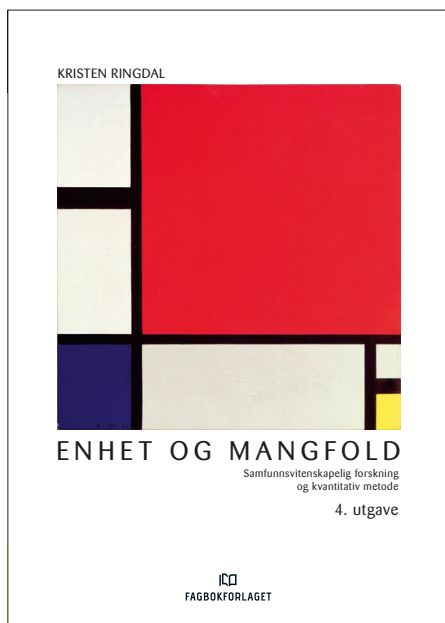


En kort innføring i Stata

Vedlegg til Kristen Ringdal: *Enhet og mangfold*, 4. utgave



Lær deg Stata (Ringdal & Wiborg, 2017) gir en mer omfattende innføring i Stata med oppskrifter på de fleste dataanalysene i *Enhet og mangfold*.

Om Stata

Stata er en programpakke for analyse av kvantitative data. Den viktigste konkurrent i Norge er SPSS som ble lansert allerede i 1960-årene, mens Stata er av nyere dato. Den første versjonen for PC ble lansert i 1985. Brukergrensesnittet har blitt gradvis utviklet fra å være rent kommandobasert til å gjøre mer bruk av menyer og dialogbokser.

Et av Statas fortrinn framfor SPSS er at nye programmer laget av andre brukere lett kan installeres ved hjelp av kommandoen `ssc install`. Denne type programmer kan en finne ved å søke på Internett eller i brukerforumet STATALIST og i *Stata Journal*.

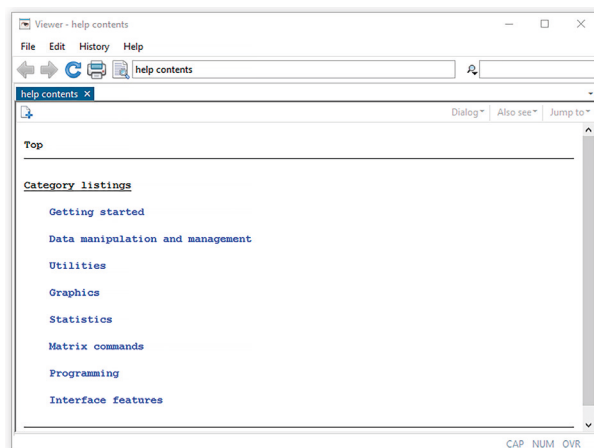
En annen styrke ved Stata er at nye statistiske teknikker blir tidlig lagt inn i Stata eller er tilgjengelig som brukerprogrammer som kan installeres i ettertid. Dette har likevel mest interesse for avanserte brukere. Alle brukere vil etter hvert sette pris på de gode mulighetene for å lage grafiske framstillinger og ikke minst kommandoen `margins` som på en enkel måte kan beregne predikerte verdier fra de fleste statistiske teknikker og framstille dem i diagrammer.

Bruk av skrifttyper

I tillegg til den vanlig teksten er kommandoer, filnavn og variabelnavn, samt utskrifter trykt i Courier New. Menyvalg og nødvendige trykk på tastaturet er gjengitt med rød skrift.

Installering av Stata

Stata installeres fra `setupStata 15.exe`. Denne kan komme på en CD eller lastes ned. Den kan starte automatisk fra CD ellers må en dobbeltklikke på filnavnet. Når du kommer til **Select Executable**, velges den variant av Stata du har lisens for. Den første gang Stata startes vil det spørres etter lisens- og aktoveringsnøkkelen. Hvis denne godtas vil du få spørsmål om du vil undersøke om, det finnes oppdateringer av Stata. Trykk **OK** og følg instruksjonene på skjermen. Dermed er det gjort.



