Data og beskrivende statistikk – Introduksjon til SPSS

7. april 2005

Tron Anders Moger

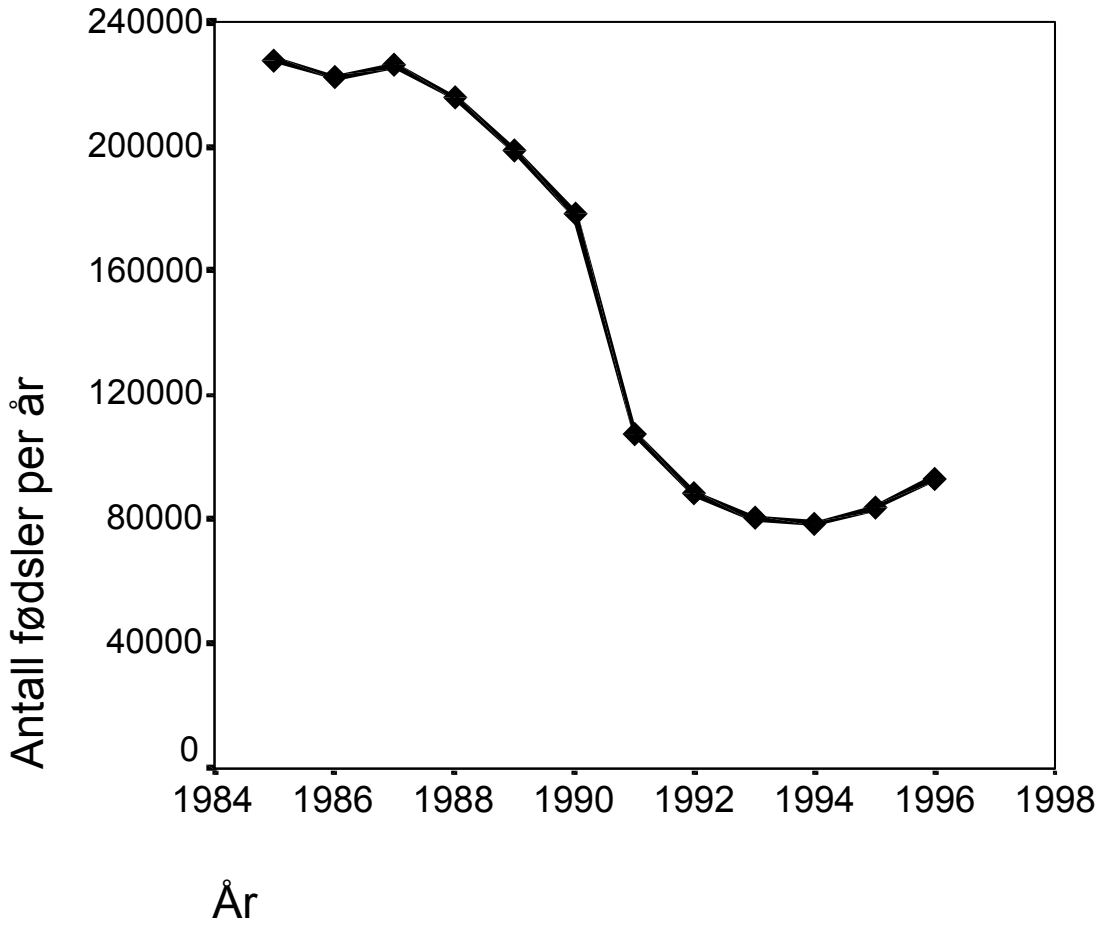
New England Journal of Medicine, Editorial, Jan. 6, 2000, p. 42-49

- The eleven most important developments in medicine in the past millennium
  - Elucidation of human anatomy and physiology
  - Discovery of cells and their substructures
  - Elucidation of the chemistry of life
  - *Application of statistics to medicine*
  - Development of anesthesia
  - Discovery of the relation of microbes to disease
  - Elucidation of inheritance and genetics
  - Knowledge of the immune system
  - Development of body imaging
  - Discovery of antimicrobial agents
  - Development of molecular pharmacotherapy

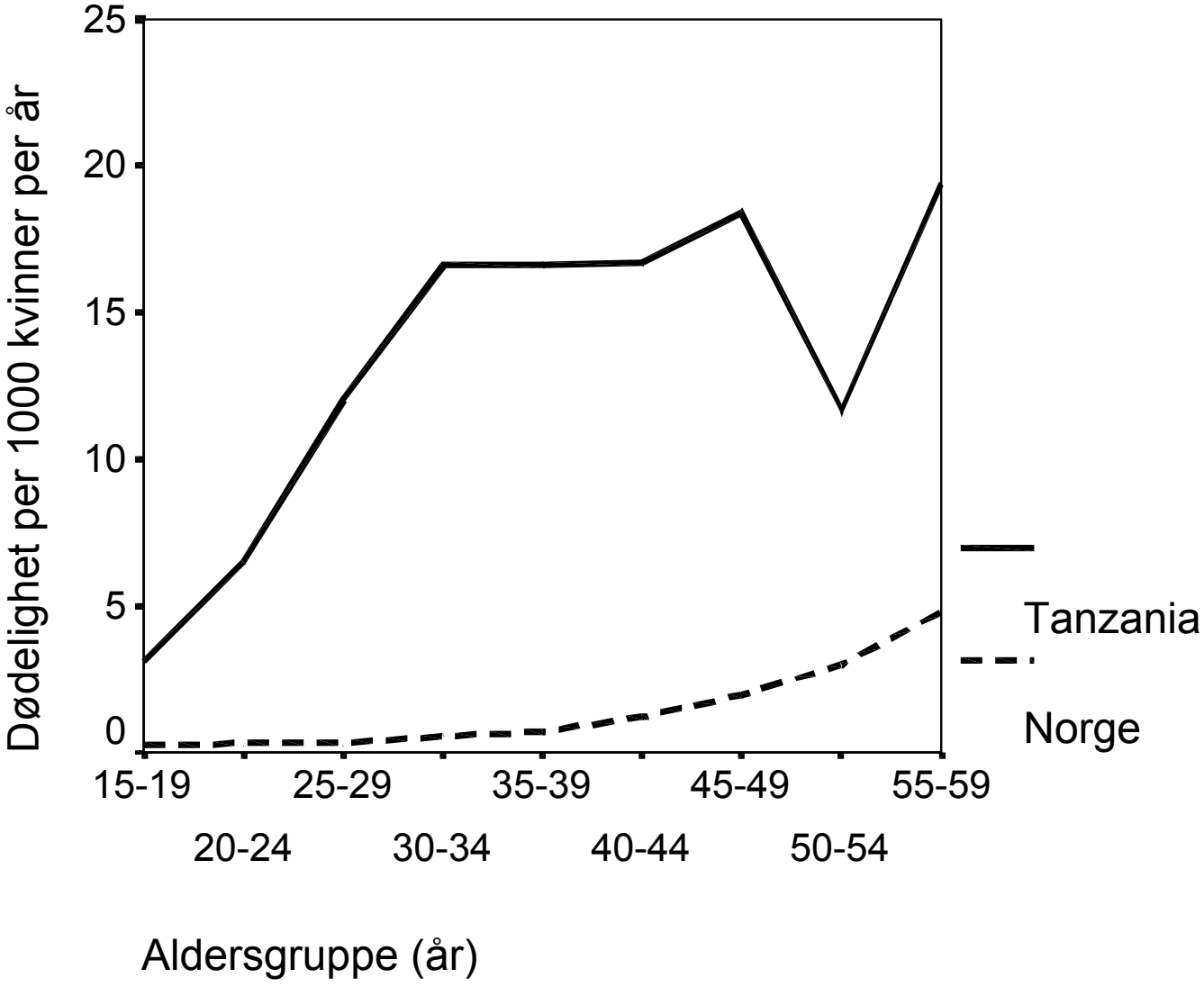
Introduksjon

- Kunnskap om verden kommer ofte via tall og data. Hvordan forholde seg rasjonelt til kvantitativ informasjon?
- Problemene i en kvantitativ tilnærming undervurderes ofte.
- Må fremme “numerical literacy” - evnen til å forstå tall og kvantitative forhold.

