningen på publiseringsnivå, 2 = {50 - 99 syss}, 3 = {20 - 49 syss}, o.s.v),
TILGANG = ordretilgang,
EFFEKT = effektuert.
Resten av variabelnavnene anses som selvforklarende.

For variable som er definert med «Labels» i SAS-datasettet, kan man velge om disse skal vises i datavinduet i tillegg til variabelnavnet.

3.1. Informasjon i datavinduet

Vi kan lese en del informasjon om datasettet fra datavinduet:

3.1.1. Størrelsen på datasettet



Tallene øverst i venstre hjørne viser at dette datasettet inneholder 287 observasjoner og hver observasjon består av 13 variable.

3.1.2. Variable

I et SAS-datasett er variablene definert som numeriske eller karaktervariable. I SAS/INSIGHT defineres de som enten nominelle variable eller intervallvariable. Alle nominelle variable blir brukt som klassevariable både ved beregninger og grafikk. For at en variabel skal fungere som vanlig tall ved beregninger, må den være definert som intervallvariabel. Variabeldefinisjonen sees over variabelnavnet i datavinduet:

Nom

Nominelle variable eller klassevariable.
Karaktervariable settes automatisk til nominelle variable.

Int

Intervallvariable eller kontinuerlig variable til beregninger.
Numerisk variable settes automatisk til intervallvariable.

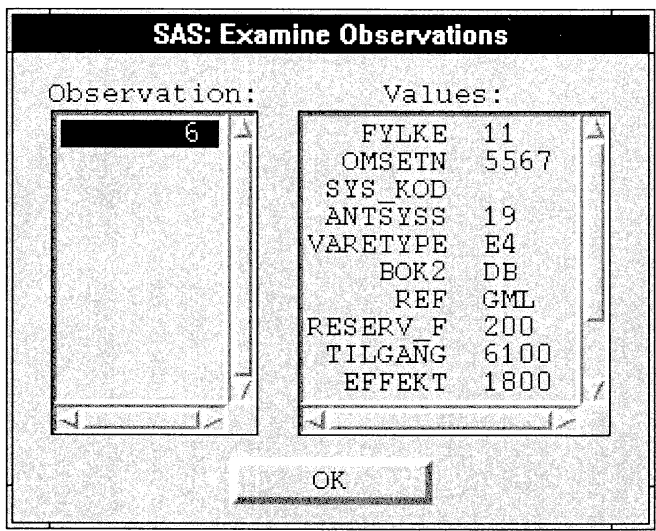
3.1.3. Observasjoner

Observasjonsnummer sammen med symbol (og farge) som brukes i eventuelle plott vises lengst til venstre i datavinduet:



Enkelt klikk på observasjonsnummer markerer observasjonen(e).
Dobbelklikk på observasjonsnummer viser alle variabelverdiene for denne observasjonen:

Figur 3.2. Variabelverdier for en observasjon



3.1.4. Missing

Manglende verdier markeres med punkt for numeriske variable og med tom celle for karaktervariable. Ved sortering regnes missing som mindre enn alle tall.

Observasjoner med missing verdi for en variabel utelates vanligvis ved beregninger og grafer med denne variabelen. Unntaket er i kapittel 8, «Distribution analyze» hvor det er mulig å velge om antall observasjoner skal inkludere missing.

3.2. Endring av data i datavinduet

Datavinduet kan brukes til å endre eller modifisere datasettet. Nye observasjoner eller nye variable kan legges til datasettet og variabelverdier kan endres. Definisjon/bruk av variable i grafer og beregninger kan endres. Datasettet kan sorteres og deler av datasettet kan trekkes ut som egne datasett. .

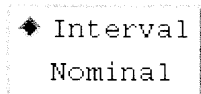
3.2.1. Menyer

Rullegardinsmenyer for filbehandling, editering og analyse sees øverst i datavinduet. Denne menyen finnes også i alle outputvinduer i SAS/INSIGHT. Menyene «Edit» har noen valg som beskrives under andre menyer i dette kapitlet. Ellers kommer vi tilbake til disse menyene i senere kapitler. Her vil vi beskrive andre menyer til å endre datasettet i datavinduet.

3.2.2. Variabel - definisjoner

SAS/INSIGHT er ikke bundet av variabeldefinisjonene i SAS-datasettet. Det er mulig å endre nominelle variable til intervallvariable og omvendt hvis dataene er slik at det gir rimelige definisjoner. Eksempel på data hvor dette kan være aktuelt, er indikatorer som er 0 eller 1.

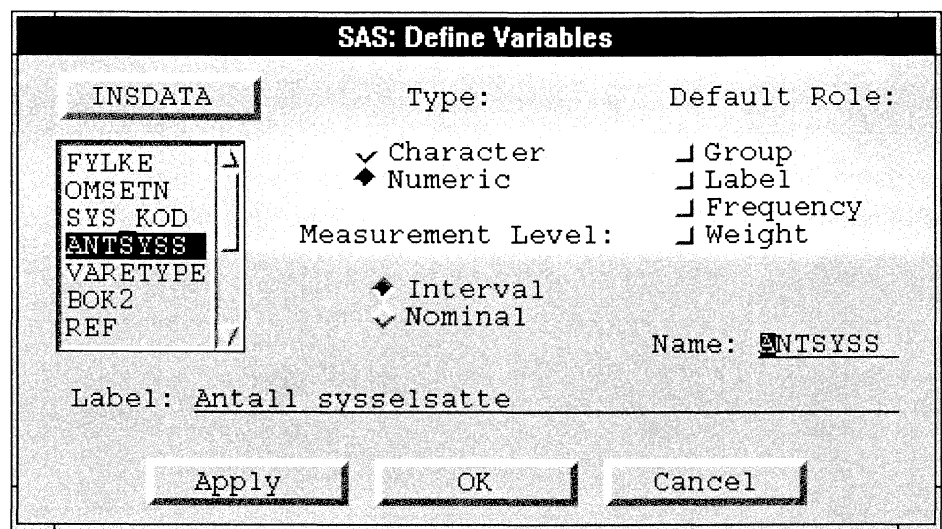
- Variable definert som numeriske variable i SAS-datasettet kan settes til nominelle- eller intervallvariable ved å klikke på «Int»- eller «Nom»-tasten



Velg «Interval» eller «Nominal». OBS! - ikke lag nominelle variable av numeriske variable som kan anta mer enn et begrenset antall verdier.

- Variable definert som karaktvariable kan omdefineres til numeriske variable hvis de er tall. Dobbeltklikk på variabelnavnet gir meny for definisjon av variable:

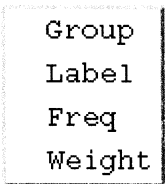
Figur 3.3. Definer variable



3.2.3. Variabel - opsjoner

Det går an å bestemme at en variabel skal ha en bestemt funksjon gjennom hele analysen. Klikk på ruten til venstre for «Nom» eller «Int» øverst i datavinduet. Du får frem menyen:

Figur 3.4. Meny for bruk av variable ved grafer og analyse



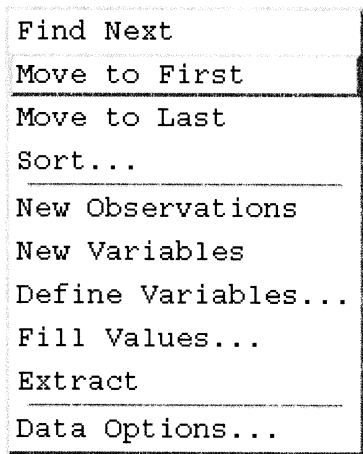
- «Group»: En variabel kan velges som fast grupperingsvariabel ved beregninger eller grafikk. Flere variable kan velges, men pass på totalt antall grupper.
- «Label»: En variabel velges som markeringsvariabel i plott, det vil si at verdien på denne variabelen vises når man markerer en observasjon. Bare en variabel kan velges her
- «Freq»: Verdien gjelder som hyppighetsmål, det vil si at observasjonen telles så mange ganger som variabelen viser. Bare en variabel kan velges og den må være heltallig.
- «Weight»: Verdien blir vekt for observasjonen. Bare en variabel kan velges.

Denne menyen kan også hentes ved å klikke høyre musetast på variabelnavnet.

3.2.4. Databearbeiding

Det er mange muligheter for å manipulere med dataene i datavinduet. Datavinduet har en spesialmeny for dette. Klikk på pilen øverst i venstre hjørne av datavinduet: , eller klikk med høyre musetast inne i datavinduet. Popup meny for bearbeiding av dataene i datavinduet kommer da frem.

Figur 3.5. Meny for bearbeiding av data i datavinduet



Menyene er langt på veg selvforklarende. Her vil vi nevne noen valg:

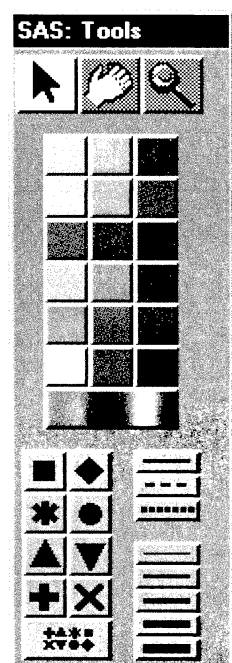
- «Move to First»: Markert variabel/observasjon flyttes til første kolonne/rad i datavinduet.
- «Move to Last»: Markert variabel/observasjon flyttes til siste kolonne/rad i datavinduet.

Hvis man ønsker å flytte variable til andre kolonner enn første eller siste kan dette gjøres ved et spesielt verktøy: Tools. Dette får du frem ved å velge

Edit → Windows → Tools

Dette er et eget vindu hvor du kan velge markør for pekefunksjonen og farger og symbol som brukes på observasjonene i plott.

Figur 3.6. Toolsvinduet med pekefunksjon, farger og symbol



Når dette vinduet kommer opp klikker du på hånden. Markøren endres til en liten hånd. Marker et variabelnavn i datavinduet med hånden, klikk og dra med musetasten nede til den posisjonen du ønsker for denne variabelen.

Markøren settes tilbake til normal igjen ved å klikke på pilen i «Tools»-vinduet.

- «Sort»:
Sorterer observasjonene etter markert variabel. Hvis variabelnavn er markert for «Sort», sorteres automatisk i stigende (ascending) ordning. Hvis variabelnavnet ikke er valgt først, kan du velge mellom stigende og avtagende (descending) ordning. Velger du flere variable å sortere etter, brukes de i den rekkefølgen de er valgt.
- «New observation»:
Legger inn ønsket antall nye observasjoner (linjer) i datasettet med manglende data som kan skrives over.
- «New variabel»:
Legger inn ønsket antall nye variable (kolonner) i datasettet med manglende data som kan skrives over.
- Nye observasjoner eller nye variable kan også skrives rett inn i datavinduet uten å gå vegen om menyer og «New observation/variable». Skriv dataene rett inn i tomme celler og flytt med «Enter»- eller «Tab»- tasten.
- «Ekstract»:
Lager et nytt datasett når en del av datasettet er markert. Ulike måter å markere på beskrives senere i dette kapitlet.
- «Data Options»:
Vise label for alle variable i datavinduet.

3.2.5. Korrigere verdier i datavinduet

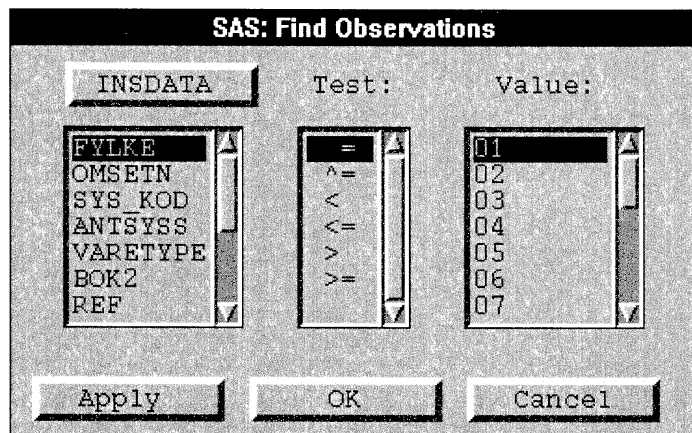
Verdier som skal korrigeres kan rettes direkte i datavinduet. Marker cellen som inneholder den gale verdien. Slett gammel verdi og skriv inn den nye. Husk «Enter» eller «Tab» avhengig av retningen du vil bevege cursoren. For å rette et enkelt tegn i en verdi, marker tegnet og skriv over. «Enter» eller «Tab». Husk å lagre datasettet hvis du vil ta vare på endringen.

3.3. Markere variable/observasjoner

Det er mange måter å markere observasjoner og variable på avhengig av hvordan datasettet ser ut og hvor mye som skal markeres.

- En enkelt markering:
 - Pek og klikk.
- Markere flere variable eller flere observasjoner som følger etter hverandre:
 - Pek, klikk og dra, eller:
 - Klikk på første variabel/observasjon og hold «Shift»-tasten nede mens du klikker på siste variabel/observasjon.
- Markere flere variable eller flere observasjoner som ikke nødvendigvis følger etter hverandre:
 - Hold «Ctrl»-tasten nede mens du klikker på hver variabel/observasjon.
- Markere alle observasjonene:
 - Klikk på antall observasjoner øverst til venstre i datavinduet.
- Markere alle variablene:
 - Klikk på antall variable øverst til venstre i datavinduet.
- Markere observasjoner med bestemte verdier:
 - Velg «Edit» fra rullegardinmenyen, deretter «Observations» og «Find», og velg variabel, testfunksjon og verdier fra vinduet i figur 3.7.

Figur 3.7. Finn observasjoner



3.4. Transformasjon av variable

Nye variable som er funksjoner av andre variable lages ved rullegardinmenyen Edit → New variable. Noen få valgmuligheter kommer da frem:

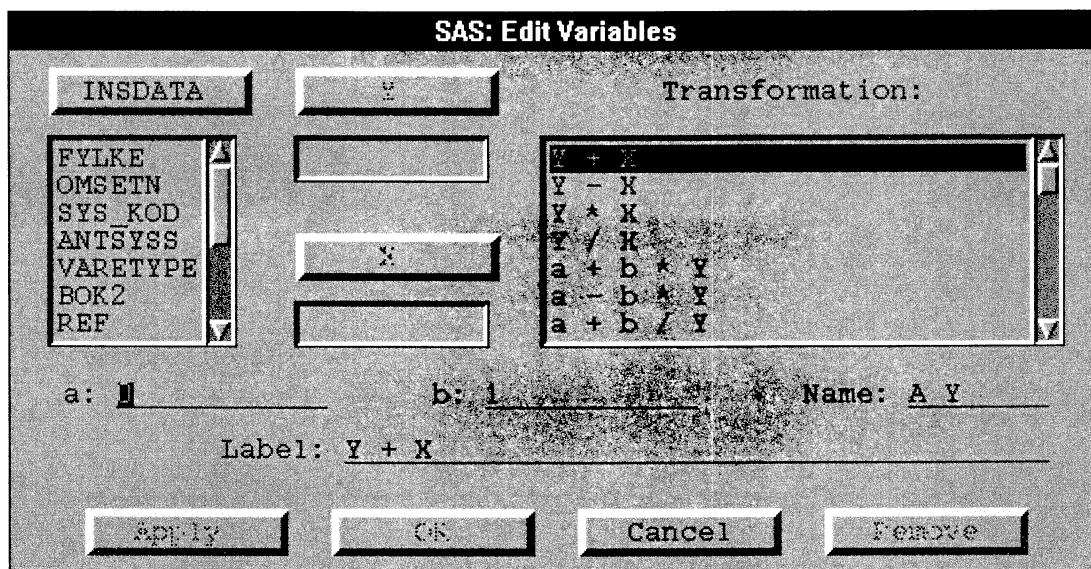

```

log( Y )
sqrt( Y )
1 / Y
Y * Y
exp( Y )
Other...

```

Ønsker du andre funksjoner, velg «Other...». Et dialogvindu viser alle funksjoner som ligger i SAS/INSIGHT:

Figur 3.8. Transformasjon av variable



Her kan du velge funksjon, variable for X og Y og verdier for a og b. Nytt variabelnavn lages automatisk av SAS/INSIGHT, men du kan endre det ved å skrive inn et annet navn.

Merk spesielt funksjonen ($a \leq Y \leq b$) som lager indikatorvariable.

Da det bare er enkle funksjoner av maksimalt 2 variable å velge blant, kan det hende at du må lage noen ekstra variable før du kan få frem ønsket variabel. Ønsker du andre funksjoner enn dem du finner ved å rulle nedover listen her, må de lages i basisversjonen av SAS.

Variable kan fjernes ved å markere variabelen og velge **Edit → Delete**.

4. Histogram/stolpediagram

Analysemuligheter i Histogram/stolpediagram:

- Vise grafisk fordeling av nominelle- eller intervallvariable
 - Antall observasjoner
 - Sum av en frekvensvariabel
- Markere for grafisk fordeling av delmengder av observasjonene

Et histogram/stolpediagram er en grafisk fremstilling av fordelingen til en variabel. Det er litt forskjell på hvordan nominelle og intervallvariable behandles under 'Histogram/BarChart'.

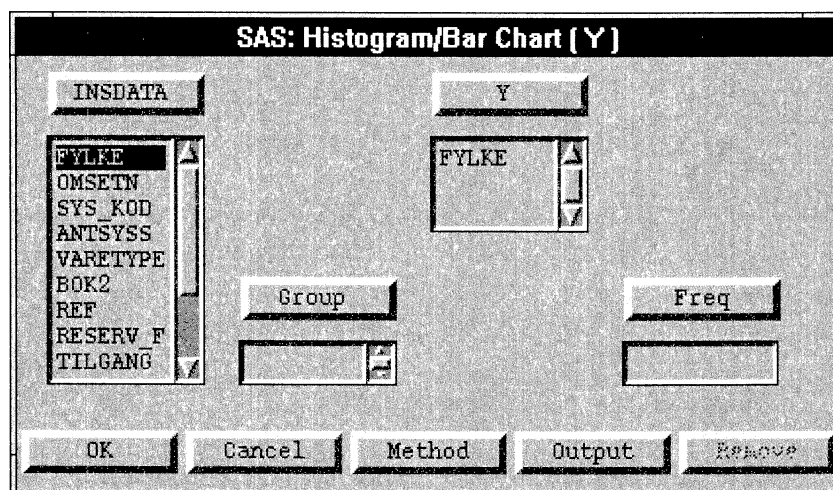
4.1. Nominelle variable

Gå til rullegardinmenyen i datavinduet, og velg

Analyze → Histogram/Bar Chart (Y).

Et dialogvindu åpnes:

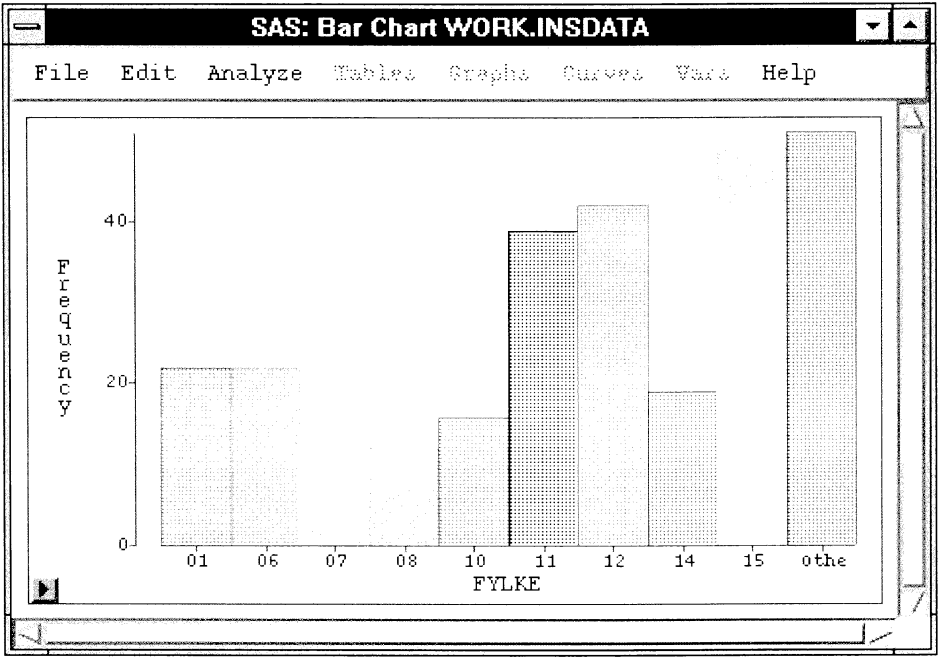
Figur 4.1. Dialogvindu for histogram/stolpediagram



Dette er det vi kan kalle et hoved-dialogvindu. Med knappene her kan vi spesifisere variable og menyvalg for diagrammet. Knappen merket «INSDATA» viser hvilket datasett vi arbeider med, rammen under viser hvilke variable som er i dette datasettet. Med Y-knappen angir vi hvilke variable som det skal lages et stolpediagram av, med «Group»- og «Freq»-knappene kan vi spesifisere gruppe- og frekvensvariable. Med en frekvensvariabel (heltallig) menes en variabel hvis verdier angir hyppighet for en annen variabels verdier. Knappene «Method» og «Output» er menyer for redigering.

Velg en nominell variabel, f.eks. **FYLKE**, og klikk på **Y**. Da er FYLKE angitt som den variabel vi vil beskrive i et stolpediagram. Klikk deretter på **OK**, og et grafisk vindu med et stolpediagram over FYLKE åpnes.

Figur 4.2. Stolpediagram av variabelen FYLKE



Diagrammet viser antall observasjoner pr. fylke, fylkene er angitt ved fylkesnummer. Det har antall observasjoner vertikalt, fylkene horisontalt, og har en standard orientering av aksene. Siste stolpe er betegnet med verdien «Other». Det betyr at denne stolpen er en samlepost, og består av fylker som har færre enn 4 % av observasjonene.

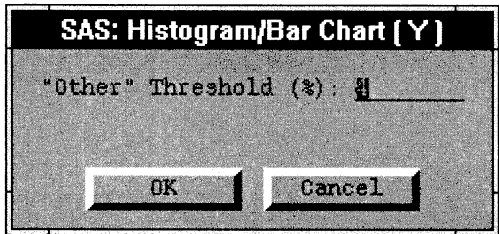
Rullegardinmenyen i grafiske vinduer er stort sett som i datavinduet. Popup-menyen finner vi ved piltast i nedre venstre hjørne i bildet, eller ved høyre musetast inne i grafikkboksen.

4.2. Menyer for redigering

Vi kan redigere stolpediagrammet på to måter, enten ved dialogvinduet i figur 4.1, eller ved høyre musetast. Bruk av dialogvinduet gir et nytt diagram, mens høyre musetast redigerer det diagrammet vi har fremme. Det er lurt å redigere ved dialogvinduet først, og deretter ved høyre musetast.

Gå til diagramvinduet og velg **Edit → Windows → Renew**. Dialogvinduet i figur 4.1 på ny, med FYLKE angitt som Y-variabel. Vi ønsker å ha med alle fylkene i stolpediagrammet. Velg **Method**, og en undermeny åpnes:

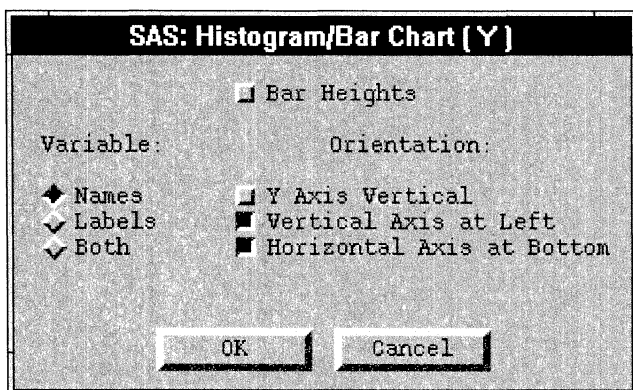
Figur 4.3. Undermeny for histogram/stolpediagram



For å få med alle fylkene endrer vi terskelen fra 4 til 0, og klikker **OK**.

Andre redigeringsmuligheter i dialogvinduet har vi ved å velge **Output**. En ny undermeny åpnes:

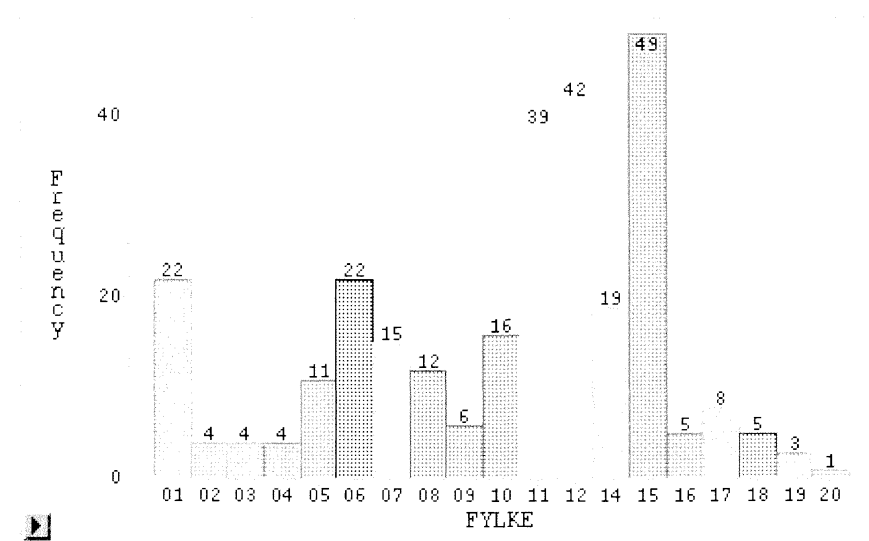
Figur 4.4. Undermeny for stolpediagram



Her kan vi redigere forhold som har med diagrammets utseende å gjøre. Ved å markere for «Bar Heights» vil antall observasjoner for hver stolpe vises i diagrammet. Under «Variable» kan vi angi om vil ha variabelnavnet eller en «label» som navn på aksene, eller begge deler. Under «Orientation» kan vi endre standard orientering av aksene. Velg **Bar Heights** → **OK**.

