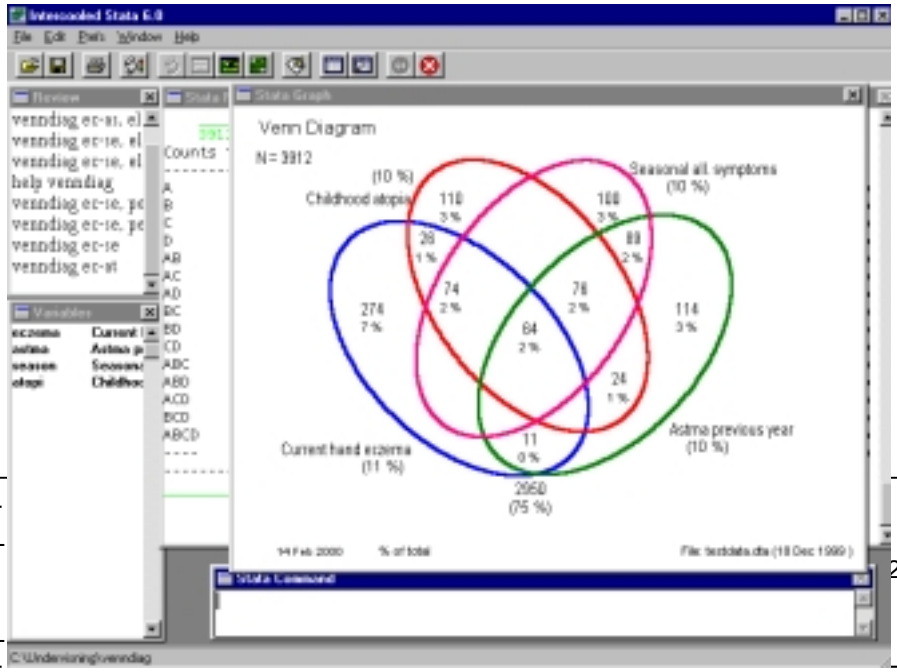
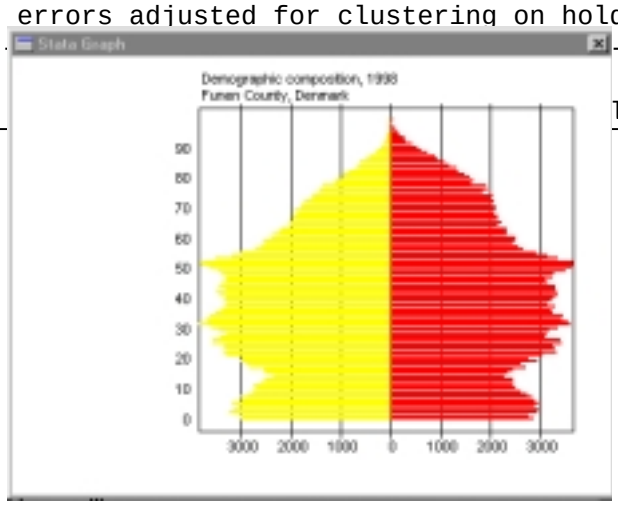
Stata-*Introduktion*




```
-
L                                                    of obs   =        240
                                                     2(1)     =      14.89
                                                     chi2     =     0.0001
L                                                    R2       =     0.0453
-------------------------------------------------------------------------
    test | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+---------------------------------------------------------------
    sr5a |   .3386243   .0980107     -3.741   0.000       .1920208    .5971564
-------------------------------------------------------------------------

->  xi: logistic test sr5b if sportn == 1 , cluster(hold)
Logit estimates                                     Number of obs   =        192
                                                    Wald chi2(1)    =       4.27
                                                    Prob > chi2     =     0.0387
Log likelihood = -126.28648                         Pseudo R2       =     0.0420
                              (standard errors adjusted for clustering on hold)
------------------------------------------------------------------------------
             |               Robust
    test | Odds Ratio   Std. Err.                                          ll
```



Jens M. Lauritsen ©

- *Forslag til arbejdsmåde*
- *Dokumentation*
- *Oparbejdning og afledte variable*
- *De mest anvendte analysefunktioner*
- *Nyttige hjælpeprogrammer*
- *Stata Technical Bulletin*
- *Tips til opsætning*

- **3. udgave 2000 (31 marts 2000)**

Stata - Introduktion .