Figur 2 «Help contents»-vinduet

Filtyper i Stata

Datafiler eller datasett som `abu89.dta` har `dta` som «etternavn», men filene blir normalt gjenkjent uten etternavnet. Do-filer er samlingen av kommandoer eller programmer som kan utføres fra denne type filer. Do-filer bør brukes av alle som trenger en dokumentasjon på det som er utført, for eksempel omkodinger og statistiske analyser. Stata grafer har etternavn `gph`. Det finnes mange eksempler, slik som histogrammer og spredningsdiagrammer, i *Enhet og mangfold*. Det er også mulig å lagre det som skrives til resultatvinduet i en log-fil, men jeg synes ikke dette er særlig nyttig. Det er bedre å kopiere det en vil beholde fra resultatvinduet til en word-fil eller liknende.

Åpne og beskrive et datasett

Filer kan åpnes og lagres ved hjelp av de to første ikonene eller ved bruk av filmenyen: **File** → **Open** og **File** → **Save**. En åpner et datasett ved først velge **File** → **Open**, deretter lete seg fram til mappen og velge det datasett en ønsker å åpne, for eksempel `abu89.dta`. Når Stata installeres følger noen datafiler som benyttes i eksempler med. Disse åpnes med kommandoen `sysuse`. Datasettet `auto` åpnes med å skrive kommandoen `sysuse auto`. Kommandoen `sysuse`

dir eller menyvalgene **File → example datasets**, gir en oversikt over denne type filer. Stata gjør det også mulig å åpne filer som er lagret andre steder, for eksempel på Fagbokforlagets nettside. Denne kommandoen ble benyttet til å åpne datasettet abu89:

```
use "https://nedlasting.fagbokforlaget.no/stata/abu89.dta", clear
```

Hvis vi ønsker å lage et nytt datasett for å registrere data direkte i Stata, må Stata først startes uten noen datafil. Deretter må «Data Editor» åpnes. Dette gjøres fra Data-menyen: **Data → Data Editor → Data Editor (Edit)**. Alternativt kan vi trykke på det andre ikonet med en blyant (rett under statistikkmenyen). Dette åpner en tom datamatrikse der vi kan skrive inn data direkte. Variablene gir automatisk navn: var1, var2 osv. Disse kan endres i egenskapsvinduet, der også verdietiketter kan legges til. Mer at det skilles mellom store og små bokstaver i variabelnavn. Det vil si at var1 og var1 er to ulike variabelnavn. Derfor anbefales det helst å benytte bare små bokstaver i variabelnavn. Hvis en fil er importert fra en annen programpakke, kan det være nyttig å sørge for at alle variabelnavn endres til små bokstaver slik:

```
rename all, lower
```

I figur 1 gir variabelvinduet en liste over variablene i filen. Det er flere måter å enkelt beskrive disse på. Menyvalgene **Data → Describe data → Describe data in memory or in a file** åpner en dialogboks som gir flere valgmuligheter. Siden filen har få variabler, er det bare å trykke på **OK**. Resultatet er utskriften nedenfor som gir en kort oversikt over variablene i datafilen.

```
. describe

Contains data from https://nedlasting.fagbokforlaget.no/stata/abu89.dta
obs:      4,127
vars:      9                      4 Jan 2018 12:38
size:     297,144                  (_dta has notes)
```

variable name	storage type	display format	value label	variable label
io_nr	double	%10.0g		IO-nummer
time89	double	%10.0g		Gjennomsnittlig timelønn 1989
ed	double	%10.0g		År utdanning
age	double	%10.0g		Alder
female	double	%10.0g		Respondentens kjønn
klasse89	double	%28.0g	klasse89	Goldthorpe klasse 1989
promot	double	%11.0g	promot	Noen gang forfremmet
fexp	double	%10.0g		Bedriftserfaring
private	double	%10.0g	private	Privat sektor

Kommandoen `describe` gjør det samme. Prøv å skrive denne i kommando-vinduet. Hvis filen har mange variabler, kan vi velge de vi vil beskrive i dialogboksen, eller i en kommando:

```
describe time89-private /* time89 til private */

describe time89 ed /* time89 og ed */
```

Siste del av linjene viser hvordan forklarende tekst (kommentarer) kan legges til en kommando.