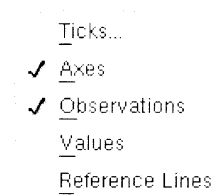
Hvis vi ellers er fornøyd med diagrammet, klikker vi **OK** i dialogvinduet. Ett nytt stolpediagram med alle fylkene åpnes, og antall observasjoner for hver stolpe vises.

Figur 4.5. Stolpediagram av variabelen FYLKE



Popup menyen

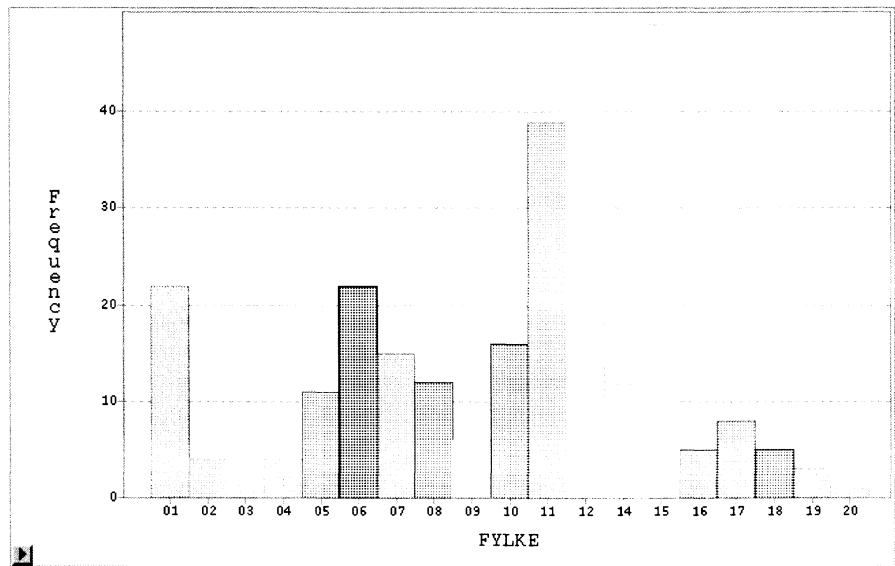
Høyre musetast inne i den grafiske boksen aktiverer en popup - meny:



Særlig nyttige funksjoner her er og «Values». «Values» har samme funksjon som Bar Heights beskrevet over. «Ticks» gjelder for intervallvariable og kan også brukes på «Frequency», her langs Y -

aksen. Endringer av «Ticks» beskrives under kap. 4.5. Referanselinjer kan legges inn i histogram eller stolpediagram og de plasseres da automatisk etter definerte «Ticks» -verdier.

Figur 4.6. Stolpediagram med referanselinjer av variabelen FYLKE



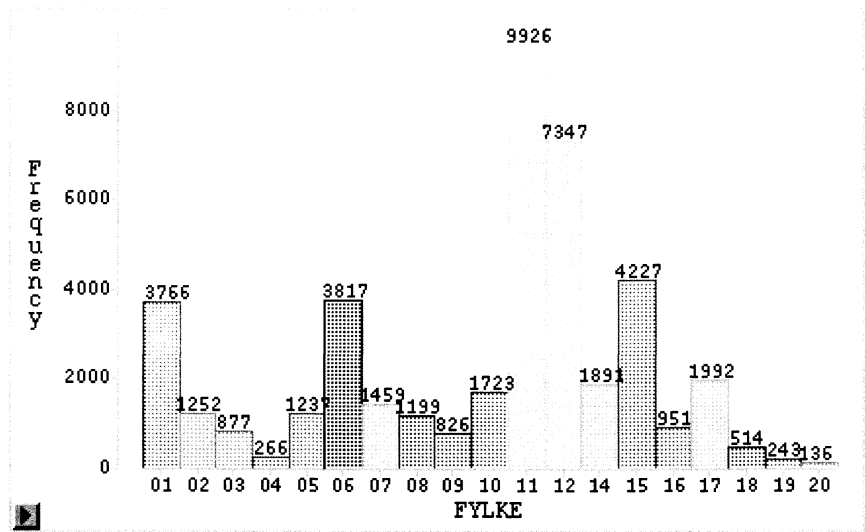
4.3. Stolpehøyde etter andre variable

I enkelte datasett vil en eller flere intervallvariable kunne ses som et størrelsesmål eller hyppighetsmål for observasjonene. Hvis vi ønsker å se hvordan en slik intervallvariabel fordeler seg etter en gruppering som er definert ved en nominell variabel, kan vi få dette frem ved et stolpediagram over den nominelle variabelen med intervallvariabelen brukt som «Freq» - variabel.

Gå til diagramvinduet og velg **Edit → Windows → Renew**.

Dialogvinduet som i figur 4.1 åpnes. Velg variabelen **ANTSYSS**, og klikk **Freq → OK**. Det åpnes et nytt stolpediagram som viser en fylkesvis fordeling av sysselsettingen.

Figur 4.7. Stolpediagram av FYLKE med ANTSYSS som frekvensvariabel

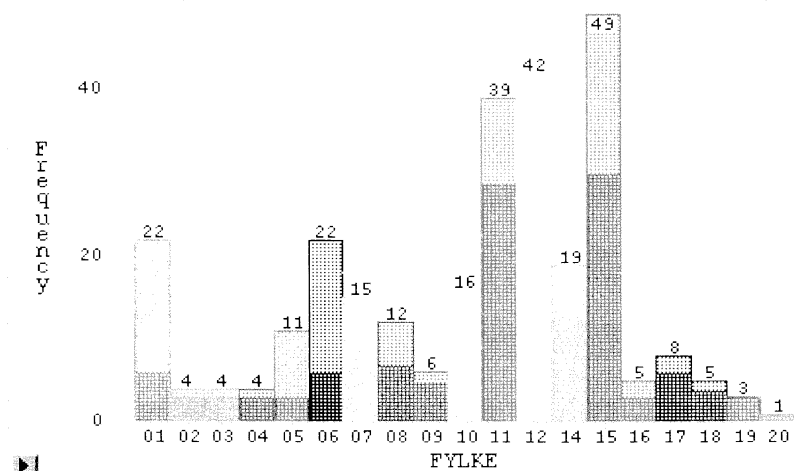


4.4. Markering

4.4.1. Markere for en variabelverdi

Ta for deg stolpediagrammet over antall bedrifter pr fylke. For å definere hvilke observasjoner som skal markeres, gå til **Edit → Observations → Find** og velg **BOK2 = DM** og **OK**. Markeringen viser nå antall bedrifter med næring DM i hvert fylke.

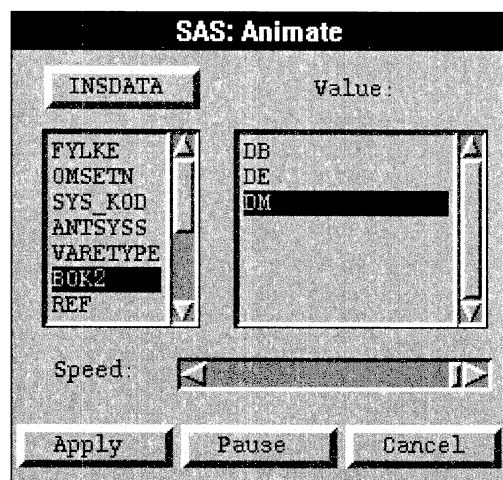
Figur 4.8 Stolpediagram over fylkene, der næring DM er markert



4.4.2. Animering

Med SAS/INSIGHT kan vi lage bevegelig grafikk. Vi kan f.eks. lage et stolpediagram som suksessivt viser den fylkesvise fordelingen av dataene innen hver næring. Gå til rullegardinmenyen, velg **Edit → Windows → Animate**, og et dialogvindu for animering åpnes.

Figur 4.9 Dialogvindu for animering



Marker BOK2 (næring) som «animeringsvariabel» og klikk **Apply**. Histogrammet animeres da ved at det enkelte næringene DB, DE og DM markeres suksessivt. Vi kan endre farten på animeringen ved å

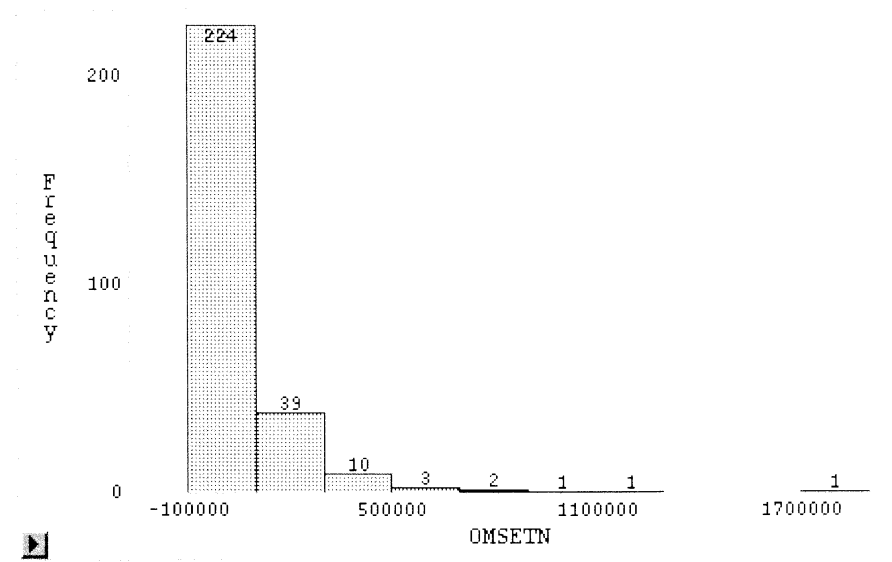
klikke på **Pause**, flytte på «Speed-slideren» og igjen klikke **Apply**. Speed avhenger av maskinhastighet og det er ikke alltid mulig å få det til å gå sakte nok.

4.5. Intervallvariable

Hurtigversjonen av et histogram/stolpediagram er som følger: Gå til datavinduet, og marker i kolonnehodet den variabel du vil ha et histogram/stolpediagram over, f.eks OMSETN.

Velg **Analyze → Histogram/Bar Chart (Y)**, og det åpnes et histogram over omsetningen:

Figur 4.10. Histogram over omsetning



Dette histogrammet er kanskje ikke helt tilfredsstillende. Venstre intervall (stolpe) overlapper negative verdier, og intervalllengden er litt for stor. Vi vil derfor redigere startverdien og intervalllengden i histogrammet.

Bruk høyre musetast inne i grafen, og velg **Ticks**. Følgende meny åpnes:

Figur 4.11. Dialogvindu for Ticks

SAS: Ticks

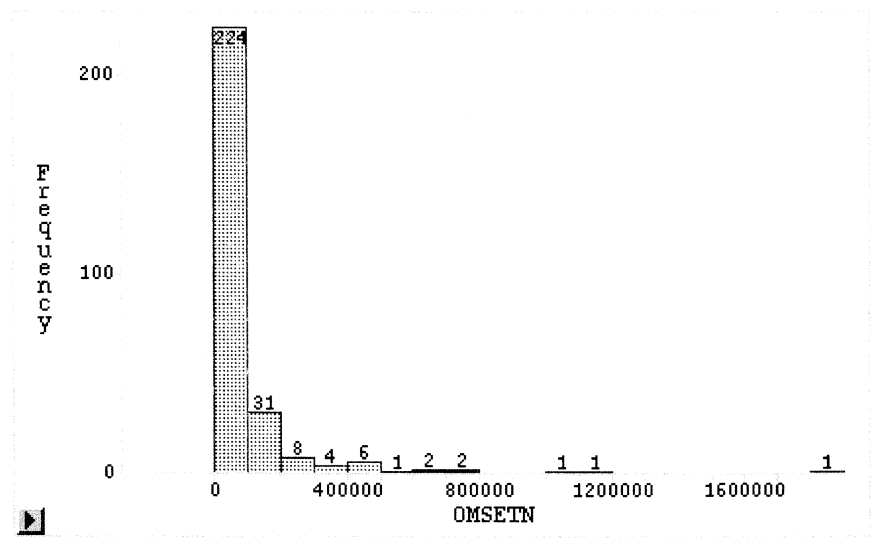
First Tick: 100000
Last Tick: 1900000
Tick Increment: 200000
Minor Ticks: 0
Axis Minimum: -200000
Axis Maximum: 1900000

OK Cancel

Vi endrer startverdien ved å sette «First Tick» lik 0, og vi endrer intervall-lengden ved å sette «Tick Increment» lik 100 000. Sluttverdien (Last Tick) er ikke så viktig, men hvis denne endres så pass på at «Axis Maximum» er lik (eller større enn) «Last Tick».

Klikk deretter **OK**, vi får vi et mer «korrekt» histogram over omsetningen.

Figur 4.12. Revidert histogram av omsetning



5. Boksplott/Mosaikkplott

Analysemuligheter i Boksplott/Mosaikkplott:

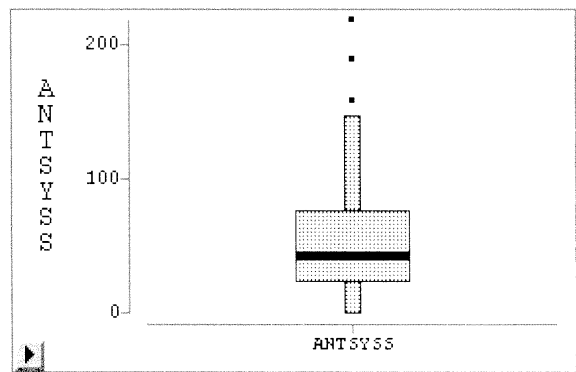
- Boksplott
 - viser grafisk fordeling av intervallvariable
 - median, kvartiler
 - gjennomsnitt, standardavvik
 - test for sammenligning av middelværdier
- Mosaikkplott
 - viser grafisk fordeling av nominelle variable
 - antall og andeler

5.1. Boksplott

Boksplott er en grafisk fremstilling som markerer median, kvartiler og utliggere i et datasett. Medianen er den markerte linjen i boksen, mens nedre og øvre kvartil markeres ved grensene for «boksen». Haler trekkes fra kvartilene til fjerneste observasjon som har avstand til kvartilen maksimum 1.5 ganger avstanden mellom kvartilene. Observasjoner utenfor dette området markeres enkeltvis som utliggere.

Et enkelt boksplott kan se slik ut:

Figur 5.1. Boksplott



5.1.1. Meny for å lage boksplott

For å få frem et boksplott brukes rullegardinmenyen: **Analyze → Box Plot/Mosaic plot (Y)**. Da kommer dialogvinduet til Box Plot frem:

Figur 5.2. Boksplott dialogvindu

Velg variabel (f.eks ANTSYSS), klikk på «Y»-knappen og deretter på «OK».

Boksplott-funksjonen forutsetter at du velger en intervallvariabel. Nominelle variable vil gi et mosaikkplott som viser antall/andel observasjoner i ulike kategorier.

5.1.2. Gruppeinndeling

Ved å markere en nominell variabel og klikke på «Group», lages uavhengige boks-plott, for hver verdi av denne gruppevariabelen. Eventuelt får vi uavhengige mosaikkplott hvis Y-variabelen i utgangspunktet er nominell variabel.

5.1.3. Sammenlignbare boks-plott

Hvis man ønsker sammenlignbare boks-plott, kan gruppevariabelen isteden defineres som X-verdi i dialogvinduet. Da kommer alle plottene i samme bilde og med samme skalering.

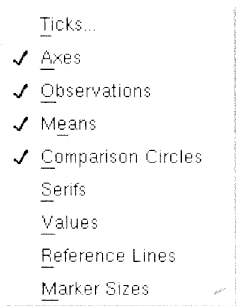
5.1.4. Defaultverdier

«Method»-knappen henter frem en meny hvor man kan endre defaultverdier for lengden på halene i boks-plott (eller minste andel observasjoner for å vises som egen gruppe i mosaikk-plott).

5.1.5. Meny for redigering av plott

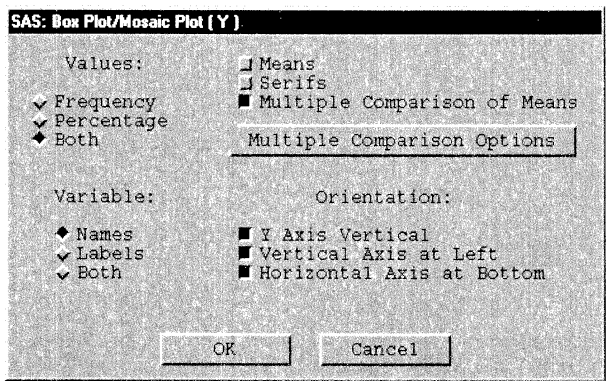
Popup menyen fra piltast nederst i venstre hjørne av plottet eller høyre musetast gir redigeringsmuligheter for boksplottet:

Figur 5.3. Boksplott-meny

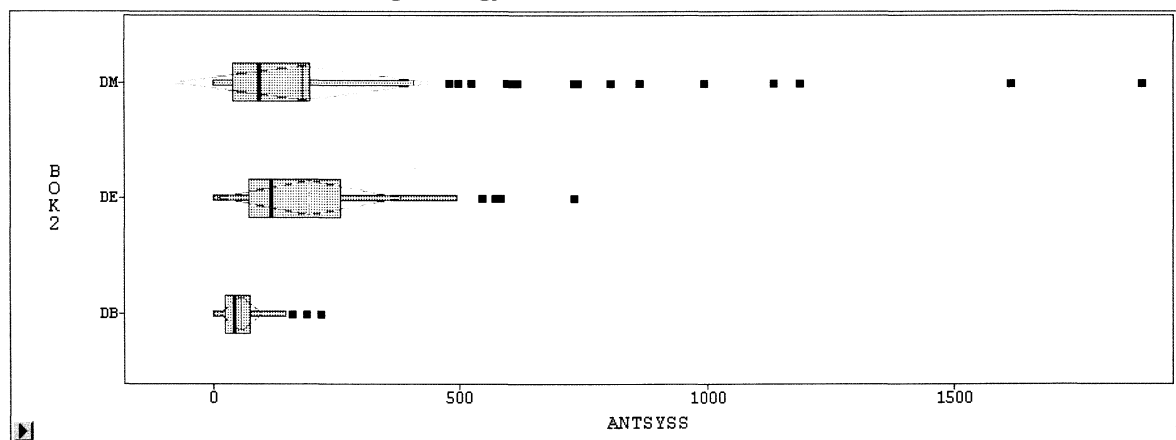


- «Means» viser grafisk gjennomsnitt pluss/minus ett standardavvik i plottet.
- «Values» viser verdier for gjennomsnitt, median, kvartiler og enden på halene i plottet.
- «Reference Lines» lager referanselinjer på samme måte som for histogram/stolpediagram.
- «Comparison Circles» gir en grafisk fremstilling av en metodestatistisk multiplisert sammenligning av gjennomsnitt. Metoden forutsetter normalfordeling og samme varians for alle observasjoner. Se kapittel 5.3.

Figur 5.4. Outputmeny



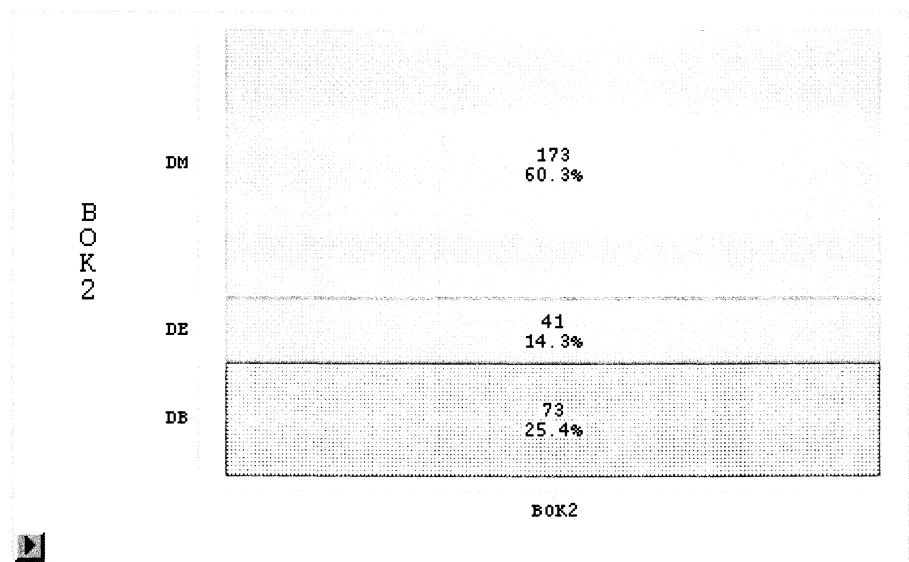
Figur 5.5. Boksplott etter næring, med gjennomsnitt



5.2. Mosaikkplott

Menyene for mosaikkplott er de samme som for boksplott. Velges nominelle variable i dialogvinduet, figur 5.2, får vi et mosaikkplott. Velges for eksempel næring (= BOK2) som Y-variabel, ser figuren slik ut etter at «Values» er markert i menyen (figur 5.3) nederst i venstre hjørne av mosaikkplottet:

Figur 5.6. Mosaikkplott

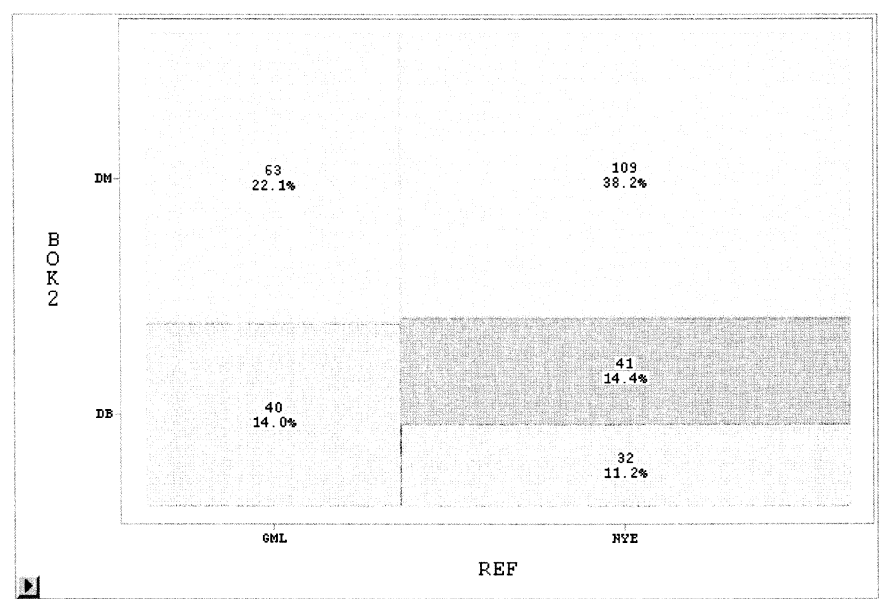


Mosaikkplottet angir flatemessig andel enheter for hver mulig verdi av valgt kjennemerke. «Values» legger til tall for antall enheter og prosentandel enheter.

5.2.1. Kombinasjon av variable i mosaikkplott.

Ved å velge en nominell variabel som Y og en annen nominell variabel som X, kommer alle kombinasjoner av Y- og X-verdier i mosaikkplottet.

Figur 5.7. Mosaikkplott etter Bok2 og Ref



5.3. Tester for sammenligning av data fra flere grupper (kategorier)

OBS!! Alle tester forutsetter uavhengige, normalfordelte variable med samme varians.

Det ligger flere testmetoder inne, defaultmetode er Tukey-Kramers metode for parvis multippel sammenligning. Testene illustreres grafisk og i tabeller.

Figur 5.8. Meny for testing

SAS: Multiple Comparison Options

Multiple Comparison Test:

Confidence Level:

☐ Pairwise t-test

☒ Tukey-Kramer All Pairs

☐ Pairwise Bonferroni

☐ Dunnett's Test with Control

☐ Hsu's Test for Best

☐ Hsu's Test for Worst

☐ 99%

☒ 95%

☐ 90%

☐ Other%

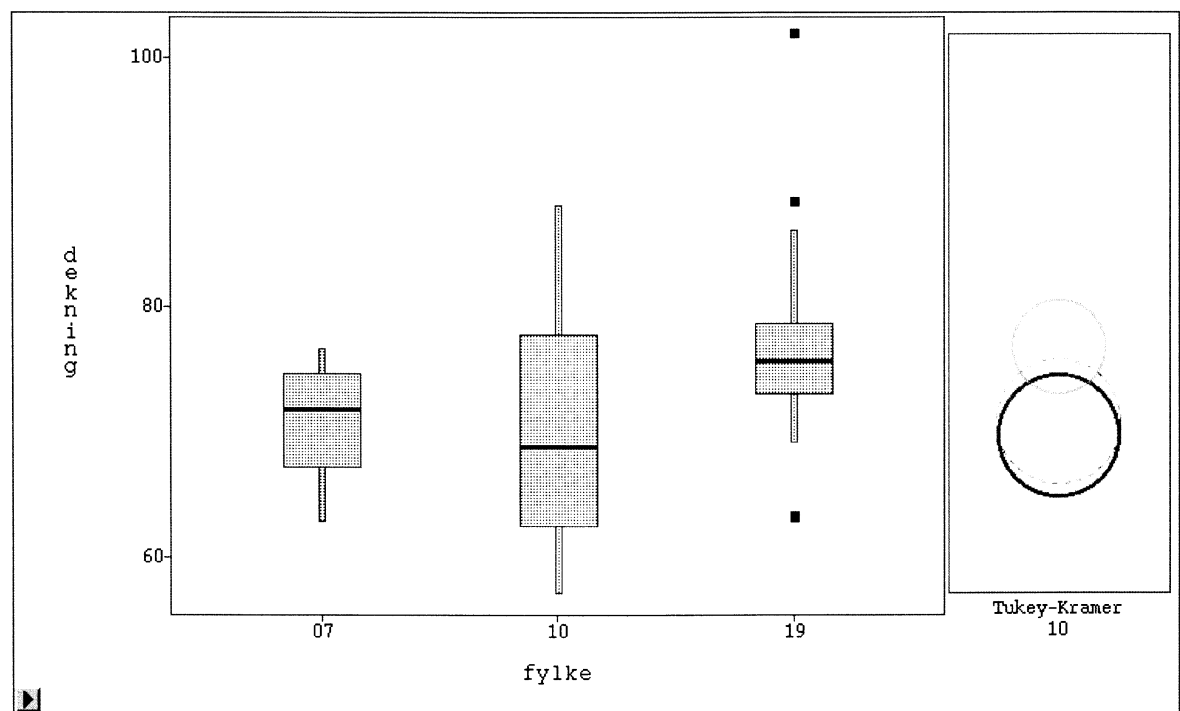
Other Level %: 80

OK

Cancel

Sammenligning av barnehagedekning i kommunene for tre fylker (Vestfold, Vest-Agder, Troms) gir følgende graf:

Figur 5.9. Barnehagedekning i 3 fylker



Sirklenes midtpunkt ligger på samme nivå som gjennomsnittsverdiene, mens sirkelens radius er omvendt proporsjonal med kvadratroten av antall observasjoner i gruppen. Når man markerer en sirkel (her den nederste - fylke 10 Vest-Agder), får denne en rød heltrukket kontur samtidig som fylkesnummeret kan leses under sirkelene. De andre sirkelene får farge og kontur avhengig av om det er signifikant forskjell i forhold til den valgte gruppe. Her blir sirkelen for fylke 07 (Vestfold) rød med stiplet kontur fordi det ikke er signifikant forskjell mellom barnehagedekning i Vestfold og Vest-Agder. Sirkelen for fylke 19 (Troms) blir blå med heltrukket kontur fordi det er signifikant forskjell mellom barnehagedekningen i Troms og Vest-Agder. Markering av sirkelen for fylke 07 - Vestfold fører til at begge de andre fylkene får rød stiplet kontur.