3.udgave, ultimo marts 2000 (kontrollér dato på forsiden)
ISBN-87-987843-0-7 (trykt udgave)
ISBN-87-987843-1-5 (Elektronisk distribueret udgave)

Den trykte udgave distribueres via Studenterbogladen ved Syddansk Universitet.
E- Mail: studenter@boghandel.sdu.dk.

Der er udarbejdet et sæt øvelsesfiler til denne note. Kan downloades fra http://www.bola.suite.dk Se også side 13.

Forbehold: Da det er et meget stort arbejde at udarbejde noter, holde styr på figurer og undersøge om øvelserne fungerer må læserne bære over med uoverensstemmelser mellem de præsenterede skærmbilleder og de billeder der kommer frem ved at udføre øvelserne. Fejl og mangler vil jeg dog gerne have oplysninger om. Den elektroniske form tilpasses efterhånden, mens der kan være – forhåbentlig mindre – uoverensstemmelser med indholdet heri.

Kommentarer til noten sendes til: JM.Lauritsen@dadlnet.dk.
Angiv venligst hvilken udgave af noten, som kommentaren knytter sig til (Se forsiden).

Nogle nyttige internet sider om Stata, biostatistik eller hjælpeprogrammer:

Statistik rutiner og uddybende forklaringer af forskellige analysetyper i Stata:
http://www.oac.ucla.edu/training/stata/ **og** http://www.stata.com/links/resources1.html

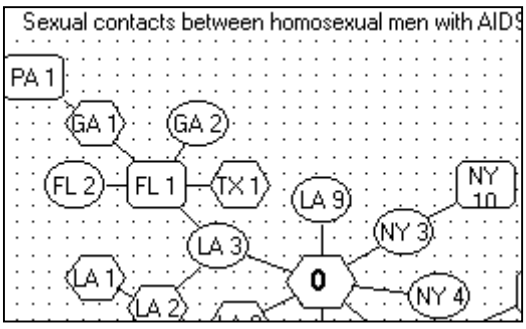
Instruktionsrutiner og forklaring af analyse af forskellige datatyper, inklusive øvelsesdata. Er baseret på EpiInfo, men principperne er generelt gyldige uanset statistikprogram:
http://www.sjsu.edu/faculty/gerstman/EpiInfo

Enkle manualer http://mkn.co.uk/help/extra/people/Brixton_Books

Gratis software til særlige formål:

EpiData til indtastning af data på en nem og pålidelige måde win95/98/NT/2000. Send en mail til adressen mailto:epidata@dadlnet.dk for nærmere oplysning. Programmet skriver datafiler, der direkte kan analyseres i Stata. (+Dbase, Excel og EpiInfo)

Epicalc 2000 (Simple beregninger-nærmest lommeregner) og et simpelt program til at tegne flowcharts og diagrammer (**EpiGram**):
http://www.myatt.demon.co.uk/



Uddrag af diagram tegnet med EpiGram. Ved endelige brug af diagrammet kopieres prikkerne ikke.

Stata - introduktion.

Forord

Hensigten med denne note er, at introducere brugen af statistik programmet Stata version 6. Noten giver anvisninger til den konkrete måde at arbejde på og kan **ikke** erstatte håndbøger, manualer og lærebøger i statistik og forskningsmetode. Tidligere udgaver af noten har angivet flere alternative metoder til arbejdet, hvilket i et vist omfang har forvirret "ikke erfarne" edb/statistik brugere. Denne udgave søger at fokusere på de muligheder der findes ved at bruge Stata, som programmet er udviklet. Dermed bliver noten også mere uafhængig af styresystem. idet Stata fungerer stort set ens uanset om du arbejder ud fra Windows, Linux eller Macintosh.

Forslag til hjælpeprogrammer er samlet i et særligt afsnit, herunder hvordan backup og lignende kan foretages.

