En introduktion til interaktiv dataanalyse med programsystemet SAS®

Poul Thyregod

september 2001 (under revision)

LYNGBY 2001

IMM

Trykt af IMM, DTU

Indhold

Introduktion

Denne note er udarbejdet til brug for undervisningen i statistik og dataanalyse ved IMM, DTU. Noten giver en ganske kort introduktion til brug af SAS-systemet til statistiske analyser. Noten er under revision med henblik på at knytte referencer til den lærebog (Petrucelli et al.), der nu benyttes ved det indledende kursus.

Den aktuelle version af SAS-systemet er version 8.1 (betegnet Nashville-versionen). Denne version erstatter den tidligere version (version 6.12) og omfatter væsentligt flere funktioner og menuvalg end version 6.12.

SAS-programsystemets historie går tilbage til begyndelsen af 1960'erne. De første versioner var bygget op omkring *datasæt* og *procedurer*, og brugergrænsefladen var den gængse for den tids databehandling: Man indlæste et *program*, som blev *eksekveret* og - hvis det gik godt - frembragte et *output* på den tids outputmedium, en linieskriver. Under eksekveringen af programmet blev der endvidere frembragt en *log* med meddelelser fra programudførelsen.

Systemet er senere udviklet således at det i dag udover at være et programsystem til statistisk analyse også er et effektivt system til administration af databaser og til ledelsesinformation. Parallelt med denne udvikling i anvendelsessigtet foregår der også en løbende tilpasning af brugergrænsefladen, således at der lægges større vægt på en peg-og-klik kommunikation end på den traditionelle programskrivning.

Version 8 af SAS-systemet har en række forskellige brugergrænseflader, herunder bl.a.

- Interaktiv statistisk dataanalyse (*SAS/INSIGHT*)
- Udførelse af *SAS-programmer*, udarbejdet af brugeren
- Peg og klik fremstilling af grafer til præsentation, (*Graph-N-Go*)
- Menustyret organisering og udførelse af analyseopgave, (*Enterprise Information System*)
- SAS/ASSIST, en peg og klik grænseflade, der giver mulighed for dataadministration, rapportskrivning, fremstilling af grafer mv.

Den store variation i brugergrænseflader, som tilbydes i de aktuelle versioner af SAS-systemet, har til konsekvens, at systemet godt kan virke lidt uoverskueligt, hvis man bare skal udføre en bestemt statistisk analyse. Systemet tilbyder nemlig en række forskellige indgange til den samme analyse.

I denne note vil vi fortrinsvis lægge vægt på beskrivelsen af *interaktive analyser*, som foretages ved en modul, benævnt SAS/INSIGHT. Afsnittene 4 og 5 på side 19 ff. gennemgår en række af de muligheder, der tilbydes for interaktiv dataanalyse. Afsnit 5 illustrerer en række af de eksempler, der er anført i lærebogen i Statistik 1 (K.Conradsen, En Introduktion til Statistik, Bind 1A og 1B).

I afsnit 6 gives en introduktion til afvikling af brugerkonstruerede SAS-programmer, bestående af *DATA-trin* og *PROCEDURER*. I afsnit 7 gives en oversigt over de forskellige procedurer, der er tilgængelige i SAS-systemet, og afsnit 8 giver eksempler på brug af SAS-systemet til bestemmelse af sandsynligheder i standardfordelingerne, herunder til bestemmelse af *p-værdi*, *signifikansniveau* og *styrke* for en række af de test, der er anført i lærebogen.

Endelig gives i afsnit 10 på side 91 en kort vejledning i eksport af tekst og grafik fra SAS-systemet.

SAS-systemets brugergrænseflade kan godt være lidt forskellig på G-databarens SUN-system, og på en PC under Microsoft Windows. Dette gælder specielt håndteringen af tekst og grafik. Det er tilstræbt, at vejledningen skal kunne dække begge opsætninger, men mindre afvigelser kan dog forekomme.

Vi henviser i øvrigt til *Hjælp-funktionen*, som indeholder en udførlig beskrivelse og dokumentation af systemet tilligemed forskellige *tutorials* med betegnelserne 'Getting started with ...'. I PC-versionen findes desuden en on-line dokumentation på en separat CD-rom.

1 SAS-systemet, struktur og dokumentation

1.1 SAS-systemet på DTU

SAS systemet er på DTU tilgængeligt i databarerne og på IMM's computersystem.

Licensaftalen med SAS-Institute tillader studerende, der følger statistikkurser ved IMM, at indlægge SAS-systemet på deres egne PC'er, så længe de er indskrevet ved DTU. Til dette formål kan man låne en CD-rom med SAS-systemet ved henvendelse til Ellen Borup, IMM, Bygning 321, rum 010.

1.2 Hjælp-funktion

Systemets Hjælp-funktion benytter den tilknyttede Net-browser (sædvanligvis *Netscape*). Funktionen aktiveres ved at klikke på `Help` i den øverste bjælke i et vilkårligt SAS-vindue.

Der fremkommer nu en række valgmuligheder

- `Sas System Help` åbner et hjælp-vindue ved hjælp af browseren.
- `Using this window` giver hjælp til det aktuelle vindue
- `Books and Training` giver mulighed for aktivering af SAS Online Doc og SAS Online Tutor
- `Getting Started with SAS Software`
- `SAS on the Web` giver adgang til *Technical Support*, *Frequently Asked Questions* og SAS-Institute Home Page

SAS System Help giver mulighed for at vælge mellem forskellige faneblade,

- `Contents`, der giver en hierarkisk oversigt over indholdet
- `Index`, der giver mulighed for at søge på index-ord, fx procedurenavne, procedure-optioner eller SAS-ord. Når man klikker på ordet i index-listen, fremkommer forklaringen i det andet hjælp-vindue.
- `Search`, der giver mulighed for at søge på *keywords*, dvs ord, der har speciel betydning i SAS-programmer og procedurer. Ved indtastning af et keyword fremkommer der en række *topics*, der vedrører det pågældende keyword.

Under `Contents`-fanebladet har man blandt andet mulighed for at vælge programeksempler (*Sample SAS Programs*) og endvidere en oversigt over de forskellige produkter (`Help on SAS Software Products`). Vælger man denne sidste mulighed, fremkommer en liste over de forskellige produkter, fx SAS-Base, SAS/STAT mfl. Når man "åbner" det pågældende produkt, fremkommer en liste med de procedurer, der er omfattet af produktet. Når man nu "åbner" en procedure, kan man læse en introduktion, samt få beskrevet syntaksen for den pågældende procedure. I afsnit 7 på side 56 ff. gives en oversigt over procedurerne.

I flere af SAS-menuerne er der desuden en knap mærket `Help`. Når man aktiverer hjælp-systemet ved denne knap, vil den højre ramme i hjælp-browser vinduet give information om den pågældende menu. (Man kan bruge ikonen mærket `Locate` øverst i browseren til at finde den hjælp-indgang i browserens venstre ramme, der indeholder det pågældende opslag).

1.3 Dokumentation

Programsystemet SAS® er et integreret programmelsystem til brug for dataadministration og behandling. Blandt de moduler, som indgår i systemet, kan nævnes programmel til håndtering af databaser og filsystemer, til projektledelse og operationsanalyse, computer performance vurdering, interaktivt matrix beregningssprog, samt programmel til statistisk analyse.

Alle modulerne i systemet bygger på den samme kerne af basisprogrammel, herunder

- SAS® language
- Base SAS® procedures
- SAS® Macro language

Producenten, SAS Institute Inc. har udgivet en omfattende samling af vejledninger i brug af systemet.

Således er der udgivet en række User's Guides med vejledning i brug af de enkelte moduler i SAS-systemet. Følgende User's Guides vedrører Version 8

- *Doing More with SAS/ASSIST Software, Version 8*

- *SAS/ETS User's Guide, Version 8*, Volumes 1 and 2
- *SAS/IML User's Guide, Version 8*
- *SAS/INSIGHT User's Guide, Version 8*
- *SAS/OR User's Guide: Mathematical Programming, Version 8*
- *SAS/OR User's Guide: Project Management, Version 8*
- *SAS/OR User's Guide: QSIM Application, Version 8*
- *SAS/QC User's Guide, Version 8*, Volumes 1, 2 and 3
- *SAS/SPECTRAVIEW software User's Guide, Version 8*
- *SAS SQL Procedure User's Guide, Version 8*
- *SAS SQL Query Window User's Guide, Version 8*
- *SAS/STAT User's Guide, Version 8*, Volumes 1, 2 and 3
- *Working with Spatial Data Using SAS/GIS Software, Version 8*

Desuden findes detaljerede *Reference Manuals* for de enkelte produkter, og *Technical Reports*, der beskriver specifikke temaer.

Endvidere er der udgivet en række problemorienterede publikationer, der går på tværs af de enkelte procedurer, og som i enkelte tilfælde kan have karakter af lærebøger i statistiske metoder. Nedenstående publikationer kan være af interesse. Publikationerne synes dog endnu ikke at være opdateret til Version 8.

- *SAS System for Regression*, der beskriver regressionsanalysemodeller.
- *SAS System for Factor Analysis and Structural Equation Mode*, der beskriver faktoranalyse og såkaldte strukturelle .
- *Statistical Quality Control using the SAS-system*
- *Survival Analysis Using the SAS System. A practical Guide*
- *SAS System for Mixed Models*
- *Categorical Data Analysis Using the SAS System*

- *Logistic Regression Using the SAS System, Theory & Applications*
- *Your Guide to Survey Research Using the SAS System*
- *Univariate and Multivariate General Linear Models: Theory & Applications Using SAS Software*
- *Applied Multivariate Statistics with SAS Software*

Hjemmesiden for SAS Institute har en oversigt over publikationer, der forhandles af SAS Institute, herunder Reference Manuals og Technical Reports.

1.4 Hjemmeside

Hjemmesiden for SAS Institute er:

`http://www.sas.com`

her findes en række informationer, henvisninger, downloads mv. Nogle af siderne er kun tilgængelige, hvis man oplyser nummer og licenshaver for sin SAS licensaftale. Licenshaveren er INSTITUT FOR MATEMATISK MODELLERING, (instituttets tidligere navn) og *Site*-nummeret er `SITE=83972003` for PC-versionen, og `Site = 0083972010` for databar versionen.

1.5 Struktur

Som nævnt i indledningen er der forskellige brugergrænseflader, men under disse grænseflader opererer systemet med SAS-procedurer, der udføres på SAS-datasæt eller tabeller. Datasættene (tabellerne) importeres fra databaser eller regneark, eller de kan frembringes og ændres i såkaldte DATA-trin, se afsnit 3.3. Output kan håndteres ved det såkaldte *Output Delivery System* (SAS/ODS), der giver mulighed for valg af forskellige udskriftsformater, herunder eksempelvis HTML-filer. I afsnit 10 angives nogle af mulighederne for eksport af tekst og grafik.

Brugeren kan kombinere disse elementer ved at udføre eksplicitte SAS-programmer, ved at sammenstykke procedurer ved hjælp af fx *Analyst*,

Enterprise Information System (EIS), SAS/ASSIST, eller udnytte dem i en interaktiv form ved brug af SAS INSIGHT (der selv er en SAS-procedure).

SAS systemet er ikke følsomt overfor forskelle mellem STORE og små bogstaver. I eksemplerne har vi dog bestræbt os på at skrive `NØGLEORD` (kommandoer mv.) med VERSALER, mens `brugerdefinerede betegnelser` som fx `variabelnavne mv.` er skrevet med `små` bogstaver.

2 Start af en session, hovedvinduer

På DTU's databarer aktiveres SAS-systemet ved at klikke med midterste musetast, hvorved der fremkommer et vindue `Applications`. I dette vindue vælger man `Statistics` → `SAS 8.01`.

På en PC aktiveres SAS-systemet ved at klikke på ikonen mærket SAS.

Der fremkommer nu fem vinduer samt en *Toolbox-bjælke* (værktøjsbjælken)..

De fem vinduer er betegnet:

Program Editor, programvindue benyttes til at skrive egne SAS-programmer

Log angiver systemets meddelelser fra kørslen af programmer

Output, SAS-systemets outputvindue indeholder tekstoutput fra kørsel af programmer

Results, resultatvinduet angiver filsystemet med resultater fra analyser, dvs fra

Explorer indeholder oversigter over biblioteker og deres indhold, samt angivelse af sas-navne paa systemets filnavne

Oprettelse af brugeromgivelser
Ved start af SAS-systemet eksekveres et lille SAS-program med navnet `autoexec.sas`, såfremt et sådant program findes i roden hos brugeren. Dette program klargører brugeromgivelserne, som fx angiver referencer til SAS-biblioteker (se afsnit 3.5). Endvidere indlæses en fil `SASUSERPREFS` fra brugerens filmappe `sasuser.800`. Denne fil angiver parametre til opsætningen af den aktuelle session. Fra filmappen `sasuser.800` hentes desuden et *katalog* med angivelse af brugerprofilen.

2.1 Menubjælken

Øverst i hvert af disse vinduer er der en *menubjælke*med mulighed for en række valg overordnede valg

File- giver mulighed for at gemme indholdet af det tilknyttede vindue (og i nogle sammenhænge også for at indlæse til vinduet)

Edit- giver mulighed for at redigere indholdet af vinduet (hvis dette er tilladt)

View- giver mulighed for at skifte mellem vinduer

Tools- giver mulighed for at operere på SAS-objekter i form af *tabeller*

Solutions- giver mulighed for at vælge mellem de forskellige analysemoduler, se afsnit 9

Help- starter hjælp-funktionen (se afsnit 1.2).

2.2 Værktøjsbjælken, (Toolbox - vinduet)

Dette vindue indeholder en række funktioner, symboliseret ved sædvanlige ikoner.

Værktøjsbjælken er tilknyttet det aktuelle vindue. Det fremgår af overskriften på værktøjsbjælken, hvilket vindue der er det aktuelle vindue. Når man skifter aktuelt vindue, fx ved at klikke på et andet vindue, skifter værktøjsbjælken automatisk.

Man kan også skifte vindue ved at skrive navnet på det nye vindue i den første, blanke rubrik i værktøjsbjælken.

Funktionerne i værktøjsbjælken virker på det vindue, som bjælken aktuelt er tilknyttet.

2.3 Menubaserede analysevinduer

Ved at vælge optionen `Solutions` i et af vinduerne, og derefter vælge optionen `Analysis`, får man adgang til en række menubaserede analyseværktøjer, se afsnit 9.

I denne note vil vi hovedsageligt beskæftige os med optionen `Interactive data analysis`, der svarer til kald af SAS proceduren, `PROC INSIGHT`. Mulighederne under denne brugergrænseflade vil blive beskrevet i afsnittene 4 og 5.

3 SAS-datasæt og biblioteker

3.1 Lidt om datastrukturer

I lærebogsfremstillinger af den matematiske statistik bruges sædvanligvis symbolerne $X, Y, \ldots$ til at repræsentere stokastiske variable (potentielle observationer) og de tilsvarende små bogstaver til at symbolisere en vilkårlig observeret værdi.

Når der optræder *kvantitative* "forklarende variable" som fx i regressionsanalysemodeller, bruger tilsvarende små latinske bogstaver (fx t) til at symbolisere værdier af den forklarende variable.

Denne notation er i overensstemmelse med de sædvanlige konventioner for opskrivning af matematiske udtryk, hvor *størrelser* eller *numeriske værdier* symboliseres ved et enkelt bogstav, eventuelt med fodtegn og indices (se fx DS/ISO 31, *Quantities and Units*).

I indledende fremstillinger af den matematiske statistik skelner man sædvanligvis mellem variable, der er udtryk for en gruppering (klassifikation af data) og variable, der er udtryk for en numerisk størrelse af den pågældende egenskab.

Eksempel 1 ***Lærebogsnotation***

Betragt fx et sæt samhørende registreringer af vægt, højde og køn for en række personer og antag, at man er interesseret i at beskrive variationen af vægten.

Her er personens køn udtryk for en klassifikation. Hvis man blot er interesseret i at beskrive en eventuel forskel i fordelingen af vægten for de to køn, ville man i lærebogsfremstillingen fx vælge symbolet X_i for vægten af den i'te person fra kvindegruppen, og symbolet Y_i for vægten af den i'te person fra mandegruppen.

Hvis man i stedet blot var interesseret i at beskrive variationen af vægten som funktion af personernes højde uden hensyn til kønnet, kunne man vælge symbolet x_i til at angive højden af den i'te person (i gruppen af alle personer) og Y_i til at angive den registrerede vægt for den pågældende person.

Var man endelig interesseret i at beskrive variationen af vægten under hensyntagen til såvel køn som højde af de undersøgte personer, måtte man benytte en dobbeltindicering til at symbolisere de enkelte observationer, hvor første index angiver kønnet ($i = 1$ for kvinde, og $i = 2$ for mand), og andet index angiver observationsnummeret i den pågældende gruppe ($j = 1, \ldots, n_1$ for kvinder, og $j = 1, \ldots, n_2$ for mænd). Notationen bliver imidlertid hurtigt uoverskuelig, når antallet af forskellige grupperinger (fx kvinde/mand, ung/gammel, gift/ugift, hårfarve, etc) vokser. □

I programmel til statistisk analyse og databehandling benyttes sædvanligvis en anden konvention. Dels kan man tillade brug af mere meningsfyldte *navne* til at symbolisere variable. Hvis fx en variabel angiver vægten af en person, kan man betegne denne variable med `vaegt` snarere end med det matematiske symbol Y, og endvidere er det mere hensigtsmæssigt at benytte samme slags notation for variable, hvad enten de er udtryk for en klassifikation, eller udtrykker en numerisk størrelse.

Eksempel 2 ***Notation i statistikprogrammel***

I den situtation, der blev betragtet i eksempel 1 ville man således lægge vægt på, at samlingen af data omfatter en række *personer*, og at der til hver person er knyttet en værdi af hver af de tre størrelser, vægt, højde og køn. Man kunne derfor bruge navnene `vaegt`, `hoejde` og `koen` til at symbolisere de tre størrelser.

Sædvanligvis knytter der sig ikke speciel interesse til observationens nummer, hvorfor dette nummer kun optræder som en mere skjult intern variabel (med det reserverede navn `_N_`). Hvis der er interesse for også at inddrage nummeret eller en anden identifikation af den enkelte observation, kan man på tilsvarende måde bruge et sædvanligt variabelnavn til at symbolisere denne identifikation.

I afsnit 5.13.1 på side 36 vil vi nærmere komme ind på, hvorledes man formulerer *statistiske modeller* ved brug af denne notation. □

3.2 SAS-datasæt, observationer og variable

Data er i SAS-systemet organiseret i de såkaldte datasæt (engelsk: *dataset*). I nogle grænseflader kaldes et datasæt for en table. Blandt andet kan output undertiden forekomme som 'tabeller' (tables). Sådanne objekter behandles eksempelvis i menubjælken under `Tools`.

Et SAS-datasæt består af et antal observationer, hvor der til hver observation er knyttet den "observerede værdi" af et antal variable. I en *Table* kaldes observationerne for rækker (rows), og de variable kaldes for søjler (columns).

Man kan se indholdet af et datasæt ved at gå hen i *Explorer-vinduet*, klikke på navnet, der angiver det relevante bibliotek (ofte `WORK`) og derefter på datasættets navn. Der fremkommer da en viewtable (se afsnit 3.4), som er et skema, hvor hver række angiver én observation, og hvor værdierne af de tilknyttede variable er angivet i kolonnerne.

Observationerne i et givet datasæt har alle tilknyttet de samme variable; evt kan de observerede værdier af en variabel have værdien uoplyst, sædvanligvis symboliseret ved et punktum.

Man refererer til de variable ved et variabelnavn. Et variabelnavn må sædvanligvis højst bestå af 32 alfanumeriske karakterer (herunder karakteren underscore (_)).

En variabel har desuden tilknyttet en række forskellige karakteristika:

- en label, som er den betegnelse for den variable, der bruges i udskrifter
- en type, som angiver om den variable er numerisk (*Numerical*), eller en tekstvariabel (*Character*)
- et format, som angiver, hvorledes den variable udskrives.

Man kan ændre disse karakteristika ved fx at klikke på navnet på den variable i en viewtable (afsnit 3.4), eller i datavinduet med datasættet i SAS-Insight (afsnit 4.3).

I analyserne kan værdiskalaen for en numerisk variabel fortolkes som målt på en intervalskala (dvs som en kvantitativ (eng: Quantitative) størrelse, hvor det har mening at udføre de sædvanlige regneoperationer på værdierne), eller blot som nominalværdier (dvs som symboler for klasser (eng: Class), hvor de numeriske værdier blot er udtryk for en kodning af nogle kvalitative størrelser, fx `Kvinde = 1, Mand = 2`). I de interaktive programmer vil en *tekstvariabel* automatisk blive fortolket som nominalværdier. I *SAS-procedurer* bruger man en CLASS sætning til at angive, at værdierne af en variabel angiver resultatet af en klassifikation. Endelig kan en variabel benyttes til at angive den vægt (eng: weight), hvormed den pågældende observation indgår i analysen (se side 23).

Observationerne er (internt) nummereret fortløbende. Observationens nummer i den aktuelle version af et datasæt har variabelnavnet `_N_`.

3.3 Indlæsning af data

3.3.1 Import fra regneark

SAS-systemet giver mulighed for import af data fra regneark og databaser. Klik på `File` → `Import Data` og følg vejledningen.

Der er mulighed for at importere data fra Microsoft Excel-regneark, fra Microfsoft Office 2000 regneark, fra Microsoft Access tabeller, fra dBase filer og fra Lotus regneark. Desuden kan man selv definere sit indlæseformat ved brug af *External File Interface*.