Grunnlaget for sirkelene ligger i data som beregnes for samtidig sammenligning av alle par kategorier som kan dannes. Man kan hente frem tabell som viser beregnede data.

6. X - Y - plott

Sammenhengen mellom to variable kan studeres ved et X - Y - plott.

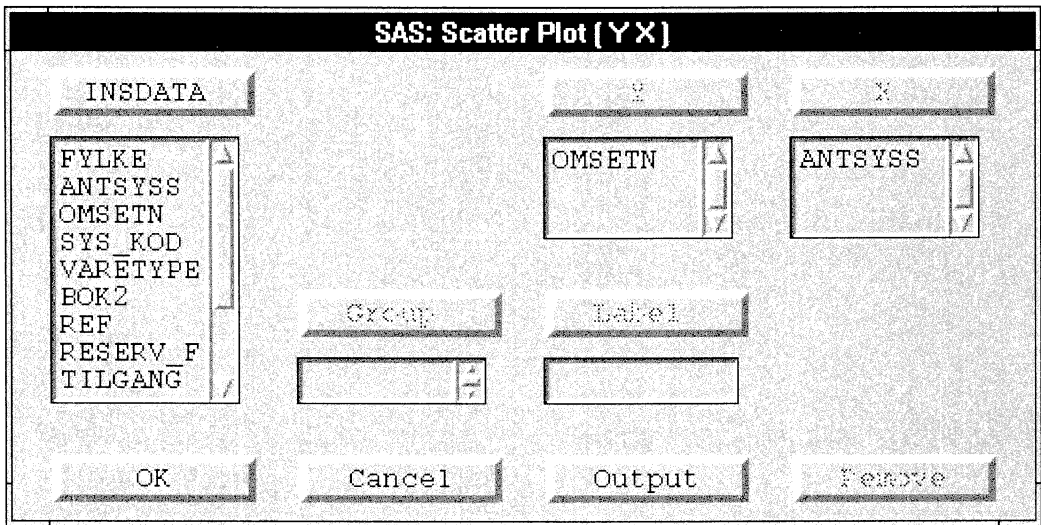
Sammenhengen mellom flere variable kan studeres gjennom parvise sammenligninger ved flere X-Y plott

Fra datavinduet eller et hvilket som helst grafisk vindu velges:

Analyze → Scatter plot (Y X).

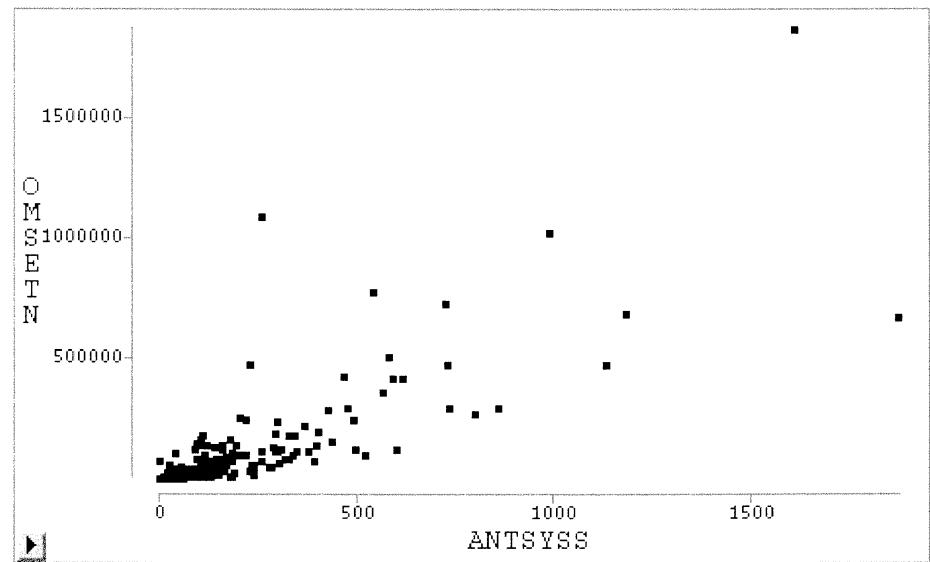
Dialogvinduet for valg av Y - og X - variabel kommer opp:

Figur 6.1. Dialogvindu for plott



Marker ANTSYSS som X - variabel, OMSETN som Y - variabel og klikk OK:

Figur 6.2. Plott



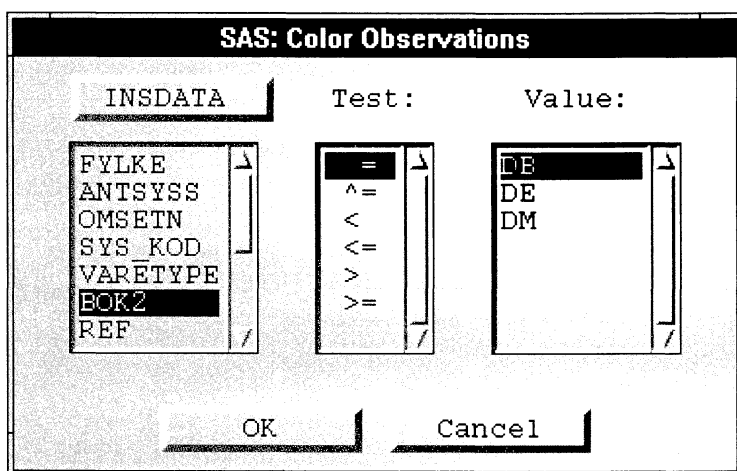
SAS/INSIGHT velger automatisk skalering langs aksene. Hvis du ønsker annen skalering, kan den endres ved pop-upmenyen nederst i venstre hjørne i plottet. Se kapittel 4.2.

6.1. Farger og symbol

Spesielt valgte observasjoner eller observasjoner med spesielle kjennemerker kan skilles ut ved hjelp av farger eller symbol i plottet. Det er ulike måter å gjøre dette på:

1. Gå til datavinduet og marker de observasjonene du vil skille ut. Hent «Tools»-vinduet og klikk på fargen eller symbolet du ønsker.
2. Gå til datavinduet. Ingen observasjoner markert. Hent «Tools»-vinduet og klikk på en farge eller et symbol. Du får da frem et dialogvindu:

Figur 6.3. Valg av observasjoner



Velg en variabel, et testsymbol og en eller flere verdier, (f.eks. «BOK2», «=» og «DB») og klikk på OK.

3. Gå til datavinduet. Marker en nominell variabel, gå til «Tools»-vinduet og velg feltet med varierende fargeskala, eventuelt varierende symboler. Dette gir ulike farger etter verdien på den nominelle variable.
4. Gå til datavinduet. Marker en intervallvariabel, gå til «Tools»-vinduet og velg feltet med varierende fargeskala. Fargen i plottet vil nå variere over en glideskala avhengig av verdien på intervallvariabelen.

6.2. Identifisering av punkter i plottet

Når man er i plott-vinduet kan man ønske å identifisere hvilken observasjon som ligger bak enkelte punkter. Man kan få frem ulik informasjon om observasjonene:

- Klikk på et punkt og observasjonsnummeret kommer da frem i plottet.
- Dobbeltklikker på punktet og du får frem et vindu med alle variabelverdiene for denne observasjonen.
- Legg inn en «Label»-variabel i plottet. Det vil ofte være aktuelt å bruke navnet til observasjonen. Her kan vi velge næring: «BOK2». Velg **Edit → Windows → Renew**, velg **BOK2 → Label → OK**. Hvis en variabel er definert som «Label»-variabel i datavinduet før plottet lages, vises verdien av denne automatisk når man klikker på et punkt.

6.3. Fjerne observasjoner fra plottet

Enkeltobservasjoner kan fjernes fra plottet ved å markere punktene og velge **Edit → Observations → Hide in Graphs**.

Hvis man ønsker plott av bare en del av datasettet, kan man velge bort observasjoner som ikke skal være med i plottet. Pass på at ingen observasjoner er markert og velg

Edit → Observations → Hide in Graphs

Et dialogvindu som i figur 6.3 kommer frem. Velg variabel, Test og Value.

6.4. Hente tilbake observasjoner i plottet

Alle observasjonene kan hentes tilbake ved å markere alle observasjonene i datavinduet (se kapittel 3.3) og velge **Edit → Observations → Show in Graphs**.

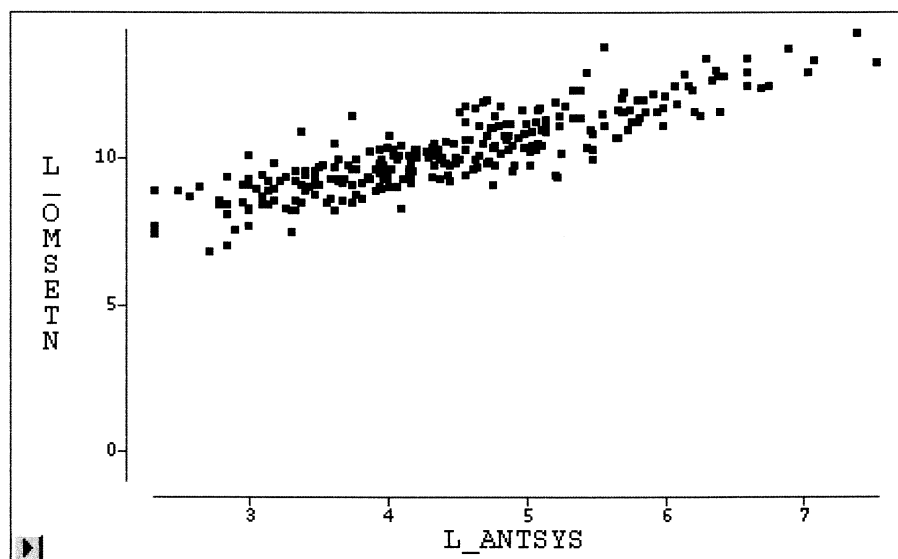
Hvis man ønsker bar noen observasjoner tilbake, pass på at ingen observasjoner er markert og velg **Edit → Observations → Show in Graphs** for dialogvinduet i figur 6.3. Velg variabel, Test og Value.

6.5. Transformasjon av variable

Man kan transformere variable direkte i plottet uten å gå vegen om datavinduet. Marker X- og/eller Y-variabelen i plottet, velg **Edit → Variables → ønsket funksjon**.

I figur 6.4 har vi laget en log-transformasjon av både omsetning og sysselsetning fra figur 6.2.

Figur 6.4. Logtransformerte variable



6.6. Punktstørrelse

Punktstørrelsen i plottene kan endres. Ligger punktene for eksempel delvis oppå hverandre kan de skilles bedre hvis punktene er noe mindre. Meny for punktstørrelse ligger under høyre musetast – ”Marker Sizes”.

6.7. Forstørring av plottet

Gå til «Tools» og klikk på forstørrelsesglasset. Markøren vises nå som et forstørrelsesglass. Gå tilbake til plottvinduet og flytt markøren rundt om på plottet. I ytterkant av plottet er forstørrelsesglasset lite mens det er større midt inne i plottet.

- Klikk med stort forstørrelsesglass forstørrer plottet.
- Klikk med lite forstørrelsesglass får plottet til å fylle vinduet.

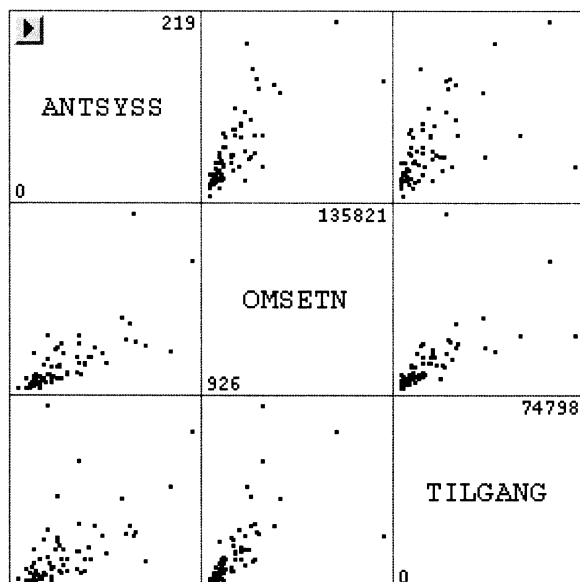
Man kan forstørre områder av plottet, for eksempel områder hvor punktene ligger svært tett. Marker området med forstørrelsesglasset ved å klikke, holde musetasten nede og dra et rektangel. Obs: dra med en gang du klikker.

6.8. Plottmatrise

Man kan markere to eller flere variable i datavinduet og så velge **Analyze → Scatter plot (Y X)**. Har man valgt to variable, får man et enkelt X - Y - plott med den først valgte variabelen langs Y-aksen.

Markeres flere enn to variable får man en symmetrisk matrise over X - Y -plott av alle kombinasjoner av to og to variable:

Figur 6.5. Plottmatrise



I diagonalen finner man navnet pluss laveste og høyeste verdi variablene har. Plottene for samme variable over og under diagonalen er speilbilder av hverandre.

I eksemplet i figur 6.5 ovenfor har vi laget en plottmatrise over sysselsetting, omsetning og ordretilgang.

7. Grafisk analyse av data

Sammenhengen mellom flere variable i et datasett kan studeres gjennom ulike grafer ved hjelp av markeringer som er gjennomgående for alle grafene

Ekstremverdier kan identifiseres

Ulike grafiske presentasjoner av samme datasett kan samlet gi god oversikt over dataene. Analysen blir den samme enten vi lager nye vinduer for hver graf eller om vi legger flere grafer inn i samme vindu. Det kan bli mye skifting mellom vinduer hvis hver graf ligger i et eget vindu. Her vises hvordan flere grafer kan legges i samme vindu og eksempler på grafisk analyse.

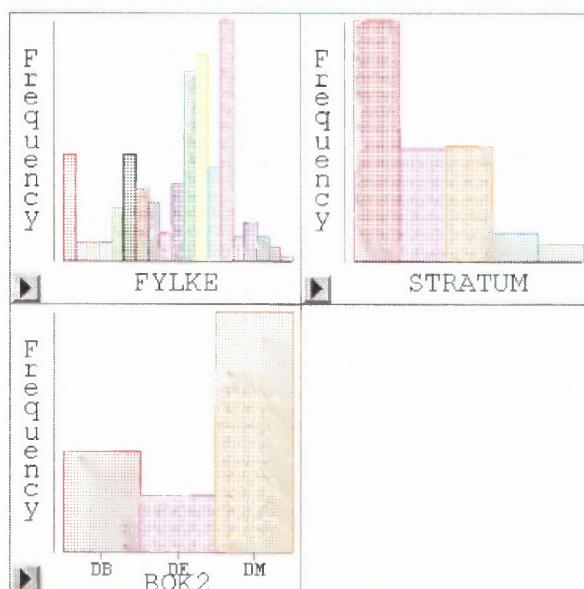
7.1. Flere grafer av samme type i et vindu

Marker flere variable i datavinduet. Nominelle og intervallvariable kan brukes, men for analysens del bør ikke alle variablene være intervallvariable. Vi får et histogram eller et stolpediagram for hver variabel ved å velge

Analyze → Histogram/Bar Chart (Y)

Vi velger her FYLKE, STRATUM og BOK2 som Y - variable og får frem figuren:

Figur 7.1. Eksempel på tre stolpediagram i samlet graf



Vi ser her at noen av stolpediagrammene mangler skalering langs aksene. Dette skyldes størrelsen på figurene og/eller plasseringen av aksene.

7.1.1. Størrelse og plassering av grafer

Størrelsen på grafene og innbyrdes plassering i vinduet kan endres på forskjellig vis.

- For å forstørre alle figurene like mye slik at de alle fortsatt synes i samme vindu, kan man bruke forstørrelsesglasset fra «Tools».
- Man kan flytte figurene i forhold til hverandre ved å markere rammen rundt figuren, klikke på en av langsiden og dra. Se figur 7.2.

- Når det er noe rom rundt rammene, kan størrelsen endres enkeltvis for hver enkelt graf. Klikk på et av hjørnene og dra til ønsket størrelse. Se figur 7.2.

Figur 7.2. Markert ramme rundt et histogram/stolpediagram

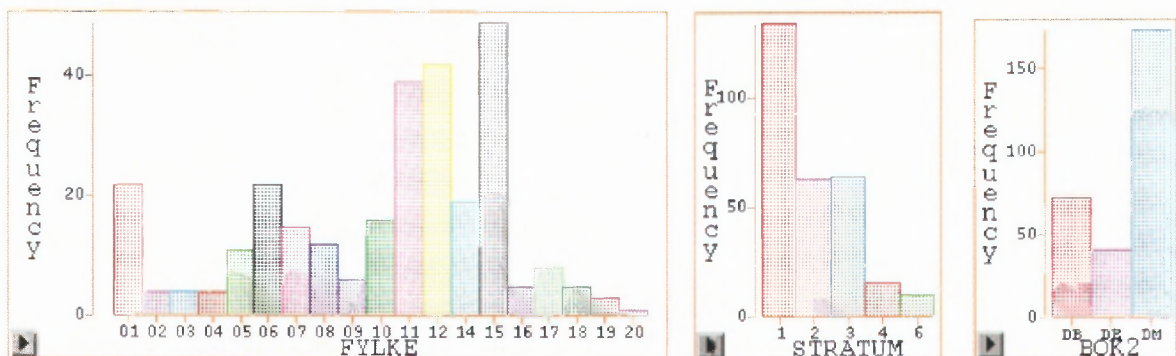


Skaleringen langs Y-aksen kan komme frem automatisk når man øker bredden på rammen rundt grafen. Tilsvarende kan skaleringen langs X-aksen komme frem når høyden på rammen økes.

Hvis skaleringen ikke har kommet frem når figurene har fått ønsket størrelse, kan man få frem skaleringen ved å flytte X- og/eller Y-aksen: Flytt markøren mot en akse til markøren vises som en liten hånd. Klikk og dra aksene i ønsket retning.

Stolpediagrammene i figur 7.1 kan for eksempel endres til de ser omtrent slik ut:

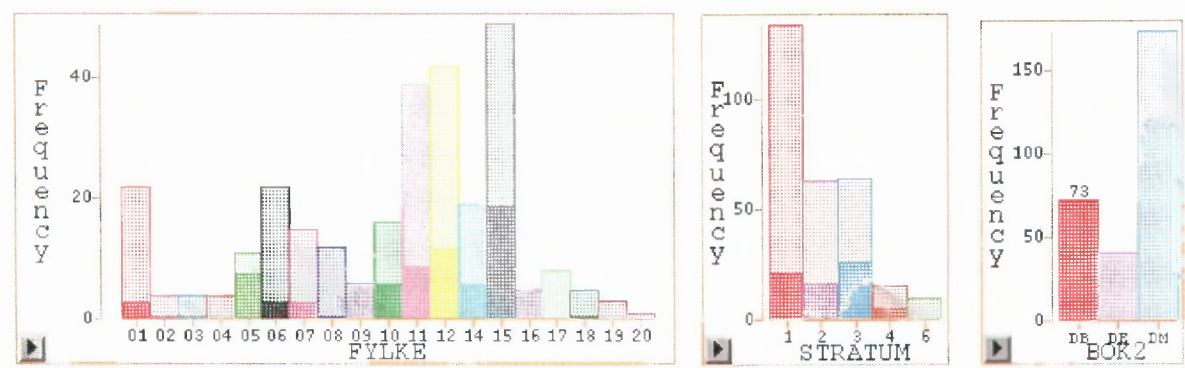
Figur 7.3. Justerte stolpediagram



7.2. Analyse ved hjelp av stolpediagram

Vi kan nå markere en variabelverdi i et av stolpediagrammene. I figur 7.4 har vi valgt næring «DB» i grafen lengst til høyre. Denne søylen får forsterket farge. Samtidig markeres med forsterket farge fordelingen til enhetene i næring «DB» i alle de andre grafene. Dette skjer enten grafene ligger i samme vindu eller om vi har laget nytt vindu for hver graf.

Figur 7.4. Næring "DB" markert i alle stolpediagrammene

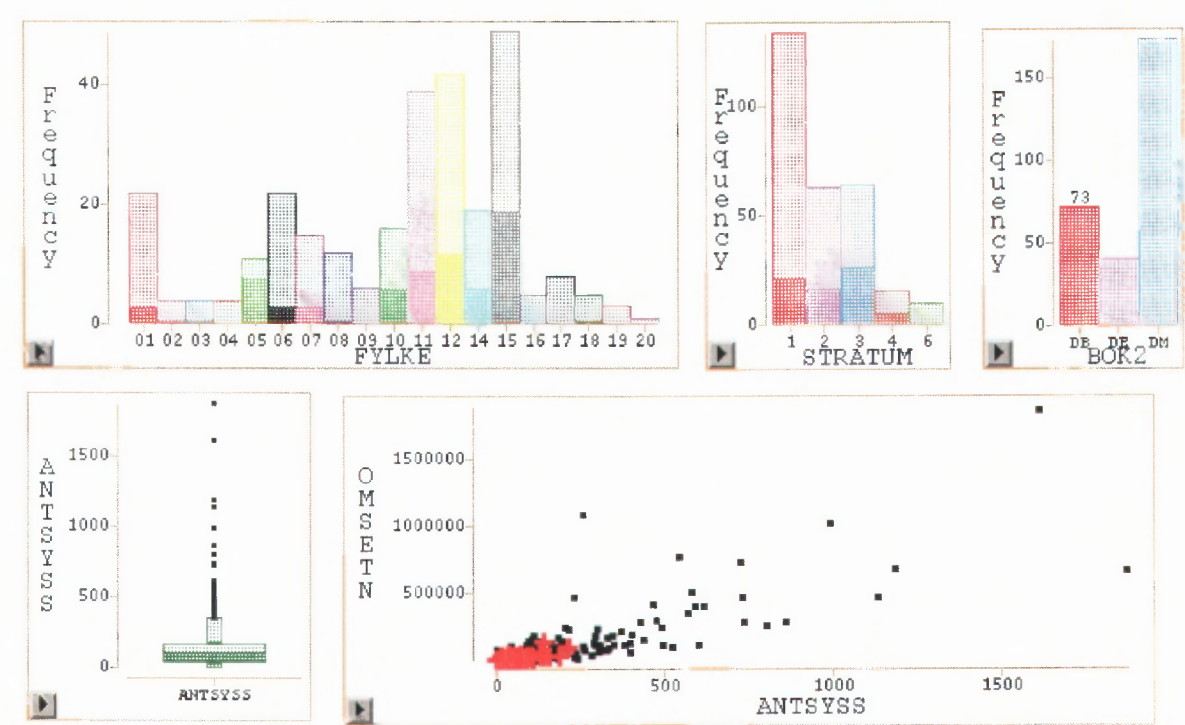


7.2.1. Legge inn flere typer grafer i samme vindu

Hvis vi nå ønsker enda en figur inn i samme vinduet, kan vi ved å klikke og dra med musen, markere et område for den nye figuren. Gå så til **Analyze** og velg graf.

I figur 7.5 har vi laget ett område hvor vi har lagt et boksplott og ett område hvor vi har lagt et X-Y - plott i tillegg til stolpediagrammene.

