Statistics Norway
Statistisk sentralbyrå

Aslaug Hurlen Foss

Grafisk revisjon av nøkkeltallene i KOSTRA

Seksjon for statistiske metoder og standarder
Emnegruppe: 00.90

Innhold

1. Hva er grafisk revisjon og hvorfor bruke grafisk revisjon?

Grafisk revisjon er å bruke grafikk til å oppdage avvikende verdier i datasettet. Den mest brukte grafikken er stolpediagram, punktdiagram og boksplot.

Ved grafisk revisjon er det mulig å oppdage feil, som ville vært vanskelig uten grafikk. Ved å plotte variabler sammen, kan grafen vise skjevheter i datasettet, som ikke ville blitt oppdaget uten grafikk. Med bruk av grafikk, er det lett å få et overblikk over fordelingen av variablene og hvilke verdier som er ekstreme. Det er også mulig å studere flere variabler på en gang. Da kan man se hvordan variasjonen mellom variabler er. Når man studerer grafer kan man med fordel bruke farger og symboler for å marker forskjeller. Den visuelle grafikken taler sterkere til oss enn hva tall eller tekst gjør.

Dette notat beskriver hvordan grafisk revisjon kan benyttes i revisjon av nøkkeltallene i KOSTRA. Nøkkeltallene i KOSTRA består hovedsakelig av forholdstall enten mellom tjenester og befolkning eller mellom tjenester og regnskapstall.

2. IT-løsninger

Valg av variabler blir gjort fra web-side som innholder nøkkeltallene. Deretter er det lagt opp til bruk av SAS og SAS-insight. En testversjon er laget.
Datasettet vil i tillegg til nøkkeltallene, inneholde kommunenummer og navn, fylke og KOSTRA-gruppe.

Praktiske tips:

- PC-SAS må være installert

- Gå inn på internettsiden fra hvor man kan lage SAS-datasett av nøkkeltallene (foreløpig ligger dette på siden: http://www.utv.ssb.no/trf/revkos/?faktaark=95254370311646)

- Merk hvilke nøkkeltall du vil studere og deretter trykk på knappen <hent SAS-datasett> SAS blir da startet og det er laget ferdig et program. Kjør SAS-programmet som kommer opp. Datasettet blir laget, det blir laget en enkel statistikk på nøkkeltallene og det starter opp SAS-insight. Dette programmet kan endres etter ønske før det blir kjørt.

Navnene på nøkkeltallene har blitt var_1, var_2 osv., med labler som har blitt kuttet på 60 tegn (som er det lengste SAS har mulighet til nå). Det er mulig selv å endre navnene etter ønske i programmet

Eksempel på hvordan programmet ser ut:

```
DATA CLASS;
INPUT kommune_nr $ kommune_navn $ fylke_nr $ kostra_gruppe $var_1 var_2
var_3;
LABEL
var_1 ='Netto driftsutgifter grunnskoleopplæring i prosent av samle'
var_2 ='Netto driftsutgifter til grunnskole og SFO i prosent av sam'
var_3 ='Netto driftsutgifter til grunnskole, i prosent av samlede n'
;
CARDS;
0101  Halden      01    K13   27.3296046458363  26.6395265502745
      26.1313512674828  22.0136390167495  4.1177122507333
```

3. Nøkkeltallstatistikk

Det er ofte fint å få en oversikt over nøkkeltallene. Derfor er det laget et program som automatisk lager statistikk over de nøkkeltallene som er valgt.
Programmet beregner:

- Antall observasjoner som det er verdier på
- Antall observasjoner som er missing
- Minimumsverdi
- Maksimumverdi
- Gjennomsnitt
- Standardavvik

Programmet kan lett bli endret til web-format eller til et format som kan lett bli tatt inn i Regneark.

```
PROC TABULATE DATA=work.class FC=' --------'  F=8.1 MISSING;
VAR var_1 var_2 var_3 var_4 var_5 var_6 var_7;
TABLES var_1 var_2 var_3 var_4 var_5 var_6 var_7, N='Antall'*F=8.
NMISS='Mangler verdi'*F=8. MIN='Minimum' MAX='Maksimum' MEAN='Gjennomsnitt'
    STD='Standard-avvik'/ROW=FLOAT;
RUN;
```

Eksempel på utskrift:

	Antall	Mangler verdi	Minimum	Maksimum	Gjennom-snitt	Standard avvik
Netto driftsutgifter pr. innbygger i kroner, kommunehelsetj	424	12	595.8	5971.1	1619.2	698.7
Netto driftsutgifter i prosent av samlede netto driftsutgif	424	12	2.6	11.1	5.0	1.2
Netto driftsutg til forebygging, skole- og helsestasjonstj	424	12	0.0	968.1	288.0	114.6
Netto driftsutg til forebygging, skole og helsestasjonstj p	424	12	0.0	9936.8	3330.5	1400.8
Netto driftsutg til forebygging, skole og helsestasjonstj p	424	12	0.0	4020.9	1284.0	527.8
Netto driftsutgifter til forebyggende arbeid, helse pr. inn	424	12	-255.6	840.8	124.7	111.4
Brutto investeringsutgifter pr. innbygger	424	12	-24.1	4878.3	47.4	270.8

