Fordelinger, mer om sentralmål og variasjonsmål

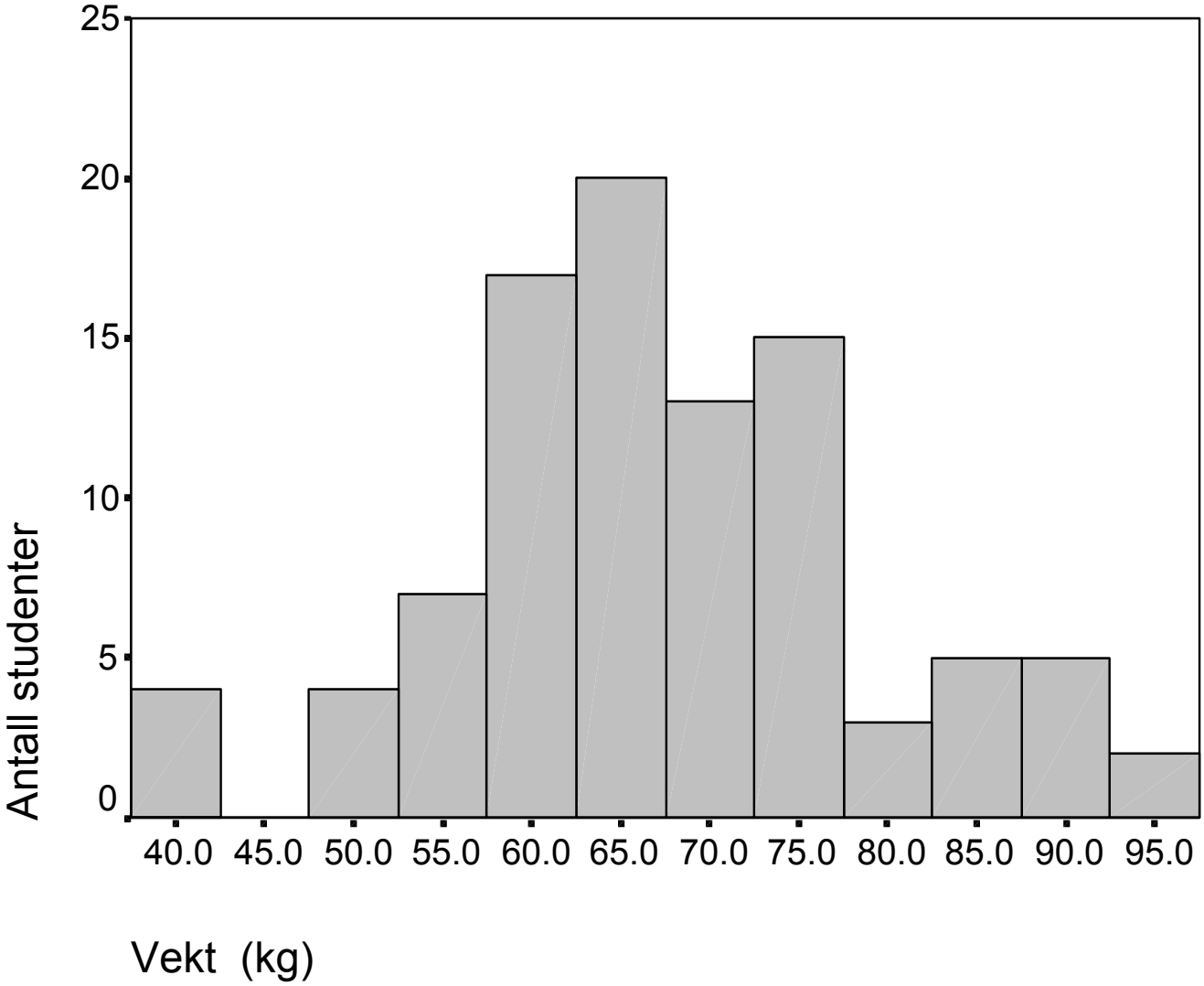
Tron Anders Moger

20. april 2005

Forrige gang:

- Så på et eksempel med data over medisinerstudenter
- Lærte hvordan man skulle få oversikt over dataene ved hjelp av *Explore* i SPSS
- Lærte om sentralmål som gjennomsnitt og median
- Variasjonsmål som standardavvik og fraktiler
- I dag skal vi knytte disse begrepene opp mot normalfordelingen