Antall fødsler i tidligere Øst-Tyskland



Dødelighet i Tanzania og i Norge



Medisinsk forskning og tall

- Medisinsk forskning, slik den utføres idag, frembringer nesten alltid tall.
- Tallene er ofte usikre
- Tallene må organiseres for at en skal forstå hva de sier
- En ønsker ofte å *generalisere* fra tallene

Statistiske data

Statistiske data kommer fra:

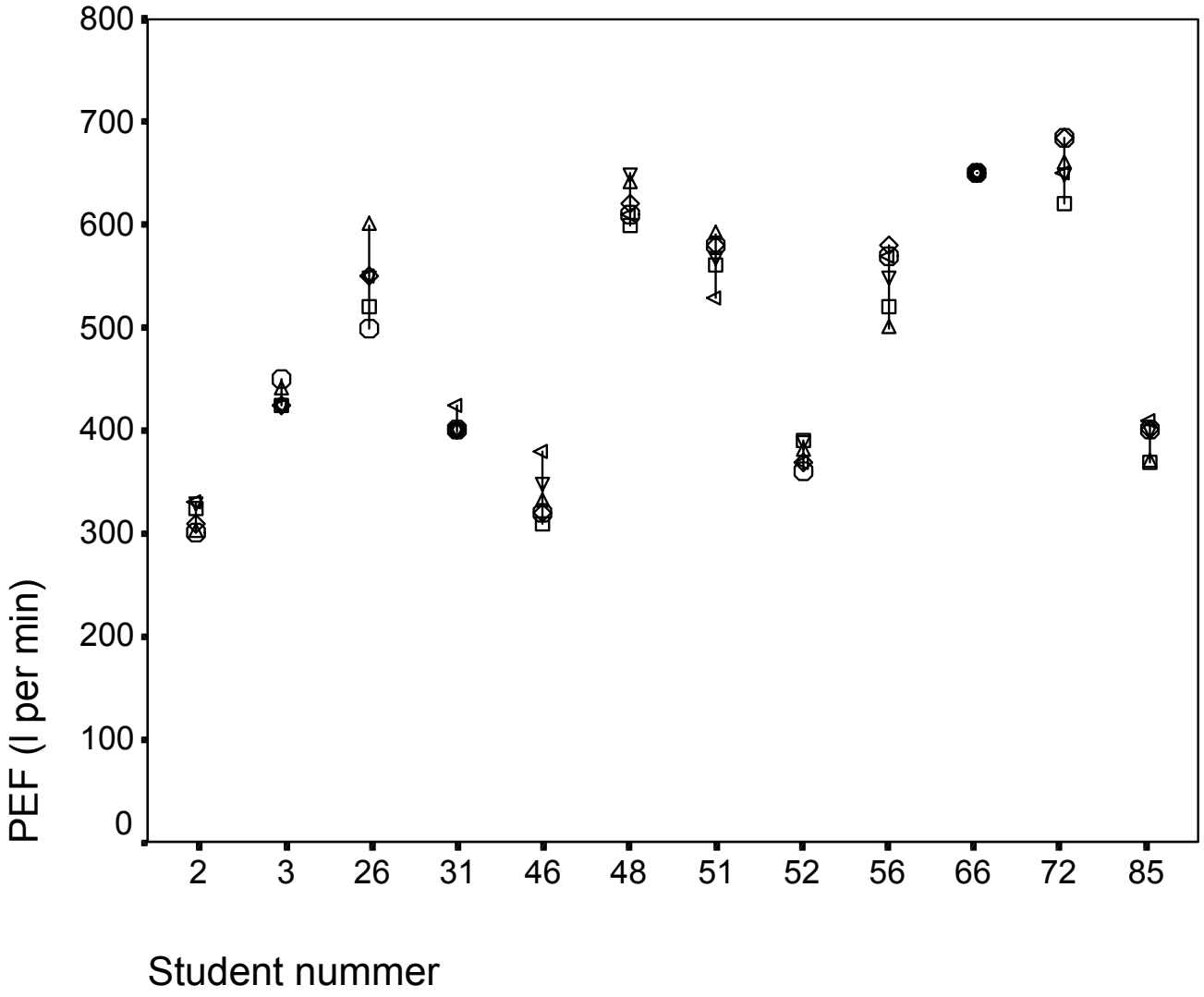
- *Måling* (kontinuerlige data) med et instrument på en skala (naturvitenskapelig eller 'mykere'). Eksempler:
  - Feber: 39.6 (Uproblematisk)
  - IQ: 116 (Problematisk)
- *Kategorisering* (kategoriske data). Eksempler:
  - mann / kvinne (Uproblematisk)
  - deprimert / ikke deprimert (Problematisk)

Usikkerhet i data

- Reliabilitet: Hvor presise er dataene? Hvor mye kan de endres hvis observasjonen gjentas?
- Validitet: Måler vi faktisk det vi ønsker å få informasjon om? Er målingen relevant?

Reliabilitet av PEF-målinger

6 målinger fra hver av 12 stud.



Reliabilitet av spørreskjema/intervju

- Undersøkelse om alkoholbruk (menn 31-50 år):
  - Gjennomsnittlig antall ganger de som sier at de har brukt alkohol siste år, oppgir at de har følt seg beruset:

- 1993 (spørreskjema): 14.1 berus. pr. år
- 1994 (MMI-intervju): 7.3 berus. pr. år

I 1994 ble det spurt om "tydelig beruset", ellers samme ordlyd.