Ved import fra regneark eller databaser skal man være opmærksom på, at SAS-systemet bruger et punktum som decimaltegn, hvor den europæiske version af mange regneark ofte bruger et komma (hvilket også er ISO-standard). Endvidere skal man være opmærksom på, at de særlige nationale bogstaver æ, ø og å ofte vil blive fortolket uhensigtsmæssigt.

3.3.2 Direkte indtastning

Man kan oprette et datasæt til direkte indtastning ved at klikke på `Tools` i menubjælken og vælge `Table Editor`. Der fremkommer nu en *Viewtable* (se afsnit 3.4), hvor man kan indtaste data. Når indtastningen er færdig, vælger man `File` → `Save As` i menubjælken, hvorefter der vælges bibliotek og datasætnavn.

Ved brug af den interaktive procedure SAS-Insight kan data direkte indtastes i et regneark, der gemmes i et datasæt (se afsnit 4.3).

3.3.3 Brug af SAS-data trin

Man kan endelig oprette et datasæt ved hjælp af et data-trin i et SAS-program, se afsnit 6.4.

3.4 Viewtable

En Viewtable er en repræsentation af et datasæt i et tosidet skema, hvor hver række angiver én observation, og hvor værdierne af de tilknyttede variable er angivet i kolonnerne. Øverst i hver kolonne står navnet på den pågældende variable. Ved at klikke med den højre musetast på variabelnavnet fremkommer en dialogboks, hvor man har har mulighed for at ændre

- Navn
- Label
- Længde (ved intern repræsentation af numeriske variable, type *Numerical*) og antal karakterer (ved tekstvariable, type *Character*)
- Format (for udskrivning)
- Informat for indtastning (kan være relevant for dato- og klokkeslets-variable)
- Type (*Numerical* eller *Character*)

3.5 SAS-biblioteker

Datasættene er organiseret i SAS-biblioteker (SAS libraries). En SAS-session opretter altid biblioteket `WORK` til de datasæt, der er lokale for den pågældende sesssion. Sædvanligvis tilknyttes også bibliotekerne `SASHELP` og `SASUSER` til en session.

Når et datasæt er tilknyttet et bibliotek, refererer man til det pågældende datasæt ved at skrive bibliotekets navn efterfulgt af et punktum og af datasætnavnet. Således refererer

```
mitbib.brugtbil
```

til datasættet `brugtbil` i biblioteket `mitbib`.

Man kan se hvilke biblioteker, der er tilknyttet den aktuelle session ved at benytte *Explorer*-vinduet, fx ved at bruge værktøjsbjælken og klikke på ikonet for *Explorer*-vinduet (et forstørrelsesglas).

Man kan se indholdet af et bibliotek ved herefter at klikke på navnet på det pågældende bibliotek.

I samme menu kan man oprette et nyt bibliotek ved at klikke på `New Library`.

Man kan endvidere oprette et bibliotek ved hjælp af en `LIBNAME`-sætning i et SAS-program eller i kommandolinien på værktøjsbjælken.

3.6 Relation til computersystemets filsystem

Selv om SAS-datasæt opbevares som filer i computerens filsystem, tillader systemet almindeligvis ikke, at man refererer til dem ved deres filnavn i computerens filsystem. Referencen foregår ved brug af datasætnavn (og evt biblioteksnavn).

Man kan heller ikke bare kopiere en fil, der indeholder et datasæt, fra ét computersystem til at andet. Såfremt man ønsker at flytte et datasæt mellem computersystemer, skal man bruge *Export*-faciliteten under `File`.

Ved andre referencer til filer i computerens filsystem skal man på tilsvarende måde knytte en forbindelse mellem et SAS-navn for filen og dens reference (sti og filnavn) i computerens filsystem. En sådan reference kan fx etableres ved brug af *Explorer*-vinduet, højreklik på `File Shortcuts` og vælg `New`, eller man kan bruge `FILENAME`-sætningen i et SAS-program.

4 Interaktiv dataanalyse

En interaktiv analysesession foregår i SAS/INSIGHT omgivelserne. Brugen af SAS/INSIGHT er beskrevet i manualen SAS/INSIGHT User's Guide, Version 8. Hjælp vedrørende disse omgivelser findes i hjælp, `Contents` under `Help on SAS Software Products`, `SAS/INSIGT`.

4.1 Start af session

Man starter en interaktiv analysesession ved at vælge optionen `Solutions` i menubjælken, → `Analysis` → `Interactive Dataanalysis`.

Alternativt kan man blot skrive `INSIGHT` paa kommandolinien i et toolboxvindue.

Såfremt man ønsker at gemme de kommandoer, der genereres under den interaktive session, etablerer man et SAS-navn for en fil (fx ved brug af *Explorer*-vinduet), og derefter skriver man

```
INSIGHT FILE=filnavn
```

på kommandolinien i toolboxen. Dette bevirker, at kommandoerne gemmes i filen med SAS-navnet `filnavn`.

Man kan endelig starte en interaktiv session fra et SAS-program i programvinduet ved kald af SAS proceduren, `PROC INSIGHT` (se afsnit 7.7).

4.2 Valg af datasæt

Ved starten af den interaktive session fremkommer et `Open` vindue, der viser de SAS-biblioteker, (`Library`), der er tilknyttet sessionen. Ved at klikke på navnet på et bibliotek, fremkommer der en oversigt over de datasæt, der ligger i det pågældende bibliotek. Biblioteket `WORK` er arbejdsbiblioteket, der indeholder de ikke-permanente datasæt, som er tilknyttet sessionen. Man åbner et datasæt ved at klikke på navnet på det pågældende datasæt og derefter klikke på knappen `Open` nederst i vinduet.

Man opretter et datasæt med et regneark klart til indtastning af data ved at klikke på knappen `New` nederst i vinduet.

4.3 Datavindue

Datasættet bliver nu vist i et `Datavindue`. Datasættet bliver vist i et skema (regneark) med én observation i hver række. Første søjle (kolonne) angiver observationens nummer (tildelt af SAS-systemet), og de følgende kolonner angiver navnene på de variable, som hører til datasættet, og værdierne af de pågældende variable for hver observation. Uoplyste værdier er angivet med et punktum.

Øverst i hver kolonne er en rubrik, hvor der står anført, hvorvidt den pågældende variabel fortolkes som målt på en *intervalskala* (`Int`), eller blot som *nominalværdier* (`Nom`). Man kan ændre fortolkningen af den variable ved at klikke på den pågældende rubrik (med venstre musetast).

Ved siden af denne rubrik er en rubrik, der giver mulighed for at vælge, at den pågældende variabel skal bruges som en `Group, Label, Freq` eller `Weight`, se side 23.

Højre musetast giver mulighed for at navigere rundt i datavinduet, samt at editere i data.

Øvelse 1 ***Datasæt med brugtbilpriser***

(Øvelsen forudsætter, at datasættet `brugtbil` er indlæst i biblioteket `mitbib`).

Vælg interaktiv dataanalyse, og vælg datasættet `brugtbil` under biblioteket `mitbib`.

Find de variable `FABR` og `ALDER`. Opfattes værdierne som målt på intervalskala eller nominalskala ?
Ret `Alder`, sådan at den bliver angivet på nominalskala.
Ret den tilbage igen til intervalskala.

□

5 Generelle valgmuligheder under en interaktiv session

5.1 Overordnede valgmuligheder

De forskellige vinduer, der optræder under en interaktiv session, har alle samme overordnede valgmuligheder, markeret i menubjælken: `File, Edit, Analyze, Tables, Graphs, Curves, Vars` og `Help`. Ikke alle valgmuligheder er dog reelt til stede i alle vinduer. De muligheder, der ikke kan bruges, er markeret med svagere skrift. De samme valgmuligheder fremkommer, hvis man klikker med højre musetast i et af outputvinduerne, knyttet til den interaktive session.

5.1.1 Forhåndsvalg af variable i datavindue

Man kan vælge variable i datavinduet ved at klikke på navnet på den variable i den øverste linie af vinduet.

Når man allerede har valgt en af de variable i vinduet, og man klikker på en ny variabel, bliver det tidligere valg annuleret. Man kan dog vælge flere variable ved at trykke på Ctrl-tasten på tastaturet samtidig med at man vælger den variable. Hvis man ønsker at vælge flere variable, som står ved

siden af hinanden, kan man blot trække musen hen over disse navne, mens museknappen er nedtrykket, eller trykke på Shift-tasten på tastaturet mens man vælger de yderste. Hvis man ønsker at annullere valget, kan man klikke på en tom plads.

Når man har forhåndsvalgt variable i datavinduet, da vil de følgende valg af analyser blive udført på disse variable, straks man klikker på den pågældende analyseform. Dette sker, uden at man bliver bedt om at bekræfte variabelvalget !

5.1.2 Valg af variable i menu-vindue

Når man vælger en de funktioner, der benyttes til dataanalyse (`Analyze, Tables, Graphs, Curves, Vars`), fremkommer en menu, der giver mulighed for at vælge variable, evt metode, og eventuelle outputvariable.

Til venstre i menuerne er en liste, der som overskrift angiver navnet på det datasæt, der betragtes, og i listen er angivet navnene på de variable, der er i datasættet.

Valg af datasæt
Man kan *vælge et andet datasæt* ved at klikke på overskriftsrubrikken, hvorefter der fremkommer et vindue med en liste over de datasæt, som er åbne.

Valg af variable
Man *vælger én eller flere variable* i det tilknyttede datasæt ved at klikke på navnet på den variable i venstre liste. Den (eller de) valgte variable bliver markeret i listen ved at den pågældende linie skifter farve.

Den (eller de) valgte variable kan nu *flyttes* over i en af de andre lister på menuen ved at klikke på listens overskrift. Herved fremkommer navnene på de valgte variable i listen.

Man kan *fjerne en variabel* fra en liste ved at *vælge* den i listen (klikke på navnet), og derefter klikke på `Remove`-knappen nederst til højre i menuen.

En række af menuerne giver mulighed for at tillægge nogle variable specielle funktioner i analysen. Man kan vælge at lade en variabel angive en *gruppering* ved at placere den i rubrikken `Group`. Den valgte analyse bliver da udført for hver værdi af den pågældende grupperingsvariable.

En variabel (kun intervalvariable) kan desuden optræde som `Freq`- eller som `Weight`-variabel. Når en variabel optræder som `Freq`-variabel, vil den pågældende observation indgå i analysen med et antal "kopier", der er lige så stort, som angivet i hyppighedsvariablen "`Freq`". Det samlede antal observationer i analysen er altså summen af værdierne af "`Freq`".

Når en variabel optræder som `Weight`-variabel, vil værdien af den pågældende observation blive vægtet med den tilsvarende værdi af `Weight`-variablen, men opfattelsen af antallet af observationer ændres ikke. Antallet af observationer er stadig blot antallet af observationer i datasættet.

Endelig kan en variabel tildeles rollen som `Label`, dvs at værdierne af den pågældende label-variable vises, når en observation vælges på et plot.

5.1.3 Valg af metode

Ved at klikke på knappen mærket `Method` kan man fremkalde en menu, der giver mulighed for at vælge mellem metoder til udførelse af den pågældende analyse. Valgmulighederne afhænger af den aktuelle menu.

5.1.4 Valg af output

Ved at klikke på knappen mærket `Output` kan man fremkalde en menu, der giver mulighed for at vælge outputvariable, outputform mv. Valgmulighederne afhænger af den aktuelle menu

5.1.5 Eksekvering af funktion i menuvindue

De valg, der er foretaget i en menu, træder i kraft når man klikker på knappen mærket `OK`.

5.1.6 Fortrydelse af valgt menu

Hvis man fortryder, at man har valgt en menu, kan man klikke på knappen mærket `Cancel`. Herved lukkes det pågældende vindue, uden at de valgte muligheder træder i funktion.

5.2 Valgmuligheder i grafer

I nederste venstre hjørne af hver graf er der en lille pil. Når man klikker på denne pil, kan man ændre på akseinddeling (tick-marks), man kan fjerne akser, ændre markører mv. (De samme muligheder fremkommer, hvis man klikker med højre musetast inde i grafen).

5.3 File funktionen

- `New` Giver mulighed for at danne et nyt datasæt ved indtastning i datavindue
- `Open ...` Giver mulighed for at åbne et nyt datasæt
- `Save` Giver mulighed for at gemme output (grafer, tabeller) i filer, der er tilgængelige fra `Results`-vinduet mv, se afsnit 10.2.2.
- `Print ...` Giver mulighed for at gemme output i en fil af brugervalgt type (fx .pdf, postscript etc.), se afsnit 10.2.1
- `Print preview`
- `End`

5.4 Edit funktionen

Edit-funktionen giver blandt andet mulighed for at

- strukturere vinduer, herunder animere vinduer
- transformere variable
- fravælge observationer, ændre labels i plot
- ændre formats (antal decimaler mv.) for udskrifter
- ændre symboler (markører) på grafer
- ændre font for tekst i grafer

Øvelse 2 ***Transformation af variable***

(Øvelsen forudsætter, at den interaktive session fra øvelse 1 stadig er åben).

Vælg `Edit`-funktionen og frembring en ny variabel `minlog`, der angiver logaritmen til prisen. □

5.5 Analyze-optioner

Under `Analyze` har man følgende valgmuligheder:

- `Histogram/Bar chart (Y)`, (Afsnit 5.6)
- `Box Plot/Mosaic Plot (Y)`, (Afsnit 5.7)
- `Line Plot (Y X)`, (Afsnit 5.8)
- `Scatter Plot (Y X)`, (Afsnit 5.9)
- `Contour Plot (Z Y X)`, (Afsnit 5.11)
- `Rotating Plot (Z Y X)`, (Afsnit 5.10)
- `Distribution (Y )`, (Afsnit 5.12)
- `Fit (Y X)`, (Afsnit 5.13)
- `Multivariate (Y's)`, (Afsnit 5.14)

5.6 Histogram

Histogram/Bar Chart menuen giver mulighed for at vælge én eller flere variable i det tilknyttede datasæt og at tegne et histogram over værdierne af de(n) valgte variable.

Vi skal ikke komme nærmere ind på de forskellige plots, men blot nævne, at man godt kan plotte flere variable på én gang, og at man med markøren kan *udvælge* bestemte observationer. De udvalgte observationer bliver da markeret på alle de aktive vinduer.

Øvelse 3 *Histogram for brugtbilpriser*

(Øvelsen forudsætter, at den interaktive session fra øvelse 1 stadig er aktiv).

Klik på `Analyze` og vælg `Histogram/Bar Chart (Y)`.

Venstre del af vinduet viser navnene på de variable, der er knyttet til sessionen. Vælg den variable `PRIS` ved at klikke på den med venstre musetast. (`PRIS` bliver nu fremhævet). Klik på knappen mærket `Y`. Herved bliver `PRIS` placeret i listen under `Y`. Klik nu endelig på `OK`

Klik nu igen på `Analyze` og vælg `Histogram/Bar Chart (Y)`, men vælg denne gang den variable `FABR` og tegn histogrammet over fabrikater.

Prøv nu at klikke på søjlen svarende til `Ford` i histogrammet. Hvad sker der med histogrammet over priser ? Kan man sige noget om, hvilket af de to fabrikater, der opnår den højeste brugtvognspris ? □

5.7 Boxplot

Et boxplot er en kompakt måde til beskrivelse af værdierne af en variabel. Plottet består af en central kasse (box), der indeholder de midterste 50 % af observationerne, fra 25 % fraktilen til 75 % fraktilen. Kassen er delt i to dele ved medianen (50 %-fraktilen). Fra hver ende af kassen udgår der en linie (smal kasse), den såkaldte pisk (engelsk *whisker*). Pisken strækker sig principielt fra 25 %-fraktilen til den mindste observation og fra 75 %-fraktilen til den største observation. Dog er længden af pisken begrænset til 1.5 gange interkvartilbredden (dvs afstanden mellem 25 % og 75 %-fraktilerne). Observationer, der eventuelt måtte ligge udenfor denne længde er markeret med et kryds. (Ved valg af metode kan man ændre denne maksimale piskelængde).

Et mosaic plot benyttes til at repræsentere en todimensional antalstabel.

Øvelse 4 *Boxplot for brugtbilpriser*

(Øvelsen forudsætter, at den interaktive session fra øvelse 1 stadig er aktiv).

Klik på `Analyze` og vælg `BoxPlot/Mosaic Plot (Y)`.

Vælg den variable `PRIS` og placer den under `Y`. Vælg den variable `FABR` og placer den under `GROUP`. Klik på `OK`.

Bemærk, at der fremkommer to boxplots; ét for hver værdi af `FABR`. De to plots er i hver sin ramme, og med hver sin akse mv.

Klik dernæst igen på `Analyze` og vælg igen `BoxPlot/Mosaic Plot (Y)`. Vælg den variable `PRIS` og placer den under `Y`. Vælg den variable `FABR` og placer den under `X`.

Klik på `OK`.

De to boxplots for de to værdier af `FABR` er nu placeret indenfor den samme ramme med samme akse mv. □

Øvelse 5 *Mosaicplot for brugtbilpriser*

(Øvelsen forudsætter, at den interaktive session fra øvelse 1 stadig er aktiv).

Klik på `Analyze` og vælg `BoxPlot/Mosaic Plot (Y)`.

Vælg den variable `PRIS` og placer den under `Y`. Vælg den variable `ALDER` og placer den under `X`. Tryk på `OK`.

Der fremkommer nu en todimensional afbildning af andelene af de to bilmærker i hver aldersgruppe. Den lodrette inddeling angiver for hver aldersgruppe andelen af de to bilmærker, og bredden af søjlerne er proportional med antallet af observationer i den pågældende aldersgruppe.

□

5.8 Lineplot

Menuen giver mulighed for at tegne `Y`-værdier mod de tilhørende værdier af `X` og forbinder dem med rette linier i observationsnummerorden.

Øvelse 6 *Lineplot af brugtbilpriser mod alder*

(Øvelsen forudsætter, at den interaktive session fra øvelse 1 stadig er aktiv).

Klik på `Analyze` og vælg `Lineplot (Y X)`.

Vælg den variable `PRIS` og placer den under `Y`. Vælg den variable `ALDER` og placer den under `X`. Tryk på `OK`.

Der fremkommer nu en graf af pris mod alder, hvor observationerne i hver aldersgruppe er forbundet i observationsrækkefølge. Billedet er ganske forvirrende, da der er flere observationer i hver aldersgruppe og der tilsyneladende ikke er nogen sammenhæng mellem observationsrækkefølge og pris.

Mens denne graf er åben, kan man finde histogrammet over fordelingen på fabrikater og *udvælge* et af fabrikaterne ved at klikke på søjlen i histogrammet. De tilsvarende observationer bliver nu markeret på linieplottet.

□

5.9 Scatter Plot

Dette valg giver mulighed for at tegne værdier af en (eller flere) `Y`-variabel op mod tilhørende værdier af en (eller flere) `X`-variable.

Øvelse 7 *Afbildning af brugtbilpriser mod alder ved scatterplot*

(Øvelsen forudsætter, at den interaktive session fra øvelse 1 stadig er aktiv).

Klik på `Analyze` og vælg `Scatterplot (Y X)`.

Vælg den variable `PRIS` og placer den under `Y`. Vælg den variable `ALDER` og placer den under `X`.

Klik på `OK`. Der fremkommer nu en afbildning af `PRIS` mod `ALDER`.

□

Øvelse 8 *Udvælgelse af observationer fra graf*

(Øvelsen forudsætter, at vinduet med afbildningen af `PRIS` mod `ALDER` fra øvelse 7 stadig er aktiv).

Aktiver vinduet med scatterplottet af `PRIS` mod `ALDER`.

Før cursoren op til punktet (7, 141800) og klik på det punkt, der er afbildet. Herved vises observationsnummeret. Bemærk at den pågældende observation fremhæves i alle de øvrige vinduer. (Se f.eks. i Datavinduet). Gå tilbage til scatterplottet, og dobbeltklik på den valgte observation. Der fremkommer nu et vindue `Examine Observation`, der viser værdierne af samtlige variable for den pågældende observation.

Fjern fremhævelsen af observationen ved at klikke på positionen ved siden af observationen.

Placer igen cursoren i punktet (7, 141800) og hold venstre tast nede, mens cursoren bevæges skråt nedad mod venstre til punktet (4, 69000), hvorefter cursoren slippes. Der dannes nu et indrammet felt, og man har udvalgt observationerne indenfor dette felt (observationerne er fremhævede i alle vinduer). Klik med venstre musetast inde i rammen, og bevæg musen mod højre. Rammen forskydes nu til højre. Hvis man venter med at slippe musetasten til rammen er faldet til ro, opnår man at flytte rammen til den nye position. Når man dobbeltklikker indenfor rammen, fremkommer `Examine Observation`-vinduet for de udvalgte observationer.

Hvis man slipper musetasten, mens rammen endnu er i bevægelse, fortsætter rammen automatisk med at køre frem og tilbage indenfor plottefeltet. Bevægelsen stoppes ved at klikke i vinduet. □

Øvelse 9 *Udvælgelse af observationer svarende til bestemte værdier af de variable*

(Øvelsen forudsætter, at vinduet med afbildningen af `PRIS` mod `ALDER` fra øvelse 7 stadig er aktiv).

Aktiver vinduet med scatterplottet af `PRIS` mod `ALDER`.

Klik på Edit, vælg Observations og vælg Find. Der fremkommer nu et vindue, hvor man kan vælge en variabel og for den valgte variabel kan man vælge et område af værdier. Vælg FABR = Ford og tryk på Apply. Man har derved *udvalgt* (selected) observationerne for biler af mærket Ford, og de bliver markerede i alle vinduer.