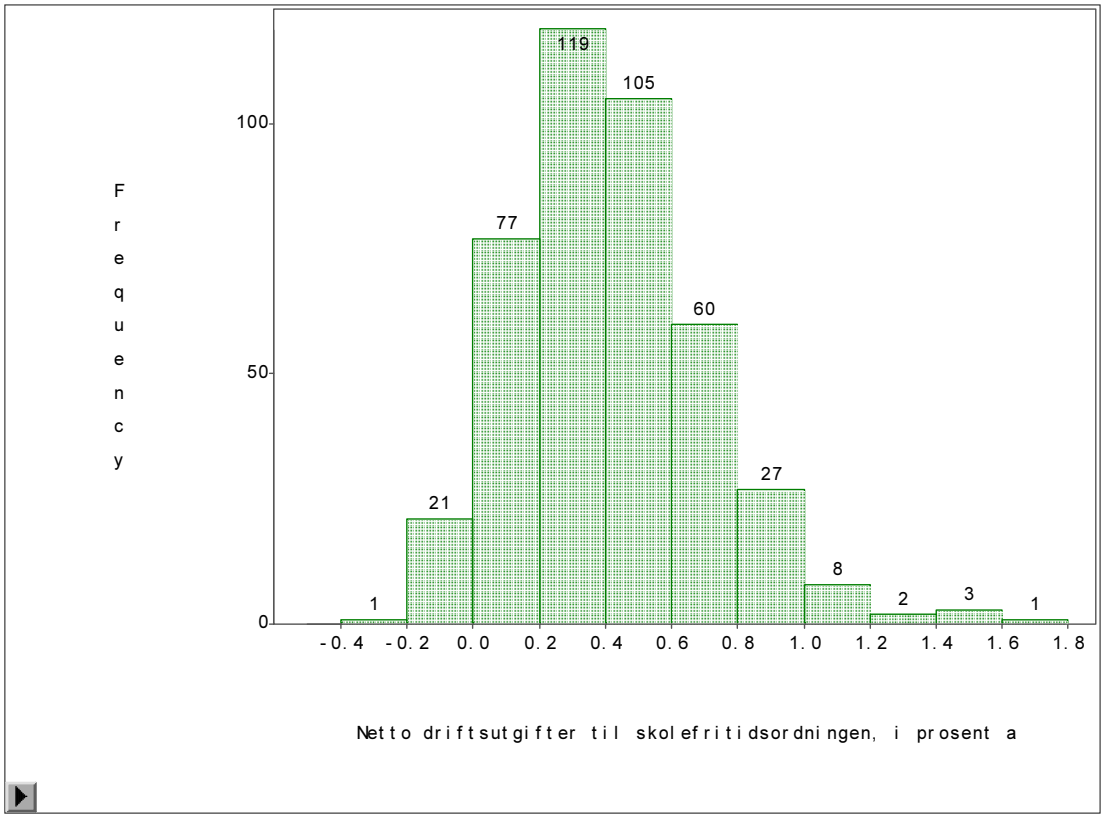
4. Grafisk revisjon

Stolpediagram - Histogram
Ved bruk av stolpediagram oppdages lett de mest åpenbare ekstreme verdiene som kan skyldes feil. I tillegg viser stolpediagrammet hvordan fordelingen til nøkkeltallet er. Av figur 1 ser vi at 22

kommuner har negative verdier på 'netto driftsutgifter til skolefritidsordningen'. Dette kan være feil og bør derfor undersøkes nærmere.

SAS-insight:
Analyze → Histogram → (variabelnavn) → y
(output → labels)

Figur 1. Histogram over netto driftsutgifter



Programkode:

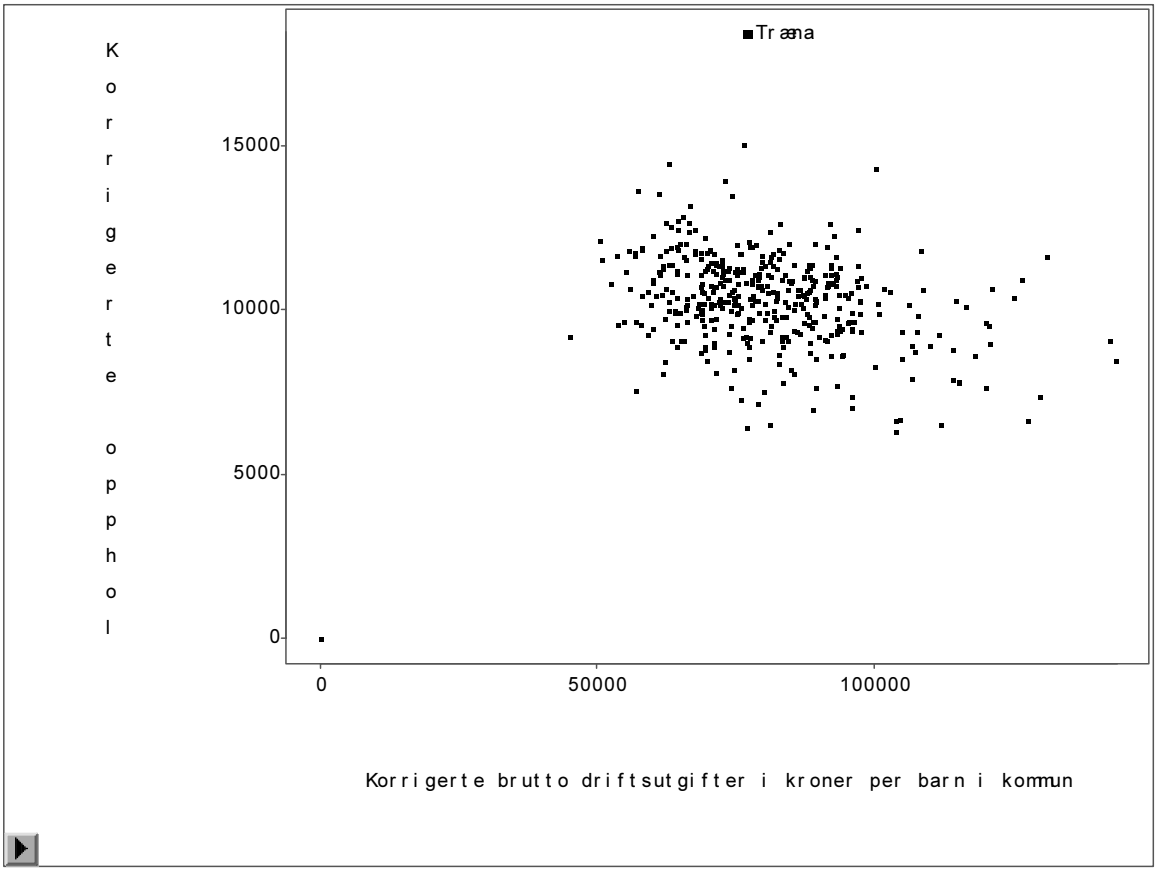
```
PROC INSIGHT data=work.CLASS ;
BAR var_1;
RUN;
```