Reliabilitet av klinisk undersøkelse

- Tatt fra Sackett et al: Clinical Epidemiology (Little, Brown and Company, 1985). Bilder av øyebunnen hos 100 pasienter vurderes av to klinikere mhp forekomst av retinopati

		Annen kliniker	
		Intet/lite	Moderat/mye
Første kliniker	Intet/lite:	46	10
	Moderat/mye:	12	32

Observert overensstemmelse:

(46+32)/100 =78%

Kilder til variasjon i data

- Laboratorievariasjon
- Observatørvariasjon
- Instrumentvariasjon
- Måleusikkerhet
- Biologisk variasjon mellom individer
- Dag til dag-variasjon hos ett individ

Generalisering

- *Utvalg:* De enheter, individer, eksperimenter som inngår i studien. Eksempler:
  - 15 pasienter med migrene
  - nevrofysiologisk studie på rotter
- *Populasjon:* Den samling av enheter etc. en ønsker å generalisere til
  - alle pasienter med migrene
  - alle gjentagelser av samme nevrofysiologiske forsøk

Begreps-par

- Utvalg
  - histogram
  - gjennomsnitt
  - andel syke
  - målt kolesterol
  - vær
- Populasjon
  - sannsynlig-hetsfordeling
  - forventning
  - risiko
  - kolesterolnivå
  - klima

Typer av data:

- Kontinuerlige data. Data som er målt på en kontinuerlig skala, f.eks. høyde, vekt, alder.
- Kategoriske data. Data som bare kan anta et endelig antall verdier, f.eks. kjønn, utdanningsnivå, alder inndelt i grupper. Eller, hvis data er samlet inn på flere sykehus, ønsker man en variabel som sier hvilket sykehus dataene er fra.

Innlegging av data i SPSS (og andre statistikkpakker):

- VIKTIG: En linje i datafilen svarer alltid til *ett* individ!
- Ny variabel opprettes enten ved og velge *Data->Insert variable* i *Data View*-vinduet, eller ved å skrive inn navnet på variabelen under *Name* i *Variable View*-vinduet
- Vanlig å ha en variabel med id-nummeret til hvert individ først
- Hvis dere mangler en måling på et individ, ikke skriv inn noe i cellen

Koding av data:

- For kontinuerlige data-variabeler skriver man inn verdiene i cellene
- For kategoriske variabeler, må man bestemme seg for en kategorisering: Eks. 0=mann og 1=kvinne, eller 0=grunnskole, 1=videregående og 2=universitetsutdannelse
- I *Variable View* kan verdiene med tilhørende definisjoner legges inn under *Values*
- Under *Label* kan dere gi mer informasjon om variabelen enn bare navnet

Beskrivende statistikk

- Tabeller
- Grafiske fremstillinger
- Sentralmål
- Variasjonsmål
- Epidemiologiske mål (insidens og prevalens, som jeg nevner kort til slutt)

Typer av grafisk fremstilling

- Histogram
- Box-plott
- Spredningsdiagram
- Insidenskurve
- Overlevelseskurve

Alder til 100 medisinerstudenter

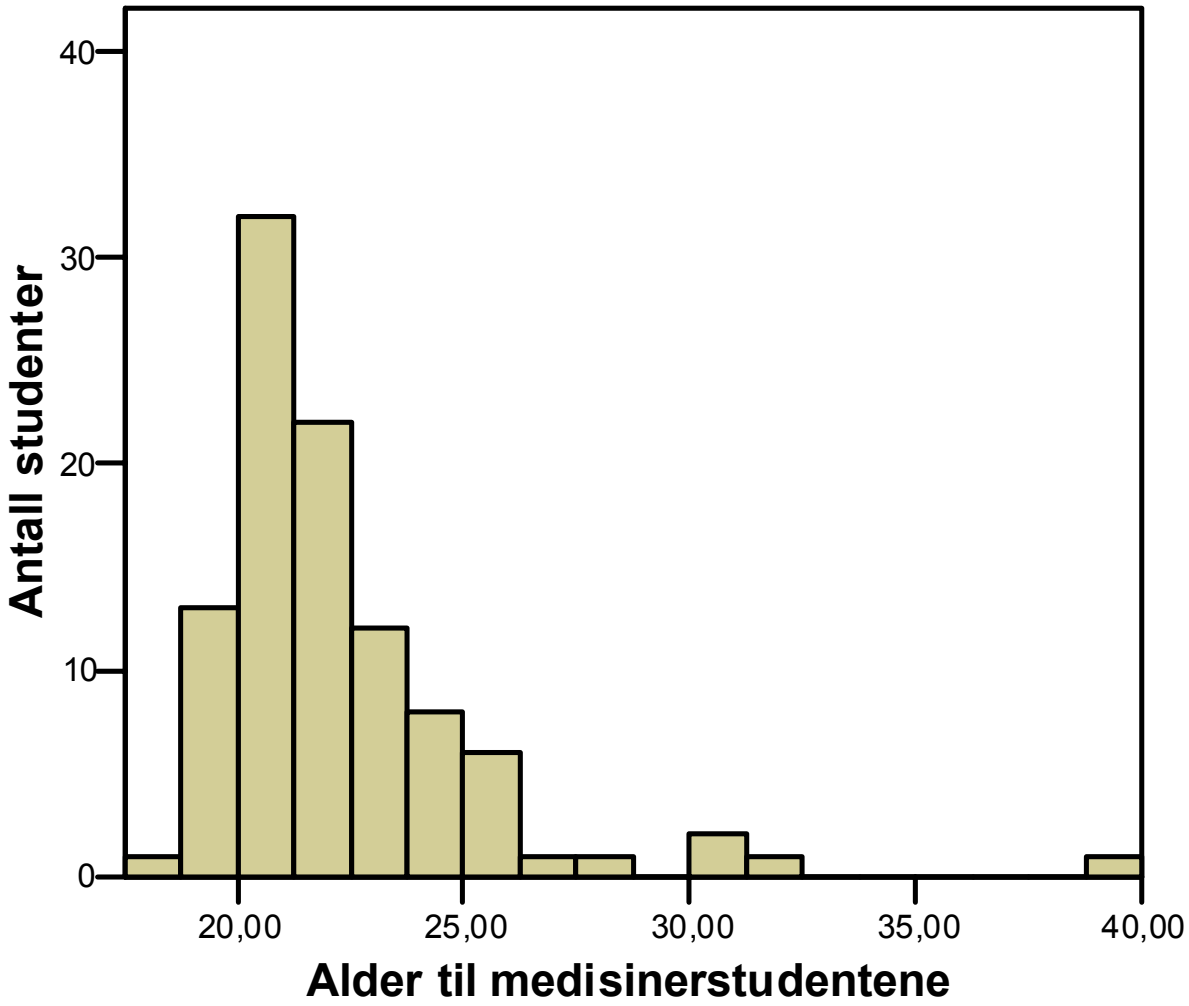
24	21	22	26	26
22	21	19	23	21
20	24	27	19	30
24	22	21	22	20
19	23	20	20	23
21	22	22	21	20
24	22	22	22	23
21	23	19	20	23
20	25	26	22	21
22	20	22	21	20
20	19	19	23	23
22	20	21	22	19
21	22	20	23	22
22	21	20	19	24
26	22	19	21	24
22	23	22	19	21
21	24	21	19	39
31	21	18	24	21
22	23	19	26	32
22	21	23	19	28

Hvordan få oversikt over dataene i SPSS? Explore!