Figur 7.5. Flere grafer



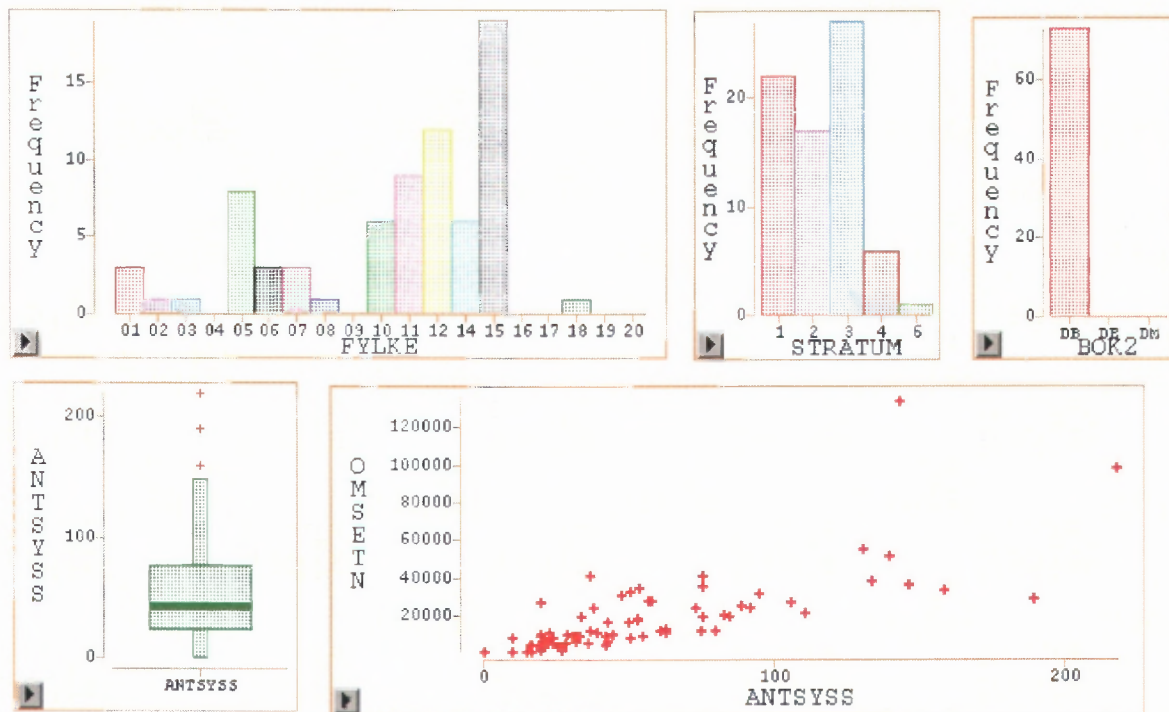
Markering av en søyle i et histogram/stolpediagram fører nå til at observasjonene som hører inn under dette området blir markert i alle grafene. Se figur 7.5 hvor næring DB er markert i stolpediagrammene som før, i boksplottet ved sterkere farge på den lavere del av boks og nedre hale og i plottet med rød farge (gråtone) og symbol '+'.

7.2.2. Selektere grupper

Vi kan lage grafer av bare enkelte grupper av observasjonene uten å gå vegen om nye datasett: Hvis man bare vil vise observasjonene bare i næring DB, kan man markere søylen for DB i stolpediagrammet og velge **Edit → Observations → Invert selection**, og **Edit → Observations → Hide in graph**.

Dette gir figur 7.6.

Figur 7.6. Næring DB



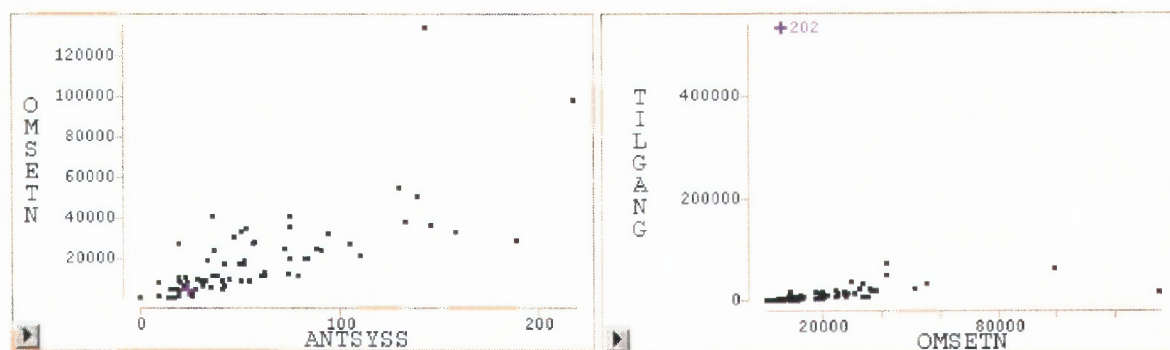
7.2.3. Markering av enkeltobservasjoner i plott

Markering av en observasjon i et plott fører til at:

- Observasjonen identifiseres med observasjonsnummer eller «label» i plottet
- Samme observasjonen markeres i alle plottene

I figur 7.7 har vi laget et plott av omsetning og sysselsetting til venstre og et plott av ordretilgang og omsetning til høyre. En observasjon skiller seg klart ut i plottet til høyre. Denne har vi markert slik at observasjonsnummeret (202) vises. For å kunne spore denne observasjonen tydeligere også i plottet til venstre, har vi lagt på farge og symbol '+' ved hjelp av «Tools». Observasjonen er nå markert med en rød (grå) + i begge plottene. Dette er en måte å identifisere og vurdere observasjoner som skiller seg ut.

Figur 7.7. Flere plott - samme observasjon



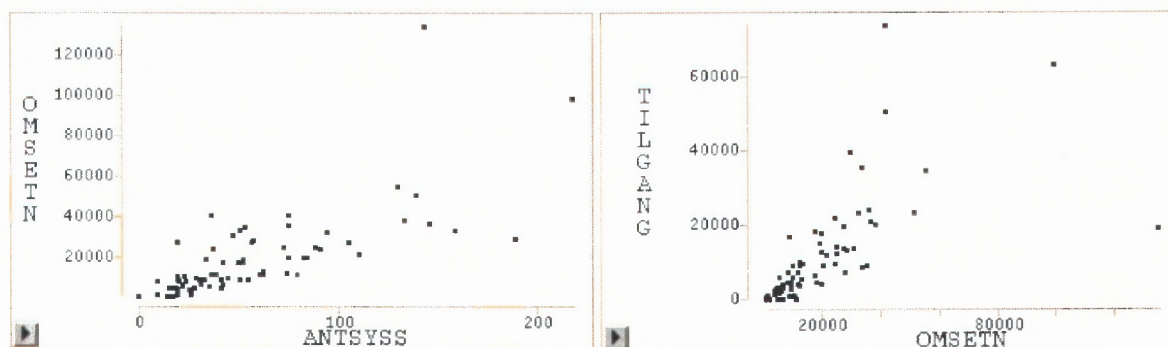
7.2.4. Utelate observasjoner

Alle grafer kan nå redigeres slik som beskrevet i kap. 4 og 5 for å få frem typiske fordelinger og sammenhenger.

Eksempelvis kan enkeltobservasjoner som fremtrer som utliggere i et plott utelates slik at plott av resten av observasjonene blir tydeligere. Marker observasjonene som skal utelates, velg **Edit → Observation → Hide in graph**.

Ved å utelate observasjon nummer 202 blir figur 7.7 til figur 7.8.

Figur 7.8. Ekstrem observasjon utelatt

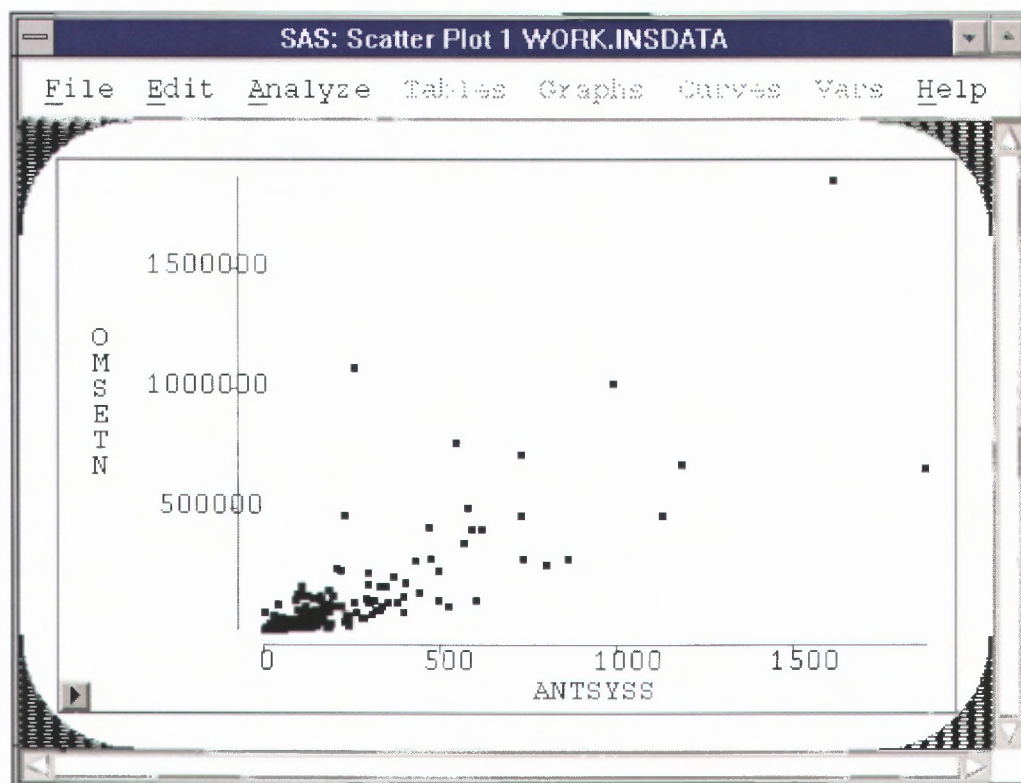


7.3. Fryse bildet

Når man fjerner observasjoner fra plottet, kan man ønske å sammenligne plottene før og etter.

Vanligvis vil alle grafer endre seg nå man redigerer en graf. For å beholde et tidligere plott må man derfor «fryse» det vinduet plottet ligger i. Fra det aktuelle vinduet velges **Edit → Windows → Freeze**. Når et vindu er frosset, markeres dette med skravering i hjørnene.

Figur 7.9. Frosset vindu



8. Analyse av en variabel

Analysemuligheter i "Distribution"

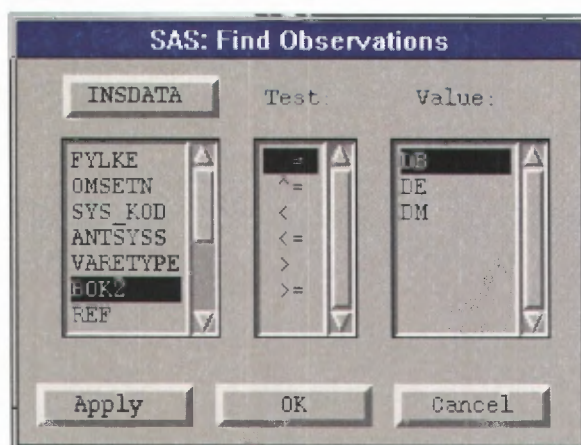
- Deskriptiv analyse av en variabel
 - boksplott
 - histogram
 - momenter (gjennomsnitt, standardavvik, variasjonskoeffisienter, etc)
 - prosentiler (median, øvre og nedre kvartil, maksimum og minimum, etc)
- Kurvetilpassing - ikkeparametrisk
- Parametrisk analyse
 - sannsynlighetsfordeling
 - konfidensintervall
 - kurver
- Teste fordelinger

Her legges hovedvekten på enkel deskriptiv analyse av fordelingen til variable. Ytterligere statistisk analyse krever kjennskap til statistiske metoder og vil bare antydes her.

Analysen utføres på en begrenset del av datasettet.

Gå til datavinduet, og velg **Edit → Observations → Find**. Sett "BOK2" = «DB», og klikk **OK**.

Figur 8.1. Finn observasjoner

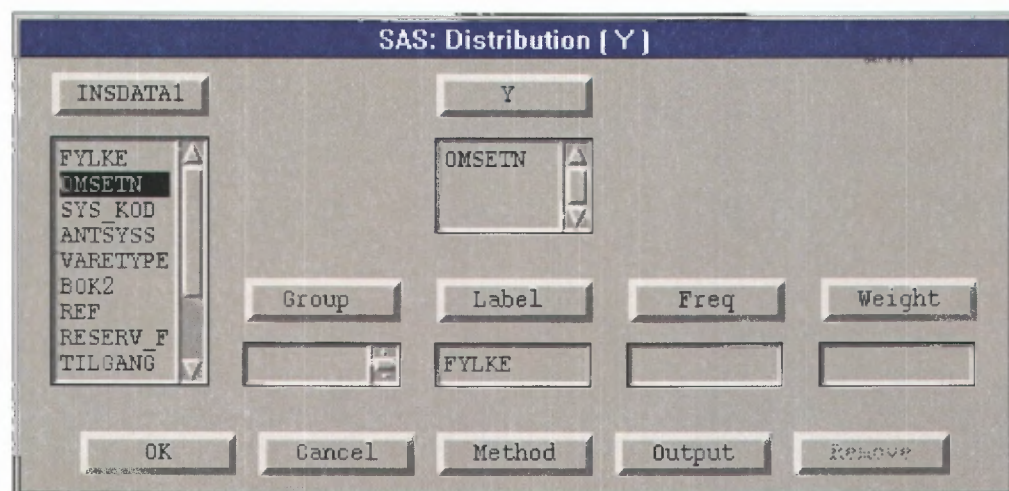


Vi har nå markert alle observasjonene som sorterer under næringsgruppe «DB». Bruk høyre musetast i datavinduet, og velg **Extract**. Da åpnes et nytt datavindu, betegnet WORK.INSDATA1, som kun inneholder observasjonene fra næringsgruppe «DB». I resten av dette kapittelet er dette datavinduet vårt arbeidsdatasett.

8.1. Fordeling

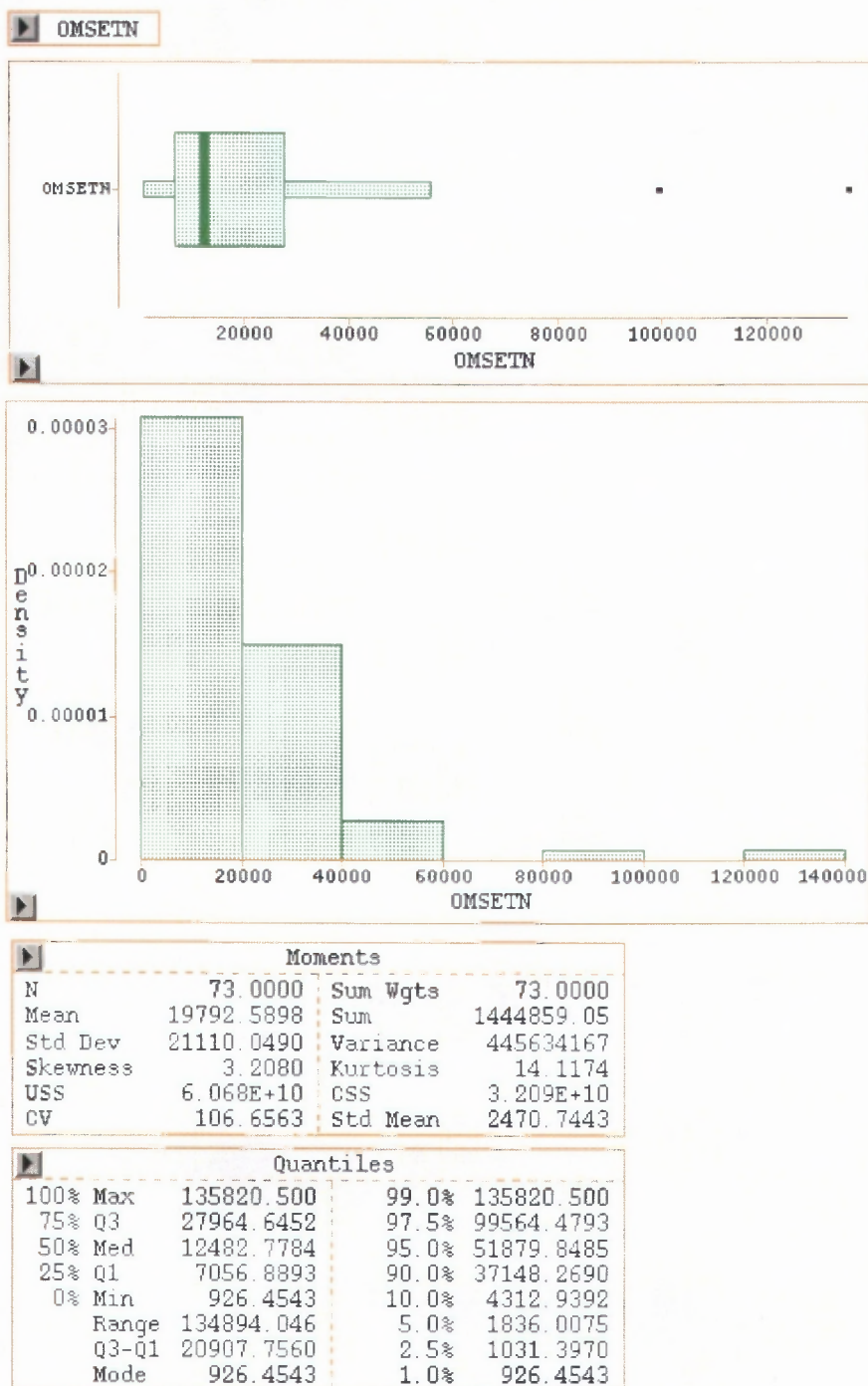
Gå til nytt datavindu, velg **Analyze → Distribution (Y)**, og et dialogvindu åpnes:

Figur 8.2. Dialogvindu for Distribution.



Velg omsetning (OMSETN) som Y-variabel og klikk **OK**. Et vindu som inneholder to grafer og to tabeller åpnes, se figur 8.3. Grafene er et boksplott og et histogram, og tabellene er en moment-tabell og en kvantiltabell.

Figur 8.3. Eksempel på Distribution av omsetning.

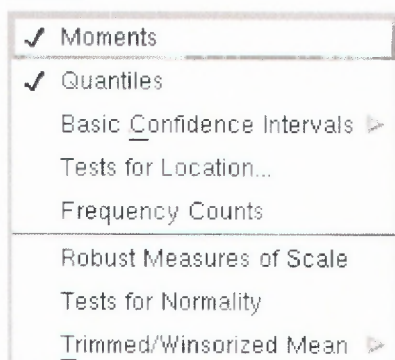


Boksplottet vises som et vanlig boksplott fra kapittel 5, mens histogrammet nå angis med stolpehøyder som tilsvarende en sannsynlighetsfordeling (det vil si at sum areal av stolpene er lik 1). Av resultatene fra Momenttabellen nevnes N (antall observasjoner med verdi på omsetning), Mean (gjennomsnitt), Std Dev (standardavvik), Skewness (skjevhet) og Kurtosis (mål for "tyngden" av halene i fordelingen).

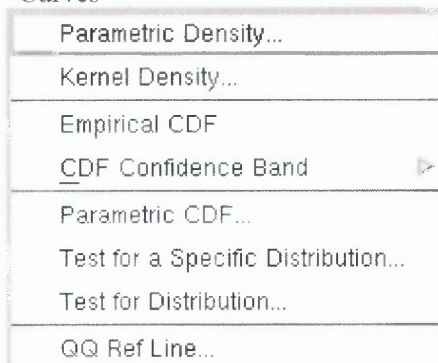
8.1.1. Menyer for redigering

- Rullegardinmenyen for dette vinduet har tre ekstra valg, det er «Tables», «Graphs» og «Curves». Med «Tables» kan vi legge inn flere tabeller, samt noen enkle hypotesetester, med «Graphs» kan vi legge inn flere grafer, og med «Curves» kan vi legge kurver inn i grafene. Nytteverdien av disse mulighetene avhenger av kunnskaper i statistisk analyse.

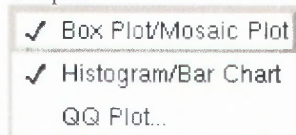
Tables



Curves

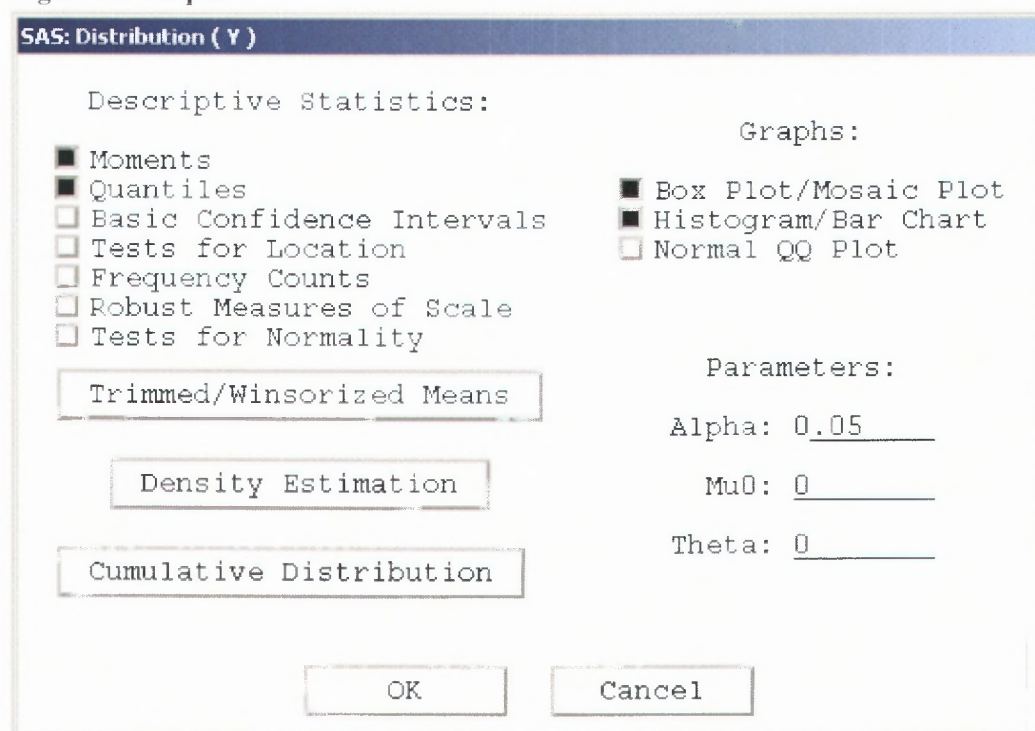


Graphs

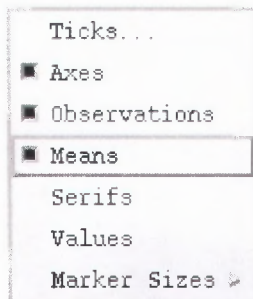


- Vi kan også velge tabeller, grafer og kurver ved Output-tasten i dialogvinduet til Distributionvinduet. Disse valgene er de samme som i rullegardinmenyen over.

Figur 8.4. Output



- I alle grafer er det popup-meny, som styres ved høyre musetast, eller ved piltast i nedre venstre hjørne av grafboksene. Høyre musetast utenfor grafene, men inne i vinduet, genererer rullegardinmenyen som popup-meny.

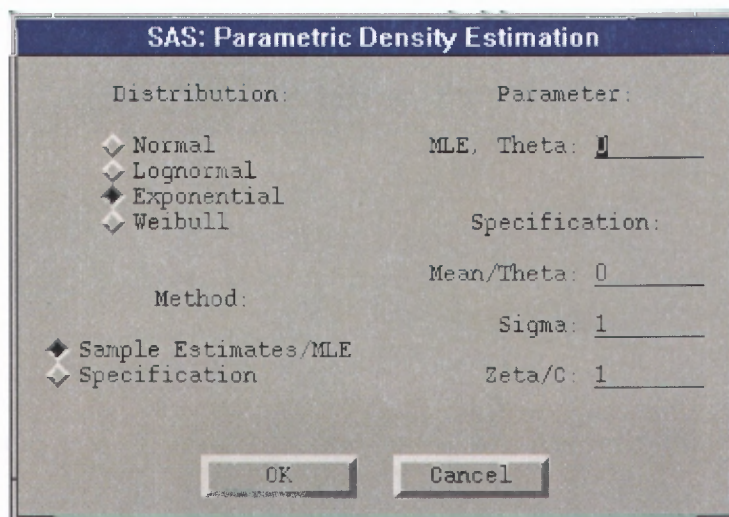


8.1.2. Utvidet envariabelanalyse.

Legg middelværdien med standardavviket inn i boksplottet. Bruk høyre musetast og velg **Means**.

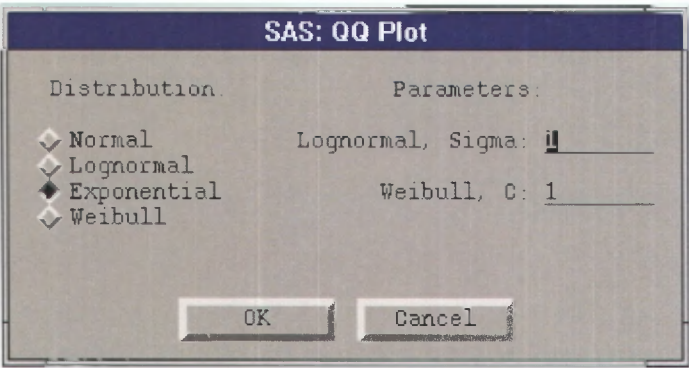
Legg en eksponentiell fordeling på histogrammet: **Curves → Parametric density → Exponential → OK**.

Figur 8.5. Dialogvindu for estimering av tetthet



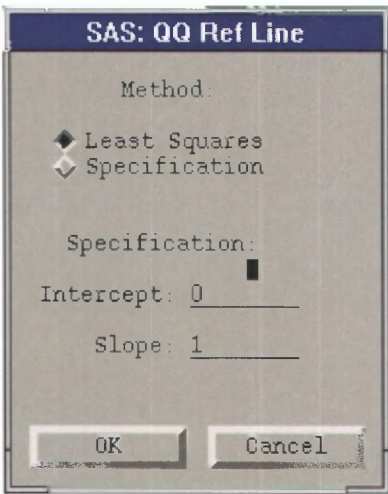
Legg inn et QQ-plott med en eksponentiell fordeling: **Graph → QQ Plot → Eksponential → OK**.

Figur 8.6. Dialogvindu for QQ-plott



Legg også inn en referanselinje for QQ-plottet: **Curves** → **QQ Ref line** → **Least Squares** → **OK**.

