

FORELESNING 3

Eivind Eriksen

JAN 26 2012

MET 3431

STATISTIKK

Plan: Deskriptiv statistikk

- ① Frekvenstabeller
- ② Histogrammer
- ③ Grafiske framstillinger
av data

} [T] 2.1 - 2.5

Kapittel 2: Å beskrive stikkprøven med tall og grafer

Deskriptiv statistikk

- Anta at vi har samlet inn data på en fornuftig måte
- Målet er å forstå dataene gjennom
 - Nøkkeltall: gjennomsnitt, standardavvik osv
 - Grafer: Paretodiagram, histogram osv
- Du bør i prinsippet kunne regne for hånd
- I praksis: bruk et statistikk program

Fem ting å vite om stikkprøven

Kap. 3
(senere)

- C Hvor er senteret i dataene?
- V Hvor mye varierer dataene?

Sentralmål : Gjennomsnitt
median
Variasjonsmål : standardavvik

Kap. 2
(1 dag)

- D Hvordan er dataene fordelt?
- O Er det noen data/observasjoner som skiller seg veldig ut?
- T Hvilken utvikling har dataene over tid?

Fem ting å vite om stikkprøven

Senter Måles med gjennomsnitt, median, modus
Variasjon Måles med kvartilavvik, standardavvik
Fordeling Histogrammer, frekvensstabeller, blad og løv diagram
Uteliggere Visuell sjekk i spredningsplott, boksdiagram
Tid Tidsrekkeanalyse

Frekvenstabeller

Frequency table

- For store stikkprøver er det nyttig å få oversikten vha. en *frekvenstabell*
- Relativt enkelt å lage slike for hånd
- Men computeren gjør dette raskere
- Det viktigste er å forstå hva en slik tabell sier om dataene

Frekvenstabell eksempel

Table 2-1	Academy Awards: Ages of Best Actresses and Best Actors																			
The ages (in years) are listed in order, beginning with the first awards ceremony.																				
Best Actresses																				
22	37	28	63	32	26	31	27	27	28											
30	26	29	24	38	25	29	41	30	35											
35	33	29	38	54	24	25	46	41	28											
40	39	29	27	31	38	29	25	35	60											
43	35	34	34	27	37	42	41	36	32											
41	33	31	74	33	50	38	61	21	41											
26	80	42	29	33	35	45	49	39	34											
26	25	33	35	35	28															
Best Actors																				
44	41	62	52	41	34	34	52	41	37											
38	34	32	40	43	56	41	39	49	57											
41	38	42	52	51	35	30	39	41	44											
49	35	47	31	47	37	57	42	45	42											
44	62	43	42	48	49	56	38	60	30											
40	42	36	76	39	53	45	36	62	43											
51	32	42	54	52	37	38	32	45	60											
46	40	36	47	29	43															

Original Data

Table 2-2

Frequency Distribution:
Ages of Best Actresses

= Frequency table
= Frekvens - tabell

(20,30]
(30,40]
...

Age of Actress	Frequency
21-30	28
31-40	30
41-50	12
51-60	2
61-70	2
71-80	2

Frekvenstabell

Intervallnotasjon

Intervaller

I flere sammenhenger (frekvenstabell, histogrammer, konfidensintervall) skal vi jobbe med *intervaller*.

- Tenk deg tallinja
- Et intervall er bare en bit av denne

Example

- $[250, 300 >]$: alle tall fra og med 250 opp til 300.
- $[250, 300]$: alle tall fra og med 250 til og med 300
- $< 250, 300 >$: det alle tall fra 250 til 300

$[10, 20) = [10, 20 >]$: alle tall fra og med 10, til 20
↑ ↑
10 er med 20 er ikke med
helteall: 10-19

Eksempel: Lønninger

Differanse: $573 - 280 \approx 300$
 Intervallbredde 50: Ca $\frac{300}{50} = 6$ intervaller

- ❶ 395 408 328 508 341 407 443 435 436 313 398 573 325 374
322 413 280 392 401 455
- ❷ Sortering: (280) 313 322 325 328 341 374 392 395 398 401 407
408 413 435 436 443 455 508 (573)
- ❸ Bestem klassebredde, f.eks. 50. Det nederste intervallet kan da være $[250, 300 >$

- ❹ Tell opp hvor mange som kommer i hvert intervall.

- ❺ F.eks. er det bare 1 person i lønnsintervallet $[250, 300 >$

$[250, 300)$	1	$[550, 600)$	1
$[300, 350)$	5		
$[350, 400)$	4		
$[400, 450)$	7		
$[450, 500)$	1		
$[500, 550)$	1		
		Totalt	20

Frekvenstabell for lønninger

Lønnsintervall	Antall/Frekvens
[250, 300 >	1
[300, 350 >	5
[350, 400 >	4
[400, 450 >	7
[450, 500 >	1
[500, 550 >	1
[550, 600 >	1

Relativ frekvens

$$\frac{1}{20} = 5\% = 0.05$$

$$\frac{5}{20} = 25\% = 0.25$$

...