Punktdiagram
I punktdiagram blir to nøkkeltall (variabler) plottet mot hverandre. Ved et slik diagram oppdages lett åpenbare avvik som kan skyldes feil og skjevheter. I tillegg blir forholdet mellom to nøkkeltall visualisert. I figur 2 ser vi de to nøkkeltallene 'korrigerte brutto driftsutgifter i kroner per barn i kommunale barnhager' plottet mot 'Korrigerte oppholdstimer per årsverk i kommunale barnehager'. Av dette plottet ser vi at nøkkeltallene sprer seg fint ut i en ellipse og viser at det ikke er noe klar sammenheng mellom disse nøkkeltallene. Det er en kommune som har oppgitt 0 på begge

nøkkeltallene, mens Træna har en svært høy verdi på nøkkeltallet 'Korrigerte oppholdstimer per årsverk i kommunale barnehager'. Hvis det var en sammenheng mellom to nøkkeltall ville det vist seg ved at plottet lagde et mønster.

SAS-insight:
Analyze → Scatter plot → (variabelnavn 1) → y → (variabelnavn 2) → x

Figur 2. Punktdiagram av to nøkkeltall

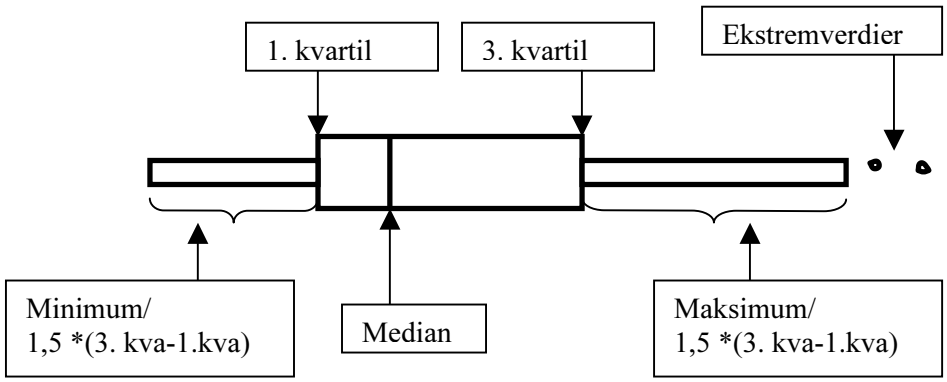


Programkode:

```
PROC INSIGHT data=work.class ;
SCATTER VAR_6
   * VAR_7/
   LABEL=kommune_navn;
RUN;
```

Boksplot
Boksplot er en figur som viser spredningen og ekstremverdier i datasettet. Boksplot kan bli brukt til å bestemme grenseverdiene til hva som er ekstremverdier.

Boksplot forklaring:



1. kvartil er verdien til den observasjonen som har 25 prosent av observasjonene lavere enn seg og 75 prosent av observasjonene høyere enn seg.

Median er den verdien som har både 50 prosent av observasjonene under og over seg.

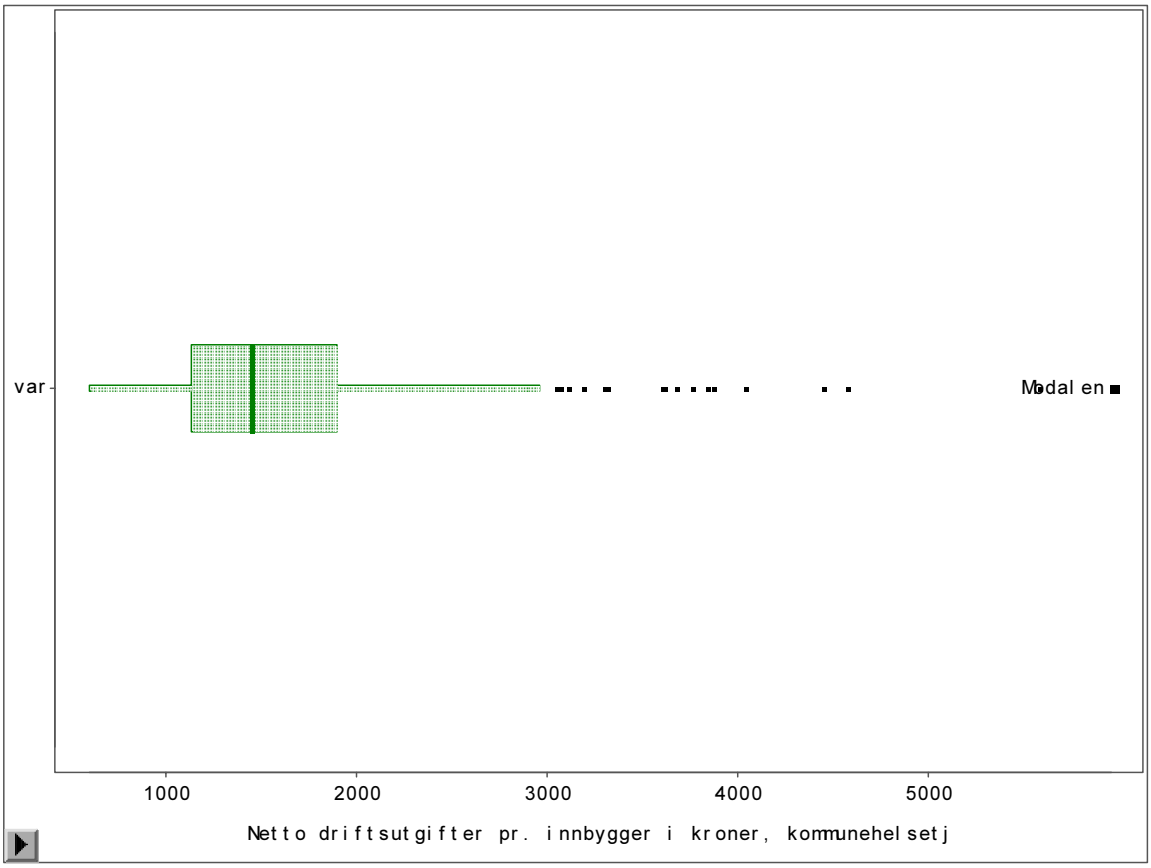
3. kvartil er verdien til den observasjonen som har 75 prosent av observasjonene lavere enn seg og 25 prosent av observasjonene høyere enn seg.