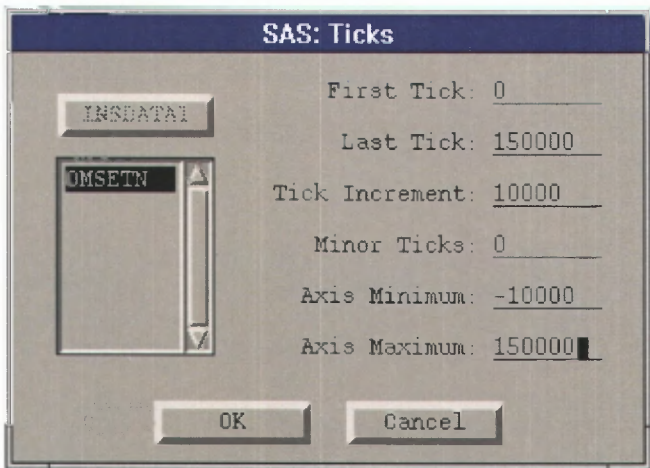
Figur 8.7. Dialogvindu for referanselinje til QQ-plott



Legg inn et 90 prosents konfidensintervall for gjennomsnittet: **Tables** → **Basic Confidence Interval** → **90 %**. Gjenta dette for 95 % og 99 %.

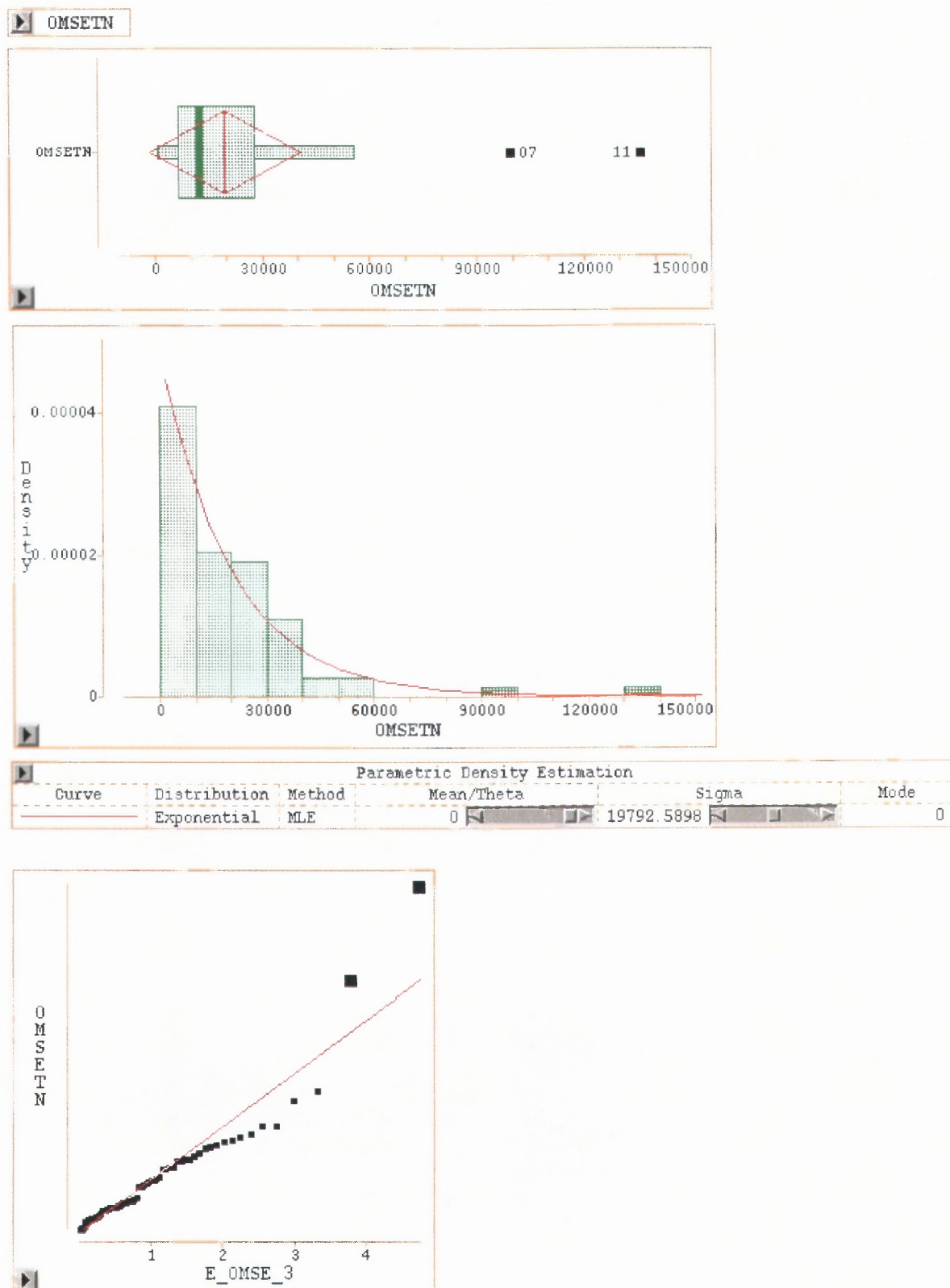
Juster aksene til passende verdier i begge grafer (sett Ticks som i figur 8.8).

Figur 8.8. Dialogvindu for startverdier til histogram

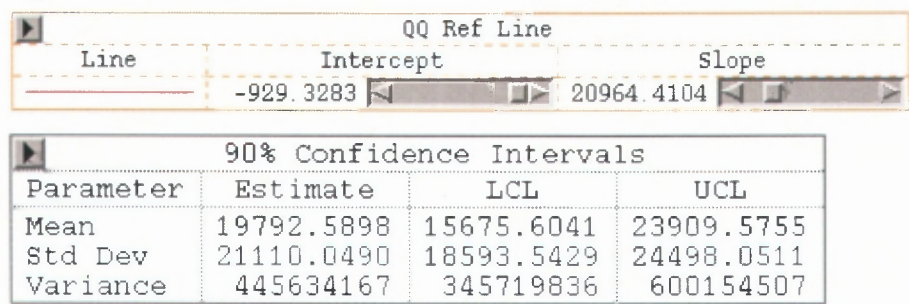


Hvis alt er riktig utført skal vi ha et distribution-vindu som i figur 8.9.

Figur 8.9. Redigert versjon av Distribution



Figur 8.9. - fortsatt



9. Multivariabel analyse

Analysemuligheter i "Multivariate"

- Deskriptiv analyse av flere variable
 - momenter (gjennomsnitt, standardavvik) for hver variabel
 - korrelasjonskoeffisienter
 - plott-matrise
- Konfidensellipser
- Analysere sammenhenger mellom variable
 - Prinsipal komponent analyse for redusert dimensjon
 - Canonical correlation analysis and maximum redundancy analysis (sammenheng mellom to sett intervallvariable)
 - Canonical discriminant analysis (sammenheng mellom en nominal variabel og et sett intervallvariable)

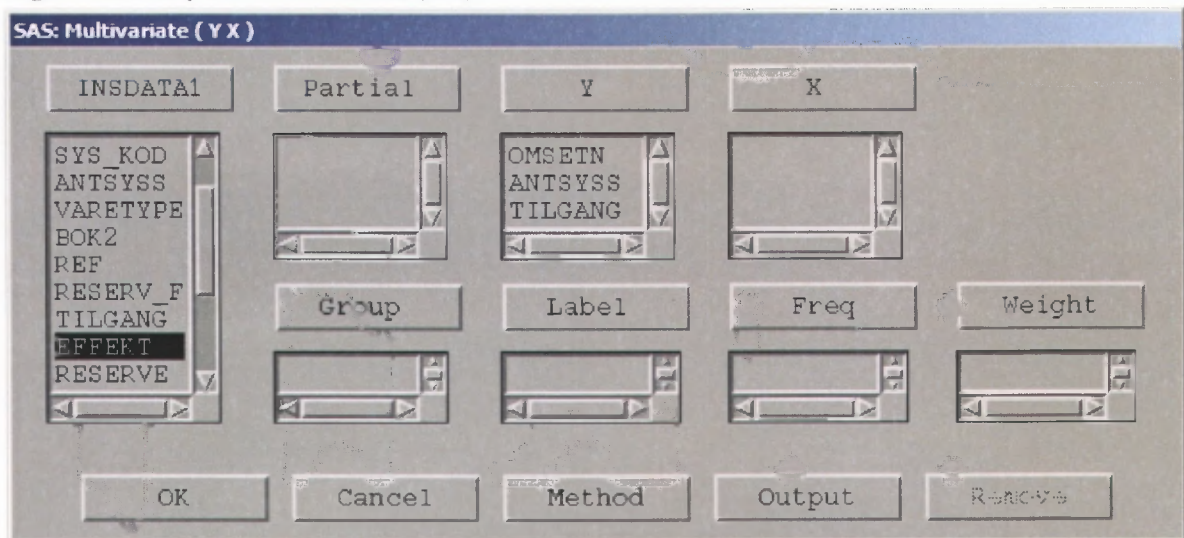
Multivariabel analyse er å studere sammenhengen mellom flere variable . SAS/INSIGHT gir flere muligheter her, blant annet Prinsipal komponent analyse. Slik analyse krever en del kunnskaper i statistisk metode, så her skal vi bare vise mulighetene for analyse av sammenhenger mellom par av variable.

Velg datasettet som bare inneholder næring DB, marker variablene OMSETN, ANTSYSS, TILGANG og EFFEKT i kolonnehodet, og velg

Analyze → Multivariate (YX)

eller bruk menyen

Figur 9.1. Meny for Multivariate (YX)



Vinduet som åpnes inneholder en tabell med enkel statistikk for hver enkelt variabel som inngår i analysen, og en tabell med korrelasjonen mellom variablene, se figur 9.1. Korrelasjonen måler graden av lineær avhengighet mellom variablene.

Figur 9.1. Enkel multivariabel analyse.

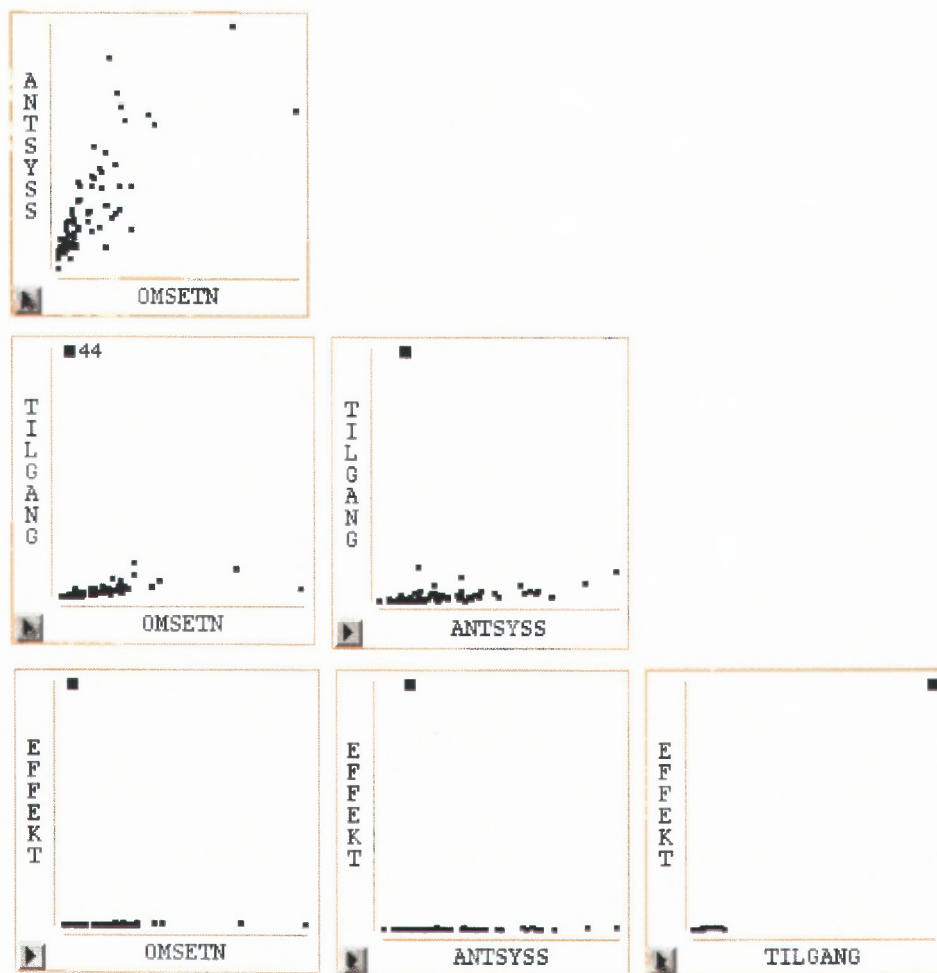
► OMSETN ANTSYSS TILGANG EFFEKT					
► Univariate Statistics					
Variable	N	Mean	Std Dev	Minimum	Maximum
OMSETN	73	19792.5898	21110.0490	926.4543	135820.500
ANTSYSS	73	56.6986	44.3003	0	219.0000
TILGANG	73	18988.7945	63874.7026	0	544080.000
EFFEKT	73	68076.3014	494508.660	0	4233872.00

► Correlation Matrix				
	OMSETN	ANTSYSS	TILGANG	EFFEKT
OMSETN	1.0000	0.7359	0.0677	-0.0602
ANTSYSS	0.7359	1.0000	0.0396	-0.0695
TILGANG	0.0677	0.0396	1.0000	0.9798
EFFEKT	-0.0602	-0.0695	0.9798	1.0000

Korrelasjonen mellom OMSETN og TILGANG har verdien 0,0677. Det kan altså se ut som om det ikke er noen lineær sammenheng mellom omsetning og ordretilgang. Samme forhold gjør seg gjeldende mellom omsetning og effektivt ordre (EFFEKT). Vi skal se at dette antakelig er feil.

Via rullegardinmenyen kan vi hente inn andre tabeller og grafer, f.eks. vil en plott-matrise være nyttig her. Velg **Graphs → Scatter Plot Matrix** i rullegardinmenyen, og en plott-matrise legges inn i øverst i analysevinduet.

Figur 9.2. Plott-matrise.



I plottet ser vi en tydelig utligger, det er observasjon nr. 44. Denne observasjonen har nokså sikkert feil i verdiene for variablene TILGANG og EFFEKT, og vil i stor grad påvirke verdiene i korrelasjonstabellen der disse inngår. Vi fjerner utliggeren fra beregningene.

Marker observasjonen og velg

Edit → Observations → Exclude in Calculations.

Utliggeren er da markert med et kryss i plottene, mens tabellene får nye verdier. Spesielt ser vi at korrelasjonen mellom OMSETN og TILGANG har øket til 0,6571, altså har vi en tydelig lineær sammenheng mellom omsetning og ordretilgang etter at utliggeren er fjernet.

Figur 9.3 Revidert enkel multivariabel analyse.

OMSETN ANTSSYSS TILGANG EFFEKT					
Univariate Statistics					
Variable	N	Mean	Std Dev	Minimum	Maximum
OMSETN	72	19985.9812	21192.9713	926.4543	135820.500
ANTSSYSS	72	57.1528	44.4398	0	219.0000
TILGANG	72	11695.8611	14146.1987	0	74798.0000
EFFEKT	72	10218.0278	12954.4856	0	63791.0000

Correlation Matrix				
	OMSETN	ANTSSYSS	TILGANG	EFFEKT
OMSETN	1.0000	0.7342	0.6571	0.6967
ANTSSYSS	0.7342	1.0000	0.5706	0.6970
TILGANG	0.6571	0.5706	1.0000	0.8106
EFFEKT	0.6967	0.6970	0.8106	1.0000

Menyen under Multivariat (YX) (figur 9.1.) inneholder mange muligheter. Man må alltid velge minst en Y-variabel. Korrelasjonsberegninger, symmetriske, som ovenfor er resultatet hvis alle variable velges som Y-variable. X brukes ved sammenligninger av to sett variable.

Videre analyse under multivariabel analyse spesifiseres ved undermenyen "Output".

Figur 9.4. Outputmenyen

SAS: Multivariate (Y X)

Descriptive Statistics:

- ☒ Univariate
- ☐ SSCP
- ☐ CSSCP
- ☐ COV
- ☒ CORR
- ☐ CORR p-values
- ☐ CORR Inverse
- ☐ Pairwise CORR

Bivariate Plots:

- ☐ Scatter Plot Matrix
- ☐ 80% Pred. Conf. Ellipse

☐ Principal Component Analysis

☐ Canonical Correlation Analysis

☐ Maximum Redundancy Analysis

☐ Canonical Discriminant Analysis

Disse valgene hører inn under analysemetoder som ikke tas med her.

10. Modelltilpassing

Analysemuligheter i "Fit"

- Regresjonsanalyse
 - enkel lineær regresjon
 - multippel regresjon
- Generaliserte lineære modeller
 - logistisk regresjon
 - variansanalyse
- Kurvetilpasning
 - parametriske modeller
 - ikke-parametriske modeller

Vi skal her presentere enkel lineær regresjon, samt vise en del muligheter i multippel regresjon.

10.1. Enkel lineær regresjon

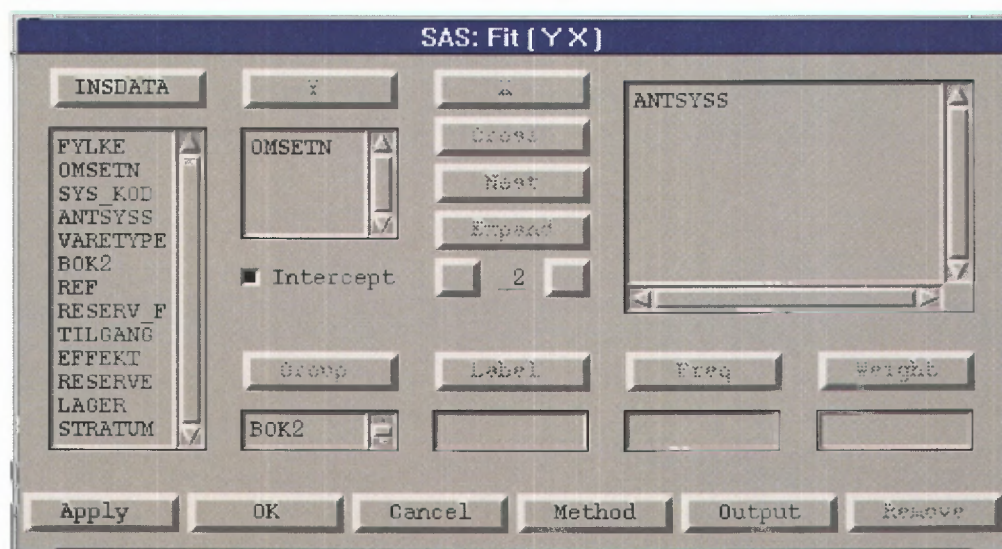
Vi skal se på sammenhengen mellom omsetning og antall sysselsatte. Vi ser på en modell på formen

$$Y = \alpha + \beta X + \varepsilon$$

der Y (respons) er omsetning, X (prediktor) er antall sysselsatte og ε (residual) er den delen av Y som ikke kan forklares av modellen. α (alfa) og β (beta) er parameterne som spesifiserer modeller, og som vi skal estimere

Gå til datavinduet, velg **Analyze → Fit (Y X)**, og et dialogvindu slås opp.

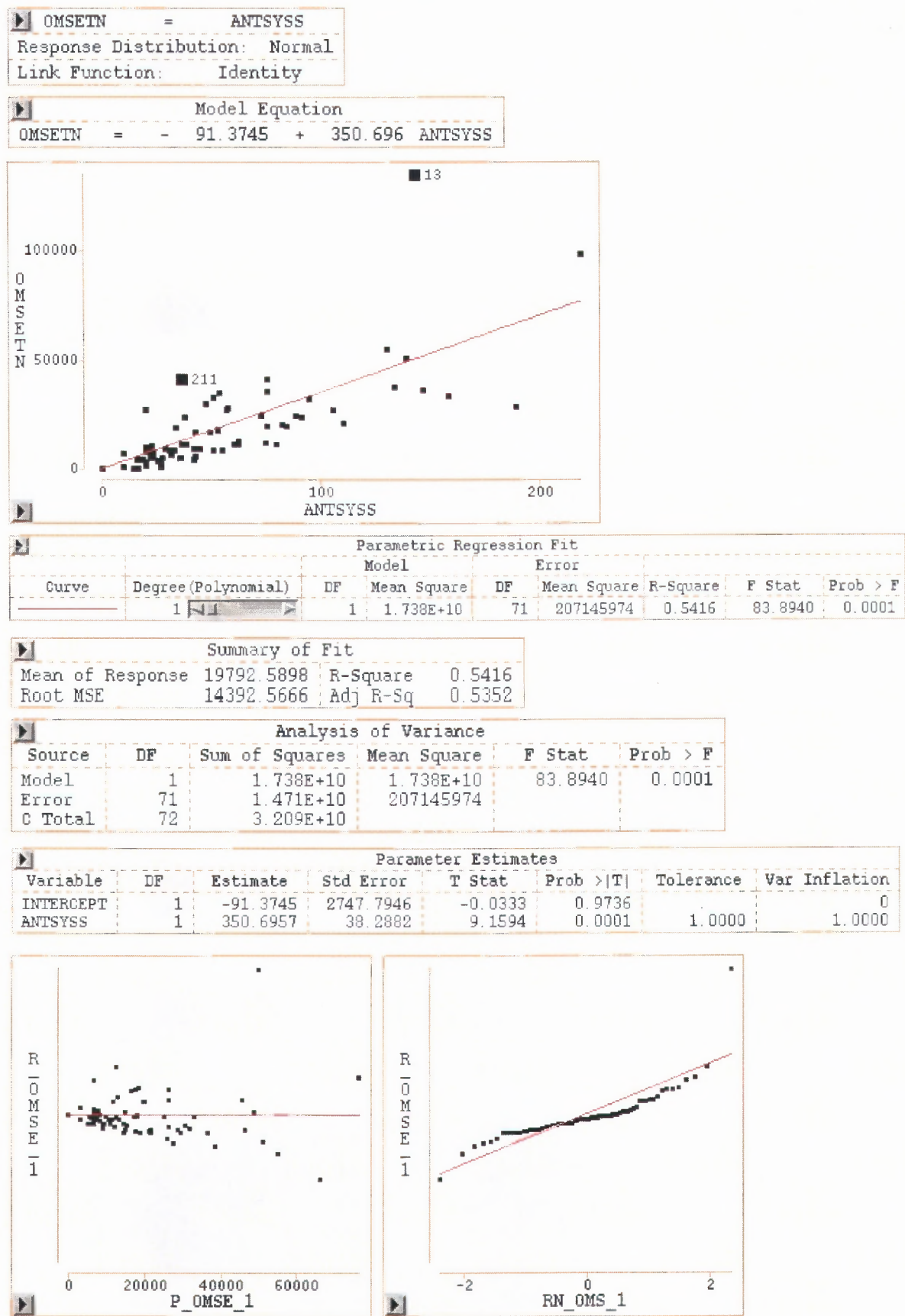
Figur 10.1. Dialogvinduet for modelltilpassing



Velg OMSETN som Y-variabel, ANTSYSS som X-variabel og BOK2 (næring) som gruppevariabel. Når det gjelder menyvalg, så angir «Method» valg av metode, og «Output» angir tabeller og grafer i analysen.

Klikk på **Output**, og ett nytt dialogvindu åpnes. Velg under «Residual Plots»: **Residual Normal Q Q** → **OK**, klikk på ny **OK**. Et vindu med en regresjonsanalyse for hver næringsgruppe åpnes.

Figur 10.2. Enkel lineær regresjon for næringsgruppe DB



Næring DB vil være arbeidsdatasettet i resten av kapitlet.

Tabellen «Model Equation» viser estimert modell (regresjonsfunksjonen). Første graf viser et X-Y-plott av omsetning på antall sysselsatte, sammen med estimert modell (regresjonslinjen). Resterende grafikk og tabeller benyttes for å analysere tilpassingen. For å vurdere estimert modell kan vi betrakte tabellene «Summary of Fit», «Analysis of Variance» og «Parameter Estimates». De to siste grafene er henholdsvis et X-Y-plott av estimerte residualer på estimert modell, og et QQ-plott av estimerte residualer med referanselinje. Disse grafene antyder om dataene er i samsvar med forutsetningene for regresjonsanalysen.

10.1.1. Revidering av regresjonsanalysen

Det er noen observasjoner som er mulige utliggere i regresjonsanalysen, to slike er markert i figur 10.2. Med utliggere menes her observasjoner som har «stor» avstand til regresjonslinjen, disse trenger ikke nødvendigvis å være utliggere i annen analyse, f. eks. boksplokk. Det kan være hensiktsmessig å fjerne slike utliggere fra regresjonen for å studere modelltilpasningen.

Markere mulige utliggere samtidig som «Ctrl»-tasten holdes nede, og velg

Edit → Observations → Exclude in Calculation.

Outputmenyen gir muligheter for konfidensintervall for forventningsverdiene/regresjonslinjen (95% Mean Confidence Curves), men om man ønsker prediksjonsintervall for enkeltobservasjoner må man gå via

Curves → Confidence Curves → Prediction 95% → OK

Dette gir en ny regresjonsanalyse, for næringsgruppe «DB» som vist i figur 10.3. Sammenlign analysen i denne figuren med figur 10.2, og observer at ny regresjonsanalyse gir en bedre tilpassing av dataene.