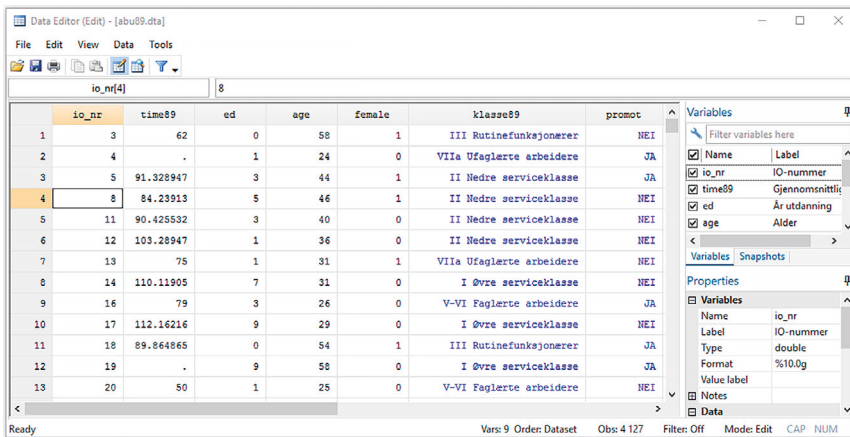
Det framgår av utskriften at datasettet inneholder «notes», det vil si notater som kan gi informasjon i tillegg til variablene. Disse vises enklest ved å skrive kommandoen `notes` som utføres ved å trykke på **ENTER**. Utskriften nedenfor viser både kommandoen og de tre notatene.

```
. notes

_dta:
1. Data fra Arbeidstaker- og bedriftsundersøkelsen 1989. Statistisk Sentralbyrå
2. Torp, H. and K. H. Skollerud (1990). Organisasjon, arbeidsmiljø og mobilitet: resultater fra Arbeids- og bedriftsundersøkelsen. Oslo, Institutt for samfunnsforskning.
3. Faye, A. (1992). Arbeidsmiljø 1989. Oslo, Statistisk sentralbyrå.
```

Se på og endre en datafil

Det seg gjøre både å se på innholdet i en datafil og gjøre endringer. Et trykk på ikonet med blyant midt under statistikkmenyen åpner vinduet i figur 1.3. Der kan vi både se på og endre data. Vi kan merke oss at for variablene som har verdietiketter slik som `klasse89`, vises bare etikettene og ikke tallkodene. Videre ser vi at «.» angir manglende data. Det er mulig å skille mellom ulike typer manglende data ved å legge til en bokstav, for eksempel «.d»



Figur 3 Stata Data Editor(Edit)

Å lage en enkel tabell i Stata

I *Enhet og mangfold* finnes det oppskrifter på hvordan tabeller lages, hvordan korrelasjonsanalyser gjøres, og hvordan en enkel varians- eller regresjonsanalyse kan utføres i Stata. Her skal vi bare gjennomgå hvordan enkle tabeller kan bestilles fra datafilen `abu89`. La oss se på fordelingen på kjønn. Velg **Statistics** → **Summaries, tables, and tests** → **Frequency tables** → **One-way table** for å åpne «tabulate1»-dialogboksen der en kan velge variabler, for eksempel `female`. Deretter trykker vi på **OK**, og frekvenstabellen for kjønn blir sendt til resultatvinduet gjengitt i utskriften.