□

Øvelse 10 *Valg af plottesymboler og farver*

(Øvelsen forudsætter, at vinduet med afbildningen af PRIS mod ALDER fra øvelse 7 stadig er aktiv).

Klik nu på Edit og vælg Tools. Der fremkommer nu et lille vindue med en række forskellige farver og grafiske symboler.

Klik på △. Der fremkommer nu en menu Mark Observations med tre lister. Én liste over variable, en anden over testmuligheder, og en tredie over værdier for den valgte variable.

Vælg den variable FABR og vælg Test "=" og vælg endelig værdien Ford.

Gå derefter tilbage til Tools-menuen og klik på den røde farve. I den menu, der nu fremkommer (Color Observations) vælger man atter FABR = Ford, hvorved alle observationer for biler af mærket Ford bliver markeret med rødt.

Vælg derefter en anden farve, og et andet symbol og marker Folkevogne med denne kombination.

Hvad mener du nu om brugtvognspriserne for de to bilmærker ?

□

Øvelse 11 *Udeladelse af observationer fra analyse*

(Øvelsen forudsætter, at datasættet Brains er indlæst i biblioteket mitbib og åbnet i SAS/INSIGHT).

Datasættet indeholder Vi vil her undersøge fordelingen af kropsvægte . kropsvægten (i kg) og hjernevægten (i g) for en række pattedyrearter. Kropsvægten er angivet i den variable Body, og hjernevægten i Brain.

Tegn et *scatter plot* af hjernevægt mod kropsvægt som i øvelse 7. Der er to observationer i øverste højre hjørne, der skiller sig kraftigt ud fra de andre. *Vælg* disse to observationer (som illustreret i opgave 8).

Klik nu på Edit i menubjælken, og vælg Observations → Hide in Graphs. Nu udelades observationerne fra grafen, og grafen ændres.

Hvis der havde været andre grafiske vinduer åbne, ville observationerne også blive udeladt fra disse grafer.

Bemærk, at Edit → Observations også giver mulighed for at udelade observationer fra beregningerne under SAS/INSIGHT. De observationer, der er udeladt, er stadig synlige i *datavinduet*, men de har fået en speciel markering ved observationsnummeret. □

Øvelse 12 *Parvis afbildning af variable i scatter plot matrix*

Datasættet fingre0*y* indeholder samhørende værdier af længden af tommel-, lang- og ringfinger samt personens køn, højde og vægt for en række studerende.

Vælg dette datasæt og vælg Scatterplot (Y X). I Scatterplot-menuen vælges nu alle de variable sex, tommel, lang, ring, vaegt, hoejd og placeres under Y. Tilsvarende placeres alle de variable under X, og der klikkes OK.

Der fremkommer nu en *scatter plot matrix*, hvor afbildningerne er organiseret i en matrix med alle parvise kombinationer af de variable. Afbildningerne er opstillet sådan at alle afbildninger i en række har samme lodrette akse, mens alle afbildninger i en søjle har samme vandrette akse. Diagonalrubrikkerne i matricen angiver *navnene* på den pågældende variable og mindste og største værdi af den variable.

Også her kan man i den enkelte graf danne et indrammet felt, en såkaldt *brush* (børste, pensel), som i øvelse 8 og bevæge den hen over grafen.

□

5.10 Contour Plot

Et kontur plot er afbildning af udglattede niveaukurver for Z-variablen som funktion af X og Y.

Ved valg af Method har man mulighed for at vælge mellem forskellige interpolationsmetoder (lineær, eller spline).

Ved at klikke med højre musetast i figuren har man mulighed for at udfylde områderne med farve og få angivet farvekoderne.

Øvelse 13 *Konturplot af brugtvognspriser*

(Øvelsen forudsætter, at den interaktive session fra øvelse 1 stadig er åben.

Klik på Analyze → Contour Plot (ZYX).

Vælg logaritmen til prisen som Z-variabel, og alder og kilometerstand som hhv X- og Y-variabel. Vælg Fabrikat som Group variabel og vælg Spline som interpolationsmetode.

Klik derefter i figuren og udfyld områderne. □

5.11 Tredimensional afbildning (roterende plot)

Valget af Rotating Plot giver mulighed for at tegne værdier af en (eller flere) Z-variabel op mod tilhørende værdier af en (eller flere) Y- og X-variable.
Endvidere kan man tilpasse en flade til Z-værdierne og farvelægge områder på fladen svarende til værdierne af en numerisk variabel ZColor.

Output-optionen giver mulighed for at vælge Rays (hvert plottepunkt er forbundet med origo), Cube (punktsværmen er indlagt i en kasse) og Depth (Markeringen af observationerne er med to størrelser symboler; store for de nærmeste og små for de fjernere).

Output-optionen giver endvidere mulighed for at vælge at tilpasse en flade Fit Surface (og Method-optionen giver mulighed for at vælge, om fladen skal tilpasses ved lineær interpolation, eller en spline-tilpasning)

I venstre side af figuren er der en bjælke, hvor man kan vælge forskellige roteringer af plottet, og en skyder, der kan ændre rotationshastigheden. Hvis man vælger hånd-ikonen i Tools-vinduet, kan man selv bestemme rotationsretningen ved at bruge håndsymbolet til at dreje punktsværmen i en bestemt retning.

Hvis man har valgt Fit Surface i output-vinduet, og man desuden har valgt en variabel for ZColor, kan man fremkalde den farvelagte flade ved at klikke med højre musetast i figuren, og derefter i Drawing Modes vælge enten Block Color eller Smooth Color. Herved bliver fladen farvelagt efter en passende farveskala for Z-værdierne. Farveskalaen vises i figurens højre side. Valget af Block Color bevirker at fladens gitternet bliver farvelagt på en sådan måde, at hvert gitterrektangel er fyldt med én farve, mens valget af Smooth Color bevirker, at fladen farvelægges i områder, der afgrænses af glatte kurver.

Øvelse 14 ***Tredimensional figur***

(Øvelsen forudsætter, at datasættet Brugtbil, aktiveret i øvelse 1 stadig er aktivt).

Klik på Analyze → Rotating Plot (Z Y X).

Vælg FABR og placer den under Z; vælg PRIS og placer den under Y og vælg endelig ALDER og placer den under X. Klik på OK. Vælg forskellige rotationsretninger i bjælken i venstre side. Vælg forskellige rotationshastigheder på skyderen nederst i bjælken.

Aktiver Tools-menuen og klik på håndikonen. Brug denne til at rotere figuren sådan at FABR kommer opad. Sæt figuren til at rotere omkring FABR-aksen og bemærk, at observationerne ligger i to lag.

Gå tilbage til Tools-menuen og klik på pile-ikonen. □

5.12 Fordelingsanalyse (Distribution)

Valget af Distribution giver mulighed for at tilpasse forskellige fordelingstyper til de observerede værdier af en eller flere variable.

Metode-menuen giver mulighed for forskellige valg af divisor ved bestemmelsen af variansestimatet.

Output-menuen giver mulighed for at vælge beregning af Momenter, kvantiler, tabel over hyppigheder, konfidensinterval for middelværdien, diverse tests for position. Endvidere kan man vælge afbildning af histogram, Box Plot, Normalfordelings fraktildiagram QQ-plot. Der kan estimeres tætheder svarende til Normal-fordeling, Lognormal fordeling, Exponentiel fordeling og Weibull fordeling.

Endvidere kan der foretages en række ikke-parametriske udglatninger af tæthedsfunktionen (kernel-estimation). Endelig kan der udføres et t-test for position.

Der kan foretages sammenligninger af den kumulerede fordelingsfunktion med de ovennævnte parametriske fordelinger, og der kan udføres test for tilpasning af fordelingen.

Endelig kan der udføres robust estimation af middelværdien ved de såkaldte *trimmed mean* og *Winsorized mean*.

Øvelse 15 ***Histogram for nittehoveddiametre***

Datasættet `eks4_7` indeholder data fra eksempel 4.7 (og 1.1) i Statistik 1 bogen. Den variable `diam` angiver nittehovedets diameter.

Øvelsen forudsætter, at datasættet er åbnet i SAS/INSIGHT.

Klik på `Analyze` → `Distribution (Y)`.

Vælg den variable `Diam` og placer den under `Y`.

Vælg `Output` og under `Graphs` vælg `Histogram/Bar Chart` og `Normal QQ Plot`. Under `Descriptive Statistics` vælg `Basic Confidence Intervals`, `Tests for Location` (og vælg `Mu0` til 13.35 under `Parameters` svarende til test af hypotesen $\mu = 13.35$). (Bemærk mulighederne for såkaldte *robuste* estimater, hhv for skalaparameter σ og forskellige `Trimmed Winsorized Means`.)

Vælg `Density Estimation` → `Parametric Estimation Normal`. Klik `OK`.

Vælg `Cumulative Distribution` og vælg `Empirical` og `Normal Distribution` som `Fit Parametric` og `Normal Distribution` som `Test Distribution`. Vælg 95 % Konfidensbånd. Klik `OK`.

Klik `OK` i `Output`-menuen

Klik `OK` i `Distribution`-menuen.

Der fremkommer nu et vindue med et histogram med en indtegnet estimeret tæthed. Under histogrammet er der et skema med de parametriske estimater `Parametric density estimation` og en skyder ved henholdsvis `Mean` og ved `Sigma`. Flyt på skyderen og se, hvorledes den tilpassede fordeling ændrer sig.

Gå ned til grafen med den kumulerede fordeling. Her er indtegnet den kumulerede fordeling for den tilpassede normalfordeling samt konfidensbåndet (Kolmogoroff-Smirnoff) bestemt ud fra den observerede empiriske fordeling. Flyt skyderen svarende til konfidenskoefficienten og bemærk hvorledes konfidensbåndet ændrer sig. Gå ned til skemaet med estimaterne for den tilpassede normalfordeling (`Fit Distribution Function`) og flyt på skyderen svarende til `Mean` og `Sigma` og se, hvorledes den kumulerede fordeling ændrer sig. □

Øvelse 16 *Histogram for afstande mellem kundeankomster*

Datasættet `eks4_6` indeholder data fra eksempel 4.6 (og 1.16) i Statistik 1 bogen. Den variable `Tid` angiver tiden mellem to kundeankomster.

Øvelsen forudsætter, at datasættet er åbnet i SAS/INSIGHT.

Klik på `Analyze` → `Distribution (Y)`.

Vælg den variable `Tid` og placer den under `Y`.

Vælg `Output` og under `Graphs` vælg `Histogram/Bar Chart`.

Vælg `Density Estimation` → `Parametric Estimation Exponential`. Klik `OK`.

Vælg `Cumulative Distribution` og vælg `Empirical` og `Exponential Distribution` som `Fit Parametric` og `Exponential Distribution` som `Test Distribution`. Klik `OK`.

Klik `OK` i `Output`-menuen

Klik `OK` i `Distribution`-menuen.

Der fremkommer nu et vindue med et histogram med en indtegnet estimeret tæthed. Under histogrammet er der et skema med de parametriske estimater `Parametric density estimation` og en skyder ved henholdsvis `Mean` og ved `Sigma`. Flyt på skyderen og se, hvorledes den tilpassede fordeling ændrer sig.

□

Øvelse 17 *Analyse af fordelingsform (tæthedsfunktion)*

(Øvelsen forudsætter, at datasættet `Brains` er indlæst i biblioteket `mitbib` og åbnet i SAS/INSIGHT).

Vi vil her undersøge fordelingen af kropsvægte for pattedyrearter. Kropsvægten (i kg) er angivet i den variable `Body`.

Indledningsvis kan man tegne et boxplot og et histogram for `body`. Hvilke af dyrearterne i stikprøven er bestemmende for histogrammets udseende ? Er elefanter mon en helt speciel pattedyreart ?

Tegn derefter et boxplot og histogram for logaritmen til kropsvægten `lbod`. Nu får man en nogenlunde symmetrisk fordeling uden afvigende observationer. Vi vil derfor undersøge om fordelingen kan beskrives ved en normalfordeling.

Klik på `Analyze` → `Distribution (Y)`.

Vælg den variable `LBOD` og placer den under `Y`.

Vælg `Output` → `Histogram/Bar Chart` → `Normal QQ Plot`.
Vælg `Density Estimation` → `Parametric Estimation Normal`. Klik `OK`.

Vælg `Cumulative Distribution` og vælg `Empirical` og `Normal Distribution` som `Fit Parametric` og `Normal Distribution` som `Test Distribution`. Vælg 95 % Konfidensbånd. Klik `OK`.

Klik OK i Output-menuen

Klik OK i Distribution-menuen.

□

5.13 Fit-optionen

Dette valg benyttes ved analyse af en responsvariabel (Y) som funktion af en eller flere forklarende variable (X)

Menuen giver mulighed for at udføre variansanalyser (svarende til at værdierne af de forklarende variable fortolkes som nominale værdier, se afsnit 4.3), regressionsanalyser (svarende til at værdierne af de forklarende variable fortolkes som målt på en intervalskala), eller kombinationer heraf (fx sammenligning af regressionsplaner). Disse modeller kaldes under ét generelle lineære modeller.

Optionen giver endvidere mulighed for at udføre analyser med ikke-lineære sammenhænge ved valg af en ikke-lineær link funktion.

Der er yderligere mulighed for valg af andre fordelinger for responset end normalfordelingen (de såkaldte generaliserede lineære modeller).

Endelig kan man vælge forskellige former for kurvetilpasning (udglatning) af relationen mellem en responsvariabel og én forklarende variabel ved forskellige parametriske og ikke-parametriske metoder.

Ved valg af Fit-optionen fremkommer en menu, der giver mulighed for valg af responsvariabel (Y) og forklarende variable (X).

Vinduet bruges til at formulere den lineære del af modellen. En eventuel anden fordeling end normalfordelingen og en eventuel anden link-funktion end identiteten vælges i Method menuen.

Under Y-rubrikken er der mulighed for at vælge, om modellen skal indeholde et intercept-led eller ej. En markering i Intercept rubrikken indebærer at intercept leddet medtages.

Hvis man allerede har *udvalgt* variable i data-vinduet (som angivet i afsnit 5.1.1), vil disse være placeret som Y- og X-variable med den første som responsvariabel (Y) og de øvrige som forklarende variable (X).

Nederste linie i fit-menuen giver mulighed for valg af Metode, dvs. fordeling, link-funktion, estimationsmetode mv og valg af Output, dvs tabeller, grafer og variable.

Apply knappen bevirker, at analysen udføres, men Fit-vinduet bevares sådan at man kan revidere den undersøgte model.

OK knappen bevirker, at analysen udføres og Fit-vinduet forsvinder.

Output fra proceduren fremkommer i et specielt vindue, der i rammen er betegnet med navnet på datasættet. Hvis der udføres flere successive analyser (ved brug af *Apply* knappen), er der også anført et løbenummer for analysen.

I afsnit 5.13.3 på side 46 beskrives en række af valgmulighederne for output. Her skal blot anføres, at sædvanligvis udskrives *modellen* (modelformlen), et skema med Analysis of Variance (eller, hvis der er valgt en anden responsfordeling end normalfordelingen Analysis of Deviance) og et skema med såkaldte Type III Tests. Skemaet, der er betegnet Analysis of Variance, angiver en meget grov opspaltning af den *totale variation* (C total) i det samlede bidrag fra de *forklarende variable* (Model), og *restvariationen* (Error).

Skemaet med Type III Tests viser for hvert led i modelformlen variationsbidraget svarende til netop dette led, idet alle øvrige led bibeholdes i modellen (dvs effekten af disse led er tilgodeset i de fittede værdier). Det kan undertiden være af interesse også at betragte skemaet med med Type I Tests. Dette skema viser variationsbidraget svarende til de enkelte led i modelformlen ved en successiv fjernelse af leddene (det sidste led fjernes først), dvs effekten af foranstående led er tilgodeset, mens en eventuel effekt af efterstående led ikke er tilgodeset.

Endelig udtegnes forskellige *residualplots*, der bør udnyttes til at vurdere, hvorvidt forudsætningerne for analysen kan antages at være opfyldt.

5.13.1 Modelformler

Afhængigheden af den (de) forklarende variable udtrykkes ved en såkaldt *modelformel*. Notationen er foreslået af Wilkinson og Rogers (G.N.Wilkinson and C.E.Rogers: Symbolic description of factorial models for analysis of variance. *Appl.Statist.* **22**, (1978), pp 392-399.)

Fortolkningen af led i modelformler afhænger af *værdiskalaen* for de forklarende variable. Hvis den variable opfattes som målt på en *nominalskala*, fortolkes den som en *klassifikationsvariabel* (`CLASS` i PROC GLM). Hvis den opfattes som målt på en *intervalskala*, fortolkes den som en *regressionsvariabel*.

Eksempel 3 *Ensidet variansanalyse*

Datasættet `eks5_1` indeholder data fra eksempel 5.1 i Statistik 1 bogen.

Den variable `Gren` angiver produktionslinien og den variable `Kali` angiver udbyttet. Da `Gren` er en tekstvariabel, fortolkes den automatisk som en *klassifikationsvariabel* (*Nominal*) i den interaktive procedure.

Placeres `Kali` som `Y` og `Gren` som `X` og klikkes derefter på `Apply` får man udført en ensidet variansanalyse.

Vælger man i output-menuen `Model Equation`, `Summary of Fit`, `Analysis of Variance/Deviance`, `Type I Tests` og `Parameter Estimates` og derefter klikker på `Apply`, fremkommer et vindue, hvor de ønskede størrelser er angivet i hver sin tabel.

Tabellen med `Summary of Fit` angiver gennemsnittet af `Y`-værdierne (`Mean of Response`), *residualspredningen*, dvs estimatet for σ (`Root MSE`), *forklaringsgraden* R^2 (`R-Square`, *Frihedsgradskorrigeret forklaringsgrad* (`Adjusted R-Sq`).

Tabellen med `Analysis of Variance` viser opspaltningen af den totale variation omkring gennemsnittet af `Y`-værdierne (`C Total`) i de to hovedbidrag, den variation, der kan tilskrives de forklarende variable i modellen (`Model`) og resten, den såkaldte *residualvariation* `Error`). Søjlen `DF` angiver *frihedsgraderne*, `Sum of Squares` angiver kvadratafvigelsessummen (SAK), og `Mean Square` angiver kvadratafvigelsessummen divideret med de tilhørende frihedsgrader.

Tabellen med `Type I Tests` viser tilsvarende opspaltningen af `Model`-variationen i bidragene fra de enkelte led i modelformlen, se forklaringen på side 36.

Endelig viser tabellen med `Parameter Estimates` estimaterne for modellens parametre. Parameterestimaterne skal ses i sammenhæng med tabellen `Parameter Information`, der viser nummereringen af parametrene (som også bruges i tabellen `Model Equation`) og fortolkningen af de enkelte parametre. I tabellen med parameterestimater angives parameterens fortolkning (som under `Parameter Information`), parameterestimatet (`Estimate`), den estimerede *spredning for estimatet* (`Std Error`), t-teststørrelsen for den hypotese, at den underliggende parameter kan antages at være nul (`t Stat`, dvs værdien af estimatet divideret med den estimerede spredning for estimatet), den tilsvarende p-værdi ved et tosidet

test (`Pr > |t|` , dvs $\mathsf{P}[|T| > |t_{obs}|]$), hvor frihedsgraderne i t-fordelingen er de frihedsgrader, der er knyttet til `Error` kvadratafvigelsessummen.

Estimatorerne er sædvanligvis *ikke uafhængige*, hvorfor teststørrelserne heller ikke er uafhængige. Man skal derfor være forsigtig med at fortolke teststørrelserne for flere parameterværdier samtidigt. En sådan samtidig fortolkning foretages bedre ved at samle de pågældende parametre i et enkelt *led i modelformlen*, hvorved deres indbyrdes forskel kan testes under ét. (Man får et udtryk for *kovariansen* eller *korrelationen* mellem estimaterne ved at vælge `Estimated Cov Matrix` eller `Estimated Corr Matrix` i outputmenuen).

Det fremgår af outputtet, at *parametriseringen* af modellen adskiller sig fra den parametrisering, der er angivet i lærebogen. Man har valgt at bruge det højeste niveau (her `D`) af den forklarende variable (`Gren`) som referenceværdi, hvorfor parameterværdien svarende til dette niveau er sat til nul. `Intercept` parameteren angiver således det forventede udbytte for gren `D` svarende til en parametrisering (i lærebogens notation)

$$\mathsf{E}[Y_{ij}] = \mu + \alpha_i$$

med $\alpha_4 = 0$, og hvor μ angiver `Intercept` parameteren, svarende til at man har valgt at udtrykke de *lineære bånd* i modellen ved at vælge niveauet svarende til gren `D` som *referenceniveau* i stedet for lærebogens bånd, $\sum n_i \alpha_i = 0$. □

Eksempel 4 *Fittede værdier og residualer*

Eksemplet forudsætter, at `Fit`-menuen fra eksempel 3 stadig er åben.

Vælger man i output-menuen `Predicted` og `Residual`, får man tilføjet to nye variable med navnene `P_KALI` og `R_KALI` til datasættet. Den variable `P_KALI` indeholder de såkaldte *prædikterede* eller *fittede* værdier, nemlig estimatet for $\mathsf{E}[Y_{ij}]$ under modellen, hvilket i dette tilfælde netop er det gennemsnitlige kaliindhold i stikprøven fra den pågældende produktionsgren. Den variable `R_KALI` indeholder *residualerne* svarende til modellen, dvs $Y_{ij} - \widehat{\mathsf{E}}[Y_{ij}]$. □

Eksempel 5 *Brug af residualer til grafisk modelkontrol*

Eksemplet forudsætter, at `Fit`-menuen fra eksempel 4 stadig er åben.