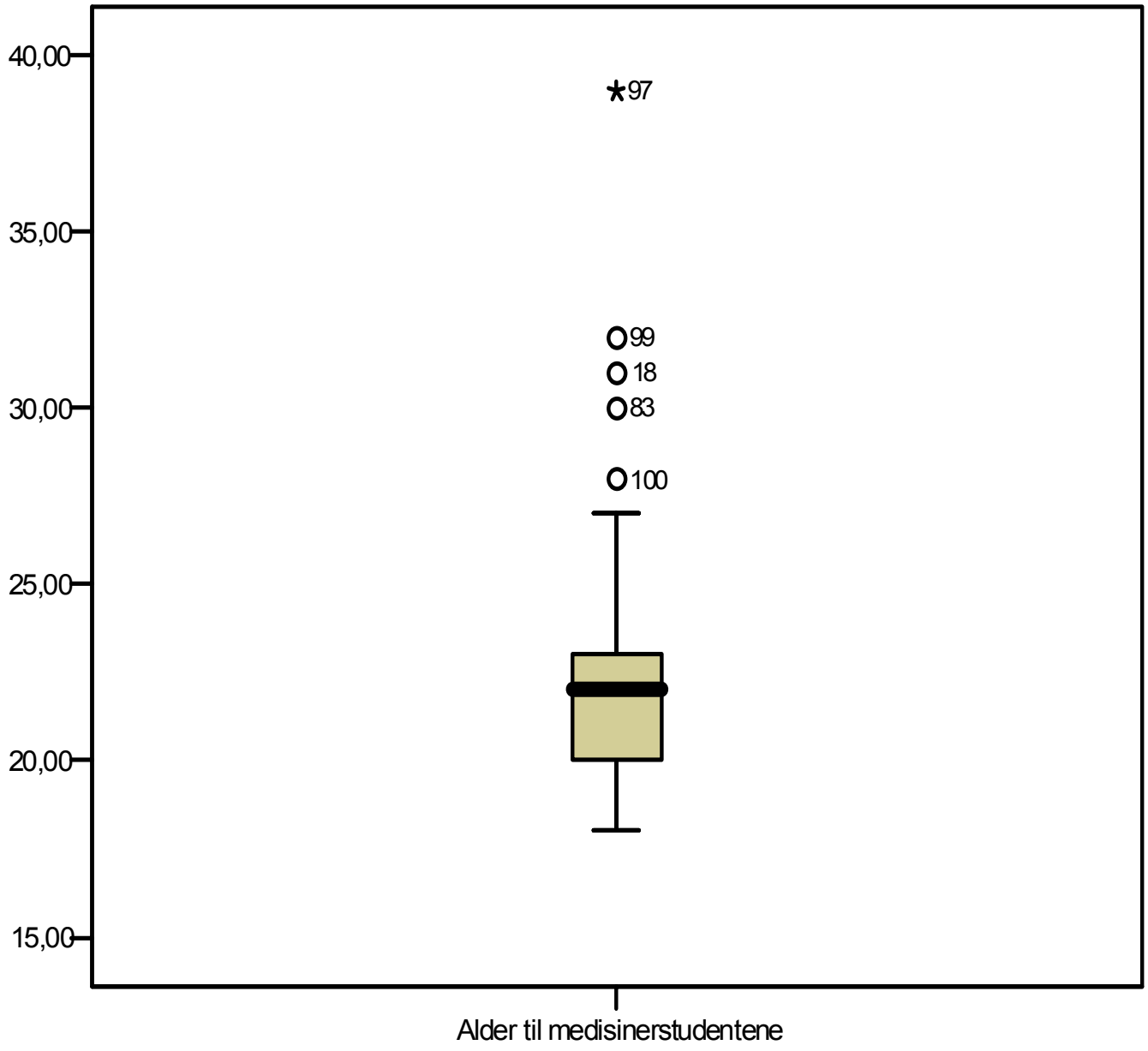
- Beskrivende analyse kan utføres på følgende måte:
  - Klikk *Analyze - Descriptive Statistics - Explore.* Merk av de relevante variablene og overfør dem til *Dependent List*. Klikk på *Plots,* fjern krysset ved "Stem and leaf" og sett i stedet et kryss ved "Histogram". Klikk på *Continue* for å forlate menyen. Klikk så på *OK* for å få jobben utført

Histogram: Fordeling av alder blant nye medisinerstudenter (n=100)

Studenter fra Med.Fak, kull H98.



Box-plott: fordeling av alder blant nye medisinerstudenter



Sentralmål

- Gjennomsnitt

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

  Studentene: 22.2 år

- Median

  Den midterste observasjonen når utvalget er ordnet i stigende rekkefølge

  Studentene: 22.0 år

- Gjennomsnittet påvirkes av ekstreme observasjoner. Medianen er robust.

Variasjonsmål

- Standardavvik

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Studentene: 3.06 år

- Fraktiler

25% fraktilen er den verdien der
25% av observasjonene er lavere
og 75% av observasjonene høyere

(I SPSS: Kryss av på *Percentiles* under *Statistics* i *Explore*)

Studentene:
25% fraktilen: 20.0 år
75% fraktilen: 23.0 år

Hva hvis man vil omkode alder til en kategorisk variabel? Recode!

- Noen ganger har man data som måles på en kontinuerlig skala, men som i praksis benyttes som kategoriske data (Eks. en måling fra 0-20, hvor de som scorer 0-10 har lav risiko, 10-15 middels risiko, 15-20 høy risiko)
- Velg *Transform->Recode->Into different variables*
- Flytt alder over til høyre i vinduet. Skriv inn navnet på den nye variabelen under *Output variable*. Klikk på *Old and New Values*. Et nytt vindu kommer opp.

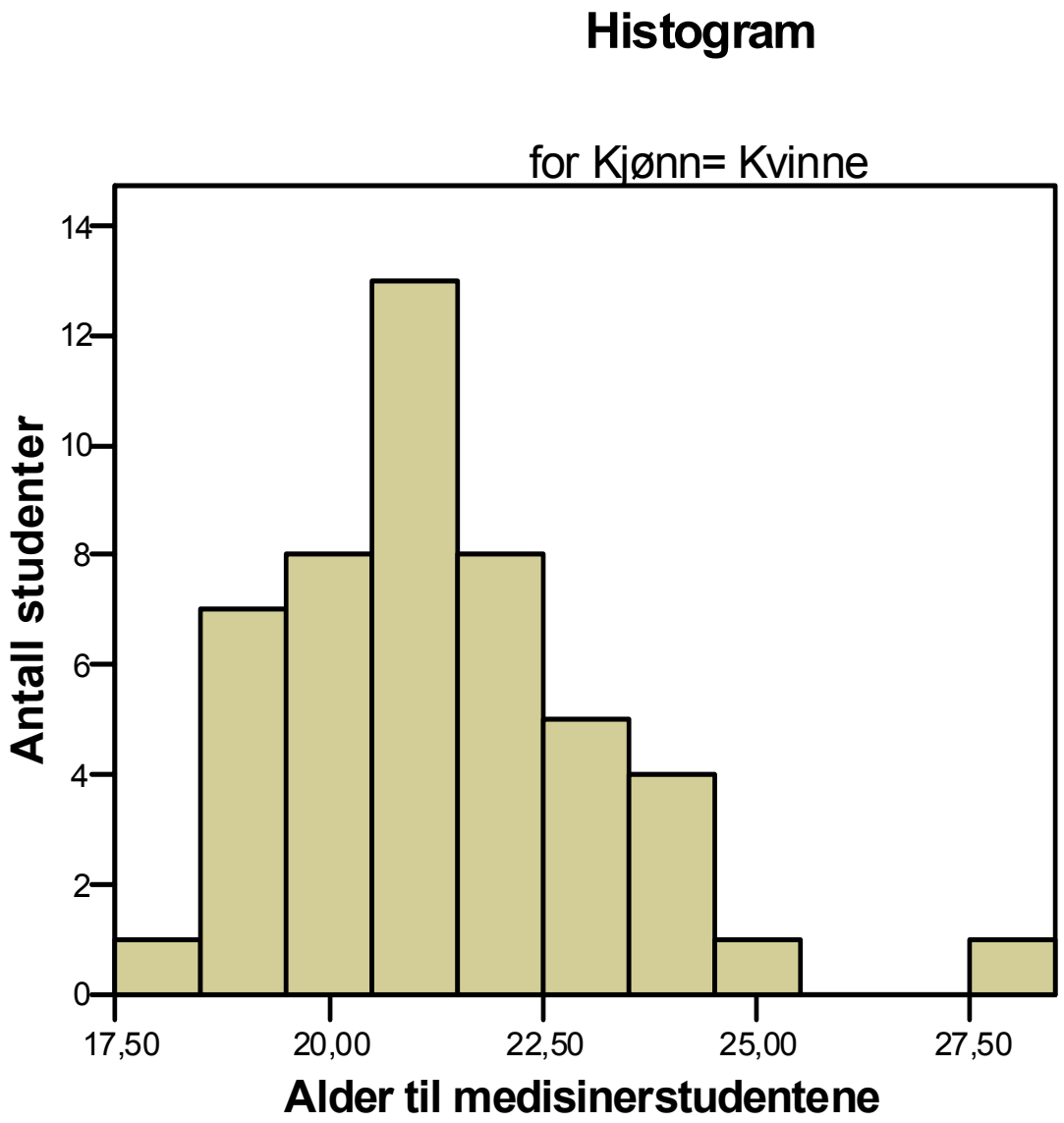
Recode forts.