Merknad

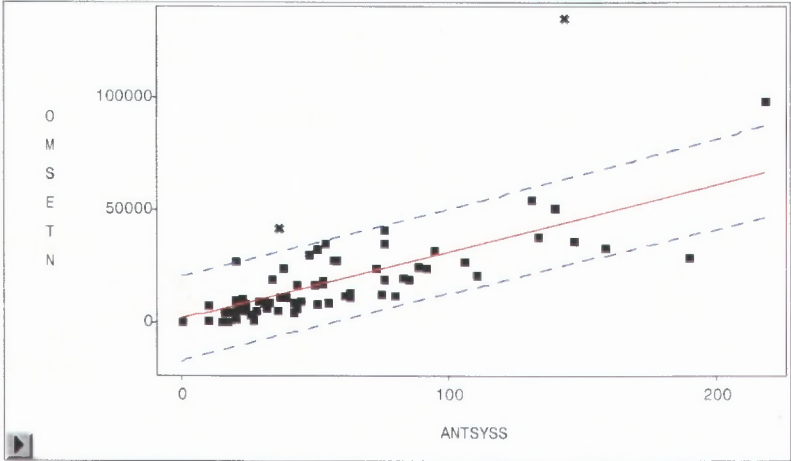
Vi skal være forsiktige med å basere en statistisk analyse på et datasett hvor «upassende» observasjoner er fjernet, men disse mulighetene kan gjerne anvendes som ledd i en innledende modellbyggerstrategi. Utliggere som ikke passer inn i en modell kan skyldes

- at modellen ikke passer
- at observasjonen er registrert med gale verdier
- spesielle observasjoner som ikke nødvendigvis skal passe inn i modellen

Figur 10.3. Revidert regresjonsanalyse

Model Equation

OMSETN	=	1222.30	+	298.207	ANTSYSS
--------	---	---------	---	---------	---------



Parametric Regression Fit

Curve	Degree(Polynomial)	Model DF	Mean Square	Error DF	Mean Square	R-Square	F Stat	Pr > F
—	1	1	1.185E+10	69	87182808.0	0.6632	135.89	<.0001

Confidence Curves

Type	Degree(Polynomial)	Coefficient	Lower Curve	Upper Curve
Prediction	1	0.9500	---	---

Summary of Fit

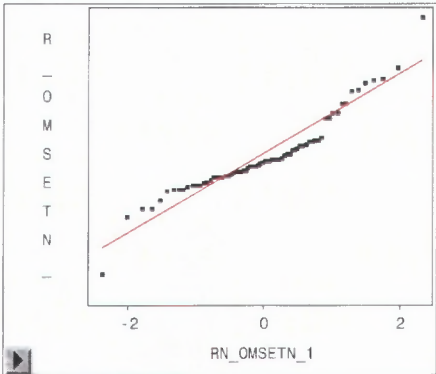
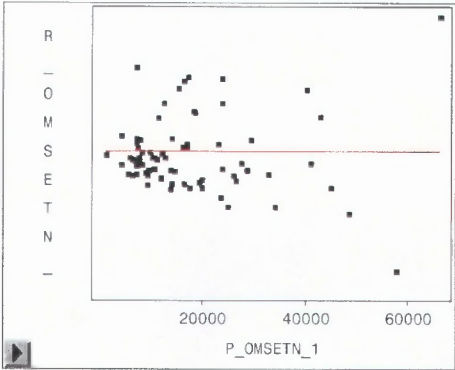
Mean of Response	17846.2832	R-Square	0.6632
Root MSE	9337.1734	Adj R-Sq	0.6584

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Stat	Pr > F
Model	1	1.185E+10	1.185E+10	135.89	<.0001
Error	69	6.016E+09	87182808.0		
C Total	70	1.786E+10			

Parameter Estimates

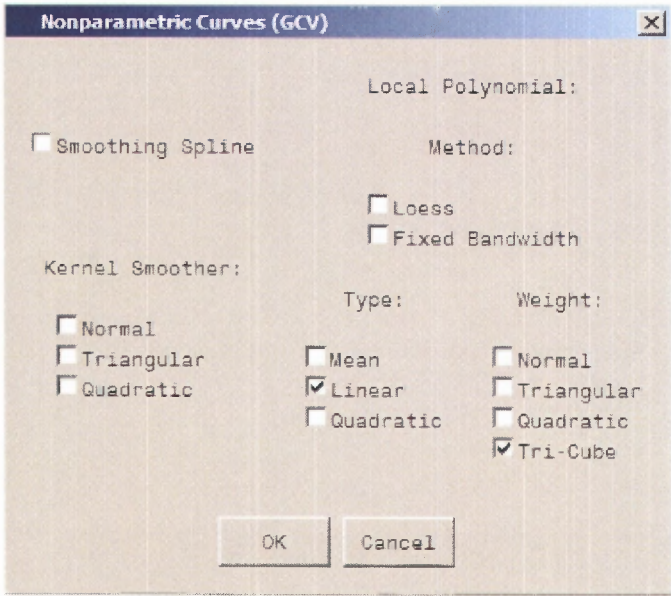
Variable	DF	Estimate	Std Error	t Stat	Pr > t	Tolerance	Var Inflation
Intercept	1	1222.3044	1805.9783	0.68	0.5008	.	0
ANTSYSS	1	298.2068	25.5811	11.66	<.0001	1.0000	1.0000



10.1.2. Ikke-parametrisk kurvetilpasning

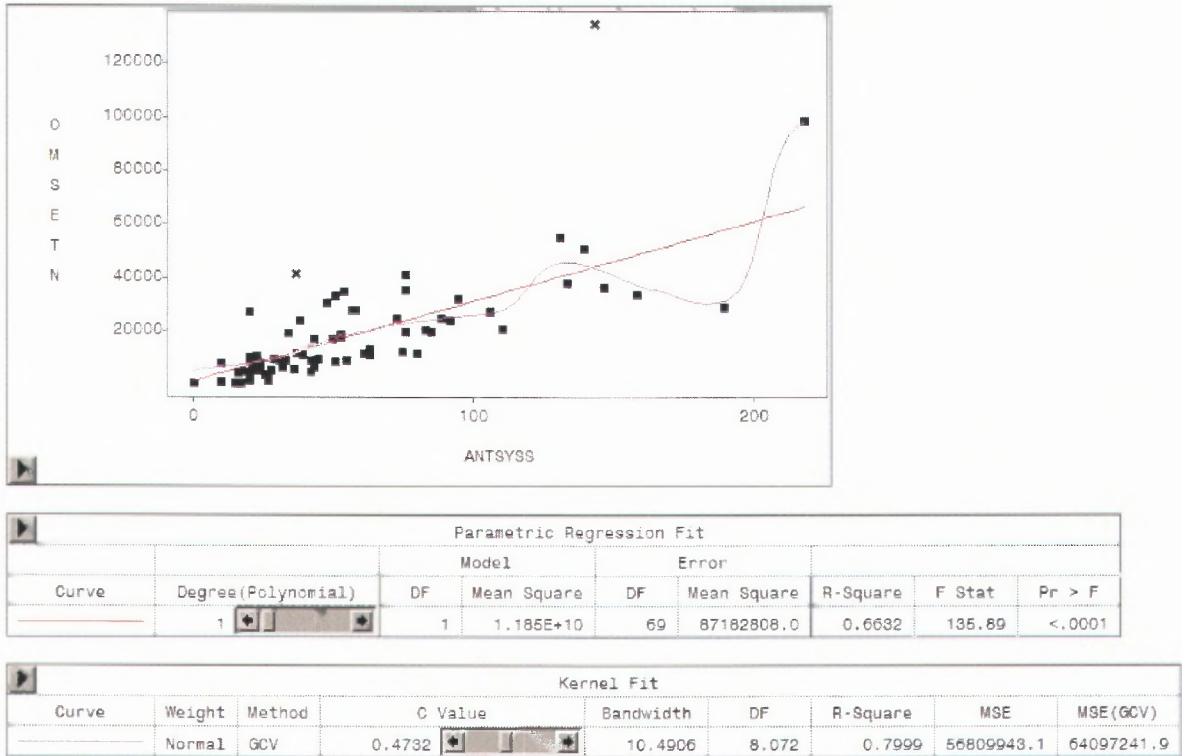
Menyer for ikkeparametrisk kurvetilpasning går via **Analyze → Fit (Y X)**. Fra **Output**-menyen velges **Nonparametric Curves**.

Figur 10.4. Meny for ikke-parametriske kurver



Her kan man for eksempel markere for **Normal** under **Kernel Smother** og få frem figur 10.5. Alternativt kunne man valgt **Curves → Kernel → OK**.

Figur 10.5. Kernel kurvetilpasning



Gliderne kan brukes til å bestemme grad av glatting.

10.2. Multippel regresjon/Variansanalyse

Vi kan ofte ha flere prediktorvariable X_1, X_2, X_3 , osv., som kan «forklare» en responsvariabel Y . Analysen som SAS/INSIGHT utfører i en slik situasjon avhenger av hvilket målenivå X 'ene har. Hvis alle X 'ene er intervallvariable kalles det ofte multippel regresjon, hvis alle X 'ene er nominelle variable kalles det bare variansanalyse. I en slik situasjon er det derfor viktig å vite «hva en vil».

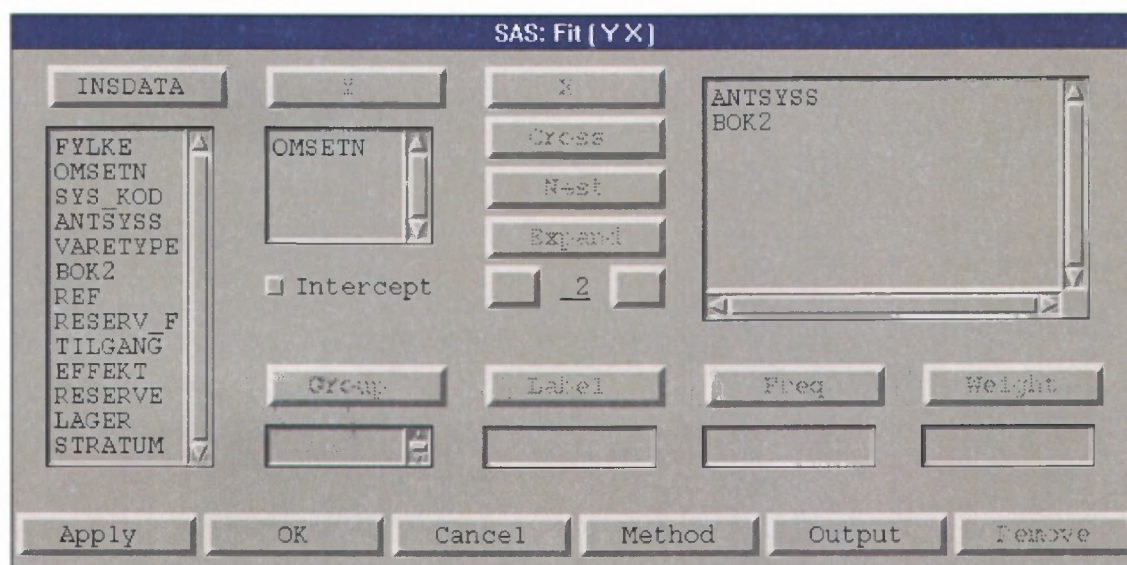
Vi skal her se på en kombinert regresjons- og variansanalysemodell. Vi skal forklare omsetning ved antall sysselsatte, som er en intervallvariabel, og næring, som er en nominell variabel. Modellen kan skrives på formen

$$Y = \alpha_j + \beta X + \varepsilon, \quad \text{for } j = 1, 2, 3$$

der Y er omsetning og X er antall sysselsatte som før. α_j , $j = 1, 2$ og 3 er parametere som angir konstantleddet som Y antas å ha for hver av næringene DB, DE og DM. Vi ser at modellen er nokså lik modellen i forrige seksjon, men vi tillater her konstantleddet å variere mellom næringene.

Gå til datavinduet, velg **Analyze** → **Fit (Y X)**, og et dialogvinduet slås opp.

Figur 10.6. Dialogvindu



Velg OMSETN som Y-variabel, velg ANTSYSS og BOK2 (næring) som X-variable. Marker ut det globale konstantleddet (Intercept), og legg på et Normal QQ-plott av residualene. Klikk **OK**, og et vindu med en kombinert regresjons- variansanalyse åpnes.

Figur 10.7. Regresjons- og variansanalyse

▶	OMSETN	=	ANTSYSS	BOK2
	Response Distribution: Normal			
	Link Function: Identity			

▶	Nominal Variable Information	
	Level	BOK2
	1	DB
	2	DE
	3	DM

▶	Parameter Information		
	Parameter	Variable	BOK2
	1	ANTSYSS	
	2	BOK2	DB
	3		DE
	4		DM

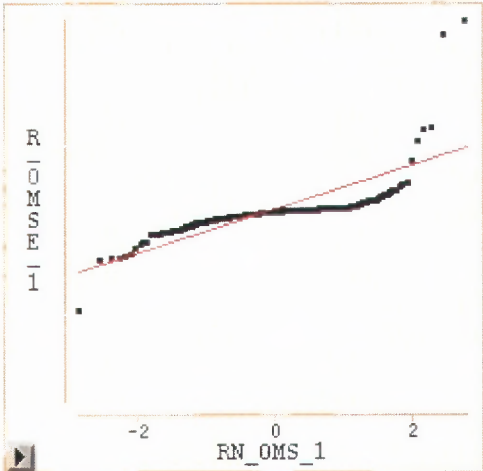
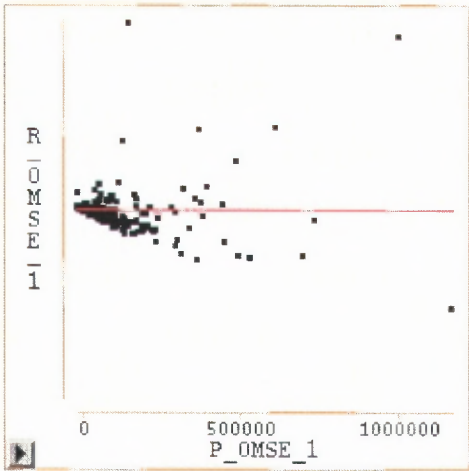
▶	Model Equation				
	OMSETN	=	638.397	ANTSYSS	- 16403.6
				P_2	+ 20574.9
				P_3	- 21375.0
				P_4	

▶	Summary of Fit			
	Mean of Response	84105.0702	R-Square	0.7100
	Root MSE	108835.456	Adj R-Sq	0.7059

▶	Analysis of Variance					
	Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
	Model	4	8.035E+12	2.009E+12	169.5823	0.0001
	Error	277	3.281E+12	1.185E+10		
	U Total	281	1.132E+13			

▶	Type III Tests					
	Source	DF	Sum of Squares	Mean Square	F Stat	Prob > F
	ANTSYSS	1	5.565E+12	5.565E+12	469.7852	0.0001
	BOK2	3	9.814E+10	3.271E+10	2.7617	0.0425

▶	Parameter Estimates								
	Variable	BOK2	DF	Estimate	Std Error	T Stat	Prob > T	Tolerance	Var Inflation
	ANTSYSS		1	638.3970	29.4538	21.6745	0.0001	0.6436	1.5537
	BOK2	DB	1	-16403.645	12847.2286	-1.2768	0.2027	0.9831	1.0172
		DE	1	20574.9141	17939.1040	1.1469	0.2524	0.8978	1.1139
		DM	1	-21375.000	10045.0666	-2.1279	0.0342	0.7029	1.4226



De tre første tabellene angir informasjon om variable og parametere som inngår i tilpassingen. Tabellen «Parameter Information» lister opp parametrene i tabellen, parameter 1 tilsvarer β , parameter 2 til 4 tilsvarer α_1 , α_2 og α_3 henholdsvis.

Tabellen «Model Equation» viser estimert modell. Notasjonen P_2, P_3 og P_4 her refererer til næringene DB, DE og DM henholdsvis. Resterende tabeller og grafikk benyttes for å analysere tilpassingen, og tolkes stort sett på samme måte som for enkel lineær regresjon.

10.3. Logistisk regresjon

Menyen "Method" under "Fit" gir muligheter for å velge sannsynlighetsfordelinger og linkfunksjoner for "Generalized Linear Models", her vist logistisk regresjon.

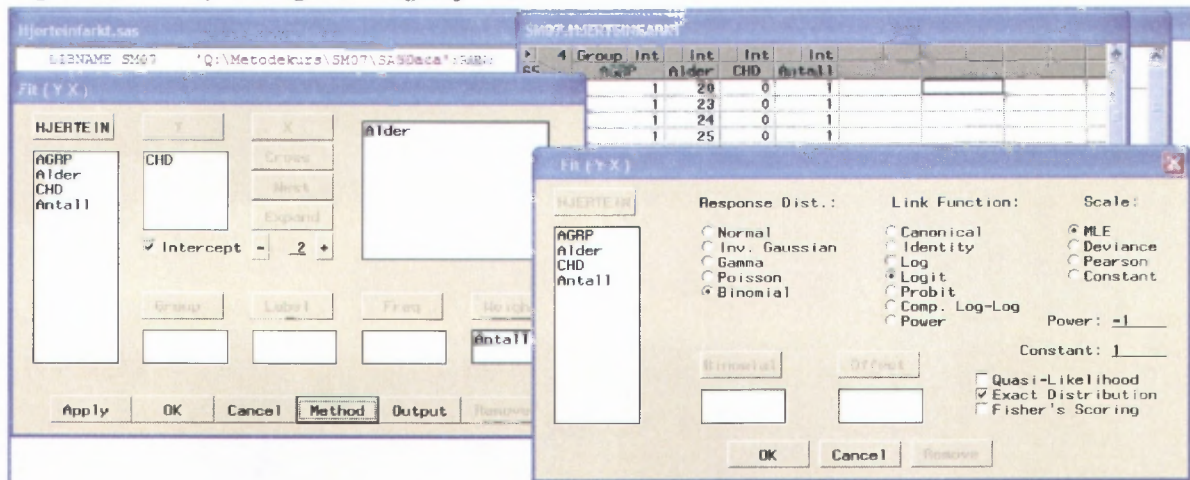
Logistisk regresjon brukes når den responsvariabelen i modellen er av den typen som bare kan ta to verdier, for eksempel "0" eller "1" og vi ønsker å studere hvordan sannsynligheten $\pi(\mathbf{x})$ for responsen "1" avhenger av en eller flere prediktorvariabelen $\mathbf{x} = x_1, x_2, \dots$. Formelen for sannsynligheten er da

$$\pi(\mathbf{x}) = \frac{\exp(\alpha + \beta\mathbf{x})}{1 + \exp(\alpha + \beta\mathbf{x})}$$

og vi skal prøve å estimere α og β . Følgende eksempel er basert på datasettet hjerteinfarkt som alle kan finne under Q:\Metodekurs\SM07\SASData.

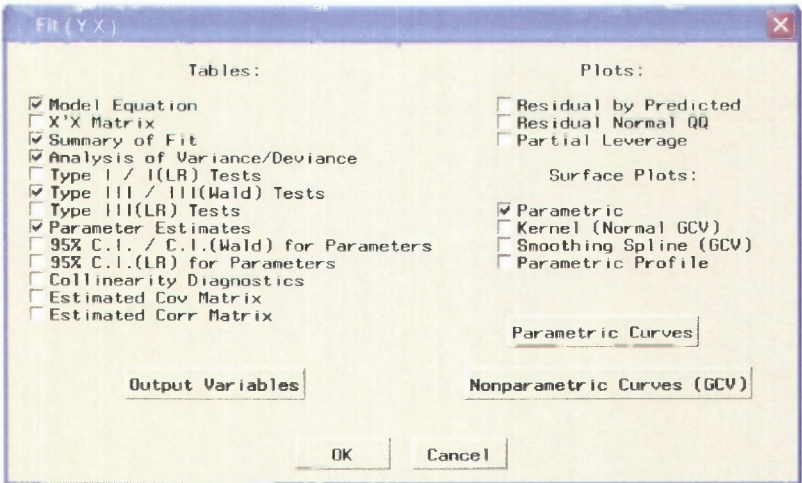
Hent datasettet frem i Insight på vanlig måte. Velg **Analyze --> Fit(Y X)**. I dialogvinduet velger du ALDER som x og CHD (= 1 for hjerteinfarkt, 0 ellers) som Y . Fjern AGRP hvis denne står i **Group** ruten. Sett Antall inn i **Weight** ruten. Klikk på **Method** og merk av binomial og logit i det vinduet som kommer opp.

Figur 10.8. Meny for logistisk regresjon



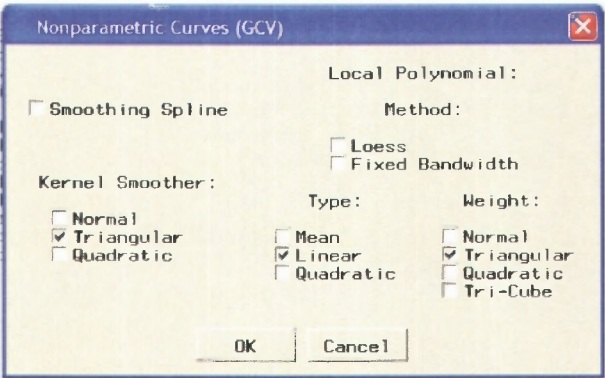
Trykk på **OK**. Ved å trykke på **Output** får dere en meny der dere kan velge hva dere vil ha ut.

Figur 10.9. Outputmeny



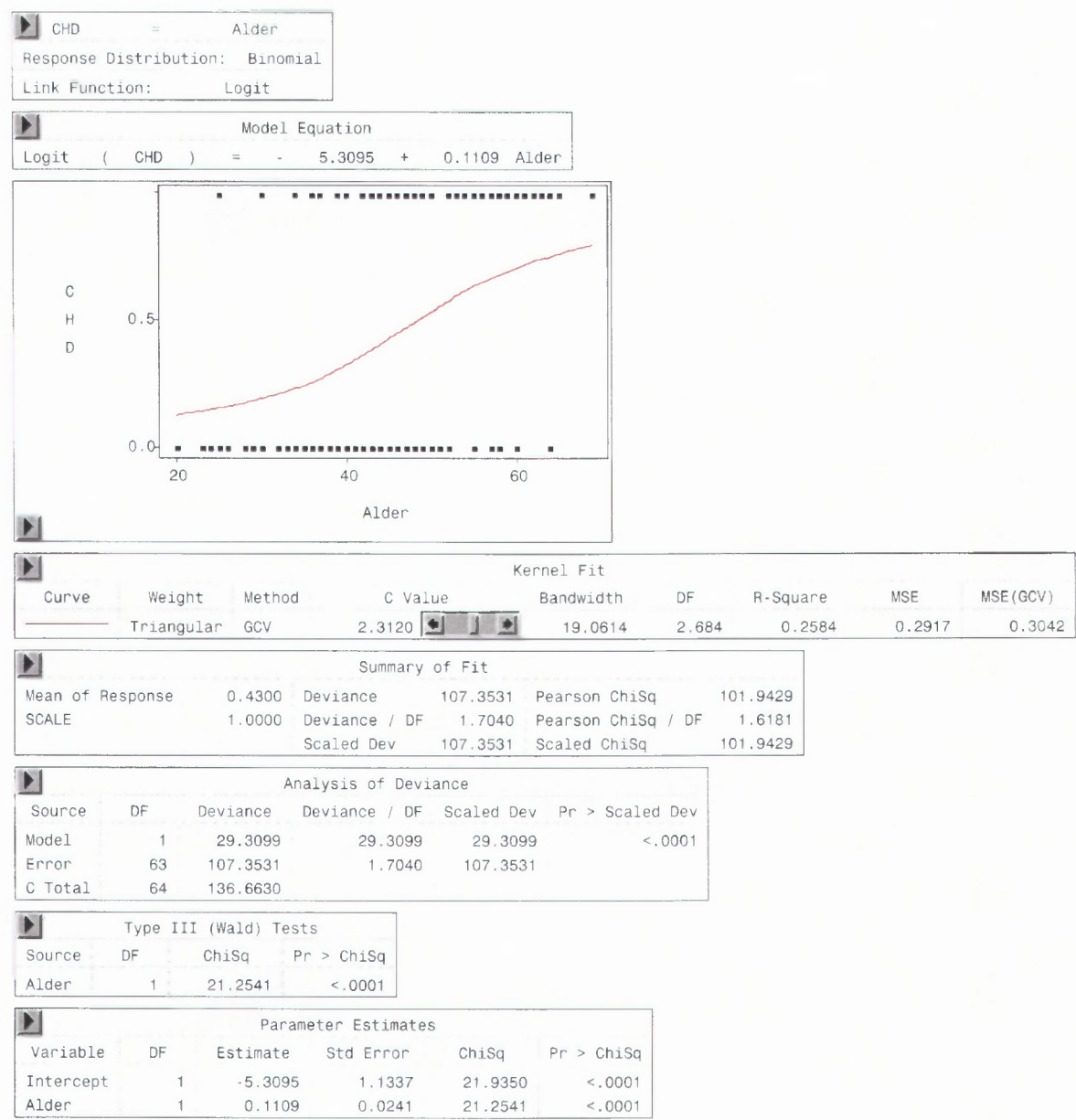
Avkrysningene i dette eksempelet er et minimumsforslag. Knappen **Parametric curves** virker ikke når en ber om logistisk regresjon. Knappen **Nonparametric Curves** virker. Hvis vi vil ha tegnet en regresjonskurve må vi fylle den ut, for eksempel slik:

Figur 10.10. Ikkeparametriske kurver



Trykk på **OK**, **OK** igjen og **Apply**. Du ser da følgende utskrift:

Figur 10.11. Logistisk regresjon



Den røde kurven er resultatet av valget i **Nonparametric Curves**. Den er imidlertid ikke helt slik den logistiske regresjonen legger opp til. Den kurven er dessverre ikke mulig å få ut i Insight. Ved bruk av PROC LOGISTIC sammen med PROC GPLOT er det imidlertid mulig. Prøv selv!

10.4. Modellkontroll

En nødvendig del av en regresjonsanalyse er kontroll av data oppfyller forutsetningene i modellen. Gjennom menyene Output og Outputvariables fra Fit-menyen kan man velge variable for slik kontroll.