Kvartilene og medianene blir funnet ved å ordne datasettet etter størrelse og deretter å plukke ut de observasjonene som deler datasettet i 25, 50 og 75 prosent.

Eksempel:
Medianen av tallene: 1, 2, 3, 4, 5, 6, 7, 8, 9 er 5 siden tallet 5 har 4 observasjoner både under og over seg.

SAS-insight:
Analyze → Box Plot → (variabelnavn) → y

Figur 3. Boksplot av nøkkeltall

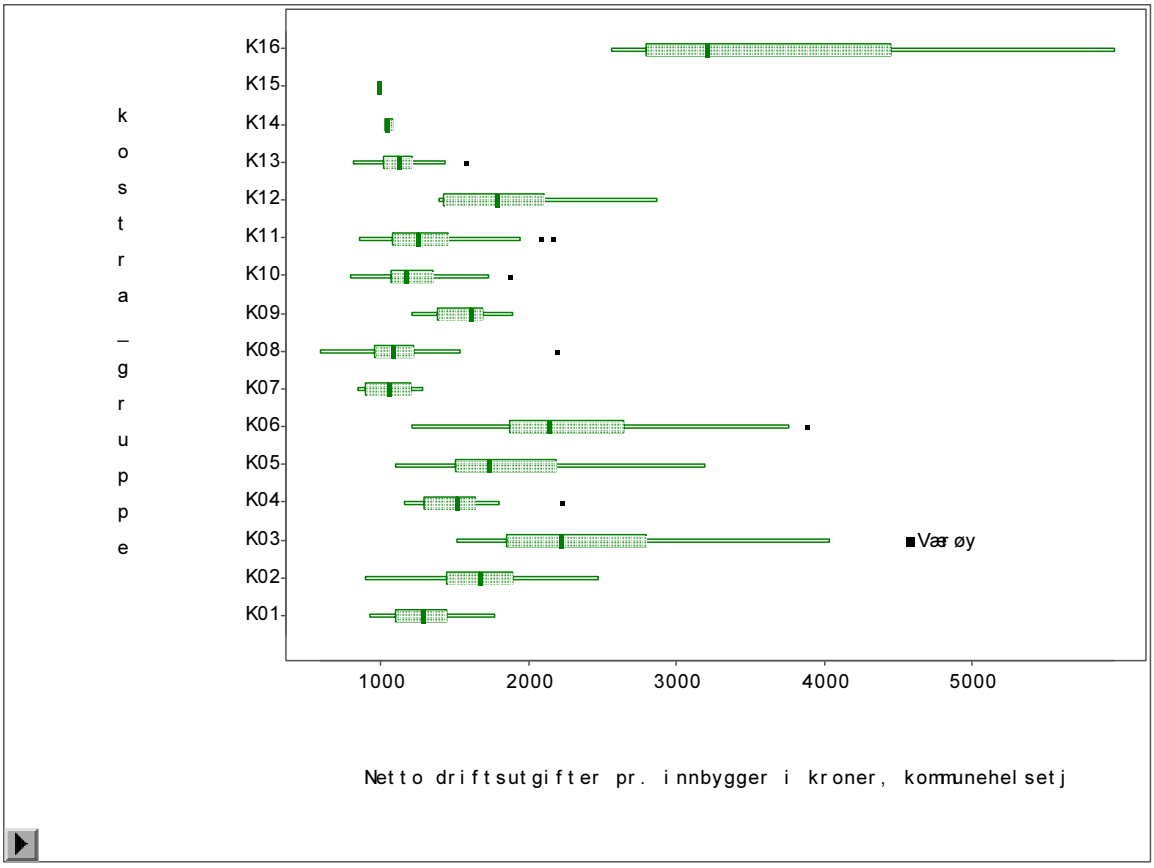


I figur 3 er det laget et boksplot av nøkkeltallet 'Netto driftsutgifter pr. innbygger i kroner, kommunehelsetjenesten. Av dette plottet ser vi at det er en del kommuner som skiller seg ut for dette nøkkeltallet. Det kan være at disse ikke er ekstremverdier, men at de hører til en gruppe med høye netto driftsutgifter. For å kontrollere dette kan vi for eksempel lage et boksplot for hver KOSTRA-gruppe. KOSTRA-grupper er en inndeling av kommuner etter befolkningsstørrelse, frie disponible inntekter og bundne kostnader.