- I det nye vinduet kan man skrive inn gamle og nye verdier for variabelen.
- Under *Old value-Range* definerer man de gamle verdiene, og under *New value-Value* definerer man de nye.
- Kan skrive inn at 0-20 år skal ha ny verdi 1, 20.1-25 år ny verdi 2 og 25.1-40 år ny verdi 3. Klikk *Add* mellom hver.
- Etter å ha trykket *Continue* og *OK*, ser man at en ny variabel har kommet inn i data-vinduet
- Etter å ha opprettet variabelen, kan man definere kategoriene under *Values* i *Variable View*

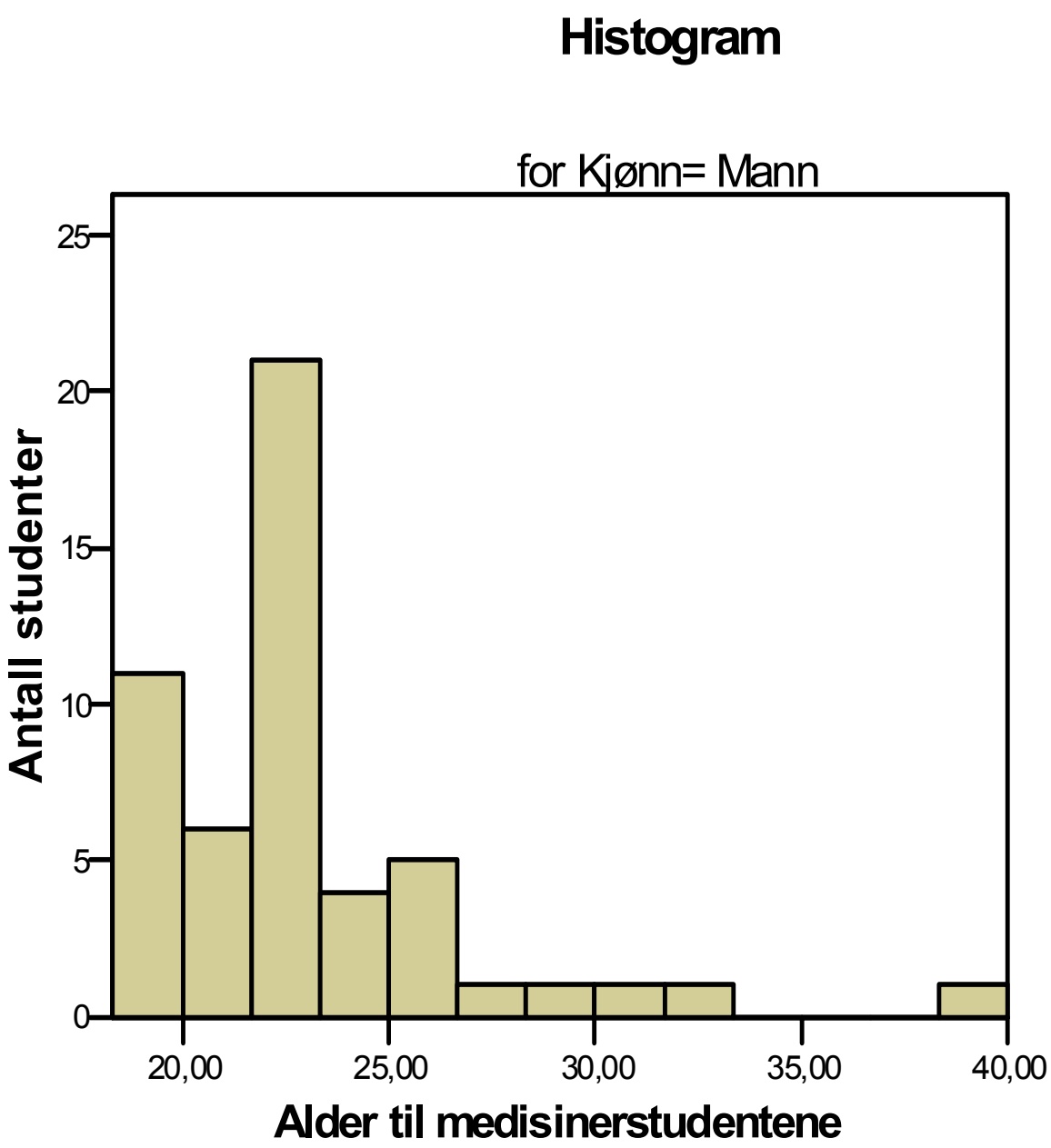
Hvordan få ut separate tabeller for en faktor, f.eks. kjønn i SPSS

- Klikk *Analyze - Descriptive Statistics - Explore*. Merk av de relevante variablene og overfør dem til *Dependent List*.
- Flytt kjønn over i *Factor List*
- Ellers som før!