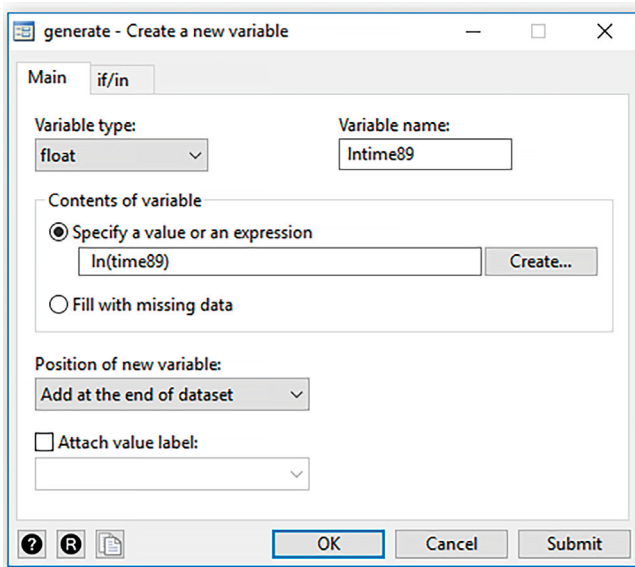
```
. tabulate female
```

Respondent ns kjønn	Freq.	Percent	Cum.
0	2,193	53.14	53.14
1	1,934	46.86	100.00
Total	4,127	100.00	

Øverst i utskriften gjengis kommandoen som ble generert av menyvalgene. Deretter følger tabellen som viser at variabelen har verdien 0 og 1. Verdietiketter mangler, men kvinner har koden 1 og menn koden 0. Den neste kolonnen viser antall menn og kvinner. Deretter følger den vanlige og den kumulative prosentfordelingen. Utskriftene har en tendens til å bli lange, og derfor er det oftest ikke hensiktsmessig å lagre filen i sin helhet. En enkel måte er å velge tabellen(e) en vil kopiere, høyreklikke, og velge **Copy as picture**. Tabellen kan så limes inn i Word. Den ser ut som på skjermen, men den kan ikke redigeres. En måte å få dette til på er å lage en tabell i Word, med tilstrekkelig antall rader og kolonner. I vårt tilfelle fungerer det bra med 4 kolonner og 7 linjer (siden de horisontale strekene krever en linje hver). Marker så hele tabellen og velg lim inn. På lignende måte kan tabeller kopieres til Excel. Noen ganger er det enklest bare å kopiere linjene med tall fordi resultatet i kompliserte utskrifter ellers kan bli rotete.

Kommandoer for redigering av variabler

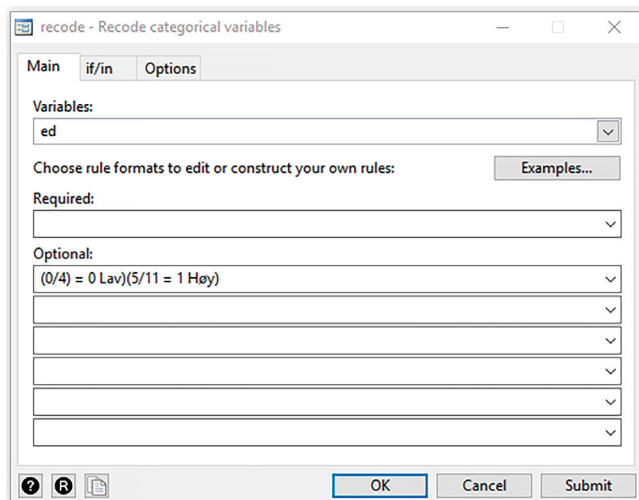
Her skal vi se på hvordan vi kan endre på variablene i en datafil. De to mest interessante kommandoene er `generate` og `recode`. Den første endrer eller lager nye variabler ved hjelp av matematiske operasjoner. La oss se hvordan vi kan lage ny variabel `lntime89`, som den naturlige logaritmen av `time89`, timelønn i 1989. Figur 4 viser hvordan «Generate»-dialogboksen må fylles ut for å lage den nye variabelen. Dialogboksen åpnes ved å velge: **Data → Create or change data → Create new variable**. For å lage uttrykket som lager den nye variabelen må vi trykke på **Create** og dobbeltklikke på **ln()** som matematisk funksjon under **Category** i **Expression Builder**, og fylle `time89` inn i parentesene. Trykk **OK** for å komme tilbake til «generate»-dialogboksen og trykk **OK** for å lage den nye variabelen. Deretter kan variabeletikett legges til ved å velge `lntime89` i Variabelvinduet og fylle inn etiketten i «Label»-linjen i egenskapsvinduet.