Boksplot i KOSTRA-gruppene
De fleste av de kommunene som så ut til å være ekstremverdi i figur 3 tilhører KOSTRA-gruppe 16, som er de 10 kommunene med høyeste frie disponible inntekter per innbygger. Siden denne gruppen har høye disponible inntekter blir det stor variasjon i hvor mye de bruker på drift av skoler. Kostra-gruppene 1, 4, 7 og 10 har lave frie disponible inntekter. Dette ser vi på figur 4 gir seg utslag i lave tall, med liten variasjon. Kostra-gruppene 2,5,8 og 11 har middels frie disponible inntekter. For disse gruppene ser vi at det er større utgifter og spredningen i tallene er større. Kostra-gruppene 3, 6, 9 og 12 har høye frie disponible midler . Disse gruppene har høyere utgifter og større spredning enn de andre kostra-gruppene med middels og lave fri disponible midler. (Gruppe 15 er Oslo, mens gruppe 14 er de andre storbyene Bergen, Trondheim og Stavanger. Og gruppe 13 er de andre store kommunene)

SAS-insight:
Analyze → Box Plot → (variabelnavn) → y → (kostra-gruppe) → x

Figur 4. Boksplot av nøkkeltall etter KOSTRA-grupper



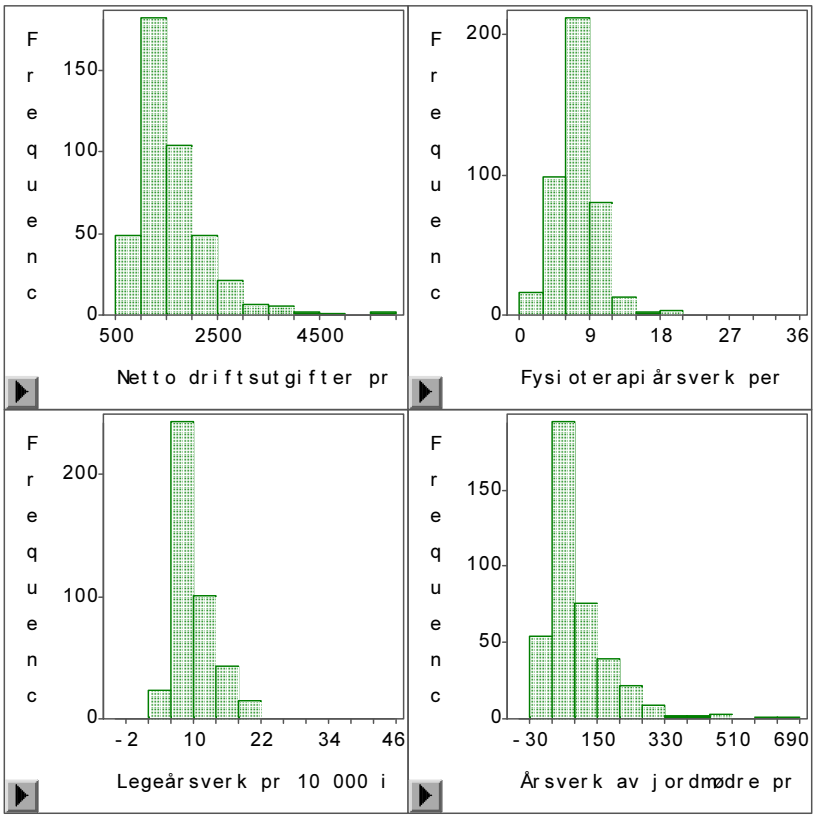
Programkode:
PROC INSIGHT DATA==work.class;
BOX var_1 * kostra_gruppe / LABEL=kommune_navn;
RUN;

Grafisk revisjon av flere nøkkeltall samtidig
Det kan være hensiktsmessig å plotte flere nøkkeltall samtidig. Enten for å få en oversikt over flere nøkkeltall samtidig eller for muligheten for å se på variasjonen mellom nøkkeltallene. Det kan imidlertid være begrensende hvor mange plot det er plass til på en side før grafene blir uoversiktlige. 1-9 nøkkeltall går ofte greit i samme figur. Fordelen med plot av flere nøkkeltall i samme graf, er at det da er mulig å studere hvordan en kommune eller en gruppe fordeler seg i forskjellige nøkkeltall. Dette kan i SAS-insight bli gjort ved å markere verdier på en av grafene og da blir disse observasjonene også markert for de andre nøkkeltallene som er plottet.

Histogram med flere nøkkeltall

SAS-insight:
Analyze → Histogram → var_1 var_2 var_3 var_4 → y

Figur 5. Stolpediagram av 4 nøkkeltall samtidig

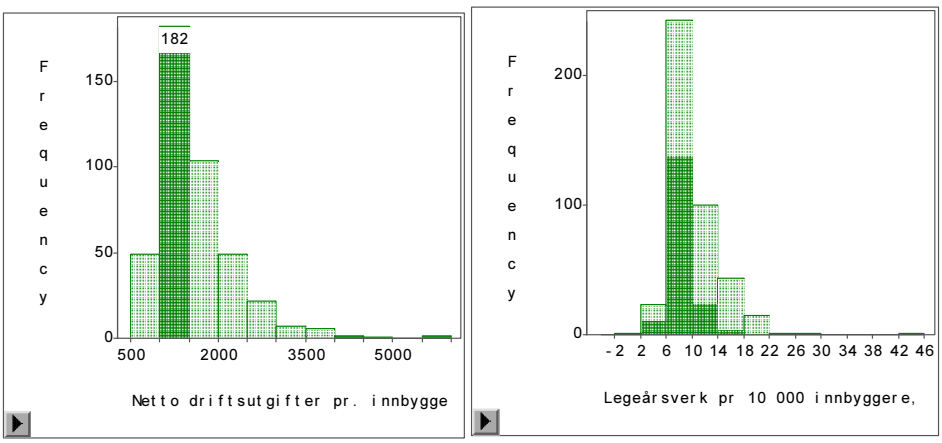


Programkode:

```
PROC INSIGHT data=work.class ;
BAR VAR_1 VAR_2 VAR_3 VAR_4  ;
RUN;
```

I figur 6 er den en søyle i det første histogrammet markert og fordelingen av disse observasjonene blir da vist i det andre diagrammet.