Output-menuen giver en række muligheder for at foretage grafisk kontrol af modellen. Sædvanligvis udføres modelkontrol ved at betragte *residualerne*, da de jo netop udtrykker variationen i data, når der er korrigeret for den systematiske indflydelse fra de forklarende variable, der er udtrykt ved modellen.

I output-menuen kan man under `Plots` vælge `Residual by Predicted`, der giver mulighed for at vurdere, om variansen ændres med den fittede værdi af responset.

Man kan også benytte andre *scatter plots* fra menubjælken til at optegne residualerne i observationsrækkefølge, eller optegne dem mod de indgående forklarende variable for at vurdere eventuelle systematiske tendenser (ikke-lineariteter mv).

Hvis man mener, at modellen giver en tilfredsstillende beskrivelse af variationen i data, kan man kontrollere normalfordelingsantagelsen ved at vælge `Residual Normal QQ`, hvorved der tegnes et *fraktildiagram* (Q-Q plot) for residualerne. (Samtidig tilføjes en yderligere variabel til datasættet, nemlig en variabel, der angiver normalfordelingsfraktilen svarende til det beregnede residual). □

Eksempel 6 *Tosidet variansanalyse*

Datasættet `eks5_4` indeholder data fra eksempel 5.4 i Statistik 1 bogen.

Den variable `Mark` angiver marken, den variable `Goedn` angiver gødningstypen og den variable `Udbyt` angiver udbyttet. Bemærk, at koden for marken er indlæst som *tekststrenge*, så selv om værdierne er `1, 2, 3, 4`, opfattes de ikke som numeriske (*Interval*), men `Mark` fortolkes som en *klassifikationsvariabel* (*Nominal*) i lighed med `Goedn`.

Først placeres `Udbyt` som `Y` og `Mark` og `Goedn` som `X`. Derefter vælger man i output-menuen `Type I tests`, hvorefter man kan klikke på `Apply`. Der bliver nu udført en tosidet variansanalyse i analogi med lærebogens eksempel. Variationsopspaltningen under `Type I tests` angiver netop opspaltningen af `Model` variationen i bidragene hidrørende fra de enkelte led i modelformlen, hhv `Mark` og `Goedn`.

Parametriseringen svarer til modellen (i lærebogens notation)

$$\mathsf{E}[Y_{ij}] = \mu + \alpha_i + \beta_j \ ,$$

$i = 1, \ldots, 4$ og $j = 1, \ldots, 5$ med $\alpha_4 = 0$ og $\beta_5 = 0$.

Additivitetsantagelsen (forsvindende vekselvirkning) kan vurderes grafisk ved at tegne en *vekselvirkningsgraf*. I `Scatter Plot (Y X)` menuen vælges `Udbyt` som `Y` og fx `Mark` som `X`. Derefter vælges plottesymboler og/eller farver således at hver gødningstype får sit symbol (se øvelse 10 på side 29). Punkterne svarende til to forskellige gødningstyper ses nu at forløbe nogenlunde parallelt. Der er således ingen tydelig afvigelse fra antagelsen om additivitet. □

Eksempel 7 *Regressionsanalyse uden gentagelser*

Datasættet `eks5_7` indeholder data fra eksempel 5.7 i Statistik 1 bogen.

Den variable `Dage` angiver antallet af dage siden justering og den produktionslinien og den variable `Fejlvis` angiver fejlvisningen.

Placeres `Fejlvis` som `Y` og `Dage` som `X` og klikkes derefter på `Apply` får man udført en regressionsanalyse.

Da `Dage` fortolkes som målt på en *intervalskala* opfattes den nemlig som en regressionsvariabel.

Parameterestimaterne i outputtet svarer til en parametrisering (i lærebogens notation)

$$\mathsf{E}[Y_i] = \mu + \beta x_i \ ,$$

hvor μ angiver `Intercept` parameteren (den fittede værdi svarende til $x = 0$). □

"Vekselvirkningsled", Krydsede led

Man krydser to eller flere forklarende variable ved at *vælge dem* i `X` rubrikken og derefter trykke på `Cross` knappen. Hvis man eksempelvis har de forklarende variable `Mark` og `Goedn` og trykker på `Cross` får man leddet `Mark * Goedn` i `X`-rubrikken. Hvis de to variable begge er *klassifikationsvariable*, fortolkes leddet som et sædvanligt vekselvirkningsled i en variansanalysemodel. Hvis den ene er en *klassifikationsvariabel*, og den anden opfattes som målt på en *intervalskala*, fortolkes leddet svarende til en særskilt regressionskoefficient for hvert niveau af klassifikationsvariablen. Hvis endelig begge variable opfattes som målt på en *intervalskala*, fortolkes leddet som en ny regressionsvariabel, hvis værdier fremkommer som produktet af de tilsvarende værdier af de to indgående variable. Tilsvarende fortolkninger gøres ved *krydsning* af flere variable.

Eksempel 8 *Tosidet variansanalyse, test for forsvindende vekselvirkning*

Datasættet `eks5_5` indeholder data fra eksempel 5.5 i Statistik 1 bogen.

Den variable `Bland` angiver blandemaskinens nummer, den variable `Knus` angiver knusemaskinens nummer og den variable `Styrke` angiver cementstyrken. (Maskinernes numre er indlæst som *tekststrenge*, så de fortolkes begge som *klassifikationsvariable*).

Først placeres `Styrke` som Y og `Bland` og `Knus` som X. Derefter *vælges* `Bland` og `Knus` i `X` rubrikken og man klikker på Cross. Nu fremkommer leddet `Bland* Knus` i `X` rubrikken

Derefter vælger man i output-menuen `Type I tests`, hvorefter man kan klikke på `Apply`. Tabellen med `Type I tests` viser nu en variationsopspaltning, der netop har testet for forsvindende vekselvirkning ud for `Bland* Knus` i analogi med lærebogens eksempel.

Parameterestimaterne i outputtet svarer til modellen (i lærebogens notation)

$$\mathsf{E}[Y_{ijk}] = \mu + \alpha_i + \beta_j + \tau_{ij} \ ,$$

$i = 1, \ldots, 3$ og $j = 1, \ldots, 3$ med $\alpha_3 = 0$, $\beta_3 = 0$ og $\tau_{i3} = 0$, $i = 1, 2, 3$ samt $\tau_{3j} = 0$, $j = 2, 3$. (Vi bemærker i lighed med eksempel 3, at man har valgt at udtrykke de *lineære bånd* i modellen ved at vælge *referenceniveauer* for klassifikationerne).

Testet giver ikke anledning til at afvise en hypotese om forsvindende vekselvirkning, hvorfor man kan vælge at inkludere den tilsvarende variation i variansestimatet. Dette gøres ved at vende tilbage til `Fit`-menuen og fjerne `Bland* Knus` fra de forklarende variable (`X`-rubrikken). Nu klikkes på `Apply` og i den analyse, der derved fremkommer indeholder `Error`-leddet nu også variationen svarende til vekselvirkningskvadratafvigelsessummen (kontroller fx ved at betragte de tilhørende frihedsgrader). □

Eksempel 9 *Regressionsanalyse med gentagelser*

Datasættet `eks5_6` indeholder data fra eksempel 5.6 i Statistik 1 bogen.

Den variable `Dag` angiver dagens nummer, den variable `Hast` angiver strømningshastigheden, og den variable `Gent` angiver gentagelsesnummeret. Endelig er der angivet en tekstvariabel, `Dagkv`, der ligeledes angiver dagens nummer, men fortolket som en klassifikationsvariabel (*Nominal*). .

Først placeres `Hast` som Y og `Dag` og `Dagkv` som X (i den angivne rækkefølge). Derefter vælger man i output-menuen `Type I tests`, hvorefter man kan klikke på `Apply`. Tabellen med `Type I tests` viser nu en variationsopspaltning, der netop har testet for linearitet ud for `Dagkv` i analogi med lærebogens eksempel.

Dette test viser ingen grund til at afvise linearitetshypotesen, hvorfor man kan vælge at inkludere den tilsvarende variation i variansestimatet. Dette gøres ved at vende tilbage til `Fit`-menuen og fjerne `Dagkv` fra de forklarende variable (`X`-rubrikken). Nu klikkes på `Apply`, og i den analyse, der derved fremkommer, indeholder `Error`-leddet nu også variationen af gennemsnittene omkring linien.

De parameterestimater, der blev beregnet for modellen `Dag Dagkv`, svarer til modellen (i lærebogens notation)

$$\mathsf{E}[Y_{ij}] = \mu + \beta x_i + \mu_i \ ,$$

$i = 1, \ldots, 10$ og $j = 1, \ldots, 3$ med $\mu_9 = 0$, $\mu_{10} = 0$. Estimaterne svarer til, at `Intercept` parameteren μ og hældningen β er bestemt så linien går igennem gennemsnittet af observationerne fra hhv dag 9 og 10, og parametrene $\mu_1, \ldots, \mu_8$ angiver den størrelse, der skal adderes til det estimerede liniepunkt for den pågældende dag for at få dagens gennemsnitshastighed.

Dette illustrerer, at man skal være varsom med at fortolke individuelle parameterestimater løsrevet fra deres sammenhæng (dvs fra de øvrige led i modellen). Estimatet, β, for liniens hældning er IKKE regressionsmodellens hældningsestimat; estimaterne for parametrene i en given (overparametriseret) model tjener kun til at *fastlægge de fittede værdier*, som jo ikke afhænger af den parametrisering, der benyttes. □

Eksempel 10 *Sammenligning af to regressionslinier*

Datasættet `malere` indeholder samhørende værdier af alder `Alder` og hjerneblodgennemstrømningsindex `Isi` for 20 malere. For hver person er desuden i den variable `Oplosn` angivet, om den pågældende har benyttet organiske opløsningsmidler (værdierne hhv `Opl` og `Ej`). (Kilde: P.Arlien-Søborg et al. *Acta Neurologica Scand.* **66** (1982), pp. 34-41).

Hjerneblodgennemstrømningsindex (ISI) benyttes ofte som udtryk for hjernens funktionsevne. Det er almindeligt anerkendt, at værdien af dette index falder med personens alder.

Man ønsker nu at vurdere, om der kan påvises forskel i ISI for disse to grupper af malere (malere, der har benyttet organiske opløsningsmidler, og malere, der ikke har brugt disse midler). Da de to grupper har forskellig alderssammensætning, er det naturligt at tage hensyn til personernes alder i analysen.

Vi vælger derfor indledningsvist at modellere data ved *to separate regressionslinier*, én for hver af de to grupper.

Vi placerer derfor `Isi` som Y og `Oplosn` og `Alder` (i den anførte rækkefølge) under X. Derefter *vælges* `Oplosn` og `Alder` i `X` rubrikken og man klikker på Cross. Nu fremkommer leddet `Oplosn* Alder` i `X` rubrikken. Derefter trykkes på Apply. Denne model svarer netop til en model med to separate regressionslinier, idet leddet `Oplosn` angiver et konstantled for hver af de grupper (da `Oplosn`

fortolkes som en klassifikationsvariabel, *Nominal*), leddet `Alder` angiver en regressionssammenhæng, og leddet `Oplosn* Alder` angiver, at modellen skal have en separat hældning for hver af klasserne i `Oplosn`.

I ovennævnte modelformel er implicit antaget, at der er samme varians omkring de to linier. Ønsker man først at undersøge en model med separat varians for hver af de to linier, udelades leddene `Oplosn` og `Oplosn* Alder` i `X` rubrikken, og i stedet placeres `Oplosn` som `Group` variabel. Herved udføres to separate regressionsanalyser.

Betragt outputtet fra modellen med leddene `Oplosn`, `Alder` og `Oplosn* Alder`. Parameterestimaterne svarer til modellen (i lærebogens notation)

$$\mathsf{E}[Y_{ij}] = \mu + \alpha_i + \beta x_{ij} + \beta_i x_{ij} \ ,$$

hvor $i = 1, 2$ angiver, om personen tilhører gruppen af ikke-eksponerede (`Oplosn = Ej`, dvs $i = 1$), eller af eksponerede (`Oplosn = Opl`, dvs $i = 2$), og x_{ij} angiver alderen for den j'te person i den i'te gruppe. Overparametriseringen er tilgodeset ved at vælge $\alpha_2 = 0$, og $\beta_2 = 0$ svarende til at gruppen af eksponerede (`Oplosn = Opl`) er valgt som referencegruppe (med intercept mu og hældning β).

Man kan nu foretage et test for parallellitet af de to linier ved at vurdere, om leddet `Oplosn* Alder` kan udelades af modellen, fx ved at betragte p-værdien svarende til dette led under `Type III Tests` (eller, da der kun indgår ét parameterestimat i dette led, får man samme resultat ved at betragte p-værdien svarende til t-størrelsen for denne parameter under `Parameter Estimates`). Man får p-værdien 0.34, og der er således ingen grund til at afvise en hypotese om at de to linier er parallelle (samme hjerneældningshastighed for de to grupper af malere).

Man vælger derfor at simplificere modellen og antage, at de to linier har samme hældning. Man *fjerner* nu leddet `Oplosn* Alder` fra `X` rubrikken og trykker igen på `Apply`. Parameterestimaterne i denne model med leddene `Oplosn` og `Alder` svarer til modellen (i lærebogens notation)

$$\mathsf{E}[Y_{ij}] = \mu + \alpha_i + \beta x_{ij} \ ,$$

hvor $\alpha_2 = 0$, og hvor α_1 angiver den lodrette afstand mellem de to parallelle linier (den alderskorrigerede forskel i hjerneblodgennemstrømningsindex for de to grupper). Man får estimatet $\alpha_1 = 7.26$ med den estimerede spredning for dette estimat, $\widehat{\sigma}_\alpha = 2.08$, og p-værdien svarende til testet (tosidet) for den hypotese, at der ikke er forskel på de to grupper ($\alpha_1 = 0$) er $p = 0.0029$. Forskellen er således signifikant på ethvert niveau større end 0.3%.

Analysen kaldes undertiden en *kovariansanalyse* (eng: *Analysis of Covariance*) fordi man korrigerer for effekten af den *Kovariate*, her alder. □

Underordnede (nestede) klassifikationer

Man frembringer nestede (underordnede) klassifikationsstrukturer ved at *vælge* dem i `X`-rubrikken, og derefter trykkke på `Nest` knappen. Antag, at man har *valgt* de forklarende variable `person` og `gent`, som begge fortolkes som *nominalværdier*, og trykker på `Nest`-knappen. Man får da leddet `gent(pers)`, der er SAS-symbolet for vekselvirkningsleddene `pers*gent`.

Udvikling af led i modelformlen, polynomial regression

Endelig giver `Expand` knappen mulighed for "potensudvikling" af led i modelformlen. Graden (potensen) er angivet i rubrikken lige under `Expand`-knappen. Graden øges ved at trykke på knappen til højre, og mindskes ved tryk på knappen til venstre for gradangivelsen. Antag, at man har valgt graden 2 og at man har valgt de variable `Form` og `Farve`, som er målt på en nominalskala. Et tryk på `Expand` knappen frembringer da leddene
`Form` `Farve` `Form * Farve`

For to variable `lengd` og `bredd`, begge fortolket som målt på en intervalskala, vil en udvikling af graden 2 frembringe leddene
`lengd` `bredd` `lengd * lengd` `lengd * bredd` `bredd * bredd`
hvor fx leddet `lengd * lengd` betyder at værdierne af `lengd`^{2} benyttes som forklarende variable, svarende til en *polynomial regressionsmodel*.

5.13.2 Metode menu

Her vælges *responsfordeling*, *link*-funktion og *estimationsmetode*for de såkaldte generaliserede lineære modeller.

Default værdien er `Response Dist.: Normal`, `Link Function: Canonical`, og `Scale: MLE`. Fit menuen kan således bruges til sædvanlige normalfordelingsmodeller (*den generelle lineære model*, GLM) uden at man behøver vælge metode.

Ideen i de generaliserede lineære modeller er dels, at man vælger at modellere en funktion, $\eta = g(\mu)$ (*linkfunktionen*) af middelværdien, μ, ved en *lineær* funktion, $\boldsymbol{\eta} = \boldsymbol{X\beta}$, af de forklarende variable. Ofte bruges den såkaldte *kanoniske link*, der fører middelværdiparameteren over på hele den reelle akse.

Yderligere tilgodeser man responsfordelingens form (variansens afhængighed af middelværdien) ved - i stedet for at vurdere tilpasningen af modellen til data ved kvadratafvigelsessummer - at måle afvigelsen mellem observerede og fittede værdier ved den såkaldte *devians*(eng.: `Deviance`), der udtrykker forskellen målt som en forskel i logaritmen til likelihoodfunktionens værdi.

Modellerne kan yderligere udvides med en såkaldt *dispersionsparameter*, der er en konstant faktor i variansen.

I metodemenuen kan vælges mellem følgende responsfordelinger

Normalfordeling (Varians afhænger ikke af μ)

Invers Gauss fordeling (Varians proportional med μ^3)

Gammafordeling (Varians proportional med μ^2)
Kan også bruges ved analyse af empiriske varianser for normaltfordelte observationer

Poissonfordeling (Varians proportional med μ)

Binomialfordeling (Varians proportional med $\mu(1-\mu)$)

Der kan desuden vælges mellem følgende linkfunktioner

Den identiske afbildning, $\eta = \mu$ (kanonisk ved normalfordelingen)

Logaritmefunktionen, $\eta = \ln(\mu)$ (kanonisk ved Poissonfordelingen)

Logit-funktionen, $\eta = \ln(\mu/(1-\mu))$ (kanonisk ved Binomialfordelingen)

Probit-funktionen, $\eta = \Phi^{-1}(\mu)$ (bruges ved binomialfordelt respons)

Komplementær log-log, $\eta = \ln(-\ln(\mu))$ (bruges ved binomialfordelt respons)

Potens-funktion, $\eta = \mu^c$, hvor c er en valgt potens. ($c = -1$ er kanonisk link ved Gammafordelingen, $c = -2$ er kanonisk link ved den Inverse Gauss-fordeling)

Vælges binomialfordeling skal Y-variablen enten være en Bernoulli-variabel, dvs med værdierne 0 eller 1, eller en variabel, der tæller antallet af forekomster af en given hændelse i et antal Bernoulli-forsøg. Hvis Y-variablen tæller antal af forekomster, skal den variabel, der angiver antallet af forsøg placeres i feltet `Binomial`.

Såfremt modellen indeholder et offset-led, skal den variable, der angiver offset-værdien, placeres i rubrikken `Offset` (forekommer sædvanligvis ved *Poisson regression*).

Rubrikken `Scale` er relateret til dispersionsparameteren σ^2. For normalfordelingen er skalaen ϕ den sædvanlige kvadratrod af variansen på den enkelte observation; for den inverse Gauss-fordeling er skalaen ϕ ligeledes $\sqrt{\sigma^2}$, og for Gamma-fordelingen er skalaen ϕ den reciprokke formparameter $\phi = 1/\alpha$. For Poisson og Binomialfordelingen er skalaen 1.

For normal, invers Gauss og Gamma-fordelingen kan skalaen estimeres ved hhv. `MLE` (Maximum-likelihood metoden), `Deviance` (devianstesteststørrelsen divideret med frihedsgraderne), `Pearson` (Pearson teststørrelsen divideret med frihedsgraderne), eller `Constant` (den værdi, der er angivet i rubrikken `Constant`).

Såfremt man vil analysere en model svarende til binomialfordeling eller Poissonfordeling med overdispersion, vælger man `Quasi-Likelihood` i estimationsmetode. Den estimerede skalaparameter er da $\sigma = \sqrt{\mathtt{V}[Y]/V(\mu)}$.

5.13.3 Output menu

Output menuen giver mulighed for valg af

tabeller over deviansopspaltning, type I og type III tests, estimater, konfidensintervaller, varians-kovariansmatricer for estimater

Residual Plots Plot af residual mod fittet værdi, fraktildiagram for residualer, plot af partiel leverage

Fit kurver For numeriske forklarende variable Forskellige parametriske og ikke-parametriske udglatninger, herunder tredimensionale plots af de tilpassede flader med farvelægning af fladen efter værdier af Y-variablen.
Hvis man har flere end to forklarende variable, tegnes fladen som funktion af de to første variable, og de øvrige variable er fastholdt på deres gennemsnitsværdi.

Hvis man kun har én (numerisk) forklarende variabel, er der mulighed for at lave en polynomial regression

Output variable En række forskellige mål til brug for vurdering af modeltilpasning og influentielle observationer, bl.a. dffits, dfbetas, prædikteret (fittet) værdi ($\widehat{\mu}$), lineær prædiktor ($\widehat{\eta}$), samt en række standardiserede og studentiserede residualer.

Eksempel 11 *Logistisk regression*

(Eksemplet forudsætter, at datasættet `Fodsl` er indlæst i biblioteket `mitbib` og åbnet i SAS/INSIGHT).

Klik på `Analyze` og vælg `Fit (Y X)`.

Vælg `DOD` som `Y`-variabel og vælg `AAR` som `X`-variabel.

Vælg `METHOD`-vindue og vælg `Binomial` som `Response Distribution` og `Logit` som `Link function`. Vælg den variable FODSL og placer den under `Binomial` (nemlig antalsparameteren i binomialfordelingen). Klik på `OK`.

Vælg `Output`-menuen. Blandt `Tables` vælges `Model Equation`, `Summary of Fit`, `Analysis of Variance`, `Type I Tests`, `Type III (LR) Tests`, `Parameter Estimates`, `95% C.I. (LR) for Parameters`. Blandt `Residual Plots` vælges `Residual by Predicted`. Blandt `Output Variables` vælges `Predicted` og `Standardized Deviance Residual`.