Vektfordeling blant 95 studenter

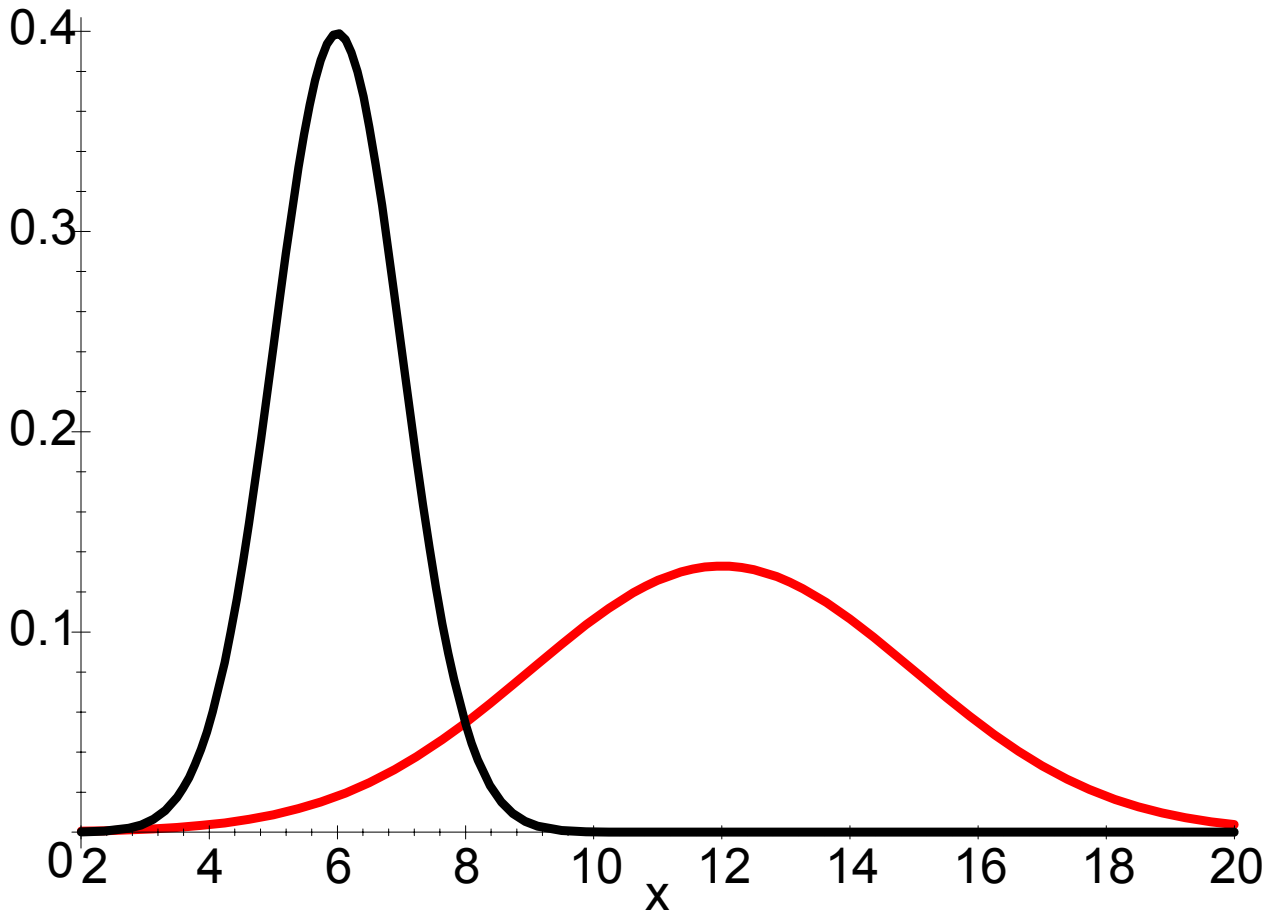


- Data om vekt samlet inn blant studenter på kull V98

Kontinuerlige data: Normalfordelingen

- Normalfordelingen
- Variasjon i målinger kan ofte beskrives med normal-fordelingen
- Fordelingen er symmetrisk og har to parametere:
  - » Forventning: μ
  - » Standardavvik (Standard deviation): σ
- Disse angir tyngdepunkt og spredning I fordelingen

Normalfordeling med liten og stor spredning σ



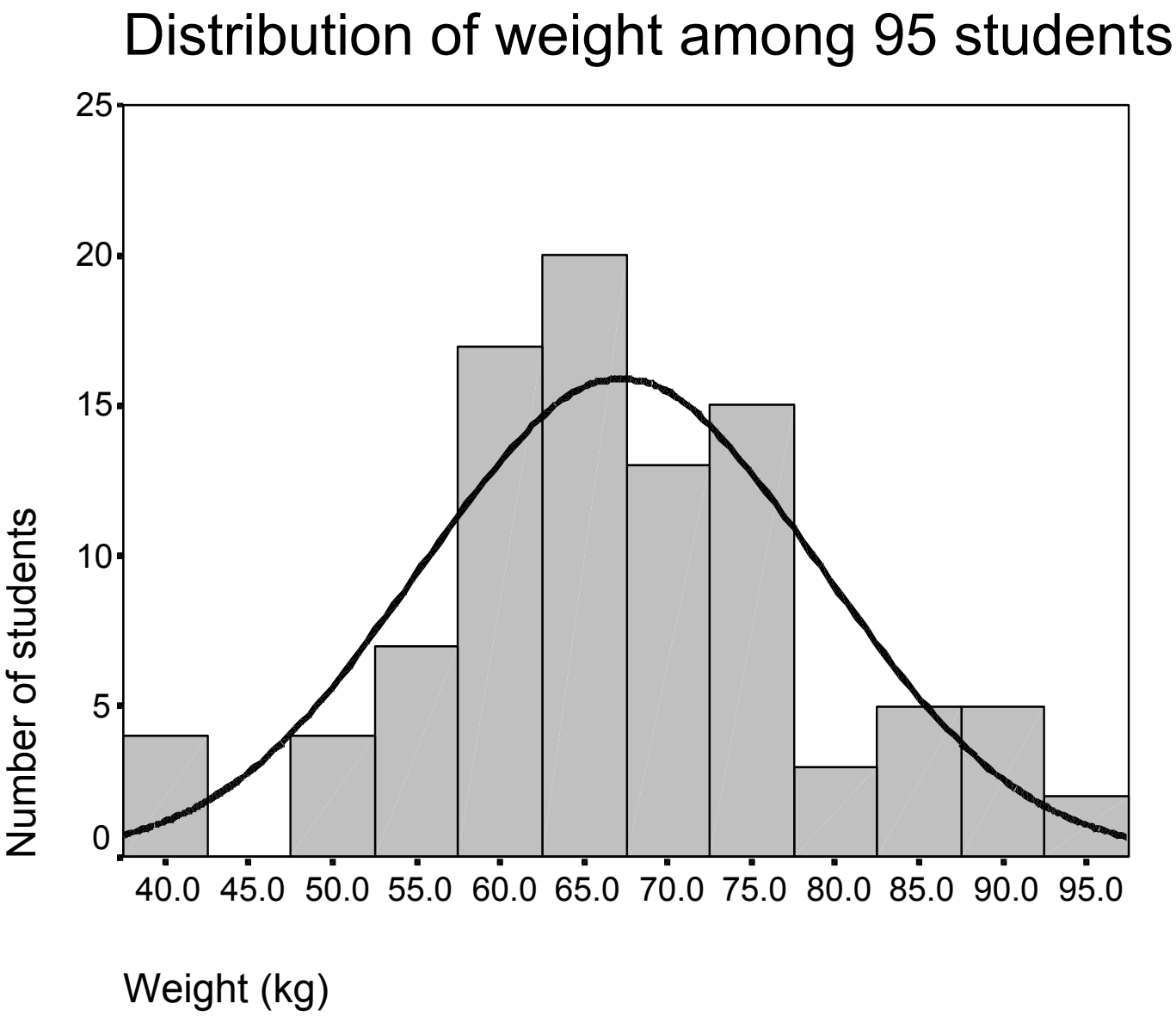
Hvorfor er normalfordelingen viktig?

- De vanligste statistiske analysemetodene for kontinuerlige data antar at dataene er normalfordelt
- Derfor må man sjekke grundig at dette faktisk er tilfellet før man gjør analysene
- I tillegg: Summer av kategoriske observasjoner er også normalfordelt
- Betyr at normalfordelingen sniker seg inn nesten uansett hva du skal gjøre