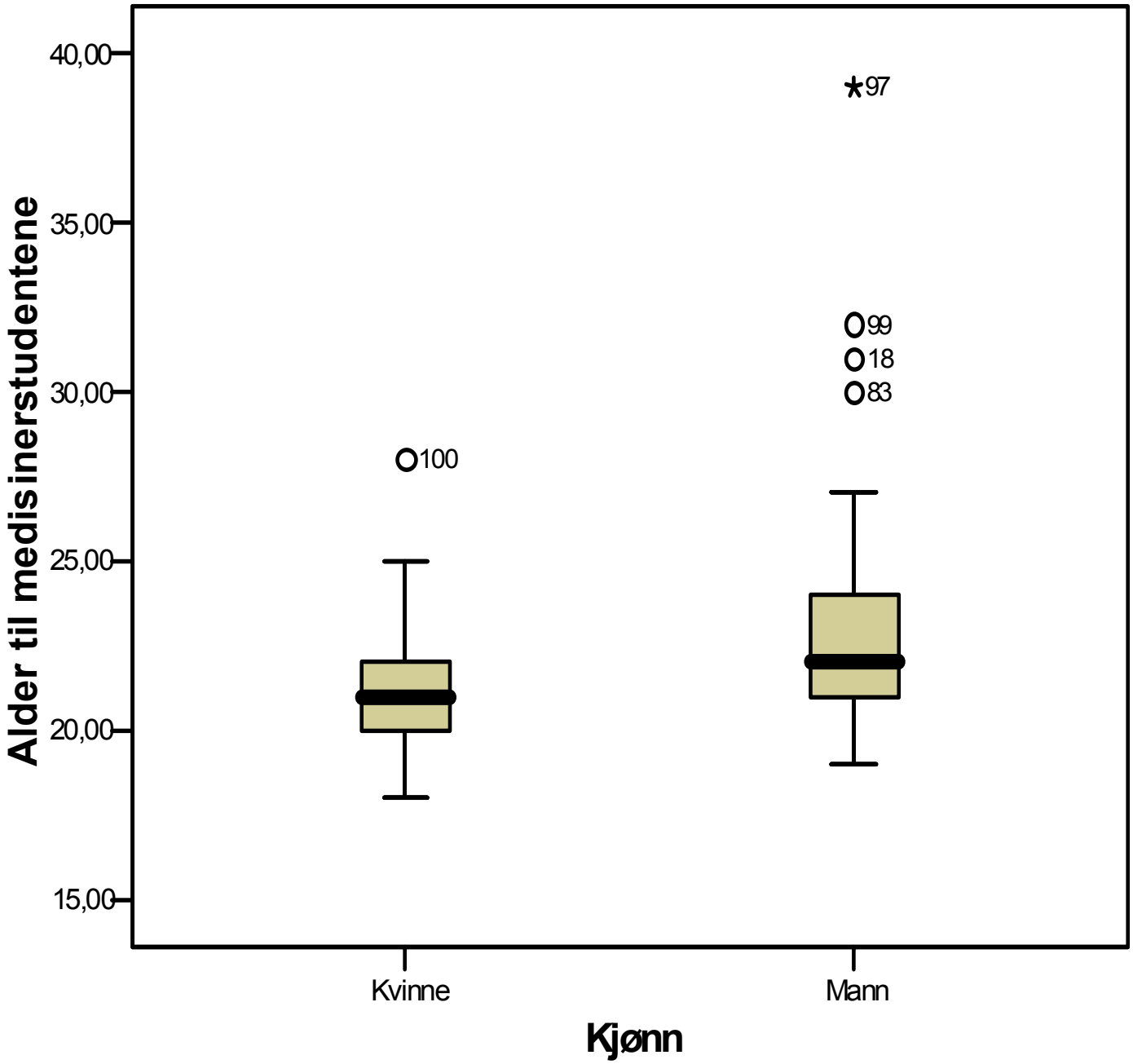
Analyser separat for kjønn



Analyser separat for kjønn

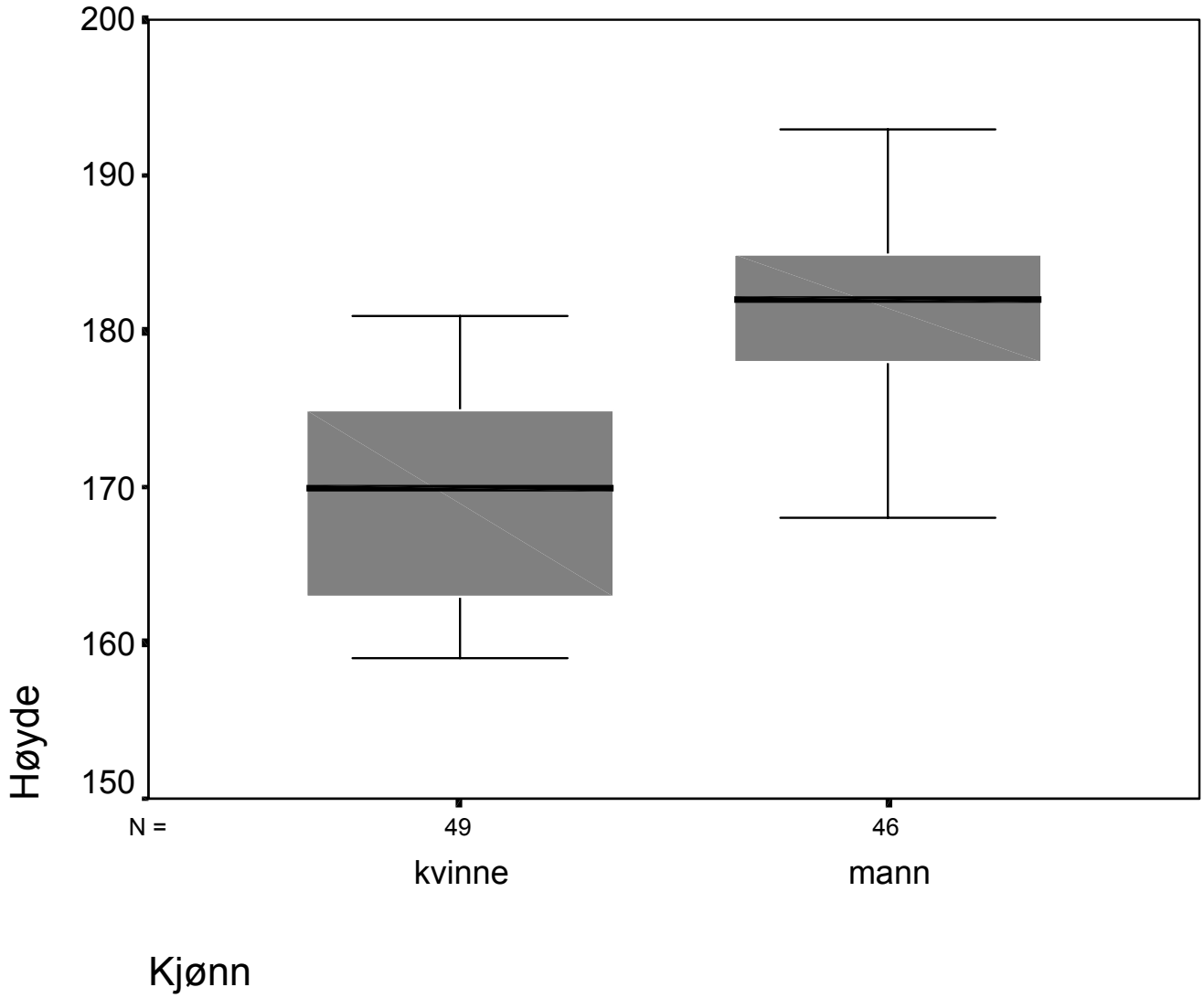


Boxplott separat for kjønn

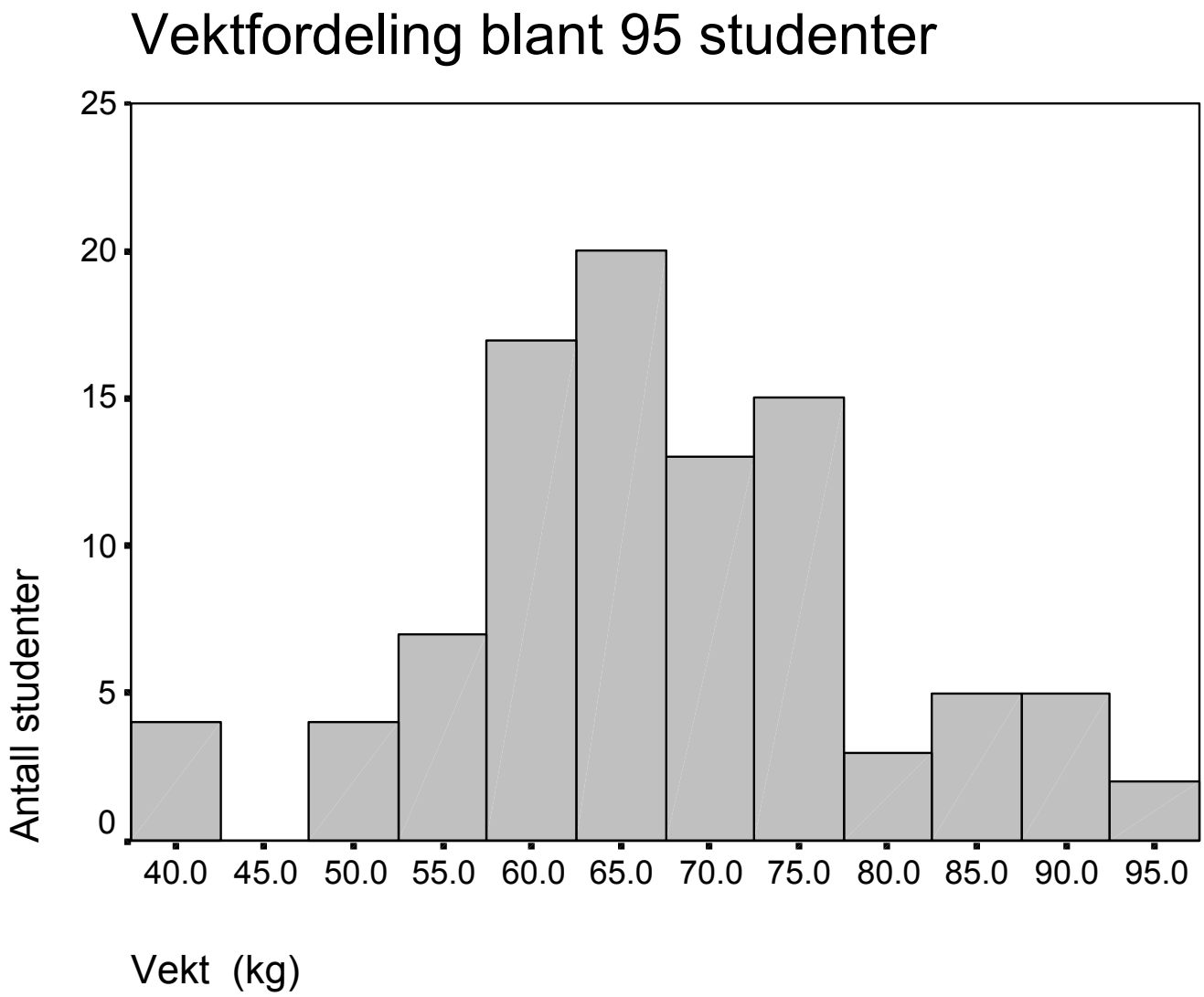


Hva hvis man bare vil se på f.eks. kvinner? Select Cases!