I `Fit`-menuen vælges nu `Apply`.

Bemærk, at der tilføjes nogle variable `R_DOD`, `P_DOD` og `RDS_DOD`, nemlig residualer, de prædikterede (fittede) værdier, samt de standardiserede deviansresidualer.

Gennemgå udskriften i det vindue, der blev genereret af `Fit`-analysen. Vurder modeltilpasningen. (99 % fraktilen i $\chi^2(26)$-fordelingen er 45.64).

Bemærk, at bjælken øverst i vinduet giver mulighed for at vælge yderligere grafer og variable (`Graphs` og `Vars`).

Betragt residualplottet. Vælg nu optionen `Analyze` i et af vinduerne, og vælg `Histogram`-optionen og tegn et histogram over værdierne af den variable `CIVST`. Dette histogram viser blot to lige store blokke svarende til de gifte og de ugifte. Klik på blokken med de gifte. Herved udvælges observationerne svarende til de gifte i alle vinduerne. Læg mærke til hvor residualerne svarende til gifte mødres fødsler er på residualet af `Fit`-analysen.

Prøv tilsvarende at udvælge de ugifte mødre og vurder placeringen af residualerne svarende til disse observationer.

Der er en klar systematik i residualerne. Residualerne svarende til de gifte mødre er gennemgående negative, mens residualerne svarende til de ugifte mødre gennemgående er positive.

Vi bør derfor inddrage `CIVST` i modellen.

Gå tilbage til `Fit`-menuen og placer `CIVST` i `X`-listen. Vælg både `AAR` og `CIVST` i `X`-listen og klik på `Cross`. Herved frembringes et led `AAR*CIVST` i `X`-listen.

Klik nu igen på `Apply`. Der blev nu frembragt et nyt output-vindue fra `Fit`-menuen, `Fit2.WORK.FODSL` og endvidere blev der frembragt nogle flere variable, `R_DOD_2`, `P_DOD_2` og `RDS_DOD_2`, nemlig residualer, prædikterede (fittede) værdier og standardiserede deviansresidualer svarende til den nye model.

Vurder modeltilpasningen for denne model. (95 % fraktilen i $\chi^2(24)$-fordelingen er 36.42).

Hvad udtrykker vekselvirkningsleddet ? Hvorfor er der ikke noget estimat for `CIVST ugift` eller `AAR*CIVST ugift` ? (Se evt. eksempel 10).

Vurder om man kan fjerne vekselvirkningen `AAR*CIVST` og beregn estimaterne under denne model. Hvad udtrykker estimatet svarende til `CIVST gift` ? Kan man fjerne leddet svarende til `CIVST` ?

Forsøg at verificere beregningen af en af de fittede værdier ved at indsætte parameterestimaterne i modelligningen. □

Eksempel 12 *Behandling af Nominal- og Intervalvariable*

(Eksemplet forudsætter, at datasættet `Fodsl` er indlæst i biblioteket `mitbib` og åbnet i SAS/INSIGHT).

Klik på `Analyze` og vælg `Fit (Y X)`.

Vælg `DOD` som `Y`-variabel og vælg `AAR` og `STATUS` som `X`-variable.

Vælg `METHOD`-vindue og vælg `Binomial` som `Response Distribution` og `Logit` som `Link function`. Vælg den variable FODSL og placer den som antalsparameter under `Binomial`. Klik på `OK`.

Vælg `Output`-menuen. Blandt `Tables` vælges `Parameter Estimates`.

I `Fit`-menuen vælges `Apply`.

Sammenlign parameterestimaterne for koefficienten til `STATUS` med estimaterne for koefficienten til `CIVST` i eksempel 11.

Grunden til at de er forskellige er, at værdierne af den variable `STATUS` opfattes som målt på en intervalskala, og derfor estimeres en regressionskoefficient svarende til denne variable. Der kan rettes på det ved at gå tilbage til `Data`-vinduet og ændre på fortolkningen af `STATUS` som i eksempel 1. Prøv det !

□

Eksempel 13 ***Animering af plots***

(Øvelsen forudsætter, at datasættet `FODSL`, aktiveret i øvelse 1 stadig er aktivt. Endvidere forudsættes at outputvinduerne fra øvelse 11 stadig er åbne.)

Klik på datavinduet og vælg de variable `R_DOD` og `P_DOD` (se afsnit 5.1.1). Klik derefter på `Edit` optionen i øverste linie i et af de vinduer, der er tilknyttet den interaktive session. Vælg `Windows` og under `Windows` vælges `Animate`. Der åbnes nu et `Animate`-vindue. I Animate-vinduet vælges nu den variable `AAR`. Flyt skyderen nederst i vinduet hen i en midterstilling og tryk derefter på `Apply`. Læg mærke til, at punkterne på residualplottet bliver fremhævet i tidsrækkefølge (årvis). Markeringerne flytter sig fra højre mod venstre i billedet.

Denne option er måske ikke så interessant i dette eksempel, hvor vi jo godt ved at de prædikterede værdier falder med stigende årstal. I andre situationer kan det imidlertid være af interesse at undersøge afhængigheden af en underliggende variabel ved en sådan animering.

□

5.14 Multivariate-optionen

Optionen giver mulighed for bivariate plots, bestemmelse af principale komponenter, diskriminantanalyse mv.

6 SAS-programmer opbygget af SAS-procedurer

6.1 Afvikling som baggrundsopgave

Et SAS-program, som ligger i en fil, kan afvikles som en baggrundsopgave. Antag, at programmet ligger i filen opgave.sas. Programmet eksekveres ved Unix-kommandoen

```
sas opgave
```


Det vil ofte være fordelagtigt at bevare sit vindue aktivt under programeksekveringen. Dette gøres ved at tilføje en *ampersand* "&" efter kommandoen, sådan at kommandoen bliver `sas opgave &` .

6.2 log-fil og print-fil

Ved eksekveringen dannes en log-fil med navnet opgave.log med meddelelser fra SAS-systemet, herunder fejlmeddelelser. Desuden dannes en print fil med navnet `opgave.lst`, med eventuelle udskrifter fra programmet.

6.3 Afvikling i interaktiv SAS-session fra programvindue

En interaktiv SAS-session påbegyndes ved kommandoen

```
sas
```

(Også her kan det være en fordel at bevare sit vindue aktivt ved at tilføje tegnet "&" efter kommandoen).

Kommandoen bevirker at der fremkommer tre vinduer

1. outputvindue
2. log-vindue
3. program editor-vindue

svarende til de tre filer, programfil, log-fil og lst-fil, der genereres ved afvikling som baggrundsopgave.

Et vindue gøres aktivt ved at klikke med venstre musetast på det pågældende vindue.

Øverst i hvert vindue er der en linie med forskellige valgmuligheder:
`File` `Edit` `View` `Locals` `Globals` `Help`
Når vinduet er aktivt fremkommer der en bjælke (ToolBox) med ikoner, svarende til det pågældende vindue. Hvis man flytter muse-pilen hen til en ikon, fremkommer en kort forklaring af betydningen af den pågældende ikon. For de tre basisvinduer (output- log- og program-vindue) er værktøjerne i ToolBoxen (`submit`, `Open`, `Save`, `Print`, `Print Preview`, `Cut`, `Copy`, `Paste`, `Undo`, `Librefs`, `SAS/ASSIST` og `Help`). Alle disse standardværktøjer kan ikke nødvendigvis bruges i ethvert vindue. Eksempelvis kan indholdet af log-vinduet ikke submittes.

I venstre side af ToolBoxen er der en såkaldt kommandolinie med plads til at skrive små kommandoer til det pågældende vindue, fx `clear`, der sletter al tekst i vinduet.

Når et SAS-program indlæses eller indtastes i programvinduet, fortolkes det under indtastningen, og de forskellige elementer i programteksten skrives med forskelligtfarvet tekst. De enkelte trin i programmet (data-trin, procedurer) er adskilt ved en vandret linie.

I venstre side af programeditoren markeres begyndelsen af et nyt trin ved symbolet `-`. Hvis man klikker på dette symbol, komprimeres visningen af det pågældende programtrin til en angivelse af den første linie, og markeringen af programtrinet i venstre side skifter til symbolet `+`.

Øvelse 18 ***Indlæsning af data med brugtbilpriser***

I biblioteket *kursnavn*/programmer ligger programfilen `brugtbil.sas`.

Indlæs filen i programvinduet.

□

6.3.1 Editering i programvinduet

Der er en række forskellige muligheder for at ændre teksten i programvinduet. Hvilke af disse muligheder, der er aktive (herunder funktionen af `Insert/Delete` tasterne), afhænger af den specifike opsætning.

Opsætningen kan ændres ved at gå ind i `Tools` → `Options` → `Preferences` under fanebladet `Edit`, hvor man bl.a. kan vælge, om man vil *overskrive* tekst i editoren.

Den udvidede editor (`Enhanced Editor`) foretager en *fortolkning* af teksten i editor-vinduet, og bruger bl.a. forskellige *farvekoder* for de forskellige elementer i de indtastede SAS-programmer. Opsætningen af fortolkeren i den udvidede editor kan ændres ved at gå ind i `Tools` → `Options` → `Enhanced Editor`. Betydningen af tastesekvenser kan ændres ved at gå ind i `Tools` → `Options` → `Enhanced Editor Keys`

6.3.2 Eksekvering af program

Hvis man allerede har en fil med programmet, kaldes denne fil ind i programvinduet ved at klikke på ikonen svarende til åbning af en fil (en åben filmappe), hvorefter der fremkommer et `Open ...` vindue, hvor man kan vælge den ønskede fil ved at dobbeltklikke på filnavnet med venstre musetast. (Man kunne også vælge optionen `File` i bjælken øverst i programvinduet.)

Når man ønsker at *eksekvere* et program i programeditoren, klikker man på ikonen svarende til `Submit` (en løbende person) i værktøjsbjælken). Programmet forsvinder fra editoren, programloggen ruller frem i log-vinduet, og eventuelt output kommer i output-vinduet. Ved eksekvering af procedurer med grafisk output fremkommer der desuden et grafisk output-vindue med graferne.

Man kunne også klikke på `Run` i menubjælken og vælge kommandoen `Submit`, eller man kunne højreklikke i programeditoren (evt. efter at have *valgt* (selected) en del af programmet) og derefter vælge kommandoen `Submit` eller `Submit Selection`)

Det er en fordel, at afslutte sit program med sætningen
`RUN ;` Dette bevirker nemlig at programudførelsen stopper efter eksekveringen af den sidste programstump. I modsat fald kan man risikere, at programmet "bliver hængende" i den sidste programstump uden at blive afsluttet.

Eksempel 14 ***Eksekvering af program i programvinduet***

Programfilen `brugtbil.sas`, der blev indlæst i øvelse 18 indeholder et SAS-program, der indlæser data fra slutningen af filen, og placerer dem i et datasæt med navnet `brugtbil`.

Programmet indlæser de variable fabrikat `fabr`, model `model`, type `type`, årgang `aar`, kilometerstand `km` og pris `pris` for en række brugte biler af fabrikaterne VW (Golf) og Ford (Escort). Programmet beregner de variable `alder` og `lpris` (logaritmen til prisen).

□

Det sidst eksekverede program kan kaldes frem igen i programeditoren ved at klikke på `Locals` i programvinduet og vælge `Recall text`.

Et program, der befinder sig i programvinduet kan lagres i det almindelige fil-system ved at klikke på ikonen svarende til `File` i programvinduet, hvorefter der fremkommer et `Save As ...` vindue, hvor man kan vælge filnavn til programfilen i unix-systemet. (Man kunne også vælge optionen `File` i bjælken øverst i programvinduet og klikke på `Save` (programmet lagres under sit gamle navn) eller `Save as` (programmet lagres under et andet navn)).

6.4 Data-trin

Data-trinet (eng: *Data Step*) er én af de fundamentale konstruktioner til oprettelse og transformation af datasæt i SAS-systemet. Eksempelvis foregår import af data til et datasæt i princippet i et data-trin.

I *SAS-programmer* indledes et data-trin med nøgleordet `DATA` efterfulgt af eventuelle *options* og et semikolon.

Data-trinet består af en række *programsætninger* (eng: program statements), hvert efterfulgt af et semikolon.

Data-trinet *slutter* med nøgleordet `RUN`, eller når der begynder en *procedure* (en `PROC`-sætning), eller når der påbegyndes et nyt data-trin (en `DATA`-sætning). I program-editoren i PC-versionen markeres afslutningen på et data-trin automatisk ved en stiplet vandret linie i editor-vinduet.

Hovedprincippet i et data-trin er, at de eksekverbare sætninger i datatrinet udføres i den anførte rækkefølge. Sætningerne udføres på den aktuelle observation i datasættet, og når man har udført den sidste sætning i datatrinet, begyndes forfra med første sætning og den næste observation.

Eksempel 15 ***Indlæsning af data direkte fra programvinduet***

Betragt programmet

```
DATA godning;
input gren $ udbyt ;
ludb = LOG(udbyt);
DATALINES;
A  11.9
A  9.0
```



```
B 15.9
B 17.2
;
PROC PRINT;
```

Programsætningen `input gren $ udbyt ;` bevirker, at der indlæses en observation fra input-mediet. Der indlæses værdien af to variable, `gren`, som er en tekstvariabel (angivet ved tegnet $) og den variable `udbyt`, som er en almindelig numerisk variabel.

I sætningen `ludb = LOG(udbyt) ;` beregnes værdien af en ny variabel, `ludb`, der bestemmes som logaritmen (den naturlige) til den aktuelle værdi af `udbyt`.

Sætningen `DATALINES ;` angiver, at her følger strømmen af observationer, der skal indlæses, én på hver linie. Strømmen afsluttes af linien med semikolon.

Data-trinet er afsluttet med `DATALINES`-sætningen

Den næste sætning er en `PROC`-sætning, og hermed er data-trinet afsluttet.

Data-trinet gennemløbes ialt fire gange, svarende til de fire observationer, der følger `CARDS` sætningen.

Såfremt de fire observationer havde været placeret i en fil i computerens filsystem, kunne man i stedet for `DATALINES`-sætningen have brugt sætningen

```
INFILE '/sti/navn' ;
```

hvor `/sti/navn` angiver sti og navn på filen i computerens filsystem.

`INFILE`-sætningen kan desuden bruges på formen

```
INFILE sasnavn ;
```

hvor `sasnavn` angiver filens navn i SAS-systemet (*File Shortcut Name*, oprettet i `Explorer`-vinduet). Bemærk, at filreferencer i computerens filsystemer angives i citationstegn (' '), mens referencer til SAS-systemets filnavne angives direkte ved det pågældende navn. □

Eksempel 16 ***Frembringelse af datasæt fra eksisterende datasæt***

Sætningen

```
SET dsnavn
```

hvor `dsnavn` angiver navnet på et eksisterende datasæt bevirker, at der indlæses en observation fra det pågældende datasæt.

Antag, at datasættet `godning`, der blev indlæst i det foregående eksempel, stadig befinder sig i biblioteket `WORK`.

Programmet

```
DATA revid;
SET godning;
kvrudb = SQRT(udbyt);
RUN;
```

frembringer et nyt datasæt med navnet `revid` med de samme observationer og variable, som datasættet `godning`, og med den yderligere variable `kvrodudb`, hvis værdi er kvadratroden af værdien af den variable `udbyt`. □

Eksempel 17 ***Brug af resultater fra PROC SUMMARY***

Antag, at datasættet `godning`, der blev brugt i det foregående eksempel, stadig befinder sig i biblioteket `WORK`.

Programmet

```
PROC SUMMARY DATA=godning;
VAR udbyt ludbyt;
OUTPUT OUT= result MEAN  =  snitudb snitludb
  VAR = varudb varludb
  N= antudb antludb ;
RUN;
DATA ttest;
SET result;
z = (snitudb - 10)*SQRT(antudb)/SQRT(varudb) ;
PROC PRINT;
VAR antudb snitudb varudb zudb;
RUN;
```

frembringer først et datasæt `result` (ved proceduren `SUMMARY`) med én observation, der indeholder værdier af gennemsnit, empirisk varians og antal observationer for hver af de to variable `udbyt` og `ludb`.

Sætningen `SET result` i det følgende data-trin indlæser én observation ad gangen fra dette datasæt (her kun én observation) og frembringer den nye variable `z`, der udtrykker teststørrelsen for den hypotese, at middeludbyttet kan være 10.

Se desuden eksempel 39 på side 84 for bestemmelse af konfidensinterval for middelværdien og eksempel 41 på side 85 for p-værdi, kritisk værdi, samt styrkefunktion for dette test. □

DATA-trinet tillader en righoldig mængde af operationer på datasæt og de variable i datasættet.

Således giver `SET`-sætningen mulighed for indlæsning af *flere datasæt*. Endvidere kan man *sammenflette datasæt* ved brug af en `MERGE`-sætning. Det anbefales at konsultere hjælp-funktionen og programeksemplerne før brugen af disse sætninger.

I afsnit 8 gives endvidere eksempler på brug af DATA-trinet til beregning af sandsynligheder og fraktiler i sædvanlige fordelinger.

7 SAS-Procedurer

Dette afsnit giver en summarisk oversigt over procedurerne i SAS-systemet.

Man kan få mere at vide om procedurerne ved at benytte hjælp-funktionen som beskrevet i afsnit 1.2, dvs klikke på `Help` og derefter tx vælge `Help on SAS Software Products`. I oversigten over procedurer omfattet af det relevante produkt, kan man nu "åbne" den ønskede procedure, hvorved man har mulighed for at læse en introduktion, samt få beskrevet syntaksen for den pågældende procedure.

Det kan desuden ofte være en støtte, at se på et af de programeksempler, som kan fremkaldes gennem hjælp-funktionen. Der findes *programeksempler* til største delen af procedurerne. Vejledning findes i hjælp funktionen under `Contents`, `Sample SAS Programs ...`.

7.1 Procedurer i SAS-Base

Basisproduktet SAS-Base indeholder en række procedurer til håndtering af datasæt og til frembringelse af simple summariske beskrivelser af fordelingen af værdierne af de variable i et datasæt.

Hjælp vedrørende procedurerne findes i hjælp, `Contents` under `Help on SAS Software Products`, `Base SAS Software`, `Using Base SAS Software`, `Working with the SAS language`, `SAS Procedures`.

SAS-base består af følgende procedurer

7.1.1 Til simple statistiske størrelser

- CHART, frembringer simple søjle- og lagkagediagrammer. Hvis der ønskes bedre grafisk kvalitet, bruges dog proceduren `GCHART` i SAS/GRAPH
- CORR, bestemmer diverse korrelationskoefficienter mellem variable
- FREQ Tabellerer (og krydstabellerer) variable i et datasæt
- MEANS, frembringer deskriptive stikprøveresultater (momenter mv) for enkelte variable i et datasæt. Resultatet kommer som sædvanligt `output`, men der er også mulighed for at placere resultatet i et datasæt. (Se også `SUMMARY` og `UNIVARIATE` procedurerne)
- PLOT, frembringer grafer ("linieskriver"-plots). Se også `GPLOT` proceduren i SAS/GRAPH, der frembringer grafer i tidssvarende kvalitet.
- PRINT, udskriver indholdet af et datasæt
- RANK, frembringer et nyt datasæt, der indeholder *rangen* for observationerne i et datasæt.
- SORT, sorterer observationerne i et datasæt efter stigende eller faldende værdier af en eller flere variable
- STANDARD, frembringer et nyt datasæt med de standardiserede værdier af de variable
- SUMMARY, frembringer et nyt datasæt med deskriptive stikprøveresultater (momenter mv) for enkelte variable i input-datasættet. I modsætning til `MEANS` og `UNIVARIATE` proceduren, frembringes ikke nogen sædvanlig udskrift.
- TABULATE, frembringer deskriptive stikprøvestørrelser og udskriver i brugerdefineret tabellarisk form
- TIMEPLOT, plotter én eller flere variable i tidsmæssig rækkefølge
- UNIVARIATE, frembringer deskriptive stikprøveresultater (momenter mv) for enkelte variable i et datasæt. (Se også `MEANS` og `SUMMARY`.)

7.1.2 Til dataadministration

- APPEND, tilføjer ét datasæt til et andet
- CALENDAR, viser data på kalenderform
- CATALOG, håndterer SAS-kataloger
- CIMPORT, importerer en *transportfil*, der er fremstillet af `CPORT`-proceduren.
- COMPARE, sammenligner indholdet af to eller flere datasæt
- CONTENTS, udskriver beskrivelse af indholdet i SAS-biblioteker
- COPY, kopierer et datasæt
- CPORT, frembringer en *transportfil* med et datasæt
- DATASETS, administrerer SAS-biblioteker
- DBCSTAB, frembringer oversættelsestabeller for tegnsæt
- DISPLAY, giver mulighed for at udføre SAS/AF anvendelser
- EXPLODE, giver mulighed for udskrivning af tekst med symboler i overstørrelse
- EXPORT, eksporterer datasæt til andre databaseformater
- FORMAT, giver mulighed for at definere udskrivnings- og indlæsningsformats
- FORMS, frembringer labels (etiketter) og andre formularer
- IMPORT, indlæser data fra andre databaseformater eller filer
- MDDB, læser og resumerer et datasæt og placerer resumeet i en *multidimensional database* (MDDB).
- OPTIONS, udskriver de aktuelle værdier af *system-options*
- PMENU, definerer *menuer*, der kan bruges i brugerdefinerede anvendelser

- PRINTTO, definerer en outputdestination for procedure- eller log-output
- REPORT, et værktøj til fremstilling af *rapporter* med simple analyser fra `MEANS` og `TABULATE`
- SQL, opsøger data i et datasæt ved brug af det såkaldte *Structured Query Language*
- TRANSPOSE, restrukturerer et datasæt ved at gøre udvalgte variable til observationer (*transponere*)
- TRANTAB, frembringer oversættelsestabeller, der kan bruges til omsætning fra én repræsentation til en anden (fx fra EBCDIC til ASCII)

7.2 Procedurer i SAS/STAT

SAS/STAT systemet består af en række procedurer til statistisk analyse. Brugen af de statistiske procedurer er beskrevet i manualerne SAS/Stat User's Guide, Version 8, Volumes 1, 2 og 3.