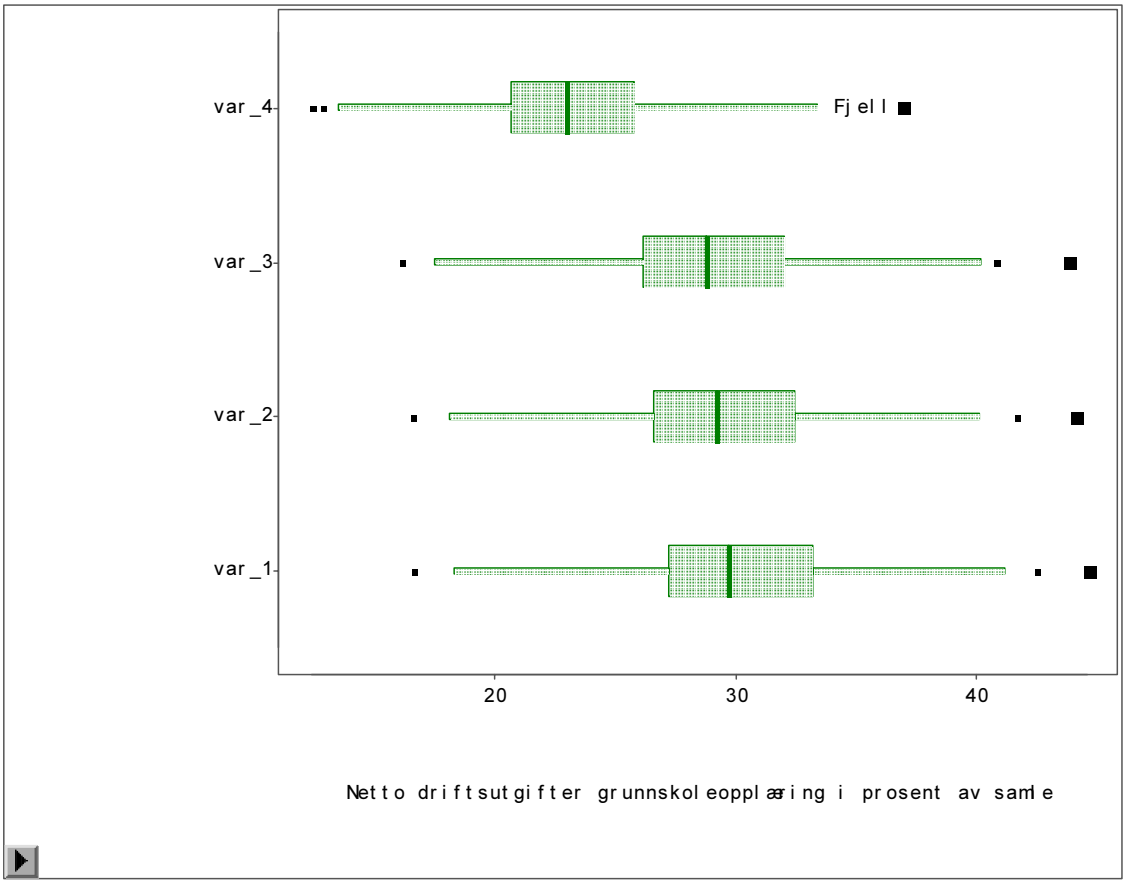
Figur 6. Stoplediagram av to nøkkeltall samtidig med en søyle markert i den første grafen



Boksplot
Ved boksplot av flere nøkkeltall samtidig får vi en oversikt over spredningen av verdiene. For at figuren skal bli tolkbar bør bare 'like' nøkkeltall lages i samme figur. Det vil si lage separate figurer for prosent og andeler. Når flere nøkkeltall plottes i samme figur kommer bare variabelnavnene fram og ikke labelene. For å få en mer forståelig bilde bør variablene ha et annet navn, dette kan endres i innlesningsprogrammet. I figur 7 er nøkkeltallene netto driftsutgifter i prosent av samlede netto driftsutgifter for henholdsvis: grunnskoleopplæring, grunnskole og SFO, grunnskole, og grunnskoleundervisning plottet. Kommunen Fjell, som er markert i plottet, skiller seg ut for alle disse nøkkeltallene.

SAS-insight:
Analyze → Box Plot → var_1 var_2 var_3 var_4 → y

Figur 7. Boksplot av 4 nøkkeltall fra samtidig



Programmeringskode:
```
PROC INSIGHT data=work.class ;
BOX VAR_1 VAR_2 VAR_3 VAR_4  ;
RUN;
```

5. Hva er en ekstremverdi?

Ekstremverdi er de nøkkeltall (observasjonene) som ligger langt fra de andre. Men hva er langt? Et mål som kan brukes er; de observasjoner som ikke ligger 2 kvartillengder fra medianen. Det vil si:

(1) Median +/- 2*(3. kvartil- 1. kvartil)

Dette tilsvarer de observasjoner som er markert som prikker i boksplottet. Dette er et robust mål på hva som er ekstremverdier. Formelen 1 ligner på det kjente 95 prosent konfidensintervallet som er gitt ved:

(2) Gjennomsnitt +/- 2*Standardavvik

Forskjellen mellom disse formlene er at konfidensintervallet (2) vil bruke ekstremverdier til å beregne grensen til det som skal settet for ekstremverdier. Dette kan føre til at ekstremverdier ikke blir funnet. I tillegg er det antatt at dataene er normalfordelte, noe de ofte ikke er.

Eksempel:
Medianen av tallene: 1, 2, 3, 4, 5, 6, 7, 8, 9 er 5 siden tallet 5 har 4 observasjoner både under og over seg. Gjennomsnittet er også 5. Hvis tallene 8 og 9 hadde ved en feil blitt registrert som henholdsvis 18 og19, da er fortsatt medianen 5, mens gjennomsnittet har blitt 7.2. Gjennomsnittet har blitt påvirket av feilen, mens medianen ikke har det. Formelen for å finne ekstremverdier som bygger på medianer og kvartiler ville ha funnet disse to som ekstremverdier, mens formelen som bygger på gjennomsnitt ikke ville ha gjort det.