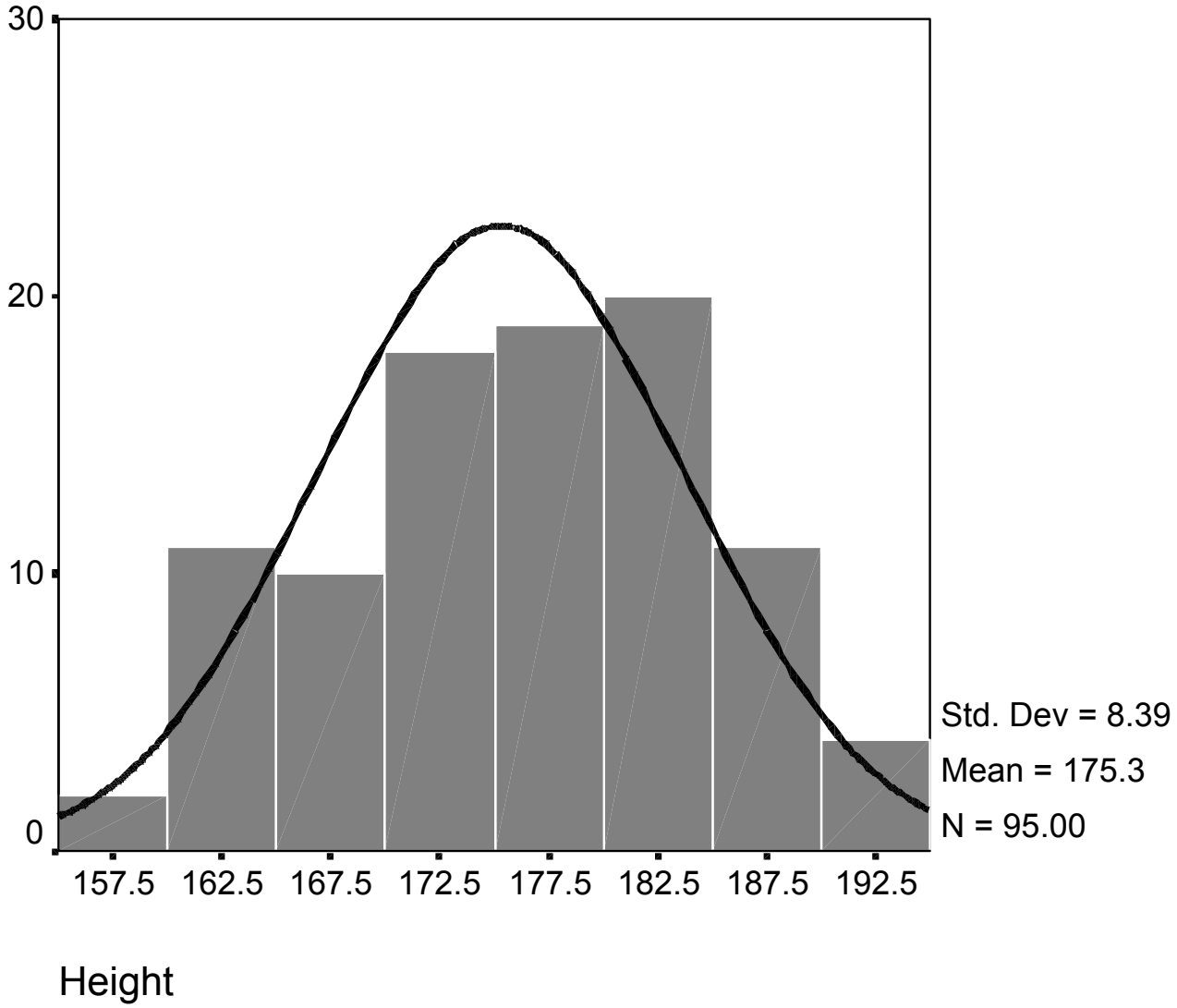
Enkel måte å se om dataene er normalfordelt på: Histogram

- Velg *Graphs->Histogram*
- Flytt variabelen du vil studere over til høyre (f.eks. vekt)
- Kryss av på *Display normal curve* og klikk *OK*.

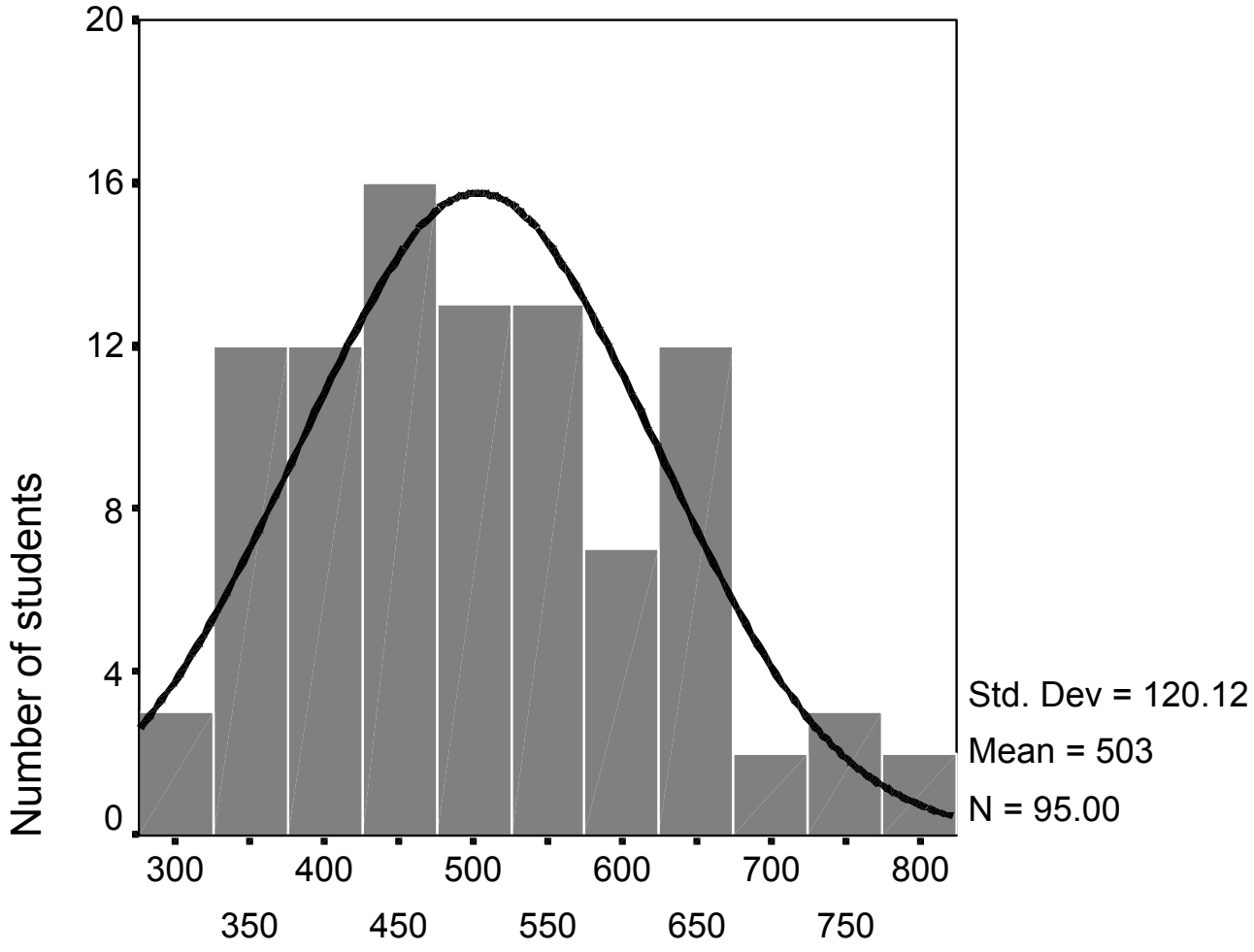
Histogram av vekt med innlagt kurve for normalfordelingen