Hjælp vedrørende procedurerne findes i hjælp, `Contents` under `Help on SAS Software Products`, `SAS/STAT`, hvorefter man vælger den pågældende procedure.

7.2.1 Procedurer for lineære normalfordelingsmodeller

Den mest generelle procedure til brug for analyse af lineære normalfordelingsmodeller er GLM-proceduren. Proceduren kan bruges til analyse af den såkaldte generelle lineære model, inklusive multivariate analyser.

Der er herudover en række procedurer, der retter sig mod specielle varianter af den generelle lineære model:

- ANOVA, udfører variansanalyse af data fra balancerede forsøg, herunder bl.a. Bartlett's test for varianshomogenitet
- GLM, analyserer den såkaldte *generelle lineære model* for normalfordelte observationer, dvs en model, hvor middelværdien modelleres som en lineær funktion af en række forklarende variable (såvel kvalitative som kvantitative). Se også `Fit`-menuen i SAS/INSIGHT (afsnit 5.13).
  Såfremt modellen er specificeret som en ensidet variansanalysemodel, kan man vælge at få foretaget Bartlett's test for varianshomogenitet
- GLMMOD, konstruerer designmatricen for en generel lineær model (GLM). (Se også TRANSREG proceduren).
- LATTICE, udfører variansanalyser og kovariansanalyser for data fra et forsøg med *gitterstruktur*
- MIXED, analyserer lineære normalfordelingsmodeller med særlig kovariansstruktur, såkaldte *Mixed* lineære modeller, herunder varianskomponentmodeller.
- MULTTEST, foretager multiple sammenligninger under hensyntagen til de særlige forhold ved multiple tests (den forøgede risiko for "falske" signifikanser). Proceduren giver blandt andet mulighed for at justere p-værdierne ved *Bootstrap resampling*.
- NESTED, analyserer data fra forsøg med hierarkisk (*nested*) struktur og tilfældige effekter
- ORTHOREG, udfører *ortogonal regression* uden beregning af $\mathbf{X}^{\mathrm{T}}\mathbf{X}$-matricen, herunder regression efter *ortogonale polynomier*.Giver bedre estimater, end GLM og REG for ill-conditioned data
- PLAN, konstruerer forsøgsdesign og randomiserede forsøgsplaner for faktorielle forsøg
- PLS, tilpasser prædiktionsmodeller (kalibreringsrelationer) ved lineære prædiktorer, herunder blandt andet ved *Partial Least Squares*-regressionsmodeller.
- REG, udfører *regressionsanalyse*. Proceduren tillader forskellige metoder for selektion af variable og producerer en række diagnostiske størrelser og plots.
- RSREG, estimerer kvadratiske responsflader.
- TTEST, foretager *t-test* i en- og tostikprøvesituationer samt i situationer med parrede observationer
- VARCOMP, analyserer den generelle lineære model med *tilfældige effekter* (varianskomponenter) (Se også MIXED proceduren)

7.2.2 Generaliserede lineære modeller

Den mest generelle procedure til brug for analyse af generaliserede lineære modeller er GENMOD-proceduren. Proceduren tillader brug af en brugerdefineret link-funktion og quasi-likelihood med brugerdefineret varians/devians. Proceduren er analog til `Fit`-menuen i SAS-Insight (afsnit 5.13).

Der er herudover en række procedurer, der retter sig mod specielle varianter af generaliserede lineære modeller:

- CATMOD, analyser modeller for kategoriske data, herunder kontingenstabeller, logistisk regression mv
- LOGISTIC, foretager logistisk regression af binære responser. Se også `Fit`-menuen i SAS-insight (afsnit 5.13).
- NLMIXED, estimation i *ikke-lineære* modeller med en blanding af systematiske og tilfældige led, herunder også modeller med anden responsfordeling, end normalfordelingen (som fx i GENMOD eller i `Fit`-menuen i SAS/INSIGHT)
- PROBIT, foretager analyse af binære og øvrige kvalitative responser som funktion af forklarende variable. Proceduren giver mulighed for valg mellem flere forskellige responsfunktioner, herunder *probit*, *logistisk*, extremværdi (gombit).

7.2.3 Analyser af multivariate data

- ACECLUS Approximate Covariance Estimation for Clustering. Bruges til at præprocesse data til analyse ved `CLUSTER` eller `FASTCLUS` procedurerne
- CALIS, modellerer såkaldt *strukturelle modeller* , dvs modeller, der modellerer relation mellem observerede variable og hypotetiske ikke-observerede *latente variable*, herunder de såkaldte *LISREL*-modeller.
- CANCORR, bestemmer kanoniske korrelationer, herunder også partielle kanoniske korrelationer
- CANDISC, foretager *Kanonisk diskriminantanalyse*

- CLUSTER, foretager en hierarkisk gruppering (*clusteranalyse*) af data i *clusters*, grupperet efter værdierne af en variabel. (Se også MODECLUS og FASTCLUS)
- CORRESP, udfører en såkaldt *Korrespondance Analyse* af faktorstrukturen i en antalstabel.
- DISCRIM, udfører en *diskriminantanalyse*, dvs bestemmer et kriterium, der kan bruges til at adskille forelagte grupper af data.
- FACTOR, udfører en *faktoranalyse*, dvs bestemmer ortogonale linearkombinationer af de variable i et multivariat datasæt.
- FASTCLUS, foretager en gruppering (*clusteranalyse*) af data i *clusters*, grupperet efter værdierne af en variabel. I modsætning til CLUSTER er grupperingen ikke hierarkisk. (Se også MODECLUS).
- INBREED, beregner koefficienter til kovarianser i et *stamtræ*
- MDS, foretager såkaldt *Multidimensional Scaling*, dvs estimerer koordinaterne i et rum af begrænset dimensionalitet for et mangedimensionalt datasæt, der er givet som en eller flere *similaritetsmatricer*
- MODECLUS, foretager en *clusteranalyse* af et datasæt baseret på en ikke-parametrisk estimation af tætheden. Se også CLUSTER og FASTCLUS.
- PRINCOMP, bestemmer principale komponenter i multivariate datasæt, dvs indlægger observationerne i et ortogonalt koordinatsystem, hvis retninger er bestemt af kovariansstrukturen i data. Se også `Multivariate`-optionen i SAS-insight (afsnit 5.14).
- PRINQUAL, bestemmer " principale komponenter" for kvalitative variable. Proceduren bestemmer lineære og ikke-lineære transformationer af variable, der giver ønskede egenskaber for kovarians- eller korrelationsmatricen for de transformerede variabel
- SCORE, proceduren multiplicerer værdier fra to datasæt, hvor det ene indeholder koefficienter og det andet indeholder de data, hvis *score-værdier* skal bestemmes ved brug af koefficienterne fra det første datasæt

- STEPDISC, udfører *trinvis diskriminantanalyse* med henblik på at bestemme den bedste delmængde af de variable til brug for at diskriminere imellem klasser i datasættet
- TREE, frembringer et *trædiagram* (dendrogram eller phenogram) med udgangspunkt i et datasæt, frembragt af CLUSTER eller VARCLUS proceduren
- VARCLUS, foretager en opdeling af et datasæt i disjunkte eller hierarkiske grupper (*clusters*) på basis af den første principale komponent

7.2.4 Tidsrækkeanalyse og analyse af spatielle data

Nedennævnte procedurer findes i SAS/STAT. SAS/ETS (se afsnit 7.3) indeholder yderligere procedurer til analyse af tidsrækker.

- KRIGE2D, foretager en todimensional *kriging*, dvs tilpasning af en responsflade til spatielt korrelerede data
- SIM2D, foretager simulation af et *random field* med specificeret middelærdi- og kovariansstruktur
- VARIOGRAM, beregner såkaldte *semivariogrammer* til beskrivelse af den spatielle sammenhæng i todimensionale spatielle data

7.2.5 Analyse af levetidsdata

Udover nedennævnte procedurer i SAS/STAT indeholder SAS/QC (se afsnit 7.6) procedurerne CAPABILITY og RELIABILITY, der også kan bruges til analyse af levetidsdata.

- LIFEREG, estimerer parametre i regressionsmodeller for levetidsobservationer, dvs positive (evt censurerede) observationer. Der er mulighed for en række forskellige parametriske fordelinger af responsvariablen (levetiden), ekstremværdi, normal, logistisk, exponentiel, Weibull, lognormal, loglogistisk og gammafordeling
- LIFETEST, foretager sammenligninger mellem levetidsfordelinger for forskellige *strata*. Der tillades ikke-parametriske estimater af den underliggende fordeling.

- PHREG, analyserer levetidsmodeller ved såkaldte *proportional hazards* regressionsmodeller, dvs hvor hazard-funktionen er proportional med funktioner af de forklarende variable. I modsætning til LIFEREG gøres ingen antagelse om den parametriske form for levetidsfordelingen

7.2.6 Forskellige ikke-parametriske og robuste metoder

- GAM, analyserer såkaldt *generaliserede additive modeller* (forsøgsvist indlagt i version 8.1)
- KDE, udfører *kernel estimation*, af en en- eller todimensional tæthedsfunktion.
- LOESS, foretager en såkaldt ikke-parametrisk estimation af en responsflade
- NLIN, estimation i *ikke-lineære* (nonlinear) modeller
- NPAR1WAY, foretager ikke-parametriske tests for positions- og skalaforskelle i en envejsklassifikation
- STDIZE, standardiserer observationer i et datasæt ved brug af *robuste estimatorer* for position og skala (fx Huber's estimat, Tukey's bivægt estimat og Andrew's wave estimat)
- TPSPLINE, bestemmer *thin-plate smoothing splines* til tilpasning af glatte responsflader
- TRANSREG, bestemmer *transformationer af variable* (inklusiv spline og andre ikke-lineære transformationer) i en række lineære modeller. Blandt de modeller, der kan benyttes, er regressionsmodeller, conjoint analyse, kanonisk korrelationsanalyse

7.2.7 Repræsentative undersøgelser

- SURVEYMEANS, frembringer estimater for populationsgennemsnit og -totaler fra *repræsentative undersøgelser*
- SURVEYREG, udfører regressionsanalyser på data fra *repræsentative undersøgelser*

- SURVEYSELECT, foretager sandsynlighedsbaseret tilfældig udvælgelse af stikprøveenheder ved *repræsentative undersøgelser* under hensyntagen til *stratifikation*, *klyngeudformning* og forskelle i udvælgelsessandsynligheder

7.2.8 Diverse procedurer

- BOXPLOT, frembringer *Boxplots*
- FREQ, foretager en tabellering, eller en krydstabellering af data
- MI, foretager multipel *imputering*, dvs estimation af manglende værdier i et datasaet
- MIANALYZE, analyserer datasæt, der har været behandlet af MI-proceduren (forsøgsvis indlagt i version 8.1)

7.3 Procedurer i SAS/ETS

SAS/ETS (Econometrics and Time Series) består af en række procedurer til analyse af data fra processer, der forløber over over tid.

Brugen af procedurerne er beskrevet i manualerne SAS/ETS User's Guide, Version 8, Volumes 1 og 2.

Hjælp vedrørende procedurerne findes i hjælp, `Contents` under `Help on SAS Software Products`, `SAS/ETS`, hvorefter man vælger den pågældende procedure.

- ARIMA, analyserer og forudsiger tidsrækkedata ved de såkaldte *ARIMA-modeller*. Modellerne inkluderer brug af transfer funktioner og interventionsdata
- AUTOREG, estimerer og forudsiger tidsrækkedata ved brug af lineære modeller for autoregressive data, herunder også *heteroskedastiske* modeller (genraliserede autoregressive betingede heteroskedastiske, GARCH-modeller)
- COMPUTAB, foretager beregning og tabellering (COMPUting and TABular reporting) ved brug af en programmerbar regnearksfacilitet

- DATASOURCE, uddrager tidsrækkedata fra flere forskellige slags datafiler og placerer dem i et datasæt
- EXPAND, *konverterer* tidsrækkedata fra én sampling hyppighed til en anden og interpolerer manglende værdier i rækken
- FORECAST, frembringer *forudsigelser* for flere tidsrækker på én gang
- LOAN, analyserer og sammenligner forskellige tilbagebetalingsmetoder for *lån*
- MODEL, analyserer tidsrækkemodeller, hvor relationerne mellem de variable er et system af en eller flere ikke-lineære ligninger
- PDLREG, estimerer *distributed lag* tidsrækkemodeller
- QLIM, analyserer modeller, hvor de afhængige variable antager diskrete værdier, eller hvor de kun observeres i et begrænset værdiområde. Proceduren omfatter logit, probit, tobit, Poisson regression og modeller for simultane ligninger, herunder modeller for diskret valg (forsøgsvist indlagt i version 8.1)
- SIMLIN, indlæser koefficienter i et sæt af lineære *strukturelle ligninger* (som fx frembragt af SYSLIN) og bruger den reducerede form til at bestemme forudsigelser
- SPECTRA, udfører spektralanalyse og kryds-spektralanalyse af tidsrækkedata
- STATESPACE, analyserer og forudsiger multivariate tidsrækkedata ved brug af *tilstandsmodeller* (State Space modeller)
- SYSLIN, estimerer parametre i indbyrdes afhængige systemer af lineære regressionsligninger
- TSCSREG, **T**ime **S**eries **C**ross **S**ection **R**egression analyserer data, der består af en kombination af *forløbsdata* og *tværsnitsdata*
- VARMAX
- X11, foretager en sæsonkorrektion af måneds- eller kvartalsserier ved en tillempning af den såkaldte X-11 metode udviklet af US Bureau of the Census

- X12, foretager en sæsonkorrektion af måneds- eller kvartalsserier ved en tillempning af den såkaldte X-12 ARIMA metode udviklet af US Bureau of the Census

7.4 Procedurer i SAS/OR

SAS/OR indeholder en række værktøjer til brug for projektledelse, optimering og beslutningsstøtte.

- ASSIGN, bestemmer den optimale løsning til *assignment problemet*
- CPM, benyttes til **P**lanning, **C**ontrolling and **M**onitoring af projekter, der forløber i kalendertid
- DTREE, frembringer et *beslutningstræ* med tilknyttede sandsynligheder og *payoffs*
- GANTT, frembringer et såkaldt *Gantt-diagram* til projektplanlægning
- LP, løser og analyserer lineære programmeringsproblemer, heltalsprogrammeringsproblemer og blandede problemer
- NETDRAW, tegner et netværksdiagram for aktiviteterne i et projekt
- NETFLOW, bestemmer *flow* i hver gren i et netværk
- NLP, løser ikke-lineære optimeringsproblemerr
- PM, interaktiv procedure til planlægning, overvågning og styring af et projekt. Proceduren løser de samme problemer som den ikke-interaktive CPM procedure
- PROJMAN, interaktivt program med en grafisk brugergrænseflade til *projektledelse*
- QSIM, modellerer og analyserer *køsystemer* ved simulation
- TRANS, analyserer og løser et *transportproblem*

7.5 Procedurer i SAS/GRAPH

SAS/GRAPH indeholder en række procedurer til fremstilling af grafisk output. Hjælp vedrørende procedurerne findes i hjælp, `Contents` under `Help on SAS Software Products`, `Using SAS/GRAPH Software`, `SAS/GRAPH Procedures` hvorefter man vælger den pågældende procedure.

Man kan endvidere interaktivt fremstille grafisk output ved brug af *Graph-N-Go* faciliteten (se afsnit 9.2).

Endvidere kan man addressere grafiske funktioner fra et `DATA`-trin ved brug af DSGI (Data Step Graphics Interface) faciliteten.

Nedenstående liste angiver procedurerne i SAS/GRAPH. For en nærmere beskrivelse henvises til *Hjælp*-funktionen.

- GANNO
- GCHART
- GCONTOUR
- GDEVICE
- GFONT
- GIMPORT
- GKEYMAP
- GMAP
- GOPTIONS
- GPLOT
- GPRINT
- GPROJECT
- GRADAR
- GREDUCE
- GREMOVE

- GREPLAY
- GSLIDE
- GTESTIT
- G3D
- G3DGRID

7.6 Procedurer i SAS/QC

SAS/QC indeholder en række procedurer til hjælp i kvalitetsstyringssammenhænge. En del af procedurerne finder dog også anvendelse i andre analysesammenhænge. Hjælp vedrørende procedurerne findes i hjælp, `Contents` under `Help on SAS Software Products`, `SAS/QC`.

Brugen af procedurerne er beskrevet i manualerne SAS/QC User's Guide, Version 8, Volumes 1, 2 og 3.

- CAPABILITY, analyser *fordelingen* af en variabel, tegner histogrammer med den tilpassede fordeling indtegnet, tegner Q-Q plots og bestemmer diverse robuste estimatorer. Der er mulighed for tilpasning ved Beta-fordeling, Exponentialfordeling, Gammafordeling, lognormalfordeling, normalfordeling og Weibullfordeling. Proceduren giver desuden mulighed for at estimere andelen uden for specifikationsgrænser og bestemmes konfidensintervaller for denne andel.
- CUSUM, analyserer data ved hjælp af et *cusum kontrolkort*
- FACTEX, konstruerer ortogonale forsøgsplaner for faktorforsøg
- ISHIKAWA, frembringer SAS-omgivelser, der giver mulighed for at tegne *Ishikawa diagrammer* (fiskebensdiagrammer)
- MACONTROL, analyser data ved hjælp af et kontrolkort for *glidende gennemsnit* (eksponentielt vægtet glidende gennemsnit (EWMA), eller almindeligt glidende gennemsnit (MA))
- OPTEX, udsøger optimale *forsøgsplaner* svarende til givne krav og begrænsninger

- PARETO, frembringer såkaldte *Pareto diagrammer*, der afbilder den relative og den kumulerede hyppighed af forskellige fejltyper
- RELIABILITY, tilpasser fordelinger til levetidsdata (herunder også censurerede data). Fordelingerne omfatter Weibull, exponential, Ekstremværdi, normal, lognormal, den logistiske og den loglogistiske.
- SHEWHART, analyserer data ved hjælp af et *Shewhart kontrolkort*, herunder ved brug af regler for runs

Vi fremhæver specielt CAPABILITY proceduren, som et simpelt værktøj til at tegne histogrammer med indtegnede estimerede tæthedsfunktioner.

Eksempel 18 ***Histogram med indtegnet normalfordelingstæthed***

Datasættet `eks4_7` indeholder data fra eksempel 4.7 (og 1.1) i Statistik 1 bogen. Den variable `diam` angiver nittehovedets diameter.

Nedenstående programstump bevirker nu, at der bliver et histogram over de observerede diametre med en indtegnet normalfordelingstæthed.

```
PROC CAPABILITY  DATA=WORK.nitter  GRAPHICS ;
      VAR diam;
HISTOGRAM / CAXES=BLACK CFRAME=CXFFFFFF
                 CBARLINE=BLACK CFILL=CXFFFF00 PFILL=SOLID
                 VSCALE=PERCENT HMINOR=0 VMINOR=0
                 NORMAL1( MU=EST SIGMA=EST W=1 COLOR=RED
                          )
     ;
   RUN;
```

□

7.7 INSIGHT-proceduren

Kald af SAS-proceduren INSIGHT bevirker at der startes en interaktiv session som beskrevet i afsnit 4.

Brugen af proceduren er beskrevet i manualen SAS/INSIGHT User's Guide, Version 8.

Hjælp vedrørende proceduren findes i hjælp, `Contents` under `Help on SAS Software Products`, `SAS/INSIGT`.

Insight-proceduren kan startes ved at skrive Programstumpen

```
PROC INSIGHT  ;
RUN;
```

i programvinduet.

7.8 Procedurestruktur og modelformler

Et procedurekald indledes altid med SAS-sætningen

`PROC` *procnavn* ;

hvor *procnavn* angiver procedurens navn. Procedurenavnet kan være efterfulgt af en række nøgleord, der angiver optioner for den pågældende procedure.

I de fleste af de ovennævnte procedurer skal man desuden specificere en modelformel. I en SAS-procedure skrives en modelformel i det væsentlige i overensstemmelse med beskrivelsen i afsnit 5.13.1 på side 36 ff. Den væsentligste forskel er, at den variabel, der i den interaktive session placeres i rubrikken `Y`, her skrives efterfulgt af et lighedstegn, hvorefter leddene, der i den interaktive session placeres i rubrikken `X`, skrives på højre side af lighedstegnet med et eller flere mellemrum imellem de enkelte led.

For eksempel angiver nedenstående programstump, at man betragter den variable `Kali` som afhængig variabel, og den kvalitative variable `Gren` som forklarende variabel. Programstumpen udfører en ensidet variansanalyse i lighed med analysen i eksempel 3 på side 37.

Eksempel 19 `PROC GLM ;`

```
CLASS Gren  ;
MODEL Kali  = Gren  ;
RUN;
```

□

Modelformlen kan sædvanligvis efterfølges af en række nøgleord, der specificerer de analyser, der skal foretages, udskrifter mv.