Figur 10.12. Fit (YX) - Output

SAS: Fit (YX)

Tables:

- ☒ Model Equation
- ☐ X'X Matrix
- ☒ Summary of Fit
- ☒ Analysis of Variance/Deviance
- ☐ Type I / I(LR) Tests
- ☒ Type III / III(Wald) Tests
- ☐ Type III(LR) Tests
- ☒ Parameter Estimates
- ☐ 95% C.I. / C.I.(Wald) for Parameters
- ☐ 95% C.I.(LR) for Parameters
- ☐ Collinearity Diagnostics
- ☐ Estimated Cov Matrix
- ☐ Estimated Corr Matrix

Plots:

- ☒ Residual by Predicted
- ☐ Residual Normal QQ
- ☐ Partial Leverage

Surface Plots:

- ☒ Parametric
- ☐ Kernel (Normal GCV)
- ☐ Smoothing Spline (GCV)
- ☐ Parametric Profile

Output Variables

Parametric Curves

Nonparametric Curves (GCV)

OK **Cancel**

SAS: Output Variables

Generalized Residuals:

- ☐ Hat Diag
- ☐ Predicted
- ☐ Linear Predictor
- ☐ Residual
- ☐ Residual Normal Quantile
- ☐ Standardized Residual
- ☐ Studentized Residual
- ☐ Partial Leverage X
- ☐ Partial Leverage Y
- ☐ Cook's D
- ☐ Dffits
- ☐ Covratio
- ☐ Dfbetas

- ☐ Deviance Residual
- ☐ Standardized Deviance Residual
- ☐ Studentized Deviance Residual
- ☐ Pearson Residual
- ☐ Standardized Pearson Residual
- ☐ Studentized Pearson Residual
- ☐ Anscombe Residual
- ☐ Standardized Anscombe Residual
- ☐ Studentized Anscombe Residual

OK **Cancel**

11. Linjeplott

Et linjeplott er et X-Y-plott med en linje trukket gjennom alle punkter fra minste til største x-verdi. Det egner seg særlig for tidsseriedata, men brukes også til å fremstille funksjoner.

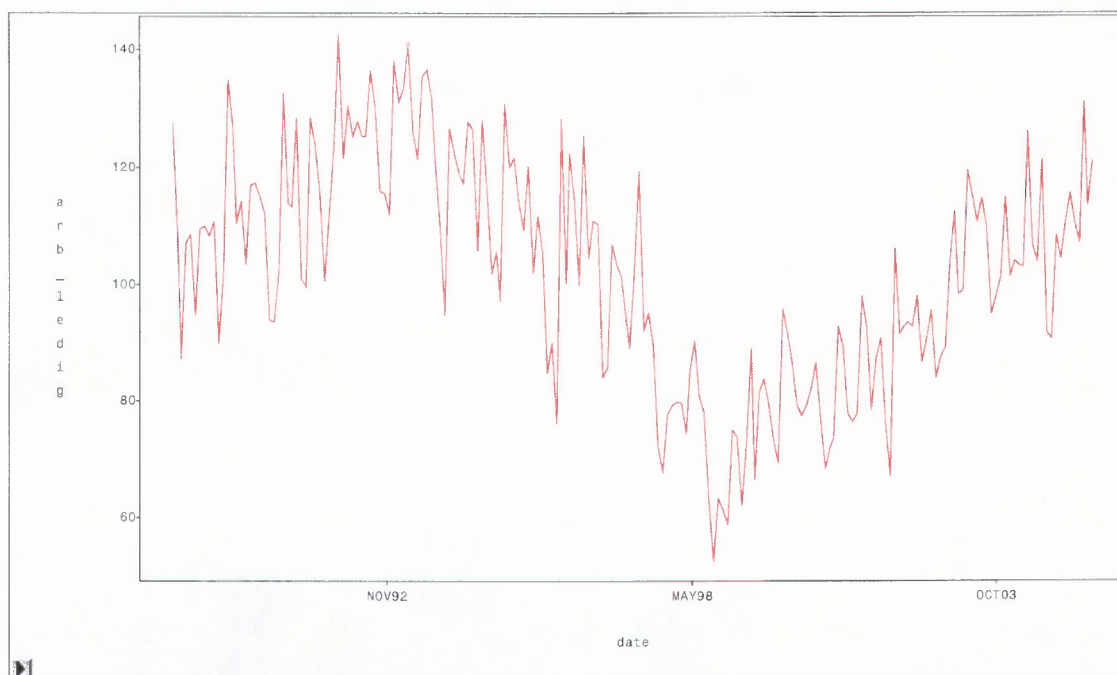
Linjeplott kan lages med en eller flere variable i samme graf

- En variabel
 - utvikling over tid
 - sesongvariasjoner
 - fordeling
- Flere variable
 - forhold mellom variable
 - sammenhenger i utvikling over tid
 - sesongvariasjoner

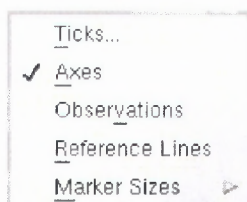
Linjeplott illustreres her av data fra AKU, antall arbeidsledige (arb_ledig) og tidsvariabel (date) som angir måned og år. Utviklingen av antall arbeidsledige får vi frem ved

Analyze → Lineplot(Y X) → arb_ledig → Y → date → X → OK

Figur 11.1. Linjeplott Arbeidsledige (1 000 personer)



Pop-up menyen gir de vanlige mulighetene for å endre grafen:

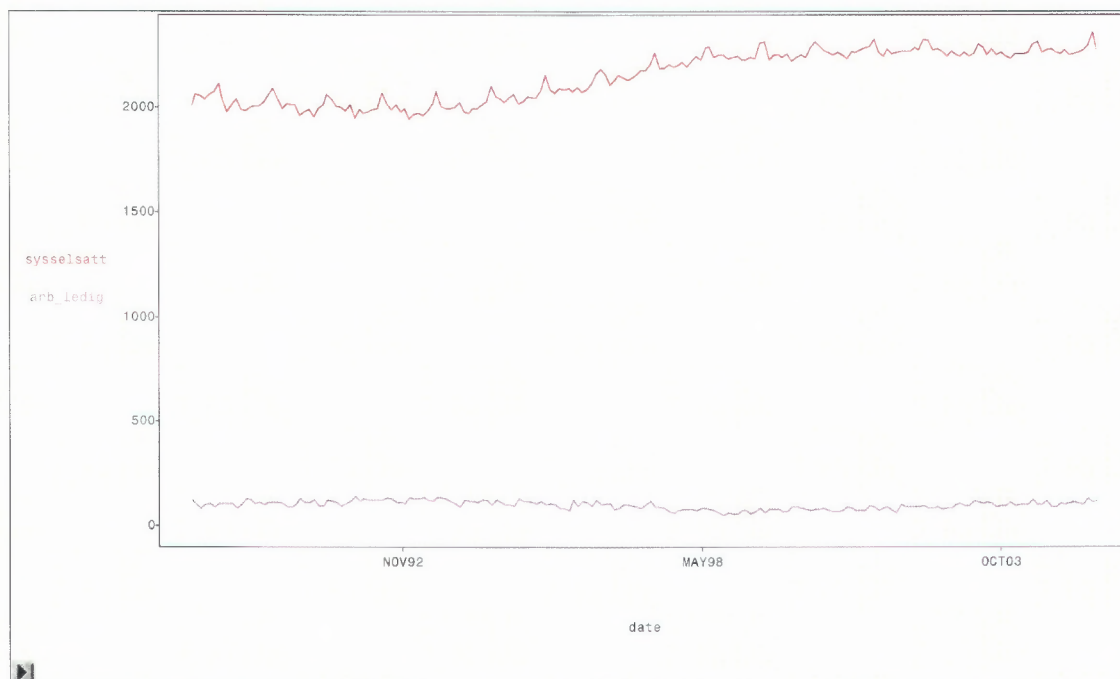


«Observations» markerer alle punktene i grafen.

Linjestil og linjefarger kan endres ved å markere variabelen, gå inn i Tools - vinduet og markere ønsket linjestil og/eller farge.

Hvis man skal ha flere variable inn i samme graf, gjøres dette ved å markere flere Y-verdier. Men alle må ha samme X - variabel og samme skala på Y-aksen. Eksempel på linjeplott med flere variable hentes også fra AKU med antall arbeidsledige (arb_ledig) og antall sysselsatte (sysselsatte) som Y-variable. Se figur 11.2.

Figur 11.2. Antall sysselsatte og antall arbeidsledige (i 1 000 personer)



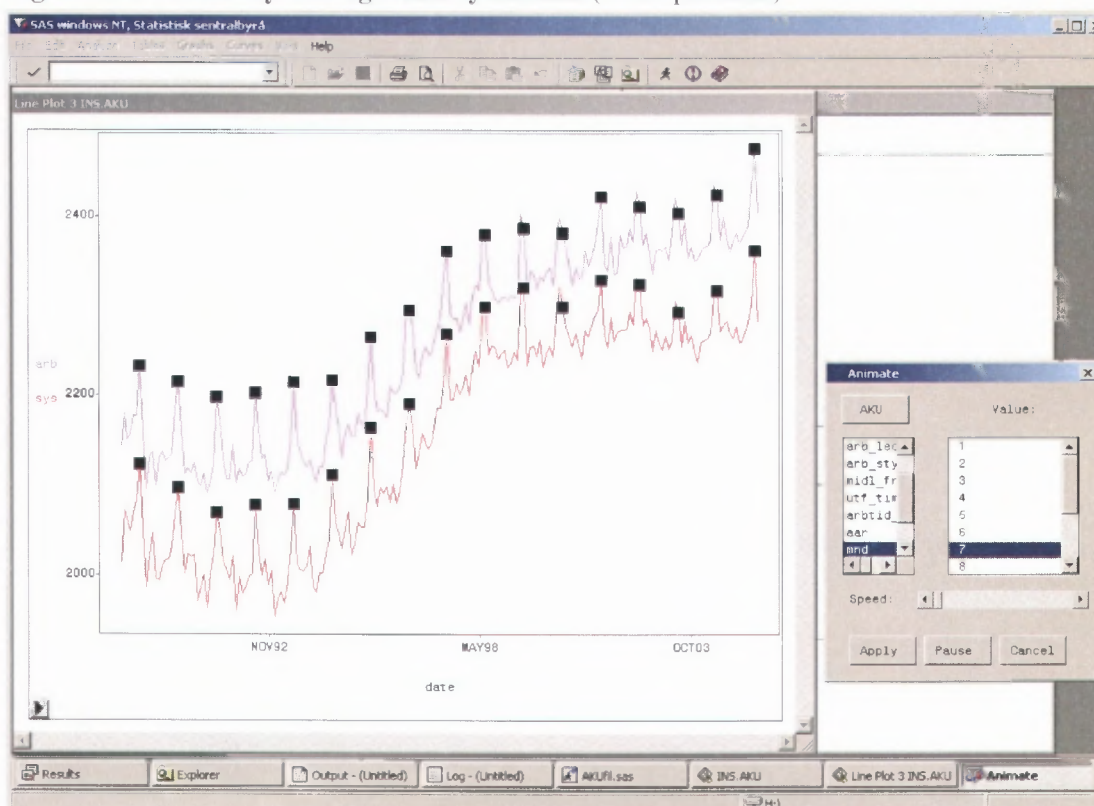
11.1.1. Animering

For tidsserier kan det være interessant å studere trend og sesongvariasjoner, og det kan vi gjøre ved å animere plottet. Velg **Edit → Windows → Animate**, og samme dialogvindu som i figur 4.9, men med andre variable, åpnes. Marker MND som «animeringsvariabel», og klikk på **Apply**. Hvis hastigheten ikke er for høy, ser vi at de enkelte månedsobservasjonene i tidsserien markeres suksessivt, se figur 11.3. Hastigheten kan justeres noe, men hvis det ikke er mulig å få den sakte nok, kan man bare markere en og en verdi av md. i "Animate"-vinduet.

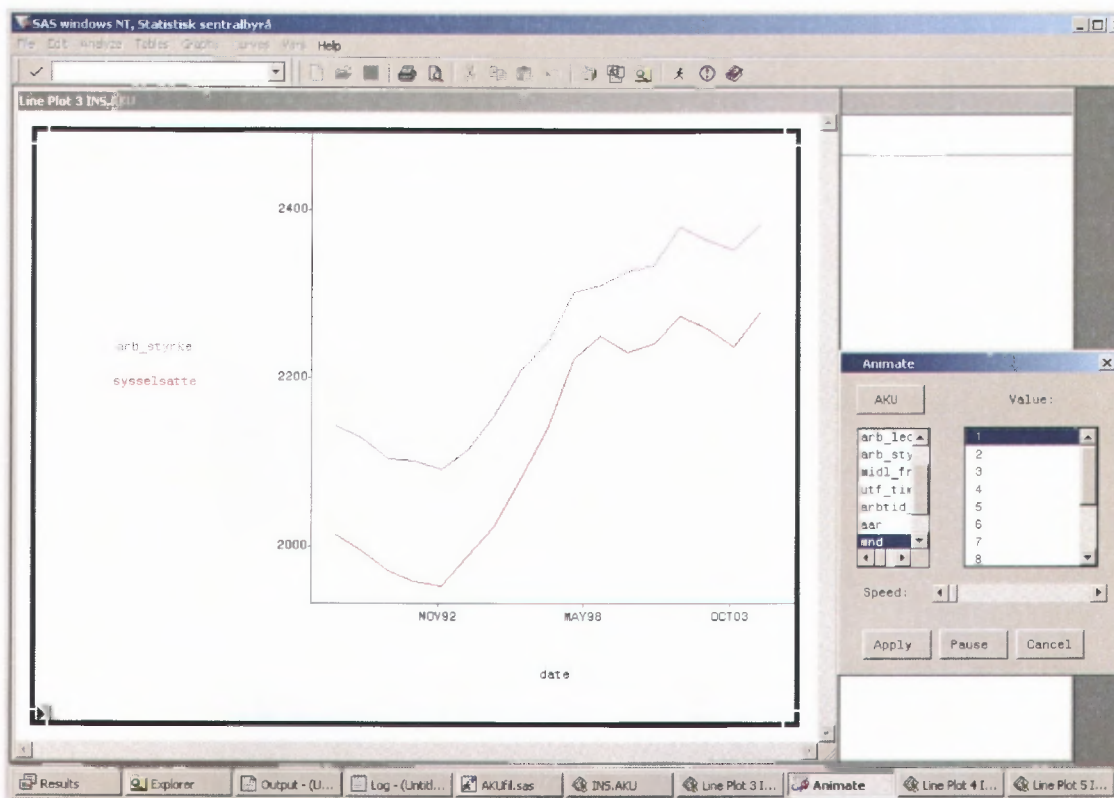
Figur 11.3. viser AKU-data med antall personer i arbeidsstyrken (arb_styrke) og antall sysselsatte (sysselsatte) som Y-variable og juli måned markert.

Hvis rammen rundt plottet er markert, får vi et linjeplott av månedsobservasjonene med en linje trukket gjennom observasjonene fra samme måned for alle årene. Det vil si at vi ser utviklingen fra år til år. Se figur 11.4.

Figur 11.3. Arbeidsstyrken og antall sysselsatte (1 000 personer)



Figur 11.4 Linjeplott med linje gjennom januarobservasjonene



12. Grafikk i flere dimensjoner

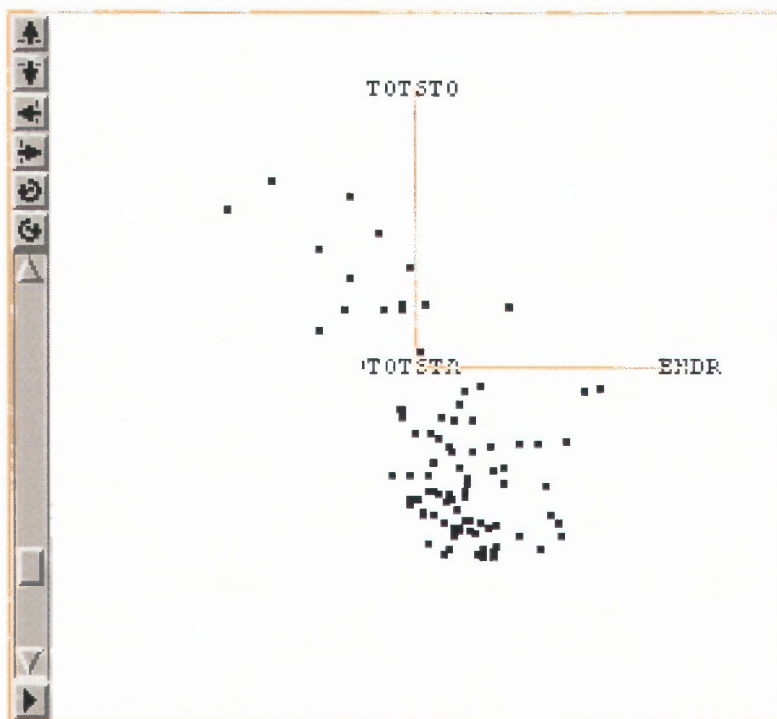
Grafisk presentasjon som dekker mer enn to dimensjoner kan i noen tilfelle vise flere sammenhenger mellom variable. Her viser vi

- roterende plott
- bruk av punktmarkering og/eller farger som en ekstra dimensjon
- kurver som nivåmarkering

12.1. Roterende plott

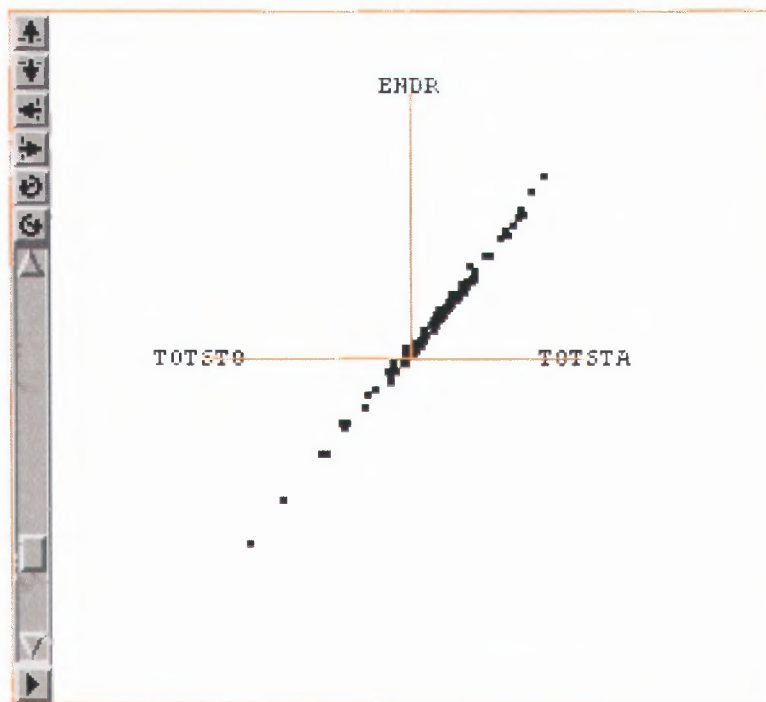
Vi velger datasettet "arb" som inneholder starter, opphør og endringer av antall arbeidsforhold, registrert på to ulike tidspunkt. Marker variablene TOTSTA, TOTSTO og ENDR i kolonnehodet i datavinduet, og velg **Analyze → Rotating Plot (Z Y X)**. Et vindu med et tredimensjonalt plott åpnes, hvor TOTSTA peker ut av skjermen, TOTSTO vises vertikalt og ENDR vises horisontalt.

Figur 12.1 Tredimensjonalt plott, sett rett ovenfra



Piltastene øverst langs venstre ramme kan benyttes til å rotere plottet. Farten på roteringen kan endres ved å stille på «slideren» under. Rotering kan også skje ved at man fører pilmarkøren ut mot kanten av grafen til den skifter til en liten hånd, klikke og dra musen og slippe musetasten mens musen beveges. Etter rotering av dette plottet kan vi ha en graf som figur 12.2 som viser at det er lineær sammenheng mellom de tre variablene i plottet.

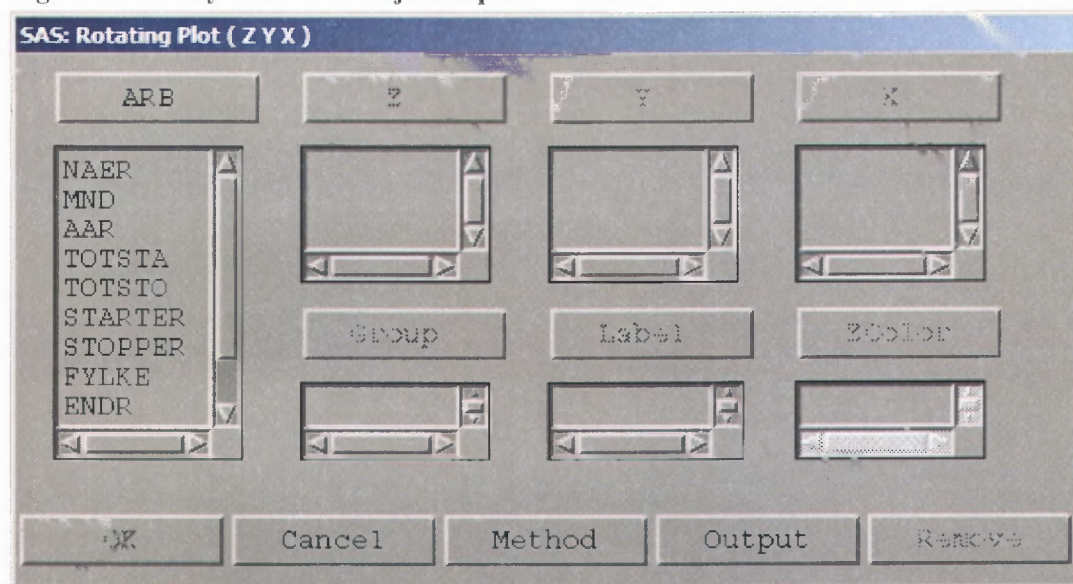
Figur 12.2 Tredimensjonalt plott fra en annen retning



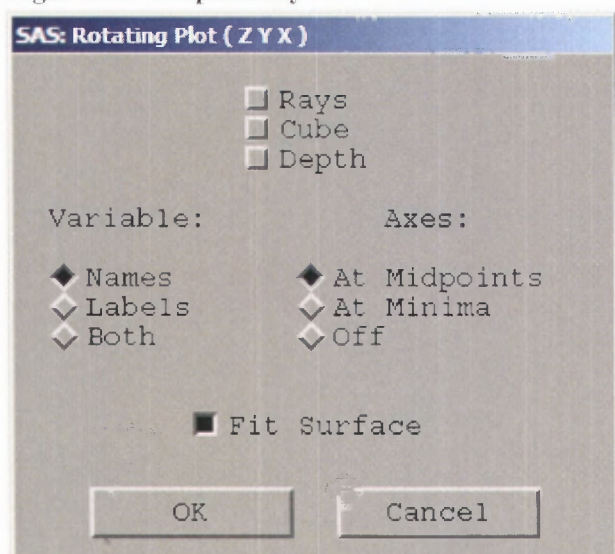
Popupmenyen gir muligheter til å legge inn flere visuelle effekter i de tredimensjonale plottene, blant annet en dybdeeffekt på punktmarkeringen og en kube rundt punktene.

Menyen for tredimensjonale plott (figur 12.3) gir også noen muligheter for å redigere plottet. Ved å velge "Output" får man menyen i figur 12.4. Valg av "Fit surface" gir figur 12.5.

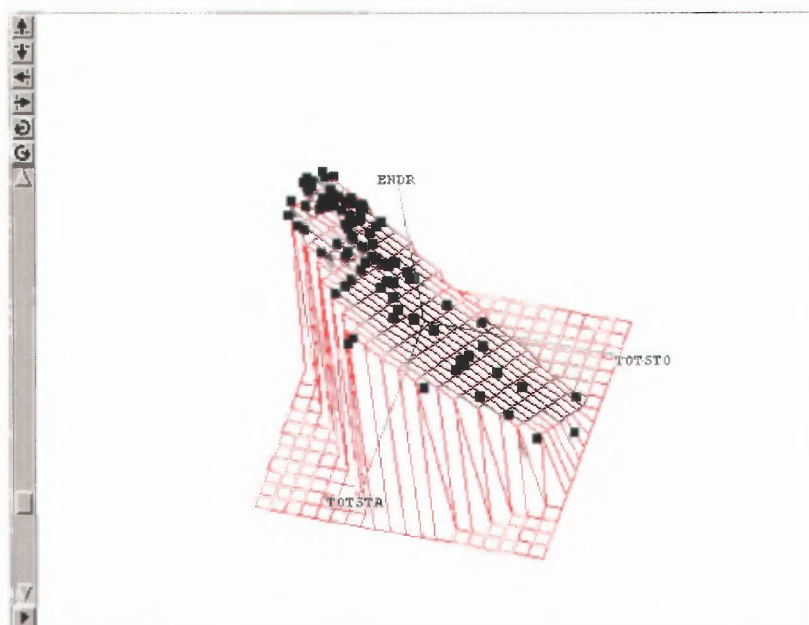
Figur 12.3. Meny for tredimensjonale plott



Figur 12.4. Outputmeny



Figur 12.5. Flatetilpasning

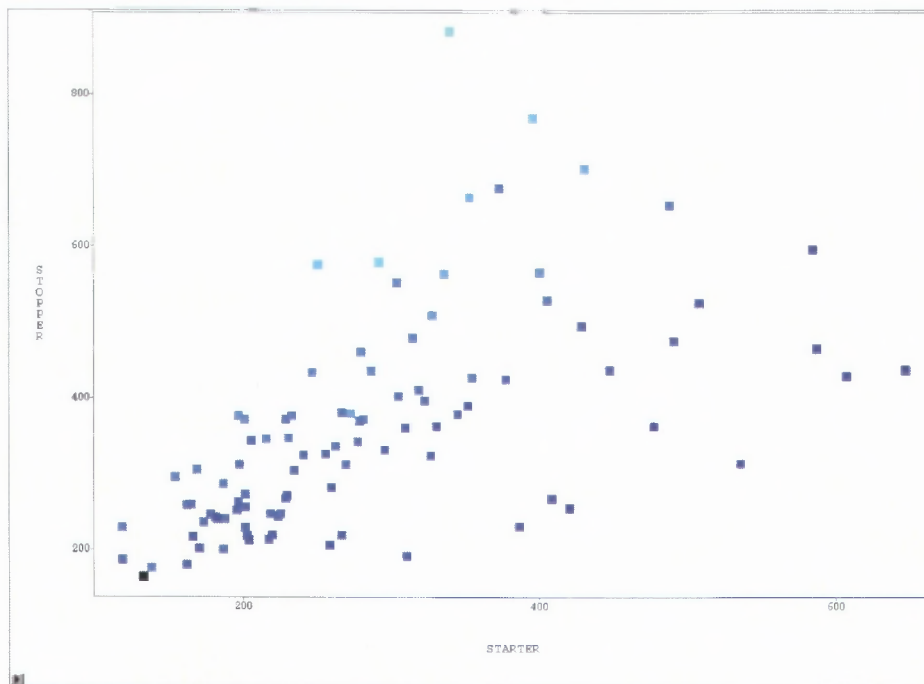


Vi kan også fargelegge etter en fjerde variabel for å få frem en «fjerde dimensjon». Felles for alt dette er at det ikke egner seg særlig godt på papir, effekten av slike plott får vi best frem på skjermen.