Fordeling av høyde for medisinerstudentene



Fordeling av lungefunksjon for medisinerstudentene

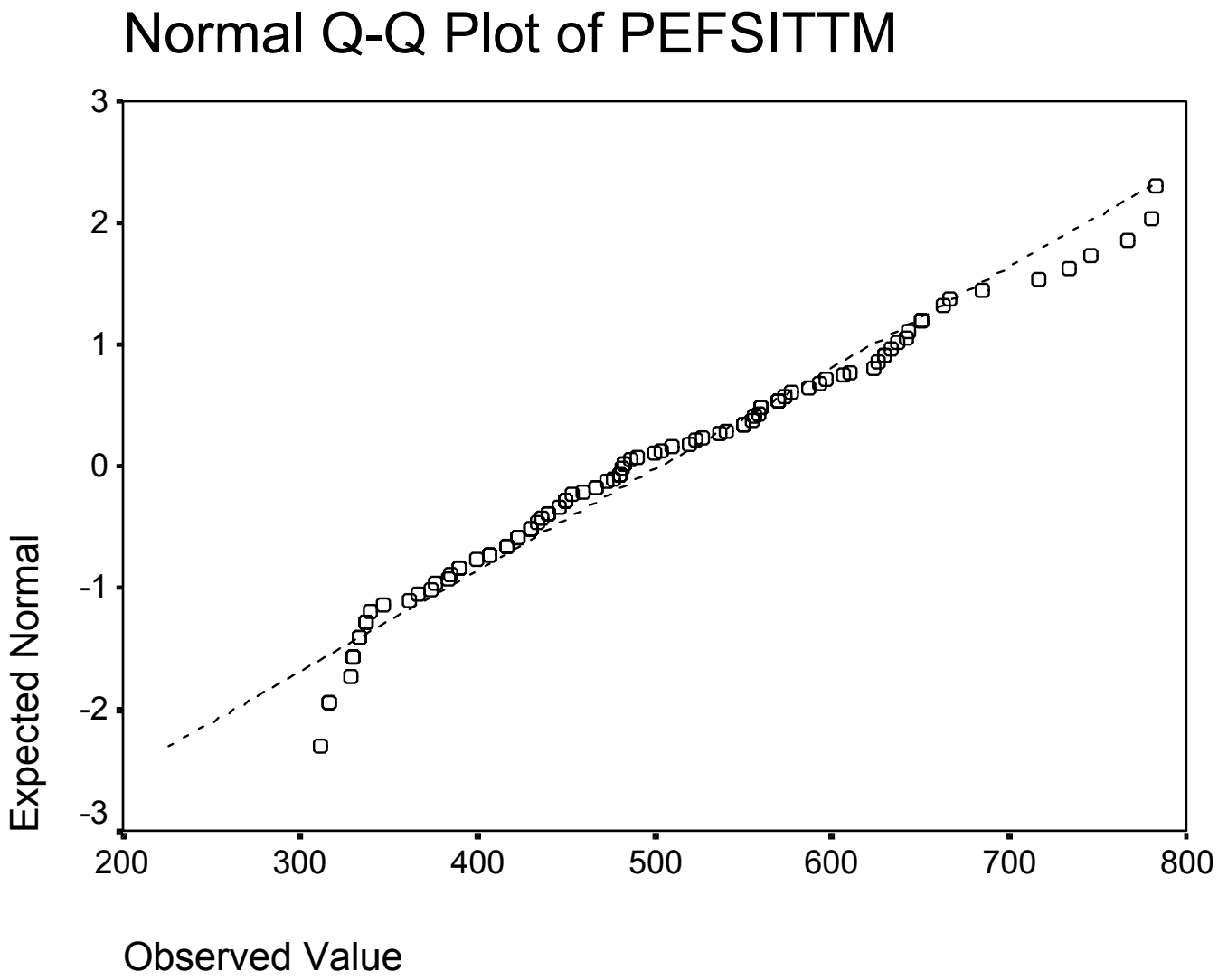


Average PEF value measured in a sitting position

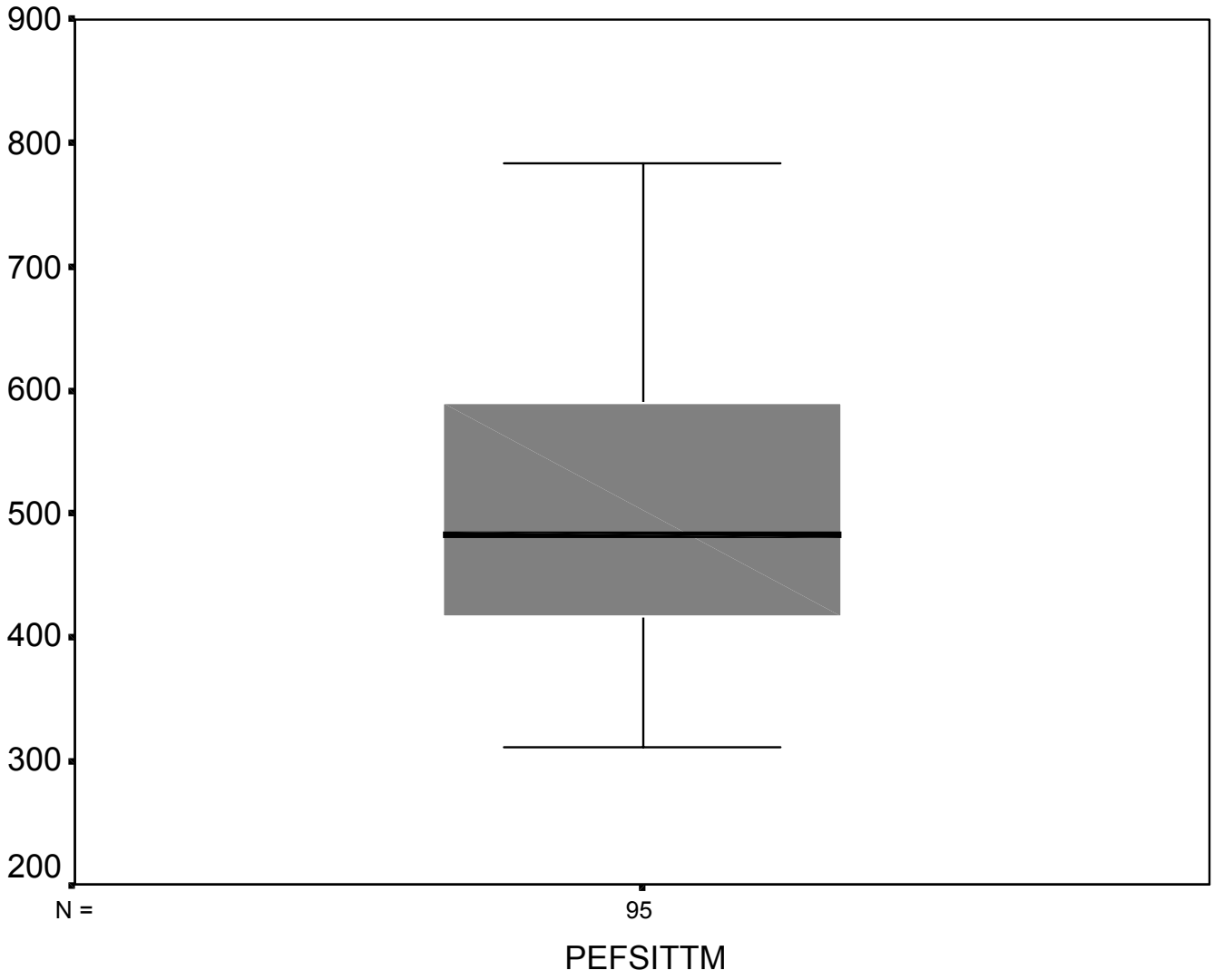
Andre måter å sjekke om dataene er normalfordelt på: Explore

- Som forrige gang: Velg *Analyze->Descriptive Statistics->Explore.*
- Flytt variabelen du vil se på over i *Dependent List* (f.eks. vekt).
- Under *Plots*, kryss av på *Normality Plots with tests*.