Figur 4 «Generate»-dialogboksen

«Recode»-dialogboksen kan benyttes til å omkode en variabel under samme navn, eller lagre endringen i en ny variabel. Variabelen `ed` er antall år ut over obligatorisk utdanning. Verdiene varierer fra 0 til 11. Dette kan bekreftes ved å lage en frekvenstabell på samme måte som for kjønn. La oss lage en ny utdanningsvariabel, `ed2`, som bare skiller mellom lav (0–4 år) og høy utdanning (5–11 år). Vi åpner dialogboksen ved å velge **Data → Create or change data → Other variable-transformation commands → Recode categorical variable**. På figur 5 er `ed` valgt som variabelen som skal omkodes. I linjen midt i figuren er omkodingen definert: $0-4 = 0$, og $5-11 = 1$. De nye verdiene har også fått etiketter. Hvis vi nå trykker på **OK** eller **Submit** vil de nye verdiene erstatte de originale, mens vi ønsker å beholde den gamle variabelen og lage en ny omkodet variabel. Dette gjøres ved å velge «Options»-fanen og skrive inn den nye variabelen, «`ed2`», under **Generate new variables**. For å utføre kommandoen trykker vi **OK**. Det er alltid lurt å kontrollere at omkodingen er korrekt. Dette kan gjøres ved å lage en krysstabell slik som vist i utskriften. I «`tabulate`»-kommandoen er den originale variabelene definert som linjevariabel og den omkodede som kolonnevariabel. Kommandoen kan også utføres

fra Statistikk-menyen: **Statistics** → **Summaries, tables and tests** → **Frequency tables** → **Two way table with measures of association**. I dialogboksen velges **ed** som **Row variable** og **ed2** som **Column variable**.



Figur 5 «Recode main»-dialogboksen i Stata inn her

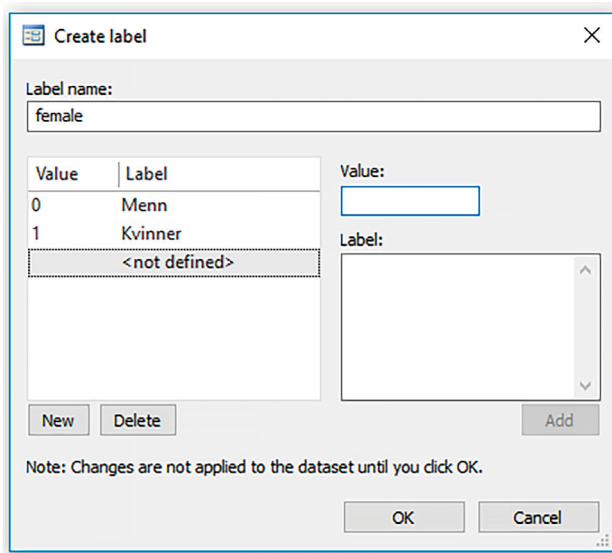
```
. tabulate ed ed2
```

År utdanning	RECODE of ed (År utdanning)		Total
	Lav	Høy	
0	839	0	839
1	1,156	0	1,156
3	1,121	0	1,121
5	0	483	483
7	0	308	308
9	0	205	205
11	0	15	15
Total	3,116	1,011	4,127