20

Relativ frekvenstabell

Noen ganger oppgir man ikke frekvensene, men bare den prosentvise frekvensen.

$$\text{Relativ frekvens} = \frac{\text{Frekvensen i intervallet}}{\text{Totalt antall}}$$

(andel, prosentvis)

Lønnsintervall	Antall/Frekvens
[250, 300 >	5%
[300, 350 >	25%
[350, 400 >	20%
[400, 450 >	35%
[450, 500 >	5%
[500, 550 >	5%
[550, 600 >	5%

Tre typer frekvenstabeller

Den siste typen er *kumulativ* frekvenstabell. Kumulativ=Oppsamlet.

Table 2-2 Frequency Distribution: Ages of Best Actresses	
Age of Actress	Frequency
21-30	28
31-40	30
41-50	12
51-60	2
61-70	2
71-80	2

Sum = 76

Table 2-3 Relative Frequency Distribution of Best Actress Ages	
Age of Actress	Relative Frequency
21-30	37% $= \frac{28}{76}$
31-40	39%
41-50	16%
51-60	3%
61-70	3%
71-80	3%

Table 2-4 Cumulative Frequency Distribution of Best Actress Ages	
Age of Actress	Cumulative Frequency
Less than 31	28
Less than 41	58
Less than 51	70
Less than 61	72
Less than 71	74
Less than 81	76

$$\begin{aligned} &= 28 \\ &= 28 + 30 \\ &= 28 + 30 + 12 \end{aligned}$$

Oppsummering: Frekvenstabeller

- Passer for numeriske (kvantitative) data (både diskret og kontinuerlig)
- Intervallbredden skal være konstant
- Intervallene skal være disjointe slik at hver ~~per~~ observasjon passer inn i ett og bare ett intervall.

Intervallbredde: Differanse mellom startpunktene til etterfølgende intervaller.

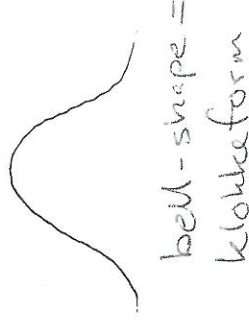
Normalfordelingen

Senere i kurset skal vi se at data ofte er *normalfordelte*.

Krav for at data skal være normalfordelte

❶ Frekvensene er lave i starten, så stiger de til en maksverdi før de avtar igjen

❷ Frekvensene er symmetriske om maksverdien



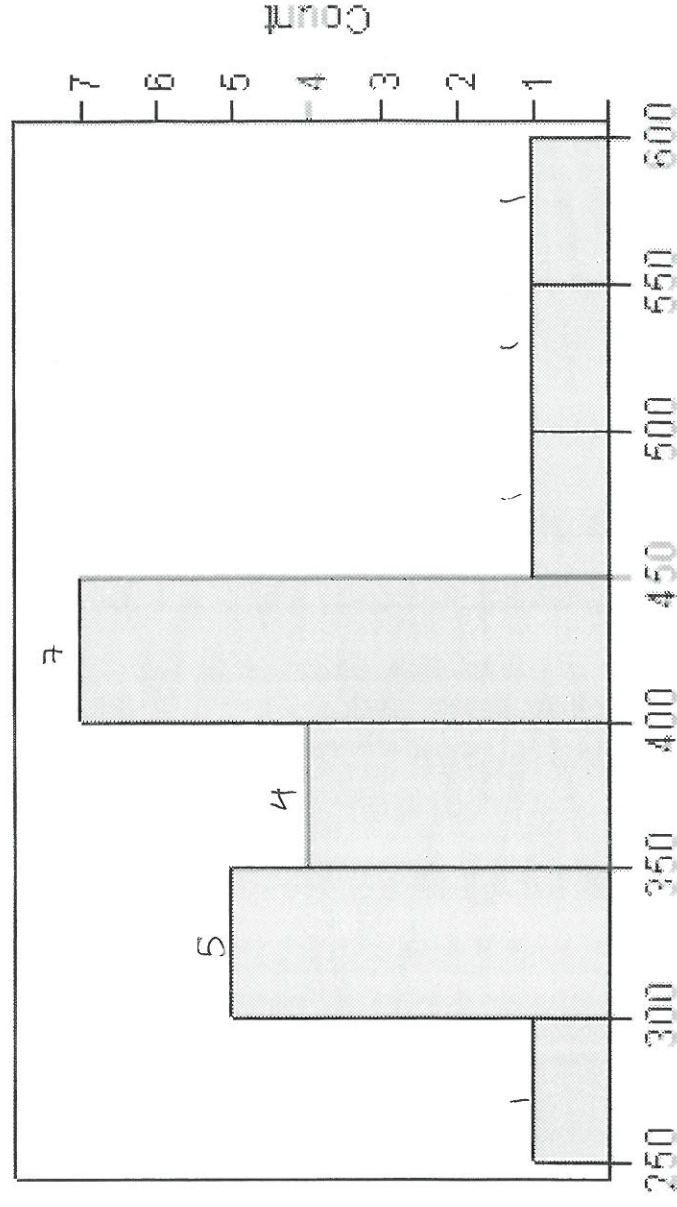
Lønnsintervall	Antall/Frekvens
[250, 300>	1
[300, 350>	5
[350, 400>	4
[400, 450>	7
[450, 500>	1
[500, 550>	1
[550, 600>	1