7.9 Kommunikation mellem procedurer og datasæt

Når bortses fra enkelte parametre i procedurekaldet, får en procedure sit input fra et SAS-datasæt. Proceduren kan levere output i `.lst`-filen, og nogle procedurer kan også levere output i specielle output-datasæt.

Som hovedregel kan kommunikationen mellem procedurer kun foregå via datasæt, hvilket ikke er særlig fleksibelt. (Den eneste mulighed for at "løfte" data ind i procedurer er via såkaldte SAS-makroer).

8 Beregning af sandsynligheder og fraktiler

SAS language, der bruges i DATA-trinet, indeholder funktioner til beregning af sandsynligheder og fraktiler i de sædvanlige fordelinger.

8.1 Beta-fordeling

Funktionen `PROBBETA`(x,a,b) returnerer værdien af den kumulerede fordelingsfunktion for en $\mathrm{B}(a,b)$-fordelt variabel i punktet x,

$$\mathrm{PROBBETA}(x,a,b) = \mathsf{P}[\mathrm{Be}(a,b) \leq x]$$

Eksempel 20 ***Brug af PROBBETA***
Programstumpen

```
DATA regn;
 px  = PROBBETA(0.2,3,4);
PROC PRINT;
RUN;
```

returnerer værdien 0.09888 □

Funktionen `BETAINV` returnerer fraktilværdien i en Betafordeling.

$$\mathrm{BETAINV}(p,a,b) = \mathrm{Be}(a,b)_p \ ,$$

hvor $\mathrm{Be}(a,b)_p$ med $0 \leq p \leq 1$ som vanligt betegner p-fraktilen i en $\mathrm{Be}(a,b)$-fordeling, dvs

$$\mathsf{P}[\mathrm{Be}(a,b) \leq \mathrm{Be}(a,b)_p] = p$$

Eksempel 21 ***Bestemmelse af fraktil i betafordelingen***

`x = BETAINV(0.001,2,4)`
returnerer værdien 0.01010 □

På grund af relationen mellem F-fordelingen og Betafordelingen, kan man bruge `BETAINV`-funktionen til direkte at bestemme konfidensintervaller for parameteren p i $\mathrm{B}(n,p)$-fordelingen.

Eksempel 22 ***Bestemmelse af konfidensinterval i binomialfordeling***

SAS-programmet

```
DATA konf;
n = 20;
x = 3;
alfa  = 0.05;
plow  = BETAINV(alfa/2,x,n-x+1);
pup   = BETAINV(1-alfa/2,x+1,n-x);
OUTPUT;
PROC PRINT;
RUN;
```

giver værdierne `plow = 0.032071` og `pup = 0.37893`, der netop er grænserne i det sædvanlige 95 % konfidensinterval for p svarende til $n = 20$ og $x = 3$. □

8.2 Binomialfordeling

Funktionen `PROBBNML(p,n,x)` returnerer værdien af den kumulerede fordelingsfunktion for en $\mathrm{B}(n,p)$-fordelt variabel i punktet x,

$$\mathrm{PROBBNML}(p,n,x) = \mathsf{P}[\mathrm{B}(n,p) \le x]$$

Eksempel 23 ***Bestemmelse af sandsynligheder i binomialfordeling***

Hvis man i et data-trin bruger funktionskaldet

`px = PROBBNML(0.5,10,4)`

returneres værdien 0.37695, der netop er $\mathsf{P}[\mathrm{B}(10,0.5) \le 4]$

Punktsandsynligheden (frekvensfunktionen) svarende til værdien x for $1 \le x \le n$ fås som

`PX = PROBBNML(p,n,x) - PROBBNML(p,n,x-1).`

Punktsandsynligheden svarende til $x = 0$ fås som

`PX = PROBBNML(p,n,0)`

□

8.3 χ^2-fordeling

Funktionen `PROBCHI(x,f)` returnerer værdien af den kumulerede fordelingsfunktion for en $\chi^2(f)$-fordelt variabel i punktet x.

Man finder p-værdien svarende til den beregnede værdi, z, af en $\chi^2(f)$-fordelt teststørrelse (ensidet test) ved

`p = 1-PROBCHI(z,f)`

Funktionen `CINV` returnerer fraktilværdien i χ^2-fordelingen.

$$\mathrm{CINV}(p,f) = \chi^2(f)_p \ ,$$

hvor $\chi^2(f)_p$ med $0 \le p \le 1$ betegner p-fraktilen i en $\chi^2(f)$-fordeling.

Eksempel 24 ***Konfidensinterval for varians***
Betragt et observationssæt $x_1, \ldots, x_n$ med modellen $X_i \in \mathrm{N}(\mu, \sigma^2)$.

Antag, at man i en aktuel situation har $n = 5$, og man har fået den empiriske varians $s^2 = 20$ (μ ukendt)

Det sædvanlige 95 % konfidensinterval for σ^2 fås da ved SAS-programmet (jvf tabellen side 266 i lærebogen i Statistik 1)

```
DATA konf;
n = 5;
s2  = 20;
alfa  = 0.05;
f  = n-1 ;
si2low = f *s2 /CINV(1- alfa/2, f);
```

```
silow  = SQRT(si2low);
si2up  =  f *s2 /CINV( alfa/2, f);
siup = SQRT(si2up);
PROC PRINT;
```

der giver værdierne `si2low = 7.1729` og `si2up = 165.146` for konfidensintervallet for variansen, σ^2, og `silow = 2.67940` og `siup = 12.8509` for konfidensintervallet for spredningen, σ.

□

Eksempel 25 ***Test af hypotese vedrørende varians***

Betragt et observationssæt $x_1, \ldots, x_n$ med modellen $X_i \in \mathrm{N}(\mu, \sigma^2)$. Teststørrelsen for test af hypotesen $\mathrm{H}_0 : \sigma^2 \leq \sigma_0^2$ mod alternativet $\mathrm{H}_1 : \sigma^2 > \sigma_0^2$ er

$$z = \frac{n-1}{\sigma_0^2}\, s^2$$

Antag, at man i en aktuel situation med $\sigma_0^2 = 16$ har $n = 5$ og man har fået den empiriske varians $s^2 = 20$. Teststørrelsen for hypotesen $\mathrm{H}_0 : \sigma^2 \leq 16$ mod alternativet $\mathrm{H}_1 : \sigma^2 > 16$ er

$$z = \frac{4}{16}\, 20 = 5.0$$

p-værdien svarende til denne værdi er

$$p = \mathsf{P}[\chi^2(4) \geq 5.0]$$

der bestemmes ved

```
DATA test;
p = 1  - PROBCHI(5.0,4);
PROC PRINT;
```

Man får `p = 0.28730`

Den kritiske værdi ved test på et 5 %-niveau fås ved dette ensidede test som

```
DATA omraade ;
krit  = CINV(0.95,4);
PROC PRINT;
```


Man får `krit = 9.48773`

Styrken af dette test, hvis $\sigma^2 = \gamma\sigma_0^2$ er (jvf oversigten side 339 i lærebogen i Statistik 1)

$$p(\gamma) = \mathsf{P}[\chi^2(4) > \frac{9.48773}{\gamma}]$$

Nedenstående SAS-program beregner styrken for dette test for værdierne $\sigma^2/\sigma_0^2 = 1, 2, \ldots, 10$ for $n = 5$ og $\alpha = 0.05$

```
DATA styrke;
n   = 5;
alfa = 0.05;
f  = n-1;
krit =  CINV(1-alfa,f);
DO gamma = 1, 2 TO 10 BY 2;
   styrke = 1 - PROBCHI(krit/gamma, f);
   OUTPUT;
   END;
PROC PRINT;
VAR gamma styrke;
RUN;
```

Programmet giver udskriften

```
 gamma   styrke
  1      0.05000
  2      0.31460
  4      0.66771
  6      0.81215
  8      0.88040
 10      0.91746
```

□

Eksempel 26 ***Bestemmelse af kritisk værdi i et χ^2-test***

Betragt et test, hvor fordelingen af teststørrelsen under hypotesen er en $\chi^2(3)$-fordeling, og hvor testet forkaster for *store* værdier af teststørrelsen, som fx ved test i antalstabeller.

Den kritiske værdi ved et test på 5 %-niveauet fås da som

```
zkrit = CINV(0.95,3),
```

der returnerer værdien 7.8147 □

8.3.1 Den ikke-centrale χ^2-fordeling

Ved tilføjelse af et yderligere argument, der angiver ikke-centralitetsparameteren, kan `PROBCHI`-funktionen desuden bruges til at angive værdier af den kumulerede fordelingsfunktion for den ikke-centrale χ^2-fordeling.

Denne variant kan bruges ved bestemmelse af *styrkefunktionen* for visse χ^2-test

Funktionen `PROBCHI(x,f,lam)` returnerer således værdien af den kumulerede fordelingsfunktion for en $\chi^2(f, lam)$-fordelt variabel i punktet x.

Ved tilføjelsen af et argument, der angiver ikke-centralitetsparameteren, kan `CINV`-funktionen tilsvarende bruges til at angive fraktiler i den ikke-centrale χ^2-fordeling.

Endelig angiver funktionen `CNONCT(x,df,prob)` ikke-centralitetsparameteren λ for den ikke-centrale χ^2-fordeling med df frihedsgrader for hvilken $prob$-fraktilen netop har værdien x.

Eksempel 27 `x = CINV(0.95,3,4.5)`
returnerer værdien 17.5046, der netop er 95%-fraktilen i den ikke-centrale χ^2-fordeling med 3 frihedsgrader og ikke-centralitetsparameteren $\lambda = 4.5$. □

8.4 F-fordeling

Funktionen `PROBF(x,f1,f2)` returnerer værdien af den kumulerede fordelingsfunktion for en $\mathrm{F}(f1, f2)$-fordelt variabel i punktet x (jvf. En Introduktion til Statistik, Bind 1 A, afsnit 1.10.4).

Funktionen `FINV` returnerer fraktilværdien x_p i en $\mathrm{F}(f_1, f_2)$-fordeling. Således vil

```
xp = FINV(0.95,2,10)
```

returnere værdien 4.1028, som er 95%-fraktilen i en $\mathrm{F}(2, 10)$-fordeling.

Eksempel 28 ***Konfidensinterval for forholdet mellem to varianser***

Betragt to observationssæt $x_1, \ldots, x_n$ og $y_1, \ldots, y_m$ med modellen $X_i \in \mathrm{N}(\mu_x, \sigma_x^2)$ og $Y_i \in \mathrm{N}(\mu_y, \sigma_y^2)$.

Antag, at man i en aktuel situation har $n = 5$, $n = 10$, og man har fået de to empiriske varianser, $s_x^2 = 20.0$ og $s_y^2 = 4.4444$, dvs $s_x^2/s_y^2 = 4.5$.

Det sædvanlige 95 % konfidensinterval for forholdet, σ_x^2/σ_y^2 mellem de to varianser fås da ved SAS-programmet (jvf tabellen side 267 i lærebogen i Statistik 1)

```
DATA konf;
n = 5;  fx = n-1;
m =  10; fy = m-1;
sx2 = 20.0 ;
sy2 = 4.4444 ; ratio = sx2/sy2 ;
alfa  = 0.05;
konflow = ratio * FINV(alfa/2,fy,fx) ;
konfup = ratio * FINV(1-alfa/2,fy,fx) ;
PROC PRINT;
```

der giver værdierne `konflow = 0.95379` og `konfup = 40.0715` for konfidensintervallet for forholdet, σ_x^2/σ_y^2 mellem de to varianser. □

Eksempel 29 ***Test for sammenligning af to varianser***

Betragt situationen fra det foregående eksempel (eksempel 28), og betragt hypotesen $\mathrm{H}_0 : \sigma_x^2 \leq \sigma_y^2$ mod alternativet $\mathrm{H}_1 : \sigma_x^2 > \sigma_y^2$. Teststørrelsen er

$$z = s_x^2/s_y^2$$

hvor s_x^2 og s_y^2 som før angiver de empiriske varianser for x'erne og y'erne.

Antag, at man i en aktuel situation har $n = 5$, $n = 10$, og at man har fået teststørrelsen $z = 4.50$.

p-værdien svarende til denne værdi af teststørrelsen er

$$p = \mathrm{P}[\mathrm{F}(4, 9) \geq 4.50]$$

Benytter man i et data-trin funktionskaldet

```
pz = 1 - PROBF(4.5,4,9)
```

returneres værdien 0.028511, der er netop er den søgte p-værdi.

Det kritiske område ved et test på niveau α er

$$z > \mathrm{F}(n-1, m-1)_{1-\alpha}$$

For $n = 5$ og $m = 10$ og $\alpha = 0.05$ fås den kritiske værdi ved funktionskaldet

```
krit = FINV(0.95,4,9)
```

der returnerer værdien `krit = 3.63309`.

Styrken af dette test, hvis $\sigma_x^2/\sigma_y^2 = \gamma$ er jvf oversigten side 345 i lærebogen i Statistik 1

$$p(\gamma) = \mathsf{P}\left[\mathrm{F}(4,9) > \frac{3.63309}{\gamma}\right]$$

Nedenstående program beregner styrken for dette test for værdierne $\sigma_x^2/\sigma_y^2 = 1, 2, \ldots, 10$ for $n = 5$, $n = 10$ og $\alpha = 0.05$

```
DATA styrke;
n    = 5;
m    = 10;
alfa = 0.05;
fx   = n-1;
fy   = m-1;
krit = FINV(1-alfa,fx,fy);
DO gamma = 1,  2 TO 10 BY 2;
   styrke = 1 - PROBF(krit/gamma, fx,fy);
   OUTPUT;
END;
PROC PRINT;
VAR gamma styrke;
RUN;
```

Programmet giver udskriften

gamma	styrke
1	0.05000
2	0.20984
4	0.49879
6	0.66868
8	0.76763
10	0.82885

□

Man kan specielt bruge FINV-funktionen til at bestemme konfidensintervaller for parameteren p i $\mathrm{B}(n,p)$-fordelingen som beskrevet i lærebogen.

Eksempel 30 ***Bestemmelse af konfidensinterval i binomialfordeling***

SAS-programmet


```
DATA konf;
n = 20;
x = 3;
alfa  = 0.05;
flow = FINV(1- alfa/2, 2*(n-x+1),2*x);
plow = x/(x + (n-x+1)*flow  );
fup  =  FINV(1- alfa/2, 2*(x+1),2*(n-x) );
pup = (x+1)*fup /(n-x + (x?1)*fup ) ;
OUTPUT;
PROC PRINT;
```

giver værdierne `plow = 0.032071` og `pup = 0.37893`.
(Se også eksempel 22 på side 73). □

8.4.1 Den ikke-centrale F-fordeling

PROBF-Funktionen kan desuden bruges med et yderligere argument, der angiver ikke-centralitetsparameteren. Funktionen `PROBF(x,f1,f2,gam2)` returnerer således værdien af kumulerede fordelingsfunktion for en $\mathrm{F}(f1, f2; gam2)$-fordelt variabel i punktet x

Tilsvarende kan FINV-funktionen bruges til at angive fraktiler i den ikke-centrale F-fordeling ved tilføjelsen af et argument, der angiver ikke-centralitetsparameteren.

Endelig angiver funktionen `FNONCT(x,dft,dfn,prob)` ikke-centralitetsparameteren λ for den ikke-centrale F-fordeling med (dft, dfn) frihedsgrader for hvilken *prob*-fraktilen netop har værdien x.

Eksempel 31 ***Bestemmelse af sandsynlighed i den ikke-centrale F-fordeling***

Når man eksempelvis i et data-trin bruger funktionskaldet

```
pz = PROBF(3.32,2,10,4.0)
```

returneres værdien 0.5923, der er sandsynligheden for at få en værdi af teststørrelsen, der er mindre end $z = 3.22$, når teststørrelsen følger en ikke-central F-fordeling med (2,10) frihedsgrader og ikke-centralitetsparameteren 4.0

Tilsvarende vil funktionskaldet

```
zp = FINV(0.95,2,10,2.5)
```

returnere værdien 8.4512, som er 95%-fraktilen i en F$(2, 10, 2.5)$-fordeling, dvs en ikke-central F-fordeling med (2,10) frihedsgrader og ikke-centralitetsparameter 2.5 □

Eksempel 32 ***Bestemmelse af styrke i ensidet variansanalyse***
Betragt modellen svarende til en ensidet variansanalyse,

$$X_{ij} \in \mathrm{N}(\mu_i, \sigma^2),\ i = 1, 2, \ldots, k;\ \ j = 1, 2, \ldots, n_i$$

Den sædvanlige teststørrelse for *homogenitetshypotesen*, $\mathrm{H}_0 : \mu_1 = \mu_2 = \cdots = \mu_k$ er

$$Z = \frac{\sum_{i=1}^{k} n_i(\overline{X}_{i\cdot} - \overline{X}_{\cdot\cdot})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X}_{i\cdot})^2/(N-k)}$$

som ved et test på niveau α skal sammenlignes med $\mathrm{F}(k-1, N-k)_{1-\alpha}$.

Styrken af dette test overfor et alternativ, $\mu_1, \ldots, \mu_k$ afhænger kun af *ikke-centralitets parameteren*

$$\gamma = \sum_{i=1}^{k} n_i(\mu_i - \overline{\mu})^2 \Big/ \sigma^2$$

med

$$\overline{\mu} = \sum_{i=1}^{k} n_i\mu_i \Big/ \sum_{i=1}^{k} n_i$$

Nedenstående SAS-program beregner styrken for dette test for værdierne $\gamma = 0, 1, 5, 10, 15$ for $k = 4$ grupper og $n_i = 6$ observationer i hver gruppe og med $\alpha = 0.05$

```
DATA styrke;
antgrup = 4;
ngrup = 6;
ntot = antgrup  * ngrup;
alfa = 0.05;
fmellem = antgrup -1;
ferror  = ntot -antgrup;
krit = FINV(1-alfa,fmellem,ferror);
DO gamma = 0, 1, 5  TO 15 BY 5;
   styrke = 1- PROBF(krit,fmellem,ferror, gamma);
   OUTPUT;
   END;
PROC PRINT;
VAR gamma styrke;
RUN;
```


Programmet giver udskriften

gamma	styrke
0	0.05000
1	0.10352
5	0.36921
10	0.66775
15	0.85020

□

8.5 Gamma-fordeling

Funktionen `PROBGAM(x,a)` returnerer værdien af den kumulerede fordelingsfunktion for en G$(a, 1)$-fordelt variabel i punktet x.

Eksempel 33 `p = PROBGAM(1,3)`
returnerer værdien 0.08030 □

Eksempel 34 Hvis $X \in \mathrm{G}(k, \beta)$, bestemmes værdien for den kumulerede fordelingsfunktion for X i punktet x ved

```
px = PROBGAM(k,x/beta)
```

□

Funktionen `GAMINV` returnerer fraktilværdien x_p i en G$(a, 1)$-fordeling

Eksempel 35 `xp = GAMINV(0.1,2.1)`
returnerer værdien 0.58419 □

8.6 Hypergeometrisk fordeling

Funktionen `PROBHYPR(NPOP,M, NSAM,X)` returnerer værdien af frekvensfunktionen for en H$(NPOP, M, NSAM)$-fordelt variabel i punktet x.

8.7 Negativ Binomial fordeling

Funktionen `PROBNEGB(p,r,x)` returnerer værdien af den kumulerede fordelingsfunktion for en NB(r,p)-fordelt variabel i punktet x.

Punktsandsynlighederne bestemmes ved at subtrahere den kumulerede fordeling i to successive argumenter
`px = PROBNEGB(p,r,x) - PROBNEGB(p,r,x-1)` for $x \geq 1$;

8.8 Normalfordeling

Funktionen`PROBNORM(x)` returnerer værdien af den kumulerede fordelingsfunktion for en N$(0,1)$-fordelt variabel (dvs fordelingsfunktionen $\Phi(x)$ for en standardiseret normalt fordelt variabel) i punktet x.

Eksempel 36 `px = PROBNORM(1.6449)`
returnerer værdien 0.95000 □

Funktionen `PROBIT(p)` returnerer p-fraktilen, u_p, i en standardiseret normalfordeling.

Eksempel 37 `q = PROBIT(0.95)`
returnerer værdien 1.6449 □

8.9 t-fordeling

Funktionen `PROBT(x,f)` returnerer værdien af den kumulerede fordelingsfunktion for en t(f)-fordelt variabel i punktet x.

Funktionen `TINV` returnerer fraktilværdien i t-fordelingen.

$$\mathrm{TINV}(p,f) = \mathrm{t}(f)_p \ ,$$

hvor t$(f)_p$ med $0 \leq p \leq 1$ betegner p-fraktilen i en t(f)-fordeling.

Eksempel 38 ***Beregning af kumuleret fordeling og fraktiler i t-fordeling***
Hvis man i et data-trin benytter funktionskaldet


```
pt = PROBT(1.860,8)
```

returneres værdien `0.95003`

Omvendt vil funktionskaldet

```
t = TINV(0.95,8)
```

returnere værdien `1.85955` □

Eksempel 39 ***Konfidensinterval for middelværdi i normalfordelingen***

Betragt et observationssæt $x_1, \ldots, x_n$ med modellen $X_i \in \mathrm{N}(\mu, \sigma^2)$.