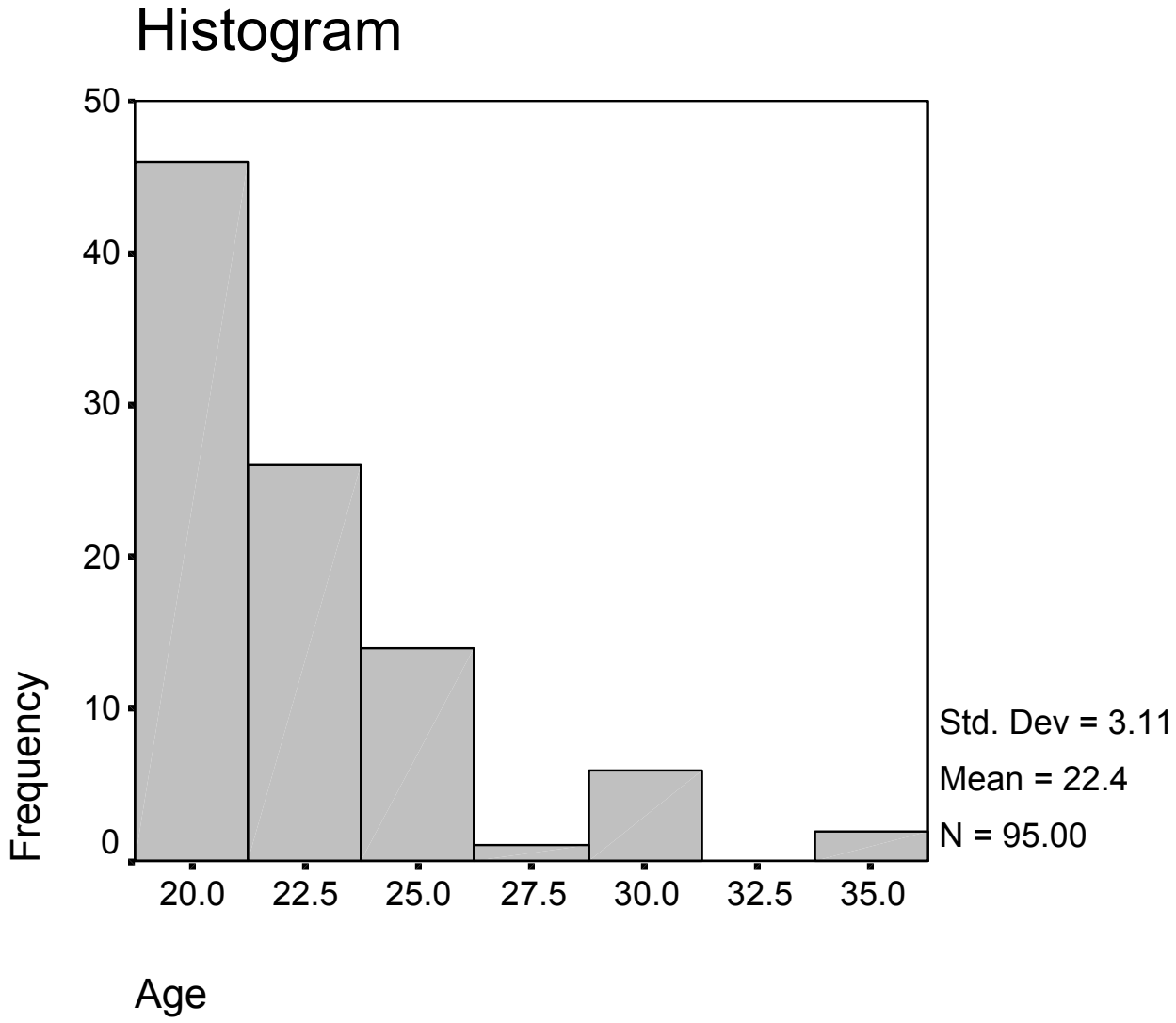
Q-Q plot for lungefunksjon



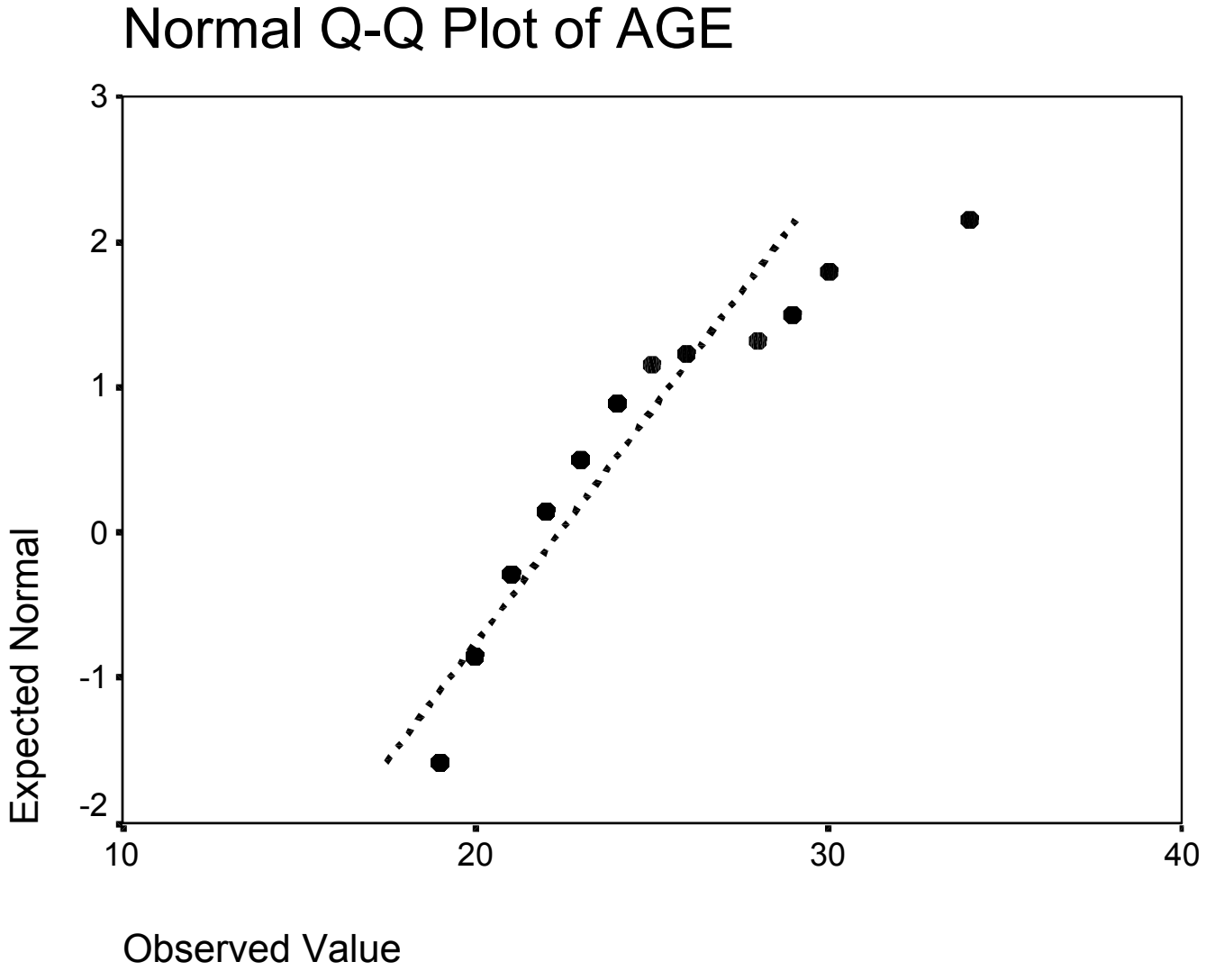
Box-plot for lungefunksjon



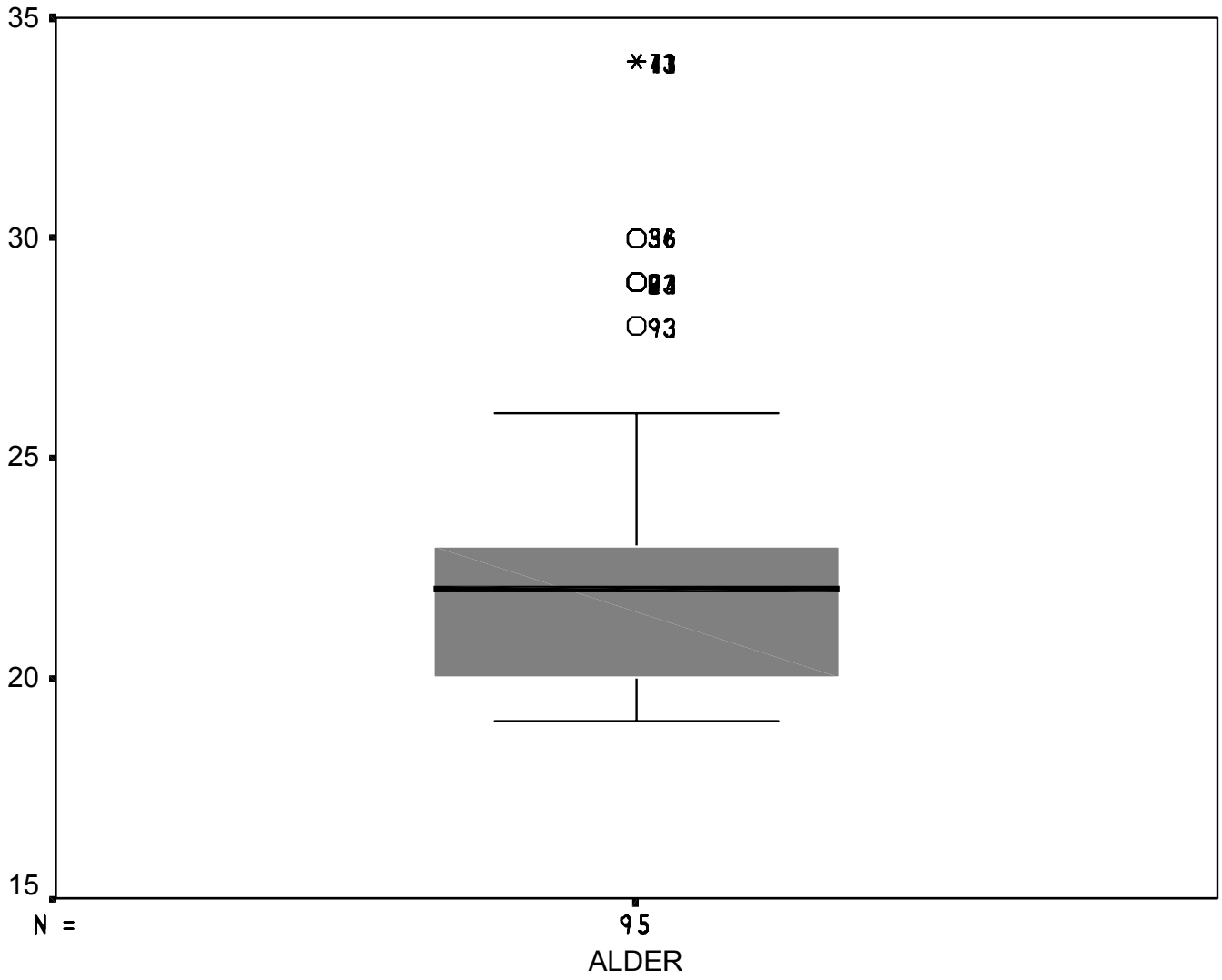
Alder for medisinerstudentene - ikke normalfordelt



Q-Q plot for alder



Box plot for alder



Tester for om lungefunksjon og alder er normalfordelt

Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
PEFSITTM	.081	95	.144

a. Lilliefors Significance Correction

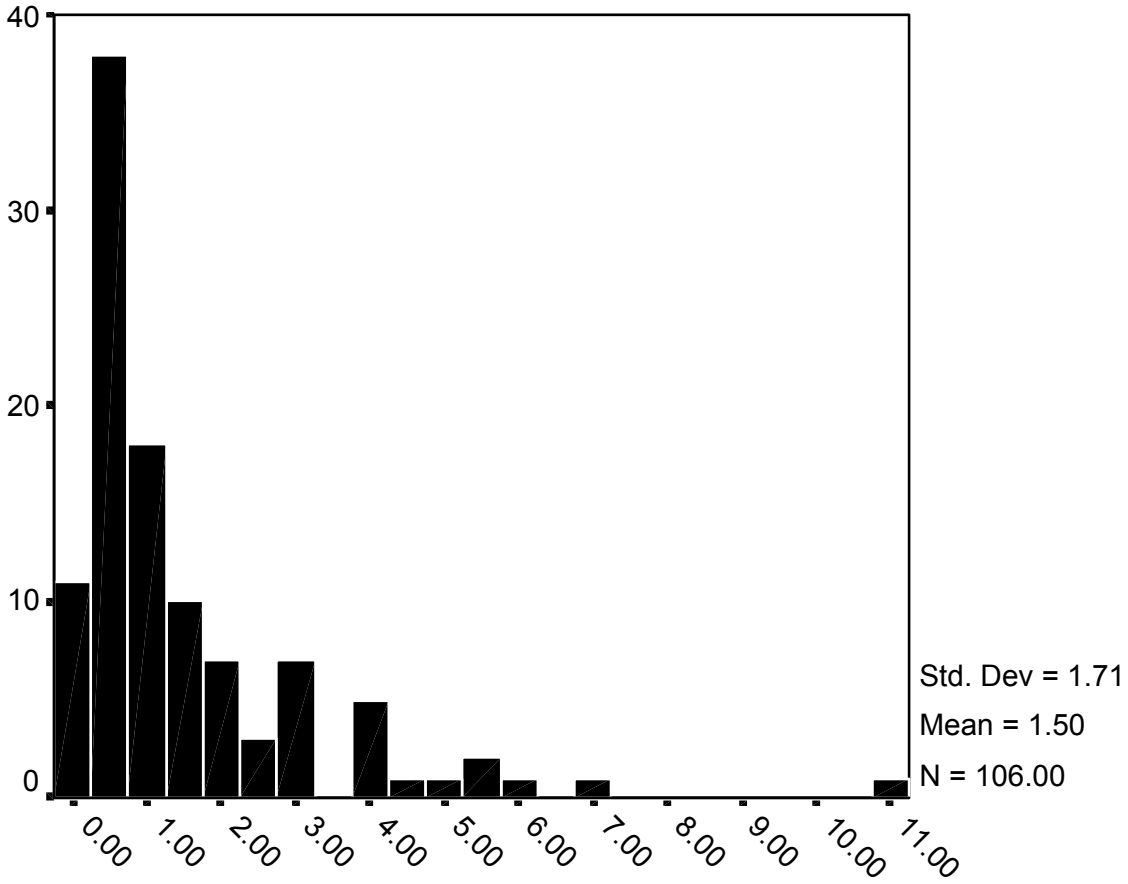
Tests of Normality

	Kolmogorov-Smirnov[a]		
	Statistic	df	Sig.
ALDER	.180	95	.000

a. Lilliefors Significance Correction

Disse vil ofte bli signifikante når det er mange data. Legg mer vekt på de grafiske metodene.