Hvor meget edb erfaring er nødvendigt for at bruge Stata effektivt ? Ikke så meget. Det er nødvendigt at mestre grundlæggende tekstbehandling, at kunne kopiere filer, oprette mapper og være i stand til at hente filer fra internet. Desuden bør du være i stand til at arbejde med pakkede filer, især hvis din PC ikke har adgang til internet.

Filen kan pakkes på forskellige måder, fx såkaldt zip, arj eller uar format. Når en fil "pakkes" sker der det, at unødig tom plads fjernes. Fx kan en halv tom side i et dokument jo ændres til kommandoen "hop til næste side". Når en sådan fil pakkes ud igen indsættes den halve tomme side blot igen. Størrelsen på pakkede filer varierer fra 10-90 % af originalfilen. Det er nyttigt at arbejde med pakkede filer, hvis man skal sende over internet fordi transmissionstiden så nedsættes, men det forudsætter selvfølgelig at modtager kan pakke ud igen.

Der findes ingen universel ***rigtig*** måde at arbejde på. Find din egen og sørg for at blive god til at arbejde på den måde. Det er vigtigt, at du arbejder ***på en reproducerbar måde***. Minimér brugen af "mus", minimér brug af menu'er og maksimer reproducerbarhed og dokumentation af analyser og arbejdsformer. Du sparer dig selv for megen statisk muskelbelastning ved at bruge tastaturkombinationer hyppigt, se side 77 og 78

Ved nyindkøb af computer foreslås følgende prioritering: 1. En bedre kvalitet af maskinen med langsommere processor fremfor en hurtigere processor på en dårligere generel kvalitet (pris er et godt udgangspunkt for kvalitet, 3 års garanti også udtryk for kvalitet). 2. Mindst 64 Mb Ram. 3. Skærmkvalitet (køb en dyrere 17' skærm, der har flere punkter pr enhed). 4. Et bedre skærmkort der har en hurtigere "refresh rate". 5. Lavest muligt støjniveau. 6. Hurtigere og bedre harddisk, evt. 4 Gb. 6. Optimalt backup udstyr. 7. Bedre stol, bord og lampe. 8. Arbejdstid til at blive god til at bruge systemet, fx kurser. 9. Hurtigere maskine, Mere Ram end 64Mb. osv..

Hvorfor Stata ?

Principielt er det ligegyldigt hvilket statistikprogram der bruges, hvis blot det kan udføre de analyser der er brug for, på en måde der er til at finde ud af og at programmet kan købes til en rimelig pris. Det kan ikke betale sig at gøre en masse ud af hvorfor det ene program er bedre end det andet. Nogle af begrundelserne er udtryk for vane, smag og behag. For mig er der følgende gode grunde til at anvende Stata som statistikprogram fremfor andre (spss, sas, systat, ...).

- Det understøtter den allernyeste udvikling indenfor biostatistik. Det omhandler stort set alle analysetyper der er relevante i biomedicin. Dvs ét program er tilstrækkeligt.
- Det anvendes af en række gode biostatistikere og epidemiologer, som har skrevet uddybende rutiner på flere områder. (Bl.a. David Clayton, Michael Hills, Hosmer, Lemeshaw)[1] Det er veldokumenteret og der er særdeles fyldige og gode eksempler på de fleste biomedicinske analysetyper i manualerne. Nye rutiner dokumenteres og skrives af avancerede brugere og distribueres gratis. Der er muligt at foretage omfattende matrix beregninger. Det håndterer datoer på en effektiv måde.
- Det arbejder hurtigt og fylder lidt (Kan være på tre disketter). En komplet installation svarende til noten fylder kun 6-10 Mb afhængig af antal ekstra rutiner der anvendes.
- Det er særdeles velegnet til analyse og beskrivelse af follow-up data, som fx overlevelsesdata. Både "simple" data med kun ind- og udgangstid og data med gentagne outcome, fx gentagne sygdomsepisoder.
- De fleste regressionsmetoder har indbygget mulighed for at tilføje en klynge (cluster) effekt. Samt anvende "robust methods".
- Når man har købt en licens er denne "uendelig". Dvs. at en typisk bruger har programmet til ejendom. I modsætning til en årlig afgift. Prisen er overkommelig (< 5000 kr) og specielt hvis man arbejder i forskningsmiljø (< 1000 kr). For køb se www.stata.com. Evt. www.sdu.dk/dou eller www.metrika.se.