Legge til verdietiketter (value labels)

Variabelen `female` (kjønn) manglet verdietiketter. Vi har allerede sett hvordan verdietiketter enkelt kan legges til «recode»-kommandoen, men framgangsmåten er mer tungvint for eksisterende variabler. I variabelvinduet må en først velge `female`. I egenskapsvinduet (Properties) er «Value labels»-feltet for `female` tomt. Merk at før vi kan gjøre endringer er det nødvendig å åpne

hengelåsen øverst i egenskapsvinduet. Et trykk på «...» åpner «Create label»-dialogboksen i figur 6. Først må «Label name» defineres. Det er lurt å benytte samme navn som variabelen. Deretter fyller vi inn «0» under **Value:**, skriver «Menn» under **Label:** og trykker på **ADD**. Gjenta for verdien «1» med «Kvinner» som etikett, trykk på **ADD**, og avslutt med **OK**. Dermed er verdietikettene definert, men det gjenstår å tilordne dem variabelen `female`.



Figur 6 «Create label»-dialogboksen

Etter at dialogboksen er lukket, må en trykke på haken i «Value label»-feltet og velge «female» som verdietikettsett. For å kontrollere at dette virker, kan en skrive denne kommandoen i Kommandovinduet og avslutte med **Enter**.

```
tabulate female
```

I den nye tabellen vises bare verdietikettene. Denne kommandoen viser bare tallkodene:

```
tabulate female, nolabel
```

Det lar seg også gjøre å legge tallkodene til verdietikettene. Den første kommandoen gjør det for alle variabler, den andre gjør det bare for kjønn, og den tredje fjerner tallkodene.

```
numlabel, add
numlabel female, add
numlabel, remove
```

Do-filer i Stata

Denne filtypen benyttes til å lagre kommandoer som kan utføres fra filen. Do-filer gjør det mulig å utføre programmer som streker seg over mange linjer. Do-filer fungerer også som en dokumentasjon av gjennomførte omkodinger og analyser og som et redskap for å gjenskape disse hvis analysefilen av en eller annen grunn går tapt.

La oss se hvordan kommandoene som er utført i dette kapitlet kan lagres i en do-fil. Først må en ny do-file åpnes. Dette gjøres ved å trykke på det første ikonet med en blyant i verktøylinjen. Dette åpner **Do-file Editor** med en tom fil som automatisk blir gitt navnet `untitled1.do`. Deretter kopierer vi de utførte kommandoene fra «Review»-vinduet. Det kan lønne seg å slette de som ikke skal med i do-filen først. Velg en eller flere linjer i do-filen og utfør dem ved å trykke på det siste ikonet med en blå trekant i verktøylinjen for å utføre kommandoen(e) («Execute selection»).

Do-filen har ingen eksempler på en kommando som strekker seg over flere linjer, men her er et eksempel i form av en omkoding av alder:

```
recode age (16/19=1 16-19) (20/29=2 20-29) (30/39=3 30-39) ///
      (40/49=4 40-49) (50/59=5 50-59) (60/74=6 60-74), gen(age6)
```

Første linje avsluttes med de obligatoriske tre skråstrekene «///» som viser at kommandoen fortsetter i neste linje.

```

* Do-fil med Stata-kommandoer
use "https://nedlasting.fagbokforlaget.no/stata/abu89.dta", clear
* Oversikt over variablene i filen
describe /* alle variabler */
describe time89-private /* time89 til private */
describe time89 ed /* time89 og ed */
* Frekvenstabell for kjønn
tabulate female
* Lage en ny variabel med generate
generate lntime89 = ln(time89)
* Omkode ed (utdanning i år) til ed2 med to verdier
recode ed (0/4 = 0 Lav)(5/11=1 Høy), generate(ed2)
tabulate ed ed2
** Definere verdietikettsettet female
label define female 0 "Menn" 1 "Kvinner"
* Knytte verdietikettsettet til variabelen female
label values female female
tabulate female
tabulate female, nolabel
* Legge tallkodene til verdietikettene for female
numlabel female, add
tabulate female
* Legge tallkodene til verdietikettene for alle variabler
numlabel, add
tabu klasse89
* Fjerne tallkodene fra verdietikettene
numlabel, remove
* Beskrivende statistikk for alle variablene
summ