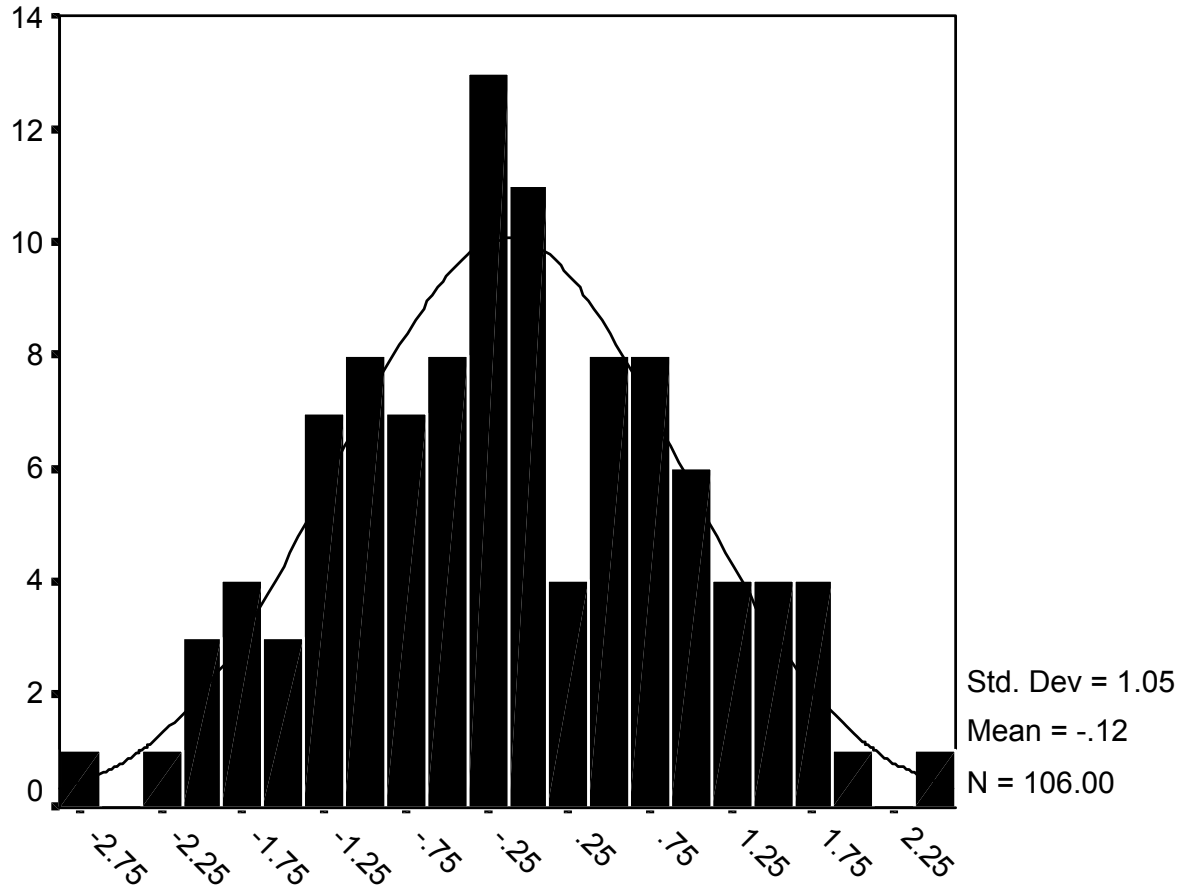
Et triks for data som er skjeve mot høyre: Transformer dataene!



SKEWED

Skjev fordeling
Eksempel, observasjonene
0.40
0.96
11.0

Log-transformerte data



Ln transformert fordeling

ln(0.40)=-0.91

ln(0.96)=-0.04

ln(11) =2.40

Kjør analysen på de ln transformerte dataene

SPSS: *transform-compute*

Vi har sett at f.eks. medisinerstudentenes vekt er normalfordelt – hva så?

- Forrige gang så vi på begreps-par:
  - » Utvalg – populasjon
  - » Histogram – fordeling
  - » Gjennomsnitt – forventning
- Poenget med statistikk er at vi ønsker å generalisere fra et lite utvalg til en hel populasjon
- Dette gjør vi ved å samle data slik at utvalget blir *representativt* (tilfeldig trekning, eller m.h.p. alder, kjønn, bosted, osv.)

Ny måte og lese tabellen og histogrammene fra *Explore* på:

- Histogrammene viser at vekten til studentene er tilnærmet normalfordelt
- Fra dette ønsker vi å konkludere at vekten i populasjonen (alle førstesemester-studenter i medisin i Norge) er normalfordelt
- Gjennomsnittet beregnet fra utvalget er et estimat på forventningen μ til normalfordelingen til populasjonen
- Standardavviket (spredningen i dataene) beregnet fra utvalget er et estimat på standardavviket σ til normalfordelingen til populasjonen
- Dette medfører også at et intervall som er 1.96 standardavvik på hver side av forventningen dekker 95% av fordelingen

Hvor kommer 1.96 fra? Standard normal-fordelingen!

- Hvis en variabel X er normalfordelt, vil $(X-\mu)/\sigma$ være standard normalfordelt.
- Standard normalfordelingen er en normalfordeling med forventning 0 og standardavvik 1
- Lærte om fraktiler sist (25% fraktil, median, 75% fraktil)
- Har også 2.5% og 97.5% fraktiler
- For standard normalfordelingen er 2.5% fraktilen lik -1.96 og 97.5% fraktilen lik 1.96

Dette gir:

$$P\left(-1.96 < \frac{X-\mu}{\sigma} < 1.96\right) = 95\%$$

$$P\left(-1.96 \cdot \sigma < X - \mu < 1.96 \cdot \sigma\right) = 95\%$$

$$P\left(\mu - 1.96 \cdot \sigma < X < \mu + 1.96 \cdot \sigma\right) = 95\%$$

- M.a.o. hvis dataene er normalfordelte, vil en ny observasjon ha 95% sannsynlighet for å havne innenfor intervallet

Gj.snitt+/-1.96*standardavvik

- Dette gir litt bedre mulighet til å fortolke om standardavviket er veldig stort eller ikke

Standardfeil (standard error)

- La oss si at du har samlet et utvalg på *n* observasjoner for å beregne et gjennomsnitt
- Samler du flere utvalg av samme størrelse får du ikke samme gjennomsnitt
- Husker fra forrige gang at standardavviket estimeres ved