SAS-program for å finne kvartiler og medianer:

```
PROC TABULATE DATA=work.class FC=' --------'  F=8.1 MISSING;
VAR var_1 var_2 var_3 var_4 var_5;
TABLES var_1 var_2 var_3 var_4 var_5,
 MIN='Minimum' MAX='Maksimum' median='Median' q1='1. kvartil, 25%' q3='3.
kvartil, 75%'
  qrange='Avstanden 3.-1. kvartil' /ROW=FLOAT;
RUN;
```

Eksempel på utskrift:

	Minimum	Maksimum	Median	1. kvartil, 25%	3. kvartil, 75%	Avstanden 3.-1. kvartil
Korrigerte oppholdstimer per årsverk i kommunale barnehager	0.0	18531.6	10482.8	9610.1	11246.2	1636.2
Korrigerte brutto driftsutgifter til kommunale barnehager	0.0	129.5	35.9	33.0	39.9	6.9

I enkelte tilfeller kan det være stor forskjell mellom KOSTRA-gruppene, i slike tilfeller kan det være fornuftig å sette en grense i hver KOSTRA-gruppe for hva som er ekstremverdi.

I tillegg til dette er viktig å bruke erfaring og kunnskap fra fagfeltet for å sette fornuftige grenser for hva som er ekstremverdier. Ekstremverdier bør undersøkes om de er feil. Det vil alltid kunne finnes observasjoner som er ekstreme, men korrekte.

Hvis man vil ha verdiene på en fil bruk prosedyren 'proc univariat' i SAS.

Det er mulig å prøve å finne ekstremverdier på flere nøkkeltall samtidig. Hvis metoden skal være robust, krever dette ganske avanserte metoder, men de finnes. F.eks. MVE eller MCD prosedyren i PROC IML. Disse prosedyrene minimerer enten ellipser eller kovariansen og finner de observasjonene som ligger langt fra de andre (Rousseeuw and Leroy: Robust regression and outlier detection.1987). Med denne prosedyren finner vi altså de kommunene med de mest avvikende nøkkeltallene i forhold til de fleste andre kommunene. Problemet med disse prosedyrene er at de ikke er særlig brukervennlige.

6. Tips til SAS-insight

1. Start av SAS-insight: Solution/analysis/Interactive data analysis

2. For å få kommunenavn istedenfor kommunenummer i plotene trykk på 'output' i valgmenyen og merk av at du vil 'label'.

3. For å se programkode i loggen for figurene som blir laget: File/save/statsments

4. For generelle tips se i notatet 2000/1: SAS/insight

5. Endring av størrelse på graf:

Det er mulig å få en større graf ved å markere rammen på figuren og så dra den ut.
Det er også mulig å angi størrelsen i en SAS-setning i proc insight: `window x y lengde bredde`
x,y angir hvor figuren skal starte. Øverste til venstre er da 0, 0 . Lengde og bredden angis i prosenter.

6. Ending av skala og stolpebredde, tics

Dette kan gjøres ved å endre verdiene til tics i grafen. Verdiene som kan endres er første verdi (first tics), siste verdi (last tics), bredden (tics increment), minste verdi (minor tics), akse minimum (axis minimum), akse maksimum (axis maximum).
I programkoden kan dette gjøres ved å skrive yaxis og de 6 verdiene som beskrevet over bak.

7. Bruk av farger :

Det kan ofte ha en sterk effekt å legge farge på de observasjonene som antakelig er feil eller ekstremverdier. For å gjøre dette i SAS-insight er det nødvendig å lage et nytt datasett, hvor det blir lagt på 'farger' på datasettet. I eksemplet under legger vi fager på de kommunene som nøkkeltall 1 (variabel 1) som har prosent som ikke er mellom 0 og 100.

```
DATA test;
 SET work.class ;
 IF 0=<var_1=<100 THEN farge='0' ; ELSE farge='1';
 IF farge='1' THEN _obstat_='1110165535    0    0';
 IF farge='0' THEN _obstat_='01101    0    065535';
 RUN;
```

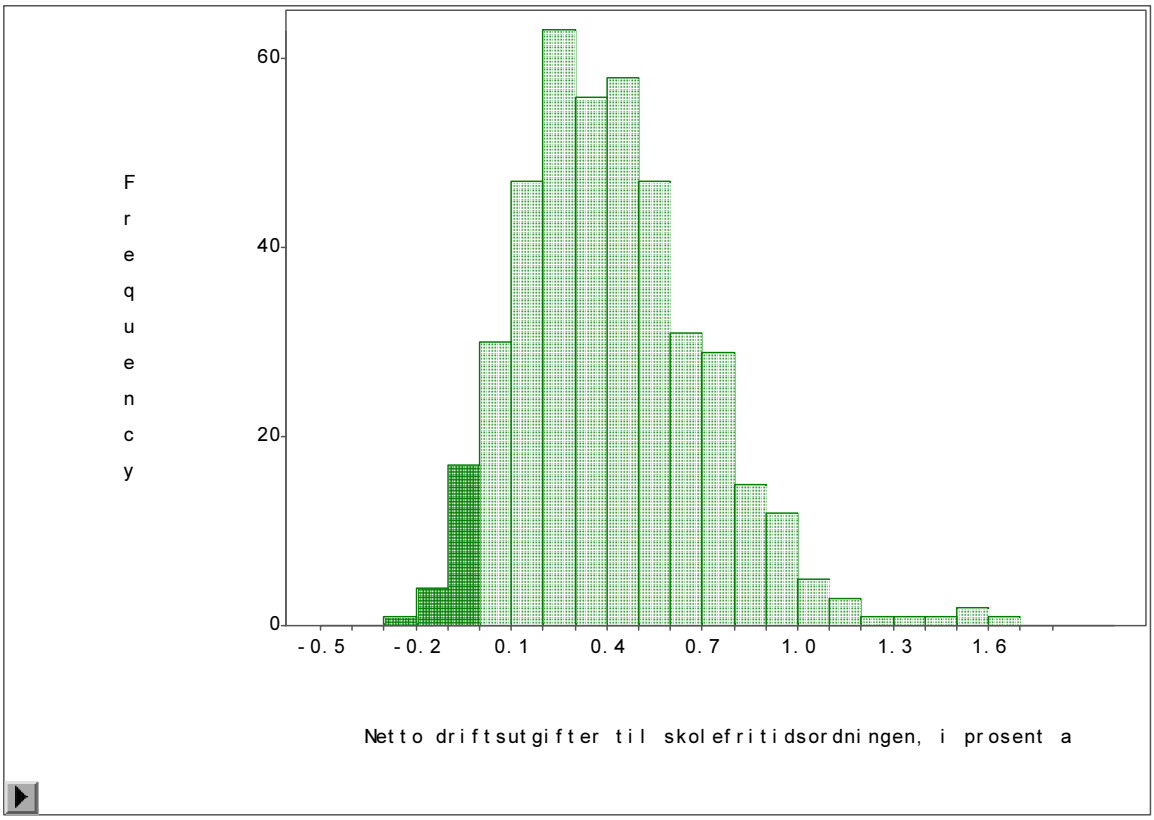
Forklaring til _obstat_
obstat består av 20 karakter hvor alle punktene betyr noe.
Karakter 1: Markerer om observasjonen er markert eller ikke.
Karakter 2: Lager om observasjonen er vist eller gjemt i grafen. '1' er at observasjonen er vist.