```

Legge til kommentarer i Stata

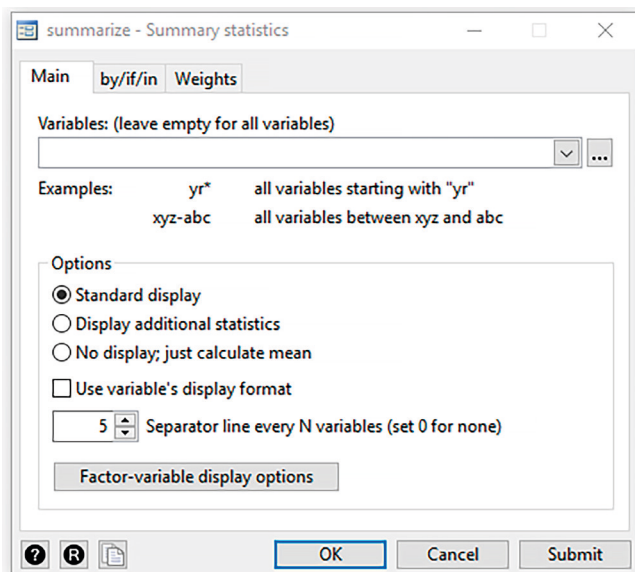
Etter at kommandolinjene er limt inn i do-filen, er det nyttig å legge til kommentarer som forteller hva kommandoene utfører. Alle linjer som starter med «*» blir oppfattet som kommentarer. Disse vises med grønn skrift, mens kommandoene har blå skrift. Kommandoer i en linje omslutes på en spesiell måte:

```
describe /* alle variabler */
```

Merk at kommentaren starter med «/*» og avsluttes med de to regnene i motsatt rekkefølge. Det er flere eksempler på begge typer kommentarer i tekstboksen.

Beskrivende statistikk

Vi skal avslutte med å se på hvordan vi kan lage en tabell med alle eller utvalgte variabler i en datafil. Den aktuelle kommandoen i Stata er `summarize` eller bare `summ`. Do-filen gjengitt i tekstboksen avsluttes med to «summarize»-kommandoer. Den første uten variabelliste gir beskrivende statistikk for alle variablene. Dette kan også gjøres fra menyene. Menyvalgene **Statistics** → **Summaries, tables and tests** → **Summary and descriptive statistics** → **Summary statistics** åpner «summarize»-dialogboksen (figur 7). Under **Variables:** kan en velge variabler. Hvis vi ikke velger noen, blir det laget en tabell for alle variablene. Dette gir den første «summaries»-kommandoen i do-filen. Den siste kommandoen i do-filen gir beskrivende statistikk for alle variablene unntatt den første. Denne er gjengitt i utskriften nedenfor.



Figur 7 «Summarize»-dialogboksen

```
. summarize time89 - lntime89
```

Variable	Obs	Mean	Std. Dev.	Min	Max
time89	3,759	90.14948	30.31473	25	343.75
ed	4,127	2.689605	2.557027	0	11
age	4,127	39.65084	12.35712	16	74
female	4,127	.4686213	.4990749	0	1
klasse89	4,042	3.021029	1.186329	1	5
promot	4,127	.3777562	.484885	0	1
fexp	4,127	.9450933	.9065842	0	4.9
private	4,127	.6118246	.4873939	0	1
lntime89	3,759	4.452624	.3076192	3.218876	5.839915