Antag, at man i en aktuel situation har $n = 5$, og man har fået gennemsnittet, $\overline{x} = 3.0$ og den empiriske varians $s^2 = 16$

Det sædvanlige 95 % konfidensinterval for μ fås da ved SAS-programmet (jvf tabellen side 266 i lærebogen i Statistik 1)

```
DATA konf;
n = 5;
xbar = 3.0;
s2  = 16; s = SQRT(s2);
alfa  = 0.05;
f  = n-1 ;
mylow = xbar - s * TINV(1-alfa/2,f)/SQRT(n);
myup =  xbar - s * TINV(alfa/2,f)/SQRT(n);
PROC PRINT;
```

`mylow = -1.96666` og `myup = 7.96666` for konfidensintervallet for middelværdien, μ. □

8.9.1 Den ikke-centrale t-fordeling

Ved tilføjelse af et yderligere argument, der angiver ikke-centralitetsparameteren, kan `PROBT`-funktionen desuden bruges til at angive værdier af den kumulerede fordelingsfunktion for den ikke-centrale t-fordeling.

Funktionen `PROBT(x,f,lam)` returnerer således værdien af kumulerede fordelingsfunktion for en t(f, lam)-fordelt variabel i punktet x.

Tilsvarende kan `TINV`-funktionen bruges til at angive fraktiler i den ikke-centrale t-fordeling ved tilføjelsen af et argument, der angiver ikke-centralitetsparameteren.

Endelig angiver funktionen `TNONCT(x,df,prob)` ikke-centralitetsparameteren λ for den ikke-centrale t-fordeling med *df* frihedsgrader for hvilken *prob*-fraktilen netop har værdien x.

Eksempel 40 ***Beregning af kumuleret fordeling og fraktiler i den ikke-centrale t-fordeling***

Hvis man i et data-trin benytter funktionskaldet

```
pt = PROBT(1.84,8,3.5)
```

returneres værdien `pt= 0.059451`

Omvendt vil funktionskaldet

```
t = TINV(0.06,8,3.5)
```

returnere værdien `t= 1.84452`, der netop er 6%-fraktilen i den ikke-centrale t-fordeling med 8 frihedsgrader og ikke-centralitetsparameter $\lambda = 3.5$ □

Eksempel 41 ***Test for middelværdien i en normalfordeling***
Betragt et observationssæt $x_1, \ldots, x_n$ med modellen $X_i \in \mathrm{N}(\mu, \sigma^2)$.

Antag, at man i en aktuel situation har $n = 5$, og man har fået gennemsnittet, $\overline{x} = 3.0$ og den empiriske varians $s^2 = 16$

Betragt hypotesen $\mathrm{H}_0 : \mu \leq 2$ mod alternativet $\mathrm{H}_1 : \mu > 2$.

Teststørrelsen for denne hypotese er (jvf tabellen side 337 i lærebogen i Statistik 1)

$$z = \frac{\overline{x} - 2}{s/\sqrt{n}} = 0.55902$$

p-værdien svarende til denne værdi er

$$p = \mathsf{P}[\mathrm{t}(4) \geq 0.55902]$$

der bestemmes ved

```
DATA test;
p = 1  - PROBT(0.55902,4);
PROC PRINT;
```

Man får `p = 0.30298`

Den kritiske værdi ved test på et 5 %-niveau fås ved dette ensidede test som


```
DATA omraade ;
krit  = TINV(0.95,4);
PROC PRINT;
```

Man får `krit = 2.13185`

Styrken af dette test, hvis $\mu = 2 + \Delta \times \sigma$ er (jvf oversigten side 337 i lærebogen i Statistik 1)

$$p(\Delta) = \mathsf{P}[\mathrm{t}(4, \sqrt{5}\Delta) > 2.13185]$$

Nedenstående SAS-program beregner styrken for dette test for værdierne $\Delta = 0.5, \ldots, 2.0$ (dvs for $\mu = 2, 2 + 0.5\sigma, 2 + \sigma, \ldots, 2 + 2\sigma$) for $n = 5$ og $\alpha = 0.05$

```
DATA styrke;
n   = 5;
alfa = 0.05;
f  = n-1;
krit =  TINV(1-alfa,f);
DO delta = 0, 0.5 TO 2.0 BY 0.5 ;
   noncent = SQRT(n) * delta;
   styrke = 1 - PROBT(krit, f, noncent);
   OUTPUT;
   END;
PROC PRINT;
VAR delta  styrke;
RUN;
```

Programmet giver udskriften

```
 delta  styrke
  0      0.05000
 0.5     0.23900
 1.0     0.57974
 1.5     0.86195
 2.0     0.97483
```

Beregningerne ved et tosidet test foregår nogenlunde analogt. Man finder således p-værdien svarende til den beregnede værdi, z, af en $t(f)$-fordelt teststørrelse i et tosidet test ved

```
p =2 *( 1-PROBT( ABS(z),f) )
```

□

8.10 Todimensional normalfordeling

FunktionenPROBBNRM(x,y,r) returnerer sandsynligheden for $X \leq x$ og $Y \leq y$ i en standardiseret todimensional normalfordeling med korrelationen r

8.11 Poisson-fordeling

Funktionen POISSON(m,x) returnerer værdien af den kumulerede fordelingsfunktion for en P(m)-fordelt variabel i punktet x,

8.12 Multiple sammenligninger

Funktionen PROBMC beregner sandsynligheder og kritiske værdier fra forskellige fordelinger, der bruges ved multiple sammenligninger.

9 Menubaserede analysemoduler

9.1 Oversigt

Ved at vælge optionen Solutions i et af vinduerne, og derefter vælge optionen Analysis, får man adgang til en række menubaserede analyseværktøjer,

3-D Visual Analysis (SAS/SPECTRAVIEW) til grafisk præsentation

Analyst SAS-Analyst (undertiden benævnt Analyst Application), system til sædvanlige statistiske analyser. De programmer (programkode) og output, der frembringes som led i analyserne, placeres i en række temporære filmapper, der er tilknyttet Analyst-sessionen. Man kan derefter - om ønsket - editere i programmerne og bringe de ændrede programmer til udførelse. Programmet giver desuden mulighed for bestemmelse af *styrken* af de udførte tests.

Design of Experiments, system til frembringelse af forsøgsplaner og analyse af de resulterende forsøgsresultater (bruger SAS/ADX)

Geographical Information System, SAS/GIS

Guided Data Analysis, system til de sædvanlige statistiske analyser (SAS/LAB). I dialogen med systemet gives forslag til analysemuligheder, endvidere gives fortolkninger af analyserne.

Interactive Data Analysis, system til interaktiv dataanalyse, bygger på SAS/INSIGHT.
Dette system er nærmere beskrevet i afsnit 4 og 5.

Investment Analysis, system til analyse og modellering af betalingsstrømme.

Market Research, system til analyse af markedsdata

Project Management, system til projektledelse

Quality Improvement, system til anvendelser i kvalitetsstyring (bruger SAS/QC)

Queueing Simulation, system til simulering af køsystemer (SAS/QSIM)

Time Series Forecasting System, system til tidsrækkeanalyse, specielt rettet mod forudsigelsesmodeller (bruger SAS/ETS)

Time Series Viewer, system til frembringelse af en række forskellige plots (autokorrelationsfunktioner, partielle autokorrelationsfunktioner, differenser mv.) af tidsrækkedata

9.2 Valg af Graph-N-Go

Man vælger Graph-N-Go omgivelserne ved at vælge Solutions → Reporting → Graph-N-Go.

Man gør et datasæt *tilgængeligt* i Graph-N-Go ved at klikke på datasæt/table ikonen øverst i venstre side af menuen. Herved fremkommer den sædvanlige oversigt over SAS-biblioteker, hvor man så kan vælge det ønskede datasæt.

Ved at klikke på ikonen for den ønskede diagramtype, fremkommer en ramme i grafvinduet. Rammen flyttes ved at flytte musen, og rammen

placeres ved et tryk på venstre musetast. Ved at klikke med højre musetast inde i rammen fremkommer en række valgmuligheder. Man vælger nu `Model`, hvorefter man får mulighed for at vælge blandt de *tilgængelige* datasæt. Navnet på det valgte datasæt fremkommer nu i rammen.

Når man nu klikker med højre musetast inde i figuren, fremkommer der en række valgmuligheder. Nogle af disse muligheder er generelle, `Copy`, `Delete`, `Move`, `Grow/Shrink`, etc., andre er relateret direkte til den pågældende graf og giver direkte mulighed for valg af x-variabel, y-variabel, plotstyle mv.

Optionen `Properties` åbner en menu, hvor man kan vælge plottevariable, plotsymboler, interpolationsform mv. for at frembringe den ønskede figur og tilpasse detaljer i figuren.

9.3 Enterprise grænseflader

SAS-systemet tilbyder en række forskellige omgivelser til dataanalyse og -præsentation. Her skal blot peges på det såkaldte *Enterprise Information System*, EIS, der er en række værktøjer til frembringelse bl.a. af *ledelsesinformation*. Enterprise grænsefladerne udvikles løbende, således forventes et nyt modul *Enterprise Guide* at komme på markedet til PC-brug (Windows) i løbet af foråret 2001.

Man vælger EIS omgivelserne ved at vælge
`Solutions` → `Development and Programming` → `EIS/OLAP Application Builder`. Der fremkommer nu en hovedmenu med ikoner for

- `Getting started` (En lille vejledning i brug af EIS)
- `Metabase`, hvor man kan danne et såkaldt *reposotory med datasæt* eller med flerdimensionale datasæt (sammensat af flere databaser)
- `Build EIS`, hvor man kan konstruere grafer
- `Applications`, hvor man kan opbygge biblioteker med tidligere opgaver
- `Setup`, der giver mulighed for at linke databaser

- `Report Gallery`, der indeholder skabeloner for grafer mv
- `Object Manager`, der giver mulighed for at finde databasefiler
- `SAS/ASSIST` , der starter SAS/ASSIST omgivelserne

Der henvises til hjælp-funktionen for yderligere vejledning i brugen af EIS-omgivelserne.

9.4 SAS/ASSIST

Vi nævner endelig SAS/ASSIST-omgivelserne, der er en peg og klik grænseflade, der er rettet mod en opgave-orienteret (task orientated) brug af SAS-systemet. Disse omgivelser bruger en række af komponenterne i SAS-systemet, som kaldes frem direkte fra SAS/ASSIST omgivelserne.

Ved start af SAS/ASSIST (fx. fra `Solutions` i File-, Edit,...-bjælken) fremkommer der ikoner for følgende muligheder

- `Data Mgmnt`
- `Report Writing`
- `Graphics`
- `Planning Tools`
- `Eis`

Her skal specielt fremhæves *Graphics*-modulen, der giver mulighed for frembringelse af simple plots (Bar Charts, Pie Chart, Plots, Maps)
Plots giver simple X*Y Plot, X*Y Plot by Category, Multiple plots.

Når man har valgt en plottype, fremkommer en menu, hvor man kan vælge `Table` (dvs SAS-datasæt), og outputform (`Graphics device`). (I Windows versionen kan output sendes til `Microsoft Windows Display`).

I rubrikken `Graphics Device` kan man vælge mellem outputformater:
`ActiveX enabled GIF Driver`, `BMP File Format` (bitmap), samt diverse

regneark- og databaseformater.

File → Save (eller Save as) giver mulighed for at placere grafen i et SAS-katalog.

File → Save last output giver mulighed for at placere grafen som et *grafisk segment* i et katalog, hvor man senere kan se den i et *grafisk vindue*.

Tasks muligheden i den øverste bjælke giver blandt andet mulighed for at vælge en række statistiske analyser.

10 Eksport af tekst og grafik fra SAS-systemet

10.1 Output fra SAS-systemet

Output fra SAS-systemet håndteres af det såkaldte *Output Delivery System* (ODS) i PROC TEMPLATE.

Output fra programmer og procedurer vises i Output vinduet (tekst output), eller i et *Grafvindue* som en graf. Sædvanligvis etableres der en reference til dette output i *resultatvinduet*.

Ved brug af den interaktive procedure SAS/INSIGHT vises outputtet i særlige vinduer, og man kan vælge at eksportere output direkte herfra, eller man kan sende det en tur rundt om systemets sædvanlige outputvinduer som beskrevet i afsnit 10.2.2 og 10.2.3.

Hvis man vil overføre output eller dele heraf til et tekstbehandlingssystem (i databarerne: Staroffice, og i Windows omgivelser fx til MS-Word), må man *eksportere* outputtet, eksempelvis som beskrevet i det følgende.

have dele af output

10.1.1 Sideopsætning, titler og fodnoter

Man kan ændre sideopsætningen ved at vælge File → Page Properties
Der fremkommer nu en menu, der giver mulighed for valg af

Title, dvs sideoverskrift

Footnotes

Page Numbers

Margens

Der kan anføres flere titellinier eller fodnotelinier.

Titler og fodnoter kan også ændres ved at skrive en TITLE - eller FOOTNOTE-sætning i programeditoren.

Under Tools → System → Log and procedure output control → Procedure output kan man se nogle af de systemoptioner, der er valgt.

Man kan ændre systemoptionerne ved at skrive OPTIONS efterfulgt af de ønskede valg (afsluttet med et semikolon) i programeditoren og submitte. Således angiver NUMBER/NONUMBER om der udskrives sidenumre, DATE/NODATE angiver om der udskrives dato. Under Tools → Options → Preferences under fanebladet General kan man se, hvilke af disse systemoptioner, der er gældende.

Detaljeret vejledning findes i hjælp-funktionen under Help on SAS Software Products → Base SAS Software → Using Base SAS Software → Setting System Options.

10.2 Eksport fra SAS-Insight

10.2.1 Udskrift af "blandet output" (grafer og tabeller)

Vælg File → Print. Der fremkommer nu en menu for valg af printer (Menuen har lidt forskelligt udseende i Windows-versionen og på Sun-systemet). Hvis den ønskede printer er markeret som den valgte, fortsættes med OK, ellers vælges Setup, hvor man kan vælge den fysiske printer, man ønsker at bruge. I G-databaren findes følgende muligheder:

- Print Postscript to disk
- gps1-302 G-databaren i bygning 302
  (HP LaserJet 4050 TN)

- `gps2-302` G-databaren i bygning 302
  (HP LaserJet 5M)
- `gps1-305` G-databaren i bygning 305
- `gps1-306` G-databaren i bygning 306
  (HP LaserJet 4050 TN)
- `gps1-308` G-databaren i bygning 308
  (HP LaserJet 4050 TN)
- `gps2-308` G-databaren i bygning 308
  (HP LaserJet 5M)

Herefter vælges Print, hvorefter man kan vælge blandt følgende muligheder for *sideopsætning*

- `Fill Page`, der fylder hver side ud
- `One per Page`, der udskriver en outputkomponent pr side
- `Titles and Footnotes`, der bevirker at de *titler* og *fodnoter*, der er valgt i sessionen, udskrives hhv foroven og forneden på hver side

Såfremt man i menuen for valg af printer (`File` → `Print`) sætter $\surd$ i feltet `Print to File` og derefter `Print`, fremkommer en oversigt over filer og mapper i computerens filsystem (Unix/Windows). Når man har valgt filnavn, klikkes på OK, og så skal man lige vælge sideopsætning som ovenfor, hvorefter udskriften placeres i den angivne fil.

Hvis man havde *udvalgt* dele af output-vinduet ved at klikke på rammen af den pågældende del med højre musetast (flere dele kan markeres ved at holde `Ctrl`-tasten nede, mens man markerer med musetasten), er det kun de udvalgte dele, der udskrives.

10.2.2 Eksport af grafer

Fra et output-vindue under SAS-Insight kan man eksportere de grafiske output ved at vælge `File`, → `Save` → `Graphics File`, hvorefter der fremkommer et vindue, hvor man kan vælge filens navn (i Unix eller Windows

filsystemet; sædvanligvis i roden, men man kan angive sti med filnavn), endvidere kan man vælge filformat

- Microsoft Windows Bitmap; `*.BMP`
- Graphics Interchange Format; `*.GIF`
- Portable Bitmap; `*.PBM`
- Adobe PostScript; `*.PS`
- Tagged Image File Format; `*.TIFF`

Desuden bliver man bedt om at vælge sideopsætning som ovenfor.

Hvis man havde *udvalgt* dele af output-vinduet, er det kun de udvalgte dele, der eksporteres.

10.2.3 Eksport af tabeller

Fra et output-vindue under SAS-Insight kan man eksportere tabeloutput ved at vælge `File`, → `Save` → `Tables`. Tabellerne (eller de *udvalgte* tabeller) bliver nu udskrevet i SAS-systemets sædvanlige output-vindue, hvorfra det kan eksporteres videre som angivet i afsnittet *Output fra SAS-procedurer*, afsnit 10.3.2.

10.3 Output fra SAS-procedurer

10.3.1 Grafisk output

Fra det grafiske output-vindue, der fremkommer som resultat af kørsel af et SAS-program med grafisk output, kan man vælge `Export as Image`, hvorved der fremkommer samme muligheder som ved eksport fra Graph-N-Go (afsnit 10.5.1), dvs et vindue, hvor man kan vælge filens navn (i Unix eller Windows-filsystemet), og hvor man kan vælge filtypen

- Graphics Interchange Format; `*.gif`
- X11 Pixmap; `*.xpm`

- X11 Bitmap; `*.xbm`
- Portable Pixmap; `*.ppm`
- Adobe PostScript; `*.ps`
- Encapsulated PostScript; `*.eps`
- Microsoft Windows Bitmap; `*.bmp`
- Tagged Image File Format; `*.tif`
- JPEG Files; `*.jpg`
- Portable Network Graphics; `*.png`

I Windows versionen kan man (hvis man bruger den opsætning, der er beskrevet i afsnit 11) kopiere grafik fra et grafvindue (og evt bare en *udvalgt* del) til clipboarded (tryk CTRL+C, eller vælg `Copy to Paste Buffer` fra `Edit` menuen. Fra clipboarded kan man nu indsætte i Windows applicationen ved at bruge `Paste` eller `Paste Special` fra applikationen (clip'et foregår ved DIB, BMP, eller WMF input).

For nærmere vejledning se hjælp funktionen, `Using SAS with your Operating System` → `Using SAS under Windows` → `Managing SAS output` → `Producing Graphics`.

10.3.2 Tekst og tabeller

SAS-procedurerne udskriver tekst- og tabeloutput i *SAS-systemets* outputvindue.

Teksten i dette vindue kan udskrives til en printer ellller en fil som beskrevet i afsnittet vedrørende udskrift af blandet output fra SAS-Insight (afsnit 10.2.1).

Teksten kan udskrives som en almindelig tekstfil (`.lst`-format) eller i *Rich Text Format* (`.rtf`-format) ved at `File`, → `Save As`, hvorefter der fremkommer et vindue med en oversigt over filer og mapper i filsystemet (Unix eller Windows). Man kan nu vælge filens navn og filtype.

I Windows versionen kan man (hvis man bruger den opsætning, der er beskrevet i afsnit 11) kopiere tekst til Windows clipboarded og derefter indsætte (`Paste`) i en Windows applikation.

10.4 Fra resultatvinduet

Klikker man med højre musetast på et element i resultatvinduet (*Results*-vinduet i venstre side af skærmen med oversigt over analyseresultater), får man mulighed for at vælge `Save As`, hvorved der fremkommer en oversigt over filer og mapper i computerens filsystem (Unix/Windows). Når man har valgt filnavn, klikkes på Save, hvorefter det pågældende element gemmes som en `.lst`-fil.

Man kan også vælge `Save As Object`, hvorved man får man mulighed for at arkivere det pågældende element som et `Output Entry` i et SAS-bibliotek.

For grafiske elementer har man kun mulighed for at vælge `Print` (med de dertil knyttede muligheder) af det pågældende grafiske segment.

10.5 Figurer frembragt af Graph-N-Go

Graph-N-Go omgivelserne kan bruges til en peg og klik fremstilling af simple grafiske præsentationer, histogrammer, Pie-charts (lagkagediagrammer), todimensionale plots, samt diverse overlay-grafer.

10.5.1 Eksport af figurer

Optionen `Export` giver mulighed for eksport af figuren som en ekstern fil (i Unix eller Windows-filsystemet), som et IMAGE entry i SAS-systemet, som en HTML-fil, eller som en source-file (programfil).

Vælges `External file` fremkommer et vindue, hvor man kan vælge filens navn (i Unix eller Windows-filsystemet), og hvor man kan vælge filtypen

- Graphics Interchange Format; `*.gif`

- X11 Pixmap; `*.xpm`
- X11 Bitmap; `*.xbm`
- Portable Pixmap; `*.ppm`
- Adobe PostScript; `*.ps`
- Encapsulated PostScript; `*.eps`
- Microsoft Windows Bitmap; `*.bmp`
- Tagged Image File Format; `*.tif`
- JPEG Files; `*.jpg`
- Portable Network Graphics; `*.png`

En række af mulighederne findes også i den *Toolbox*, der hører til Graph-N-Go vinduet.

10.6 Export som mail

I `Tools` → `Options` → `Preferences` under fanebladet `General` har man mulighed for at vælge at maile indholdet af det aktuelle vindue som enten en tekstfil, eller som en fil i RTF-format.

11 Opsætning af Windows brugergrænseflade

Hvis man under Windows vil have sit grafiske output vist i Windows systemet, skal man benytte en særlig *device driver* `WIN`. Denne aktiveres ved følgende SAS-sætning i programeditoren

```
GOPTIONS DEVICE=WIN ;
```

(Såfremt man havde valgt at vise output som HTML-filer, vil grafikken blive konverteret til en `.gif`-fil.)

Under `Tools` → `Customize`, kan man tilpasse værktøjsbjælkerne. Vælges fanebladet `Customize` har man mulighed for at tilføje eller fjerne ikoner i værktøjsbjælken. Specielt kan man i PC-versionen (Windows) tilføje elementerne `WCUT` og `WCOPY`, der henholdsvis *flytter* (`cut`) eller *kopierer* en

markeret *tekst* til Windows clipboard, hvorfra den så kan indsættes `paste`) i andre Windows applikationer.

Indeks