12.2. Farger som en ekstra dimensjon

Som eksempel vises et todimensjonalt X-Y – plott av starter og stopper av arbeidsforhold hvor størrelsen på endring er lagt inn som en tredje dimensjon ved at fargen på punktene blir mørkere jo høyere verdi det er på endringen.

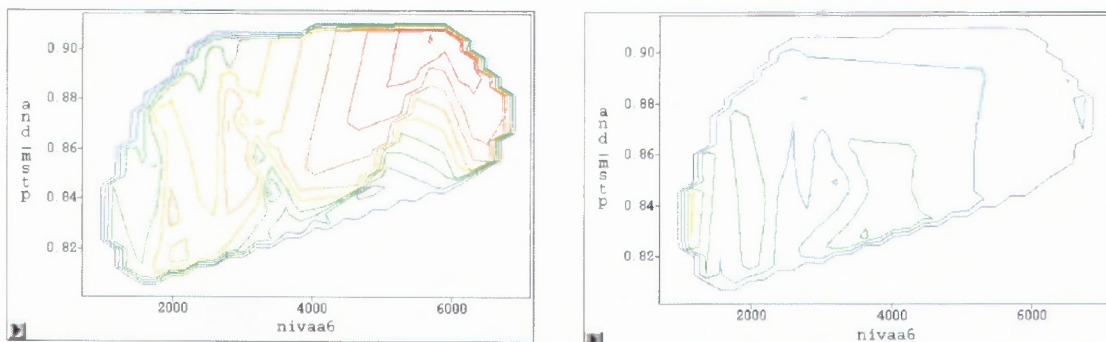
Figur 12.3. X - Y plott av STARTER og STOPPER



12.3. Konturplott

Menyen 'Contour Plot (Z Y X)' gir en grafisk fremstilling av forhold mellom 3 variable. Dette er en todimensjonal fremstilling av tre dimensjoner idet kurver (interpolert) følger nivået på Z i et X-Y diagram. Grafen fungerer best om X- og Y-verdiene fordeler seg over hele planet. Figur 12.4 viser to eksempler på konturplott med ulike Z-variable hvor X og Y variablene er de samme.

Figur 12.4. Den tredje dimensjon er henholdsvis "Gjennomsnittlig antall" og "Kostnad"



13.Lagring

De viktigste lagringsrutinene i SAS/INSIGHT er lagring av:

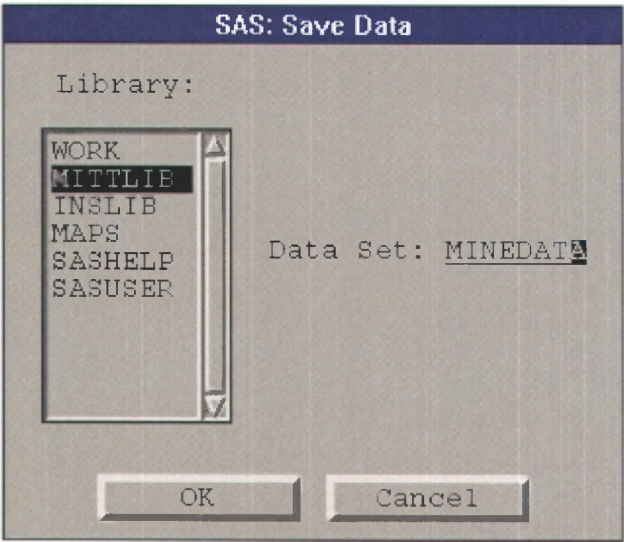
- data
- grafer
- tabeller
- hele SAS/INSIGHT kjøringen

13.1. Lagre data som et SAS-datasett

Hvis vi vil ta vare på editering som er utført på et datasett, må vi lagre dataene på nytt. All editering som f. eks. endrede verdier, fargelegging, labels, nye variable, gruppevariable osv., vil da lagres sammen med dataene.

Gå til rullegardinmenyen, velg **File → Save → Data**, og et dialogvindu åpnes:

Figur 13.1 Lagre Data



Velg et library, f. eks. «MITTLIB», - dette libnavnet må være definert på forhånd. For at datasettet skal være tilgjengelig ved en ny SAS-oppstart må vi velge et annet library enn «WORK». Bestem et navn for datasettet, f. eks «MINEDATA», og klikk **OK**. Dataene lagres da fysisk til eksempelvis filen «MINEDATA.sas7bdat» på den katalogen som er definert ved libnavnet «MITTLIB».

Endringer som er gjort i datavinduet lagres. For eksempel endrede verdier, farge på punkter eller label. Datasettet får en ny variabel, `_OBSTAT_`, som har 20 posisjoner. Posisjonsnummer viser:

Posisjonsnummer		Forklaring
1	-	markert eller ikke
2	-	vises eller skjules i grafen
3	-	inkludert eller ekskludert i beregningene
4	-	label eller ikke
5	-	symbol i plott
6 – 20	-	farge (5 posisjoner for hver av fargene rød, grønn, blå)

13.2. Lagre grafer og tabeller

Grafer og tabeller kan lagres både som grafiske kataloger og som grafiske filer, men det enkleste er å kopiere grafer og tabeller til Word og lagre Word-dokumentet (se kapittel 14.2). Tabeller kan også lagres fra Output-vinduet i SAS.

For dem som behersker SAS/Graph, er grafiske kataloger en fin måte å lagre både figurer og tabeller. Da kan de hentes inn i SAS/Graph, og redigeres via de mulighetene som ligger der.

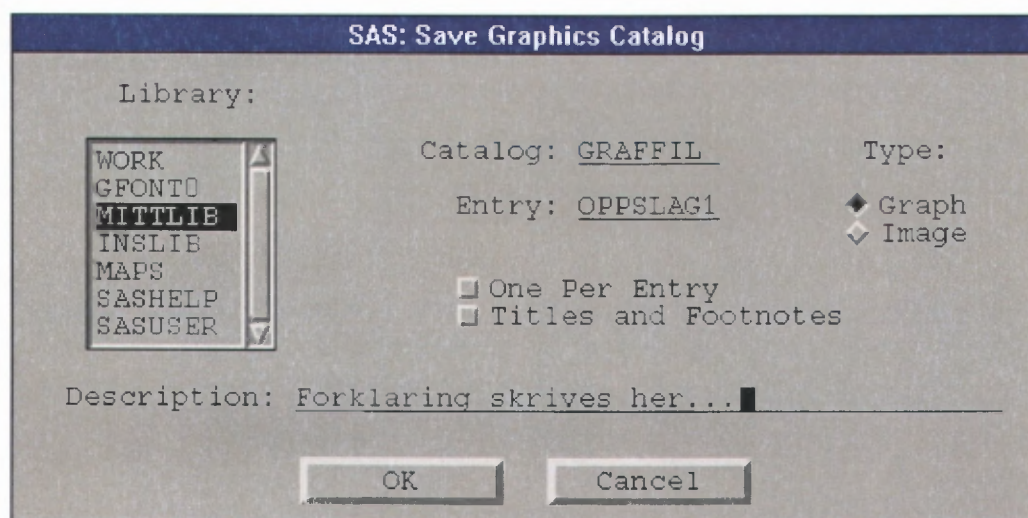
Grafiske filer er nok den enkleste og mest oversiktlige måten å lagre grafikken og eventuelle tabeller på. Grafikk som er lagret som grafiske filer kan senere lett hentes inn i Word ved **Sett inn - Bilde**.

13.2.1. SAS/Graph-katalog

Merk de grafer og tabeller som skal lagres ved å klikke på de respektive rammene med kontrolltasten nede. Hvis vi ikke merker noe, vil kun synlig innhold i vinduet lagres. Hvis alle grafer og tabeller i vinduet skal lagres, markeres de ved **Edit → Windows → Select All**.

Gå til rullegardinmenyen, velg **File → Save → Graphics Catalog**, og et dialogvindu åpnes:

Figur 13.2 Lagre som SAS/Graph-katalog



Velg et library, f. eks. «MITTLIB». Bestem et navn for SAS/Graph katalogen (Catalog) som grafikken skal lagres i, f. eks «GRAFFIL». Bestem et navn for grafikk-oppslaget (Entry) som skal lagres, f. eks «OPPSLAG1». Klikk **OK**, og et bilde med merkede grafer og tabeller lagres fysisk til SAS/Graph-katalogen (filen) «GRAFFIL.SCT01», under oppslaget «OPPSLAG1».

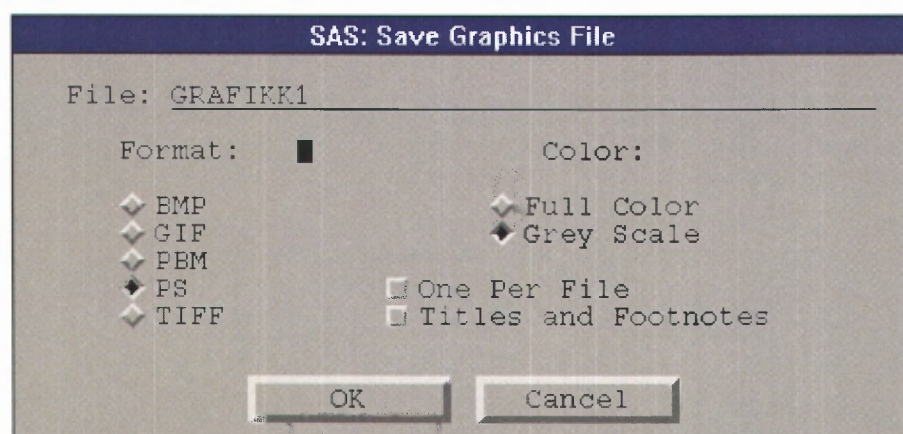
Hvis vi lagrer for mange grafer og tabeller inn i et enkelt oppslag, vil de lagrede objektene bli nok så «små». I en slik situasjon kan det være en fordel å markere for «One Per Entry» i menyen over. Da havner hver enkelt graf og tabell i egne oppslag, i den oppgitte SAS/Graph-katalogen.

13.2.2. Grafisk fil

Merk de grafer og tabeller som skal lagres. Hvis vi ikke merker noe, vil kun synlig innhold i vinduet lagres. Hvis alle grafer og tabeller skal lagres, velger vi **Edit → Windows → Select All**.

Gå til rullegardinmenyen, velg **File → Save → Graphics File**, og et dialogvindu åpnes:

Figur 13.3. Lagre som grafisk fil



Velg et filnavn, f. eks «GRAFIKK1», velg et passende format, f. eks. «PS» (postscript) eller «BMP». Klikk **OK**, og et bilde med merkede grafer og tabeller lagres til filen «GRAFIKK1.PS» eller «GRAFIKK1.BMP».

Hvis vi prøver å lagre for mange grafer og tabeller i en enkelt fil, vil lagringen avkortes. For å unngå dette kan vi markere for «One Per File» i menyen over. Da havner hver enkelt graf og tabell i egne filer.

13.2.3. Lagre tabeller til SAS/Output vinduet.

Gå til rullegardinmenyen, velg **File → Save → Tabela**s, og alle tabeller i SAS/INSIGHT-analysen skrives til SAS/Output. Fra Output-vinduet kan denne output'en lagres til fil. Det er også muligheter for å lagre dette som et SAS-datasett.

Tabeller kan skrives til SAS/Output-vinduet som en fast rutine. Velg da **File → Save → Initial Tabela**s ved oppstart av SAS/INSIGHT.

13.3. Kjøre og lagre SAS/INSIGHT rutiner

SAS/INSIGHT analysen som vi kjører kan lagres som såkalte SAS/INSIGHT statements i en tekstfil. Dette kan være nyttig både for å få en oversikt over hva vi har gjort, men også for senere batch-kjøring. Det vil si at vi kjører en ferdig oppsatt SAS/INSIGHT rutine. Vær oppmerksom på at det kun er analysen som lagres, ikke editeringen.

For å lagre eller kjøre på denne måten, må vi starte SAS/INSIGHT med «proc insight» fra SAS-editoren og ikke med rullegardinmenyen som beskrevet i kapittel 2. Gå til SAS/Program Editor, og skriv inn følgende sekvens:

```
filename lagre '$HOME/fill.txt';
proc insight file = lagre;
run;
```

Gå til rullegardinmenyen for SAS/Program Editor, velg **Run → Submit**, og SAS/INSIGHT startes. Åpne et datasett, og utfør en ordinær SAS/INSIGHT analyse. Avslutt SAS/INSIGHT, og analysen lagres til filen «fill.txt». Den kan for eksempel se slik ut:

```
/* SAS/INSIGHT Software */
OPEN INSLIB.INSDATA;
BAR OMSETN ANTSYSS;
BOX OMSETN
DIST OMSETN ANTSYSS;
```

For å kjøre denne rutinen skriver vi følgende sekvens i SAS/Program Editor:

```
filename kjor '$HOME/fill.txt';
proc insight infile = kjor;
run;
```

Velg **Run → Submit**, SAS/INSIGHT startes og utfører statementene i filen «fill.txt».

13.4. SAS – kommandoer for automatisk SAS/Insight - analyse

Dette er en praktisk måte å lage automatiske analyser som ofte gjentas. De spesielle Insight-statements tilsvarer analysealternativene i rullegardinmenyen:

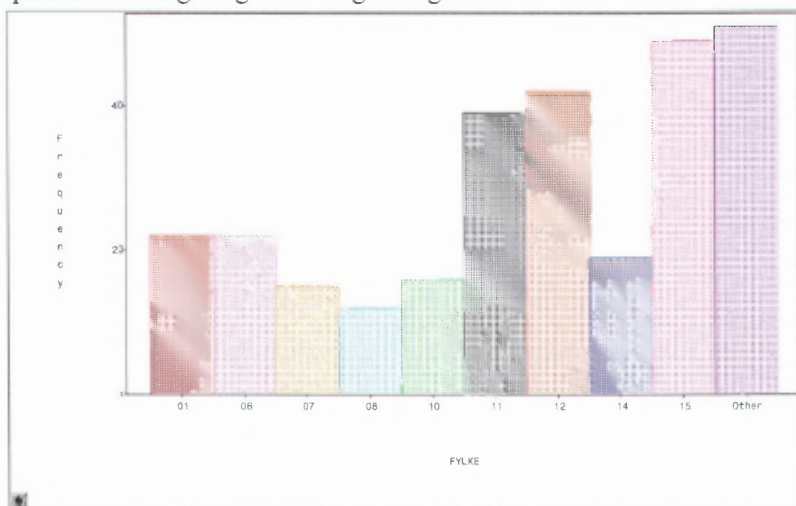
```
Proc insight

Statements:  Bar
              Box
              Line
              Scatter
              Contour
              Rotate
              Dist
              Mult
              Fit
              Tables
```

Eksempel

```
proc insight data = ut.insdata;  
bar fylke;  
run;
```

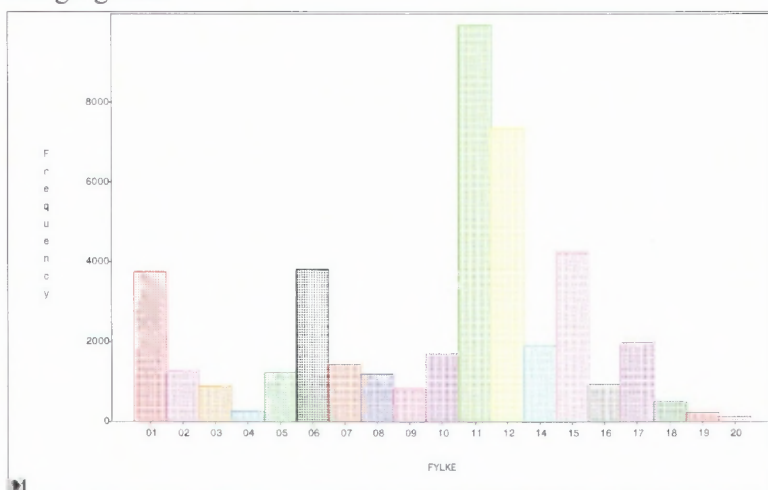
åpner SAS-Insight og viser følgende graf:



Det er mulig å legge inn 'options' for å justere grafen:

```
proc insight data = ut.insdata;  
bar fylke/freq = antsyss other = 0;  
run;
```

som gir grafen



OBS!!

SAS-Insight må lukkes før man eventuelt kan kjøre `proc insight` på nytt.

Avslutning med `run` holder SAS-Insight åpen for videre analyse, mens avslutning med `quit` lukker SAS-Insight.

14. Utskrift

Man kan skrive ut grafer og tabeller

- Direkte fra SAS
- Etter overføring til Word

Word gir de peneste utskriftene

Vi skal beskrive to metoder for å skrive ut resultater fra SAS/INSIGHT, det er direkte utskrift og utskrift via Word. Begge metodene må tilpasses brukerens eget oppsett på datamaskinen. Det er også mulig å skrive ut lagret analyse fra en SAS/Graph katalog eller en grafikkfil som beskrevet i kapittel 11.

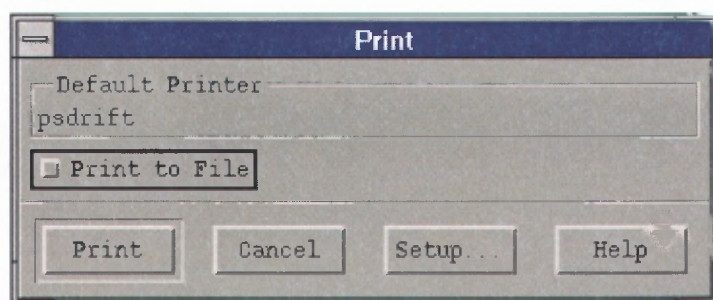
Hvilken utskriftsmetode man skal bruke, avhenger av hva vi skal bruke utskriftene til. For egen dokumentasjon av analyse kan det være greit å ta direkte utskrifter, men for å få analysen inn i dokumenter, må man overføre grafer og tabeller til Word. Word gir uansett de peneste utskriftene.

14.1. Direkte utskrift.

Merk de grafer og tabeller som skal skrives ut. Hvis vi ikke merker noe vil kun synlig innhold i vinduet skrives ut. Hvis vi vil ha ut all analyse, så velg **Edit → Windows → Select All**.

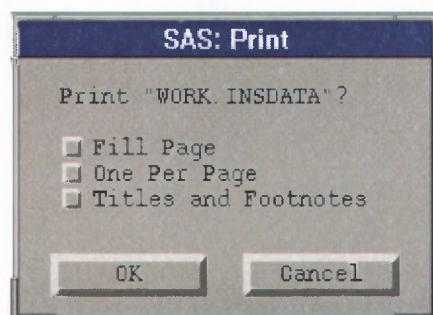
Gå til rullegardinmenyen, velg **File → Print**, og et dialogvindu åpnes. Her er det ulike varianter avhengig av om man kjører SAS fra Unix eller PC:

Figur 14.1. Print fra Unix.

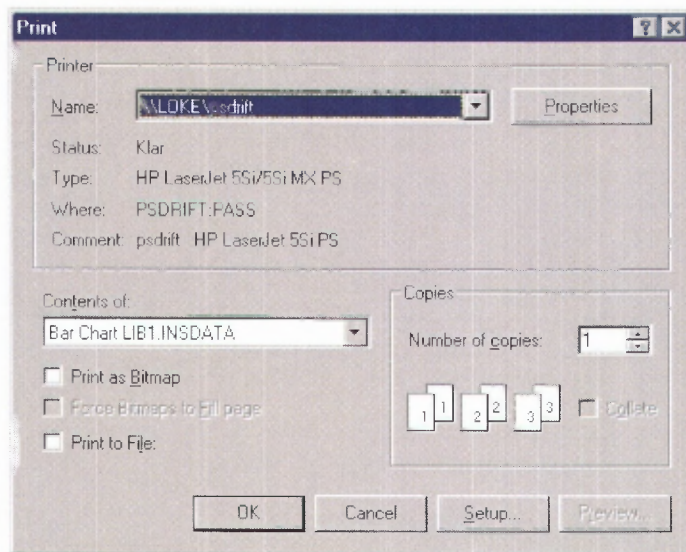


Klikk **Print**, og et nytt dialogvindu åpnes:

Figur 14.2. SAS: Print.



Figur 14.3. Print fra PC



Klikk **OK** her og samme dialogvindu som figur 14.2 åpnes.

Klikk **OK**, og utskriften sendes til oppsatt (default) printer.

Hvis vi har et «passende» antall grafer og tabeller som skrives ut, f. eks. et Fit-vindu, kan det være en fordel å markere for «Fill Page». Utskriften vil da fylle ut arket. Hvis vi skriver ut for mange grafer og tabeller, risikerer vi at utskriften avkortes. Vi kan da markere for «One Per Page», og hver enkelt graf og tabell sendes til egne utskrifter.

14.2. Utskrift via Word

Det er som oftest mest praktisk å skrive ut tabeller eller grafikk fra SAS/INSIGHT via Word, enten fordi det skal inn i et notat eller fordi vi får penere grafer på den måten.

Det er dels ulike prosedyrer for kopiering fra PC og ulike Unix plattformer.

Fra PC er fremgangsmåten slik:

- marker grafen
 - Hvis du skal kopiere hele vinduet er det ikke nødvendig å markere
 - hvis grafen ikke fyller vinduet må du eventuelt kutte bildet i Word
 - Er grafen større enn vinduet, må den markeres
- Edit → copy
- Gå til Word og lim inn med: Lim inn utvalg → Bilde (utvidet metafil)
- Juster størrelsen på figuren i Word med: Format → Bilde → Størrelse

Fra UNIX er det enda enklere:

- Gå til det vinduet du vil ha utskrift av. Klikk på menyknappen i øvre venstre hjørne av vinduet. Den ser slik ut:



- Velg **Edit** → **Copy Rectangle** i menyen som slås opp. Marker med venstre musetast det område i vinduet som du vil kopiere, dra og slipp.
- Gå til Word og lim inn med ved å klikke på Lim inn - symbolet. Størrelsen på figuren kan redigeres i Word med: Format → Bilde → Størrelse.

Grafer kan også lagres som grafiske filer, overføres til PC og hentes inn i Word (se kapittel 13.2). Alternativt til filoverføring kan grafiske filer mailes fra Unix.

15. Tips

Det er mange veier som fører til samme resultat. Her samles noen ideer til redigering av grafikk

Grafer kan dreies og speilvendes

- klikk på et hjørne og dra tvers over (eventuelt på skrå) av grafen

Endring av fargeskalaen på flerfargefeltet i Tools

- Klikk og dra en ny farge fra fargerutene over og ned til ønsket ende av flerfargefeltet

Farger kan brukes for å få med en ekstra dimensjon i plott

- Legg inn en fargeskala fra lyst til mørkt av samme farge i flerfargefeltet i Tools. En intervallvariabel (tall) markeres med dette fargefeltet og lave verdier vil få lyse markeringer mens høye verdier får mørke markeringer.

Markering av et område på X-Yplott

- Et rektangulært område markeres innen et plott ved klikk og dra. De samme enhetene markeres samtidig i alle andre plott.
- Det går an å markere flere rektangulære områder på samme graf ved å bruke 'Shift'-tasten ved etterfølgende markeringer
- Området kan utvides/endres ved å plassere markøren i et hjørne av rektangelet, klikke og dra.
- Området kan flyttes ved å plassere markøren inne i rektangelet, klikke og dra.
- Området beveger seg når man slipper musetasten mens rektangelet er i bevegelse.
- For å studere bare de markerte observasjonene kan man fjerne alle observasjonene ved hjelp av pop-upmenyen, markere og eventuelt sette rektangelet i bevegelse

Håndmarkøren fra Tools kan brukes til å justere stolpene i histogram

- Klikk med hånden nær horisontal akse øker vidden på stolpene, klikk med hånden høyt over horisontal akse reduserer bredden på stolpene
- Man kan bevege hånden over grafen til histogrammet får en ønsket form

Frekvens (Freq) og vekt (Weight)

- Freq – variable må være heltallig. Desimaler fjernes
- Weight – variabelen trenger ikke være heltallig, men er bare tilgjengelig under **Distribution**, **Fit** og **Multivariate**

B Returadresse:
Statistisk sentralbyrå
NO-2225 Kongsvinger

Statistisk sentralbyrå

Oslo:

Postboks 8131 Dep

NO-0033 Oslo

Telefon: 21 09 00 00

Telefaks: 21 09 49 73

Kongsvinger:

NO-2225 Kongsvinger

Telefon: 62 88 50 00

Telefaks: 62 88 50 30

E-post: ssb@ssb.no

Internett: www.ssb.no



Statistisk sentralbyrå
Statistics Norway