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

- Standardavviket til gjennomsnittet kalles standardfeil og estimeres med

$$\text{Standardfeil} = \frac{s}{\sqrt{n}}$$

Eksempel på hvordan standardfeilen blir mindre når datamengden øker

n	Standardfeil
1	5,00
10	1,58
40	0,79
100	0,50
400	0,25

Konfidensintervall

- Hvis observasjonene er normalfordelt, kan du være 95% sikker på at det sanne gjennomsnittet (forventningen) ligger i intervallet

  Gj.snitt+/-1.96*standardfeil
- Dette kalles et 95% konfidensintervall
- Kunne like gjerne hatt et 90% eller 99% konfidensintervall, men 95% er vanligst

Konfidensintervall i SPSS

- *Analyze- Descriptive Statistics - Explore*

Descriptives

			Statistic	Std. Error
WEIGHT	Mean		70,7000	2,6548
	95% Confidence Interval for Mean	Lower Bound	65,1435	
		Upper Bound	76,2565	
	5% Trimmed Mean		70,5000	
	Median		69,0000	
	Variance		140,958	
	Std. Deviation		11,8726	
	Minimum		50,00	
	Maximum		95,00	
	Range		45,00	
	Interquartile Range		10,0000	
	Skewness		,427	,512
	Kurtosis		,263	,992

Egenskaper ved konfidensintervall

- Statistikk handler ikke bare om å finne gjennomsnitt, man vil også si noe om usikkerheten til gjennomsnittet.
- Siden konfidensintervallet avhenger av *standardfeilen*, ser vi at jo mer data vi samler inn, jo smalere blir konfidensintervallet
- Dette betyr at vi kan med svært stor grad av sikkerhet si hvor forventningen i populasjonen ligger ved å samle inn mye data.
- Dette er positivt.

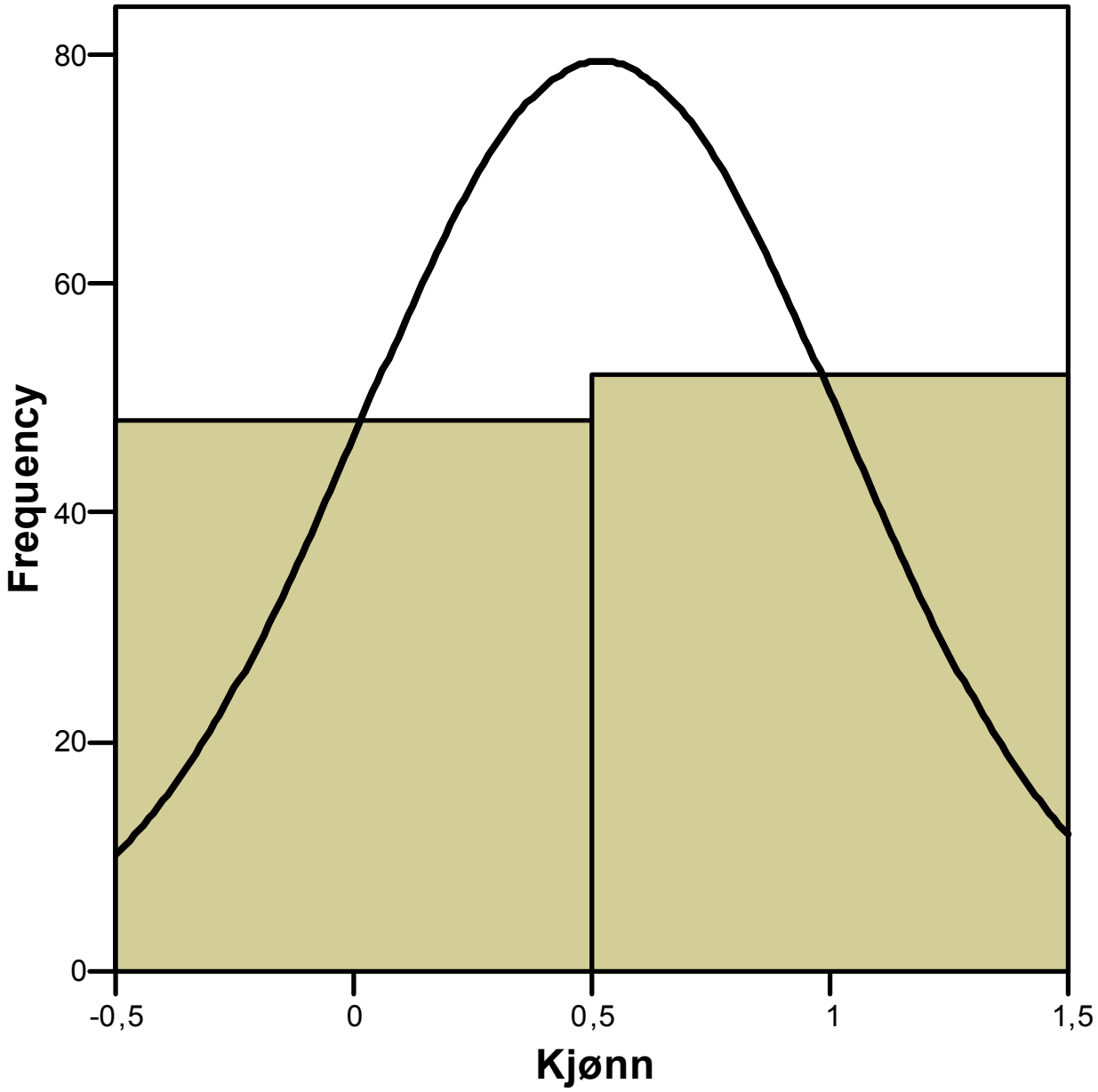
Egenskaper ved konfidensintervall forts.

- Har så langt bare sett på data som er normalfordelt
- Skal se senere at normalfordelingen kommer inn også for mange andre typer data.
- Dette betyr at dere ofte vil se konfidensintervall som baseres på +/-1.96*ett usikkerhetsestimat
- Her gjelder det samme: Jo mer data, jo mindre usikkerhet i konklusjonene deres!

Eksempel på at normalfordelingen kommer inn andre steder

- Histogram over kjønn for medisinerstudentene ser ikke særlig normalfordelt ut
- Kjønn kan antas å følge en annen statistisk fordeling: *Binomisk fordeling*, karakterisert ved
  - » Uavhengige forsøk/observasjoner
  - » To mulige utfall
  - » Sannsynligheten for utfallene forandrer seg ikke fra forsøk til forsøk

Histogram over kjønn



Data som er binomisk fordelte kan likevel tilnærmes ved normalfordelingen!

- Sannsynligheten for det ene utfallet kalles *p*
- Sannsynligheten for det andre utfallet blir da 1-*p*
- Anta at du har n forsøk/individer
- Forventningen til en binomisk fordeling er *np* og standardavviket er $\sqrt{np(1-p)}$
- Har da at

$$\frac{X-np}{\sqrt{np(1-p)}} \approx \text{Standard normalfordelt}$$

Hvorfor gjelder dette?

- *Sentralgrenseteoremet* sier at summer av mange observasjoner, hvor ingen av dem er dominerende, er tilnærmet normalfordelt
- For kjønn i eksemplet svarer dette til summen av mange 0-er og 1-ere
- Får på vanlig måte standard normalfordelingen ved å trekke fra forventningen og dele på standardavviket.
- Kommer tilbake til dette under Kategoriske data - tabellanalyse