- Velg *Data->Select cases*. Kryss av på *If condition is satisfied*, og trykk på *If*-knappen
- Et nytt vindu kommer opp. Flytt kjønn over til høyre og tilføy =1 (hvis kvinner er kodet som 1)
- Trykk *Continue*



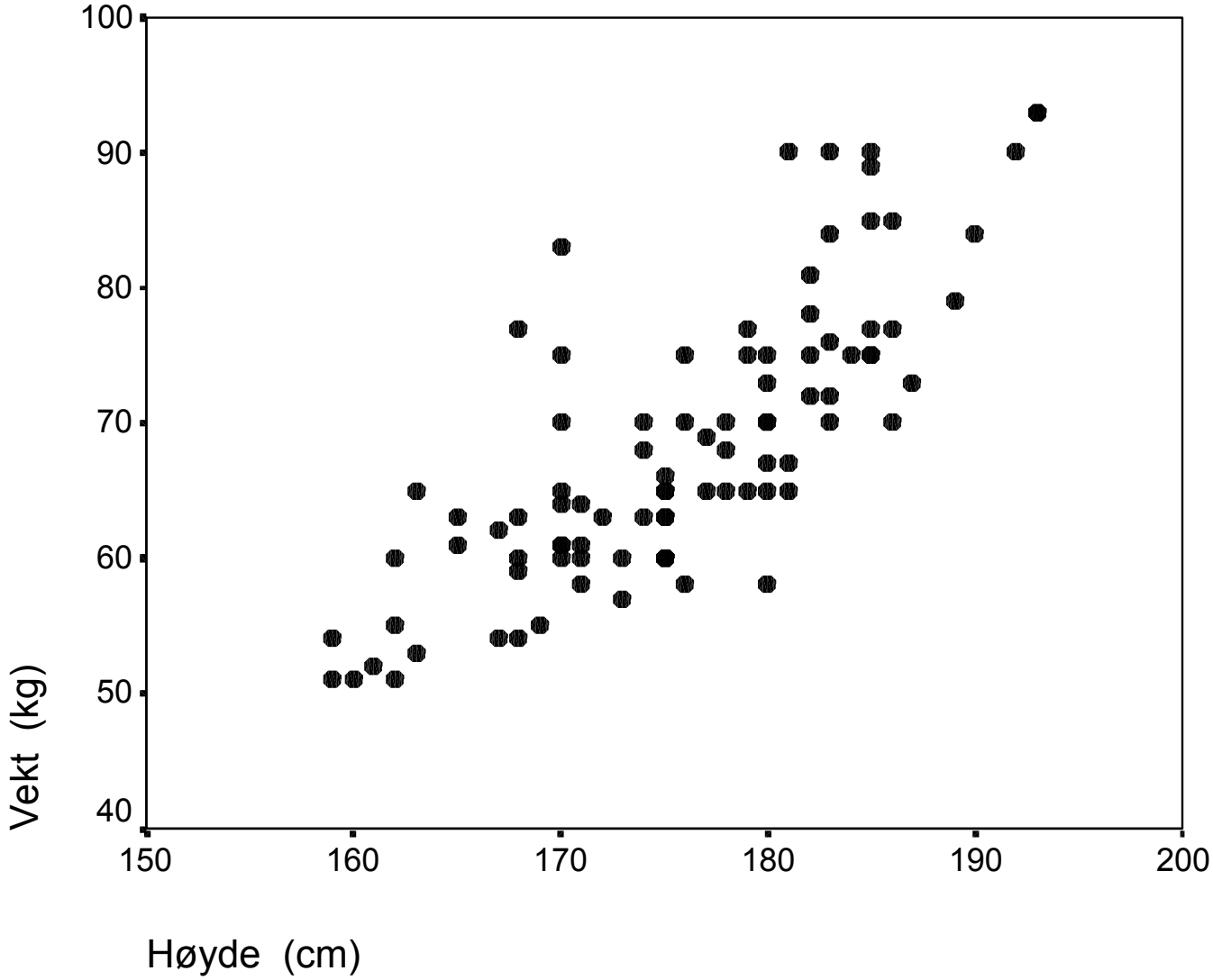
- Box-plott for sammenlikning av høyde blant menn og kvinner. Data fra kull V98 (n=95)