Der er selvfølgelig også nogle begrænsninger i Stata. Der kunne godt være flere eksakte test, fx i visse non-parametriske test.

---

[1] Clayton D, Hills M. Statistical Models in Epidemiology. Oxford, OUP, 1993. Analyser fra bogen er uddybet i: Clayton D, Hills M. Analysis of follow-up studies with Stata 5.0 STB Reprints Vol 7: 253-268 . Clayton D, Hills M. Analysis of case-control and prevalence studies STB Reprints Vol 5: 227—233. Levy & Lemeshow Sampling of Populations: Methods and Applications. Wiley, 1999. Hosmer & Lemeshaw: "Applied Survival Analysis" 1998.

General Commands in Stata

clear	Clear data from memory
display	Display values from functions
do	Execute commands
exit	Exit Stata
help	Find help
log	Echo output into a file
save	Save data for future use
use	Read a copy of a Stata file with data into memory
now	Start do file with automatic log (part of ***kursus.zip)***
webseek	Find material on internet

Data manipulation and Management

collapse	aggregate data into a table
count	count number of observations
describe	which data are in memory
destring	change string variable to numerical
drop	eliminate variables or observations
encode	create numeric variable from string
expand	duplicate observations
format	specify display format
generate	create new variable
infile	read fixed format data
input	add data via keyboard
label	descriptive text (values, vars., files)
list	list some variables
listjl	list id variable (also by group)
merge	combine files (check also mmerge)
move	change sequence of variables in file
mvdecode	recode to missing (e.g. from 9 to .)
mvencode	recode from "." to value
notes	add documentation to data + variable
order	change sequence of variables
outfile	write data in fixed format
pattern	indicate patterns of missing
query	show set options
quietly	do a command withou showing output
recode	recode numeric
rename	change name of variable
replace	change contents of variable
set	set general options
sort	sort observations by variables.

Other useful

venndiag	Venn diagrams and creation of coupled variables
hbar	Horisontal bar charts. E.g. Population Pyramids
cdf	Cumulative plot of continous var.
outreg	Regressionoutput to publication
qnorm	Q normal probability plot

Descriptive Statistics

list	list values of variables
graph	graph Y_1 Y_2 Y_3 X , xlabel ylabel xscale t1(" ") b1(" ") title(" ") Saving(" ")
hist	histogram
summarize	display summary statistics
table	multiway tables
tabulate	one and two way frequency
etab	One variable against many
codebook	Simple standard tables

General Statistics

anova	Analysis of Variance
correlate	Correlation
oneway	One-way analysis of variance
ranksum	Wilcoxon ranksum test
tabulate	crosstables incl. tests of homogeneity, gamma, exact r*c test
table	Summary statistics in tables
ttest	Mean comparison for small samples

Cohort / survival analysis

stmh	Manten-Haenzel rates
poisson	poisson regression
strate	table of rates
stset	define survival time rates
staalen	Aalen cumulative hazards
stcox	Cox regression (+stphtest)
stsplit	expand data according to a lexis diagram
sts graph	Kaplan Meyer survival functions in graph
sts test	Logrank test

Case Control / Cross Sectional analysis

mhodds	Mantel Haenzel Odds Ratios
partgam	Partial Gamma Coefficient
tabodds	Table of Odds

Regression Models

clogit	conditional logistic regression
logit	logistic regression (output on log scale)
logistic	logistic regression (output on antilog scale)
lrtest	likelihood ratio test
poisson	poisson regression
predict	obtain prediction and indicators of modelfit
regress	linear regression

Categorial variables- "dummy" indicator:
xi: regress Y i.varname

```
sort X              /* strafify on X */
by X: command variables, options
for var variables: command X ....
```