Karakter 3: Lager om observasjonen er inkludert eller ekskludert i beregningene. '1' er inkludert i beregningene.
Karakter 4: Lager om observasjonen skal ha et merke (Label) eller ikke . '1' For at merke er vist.
Karakter 5: Lager hvordan observasjonen er markert:
1 firkant
2 pluss
3 sirkel
4 diamant
5 x
6 treangel vendt opp
7 treangel vendt ned
8 stjerne
Karakter 6-20 farger rød, grønn og blå på hver 5 karakterer. Verdi fra 0-65535.

Programkode

```
PRoc insight data=test;
window 0 0 100 100;
bar VAR_1   /other=0 yaxis=-1.5 1.7 0.1 0 -2 2;
RUN;
```

Figur 8. Stolpediagram med markering av ugyldige verdier



De sist utgitte publikasjonene i serien Notater

2003/44 L. Østby: Innvandring fra nye EU- land; fortid, nåtid og mulig framtid. 44s.

2003/45 T. Dale, H. Høie og A-K.Johnsen: Evaluering av "Naturressurser og miljø" 30s.

2003/46 L. Solheim: Foreløpige landstall i KOSTRA. Prinsipper, metoder, produksjon og eksemper. 76s

2003/47 A. Hurlen Foss: kvaliteten i boligdelen av Folke- og boligtellingen. 32s.

2003/48 E. Siig Meen og O. Rognstad: Jordbrukstelling 1999- dokumentasjon. 105s.

2003/49 L.Rogstad: Statistiske temakart og X-Map. 32s.

2003/50 E. Holmøy: Velferdsregnskap - et mulig teoretisk rammeverk.35s.

2003/51 C. Wiecek: Undersøkelse om fremtidsplaner, familie og samliv. Dokumentasjonsrapport. 59s.

2003/52 KOSTRA: Arbeidsgrupperapporter 2003. 153s.

2003/53 A. Haglund: Rapport fra arbeidsgruppa om forslag til arbeidsdeling mellom Brønnøysundregistrene (BR) og Statistisk sentralbyrå (SSB). 40s.

2003/54 E. Eng Eibak: Forventningsindikator - konsumprisene. Mai - november 2003. 19s.

2003/55 G. Daugstad: Levekår for ungdom i større byer. 80s.

2003/56 A. Vedø og D. Rafat: Sammenligning av utvalgsplaner i AKU. 17s.

2003/57 L. Belsby: Frafall og vekter i Tidsbruksundersøkelsen 2000-2001. 20s.

2003/58 L.Belsby: Vekter i Forbruksundersøkelsen. 28s.

2003/59 M. Mogstad og L.C. Zhang: På veien fra familie- til husholdningsregister. En metode for prediksjon av samboere uten barn .53s

2003/60 A. Vedø og D. Rafat: Redigering av husholdningsfilen fra Kvalitetsundersøkelsen. 13s.

2003/61 M. Mogstad: Analyse av fattigdom basert på register- og folketellingsdata. 75s.

2003/62 T. Eika og J.A. Jørgensen: Makroøkonomiske virkninger av høye strømpriser i 2003. En analyse med den makroøkonometriske modellen KVARTS.16s

2003/63 B. Mathisen: Flyktninger og arbeidsmarkedet 4. kvartal 2001. 32s.

2003/64 E. Røed Larsen og D.E. Sommervoll: Til himmls eller utfor stupet? En katalogisering av forklaringer på stigende boligpriser. 31s.

2003/65 P.E. Tønjum: Tilbakemelding/ dokumentasjon av prosjektet: Avstemming av KNR mot nye årstall ifølge tallrevisjonen.43s.

2003/66 B.A. Holth: Arbeids- og bedriftsundersøkelsen 2003. Dokumentasjon. 67s.

2003/67 H. Tønseth: Kommuneale helseforskjeller -de finnes, men kan de måles? 15s.

2003/68 T.M. Normann: Omnibusundersøkelsen mai/juni 2003. Dokumentasjonsrapport. 50s.

2003/69 KOSTRA (Kommune- Stat- Rapportering) Rutinebeskrivelse og dokumentasjon. 60s.

2003/70 E. Holmøy og B. Strøm: Fordeling av tjenesteproduksjon mellom offentlig og privat sektor i MSG-6. 25s.

2003/71 J.K. Dagsvik: Hvordan skal arbeidstilbudseffekter tallfestes? en oversikt over den mikrobaserte arbeidstilbudsforskningen i Statistisk sentralbyrå. 67s.

2003/72 A. Steinkellner: Inntektsstatistikk for personer og familier 1999-2001. Dokumentasjon av datagrunnlag og produksjonsprosess. 43s.