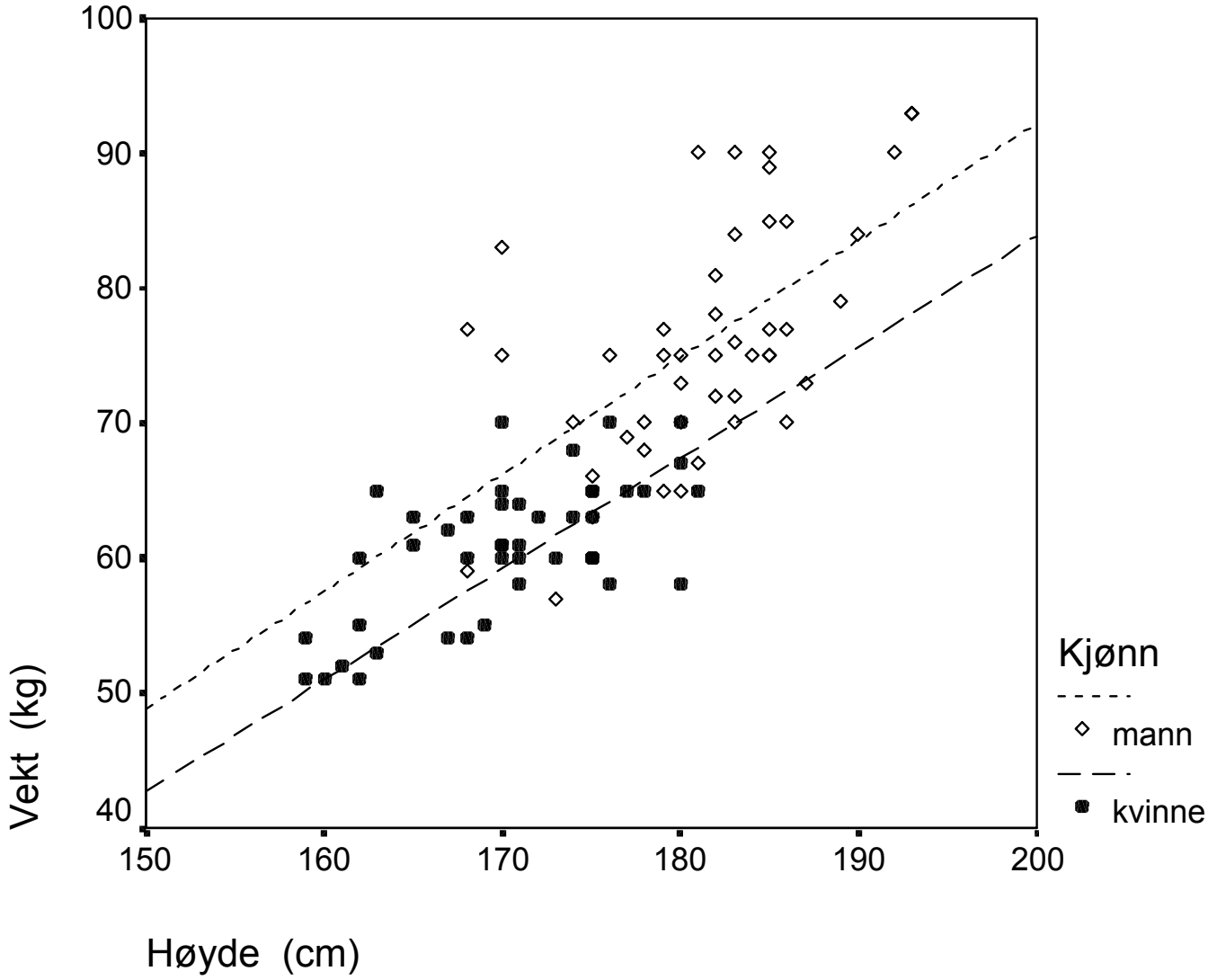
- Data om vekt samlet inn blant studenter på kull V98

Hvordan se sammenhengen mellom to kontinuerlige variabeler i SPSS: Spredningsdiagram!

- For å lage spredningsdiagram, klikk på *Graphs - Scatter - Define*. Plukk ut de to variablene som skal være på Y-aksen og X-aksen henholdsvis
- Hvis du ønsker å skille mellom gruppene, kan du overføre grupperingsvariabelen til *Set Markers by*
- Et spredningsdiagram kan redigeres ved å dobbeltklikke på diagrammet. Ved å dobbeltklikke på datapunktene i redigeringsmodus og trykke høyre musknapp, kan du legge inn en rett linje for totalen "*Fit line at total*", eller for hver undergruppe "*Fit line at subgroups*" hvis det er flere grupper



- Spredningsdiagram for vekt mot høyde. (n=95)



- Spredningsdiagram av vekt mot høyde. Innlagte regresjonslinjer for menn og kvinner

Hva hvis man vil lage en ny variabel med f.eks. BMI?

- Har høyde og vekt for studentene. Vil ha en variabel med BMI.
- Velg *Transform->Compute*. Skriv inn navnet på den nye variabelen under *Target variable*.
- Under *Numeric expression*, skriv inn (Vekt)/(Høyde/100)^2
- Forutsetter at kodingen er som i eksempelet
- Trykk OK. Ser at en ny variabel oppstår i datafilen.

Deskriptiv statistikk for kategoriske variabler

- Lite meningsfylt å oppgi gjennomsnitt for variabelen kjønn
- Vil heller se hvor mange % kvinner og menn som er i materialet
- *Analyze->Descriptive Statistics ->Frequencies*
- Flytt variabelen du vil studere over til høyre i vinduet

Kort om to epidemiologiske mål: Prevalens

- Andel av befolkningen som lider av en bestemt sykdom

Eksempel: Forekomst av tarmkreft

Antall personer i live med tarmkreft 31.12.1995: 16 861

Prevalens

$$\frac{16861}{4390000} = 38.4 \text{ pr. } 10\,000 \text{ innbyggere}$$

Epidemiologiske mål: Insidensrate

- Andel nye tilfeller pr. år

Eksempel: tarmkreft

Antall nye tilfeller i 1995: 3034

Insidensrate:

$$\frac{3034}{4390000} = 6.9 \text{ pr. } 10\,000 \text{ innbyggere pr. år}$$

Insidens av malignt melanom blant kvinner i Norge