Ser ikke ut til å være normalfordelte. Ikke symmetrisk.

Histogrammet gir et bilde av frekvenstabellen

Lønnsintervall	Antall
[250, 300 >	1
[300, 350 >	5
[350, 400 >	4
[400, 450 >	7
[450, 500 >	1
[500, 550 >	1
[550, 600 >	1

frekvenstabell



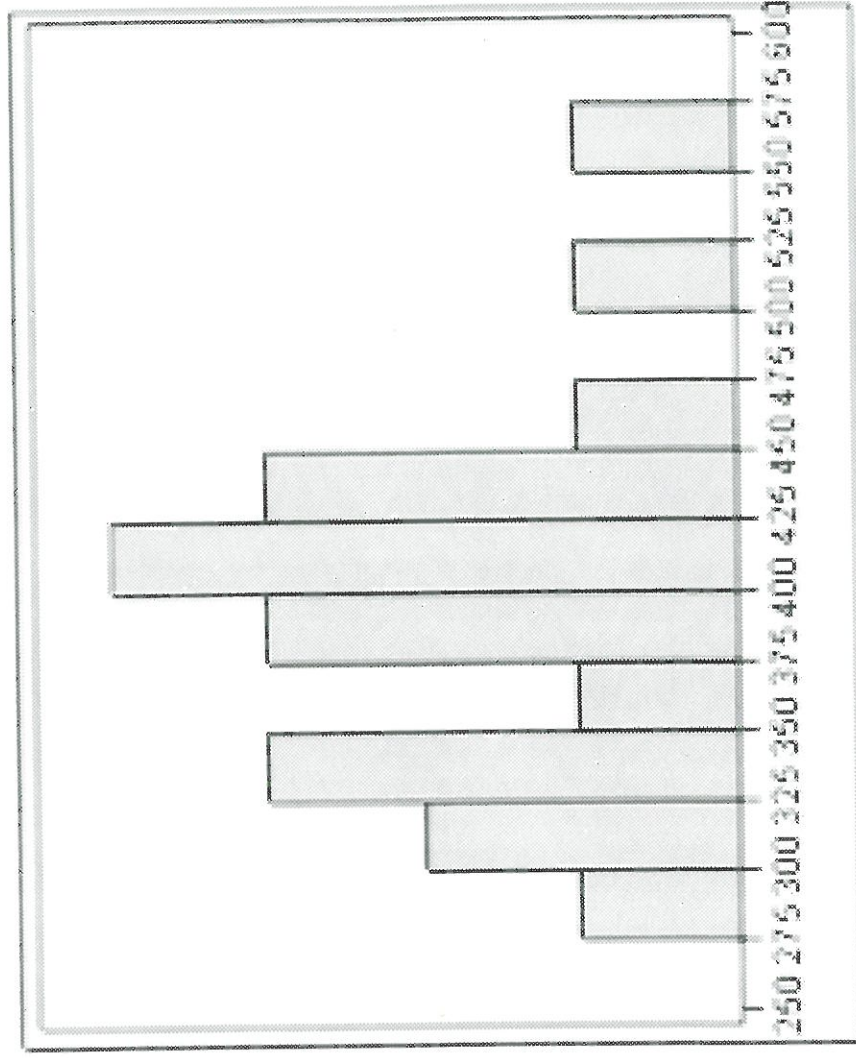
histogram

- alle stolper like bredde
- høyden svarer til frekvens til tilsvarende intervall
- Stolperne står inntil hverandre
- en stolpe = et intervall

Intervallbredde

Bredden på stolpene

Valg av intervallbredde bør gi et 5-15 stolper. Her ser du samme data med bredde 25 istedenfor 50. Ble det bedre?



Histogram

Histogram

Histogrammet er en viktig type graf som viser hvordan dataene er fordelt.

antall (frekvens)

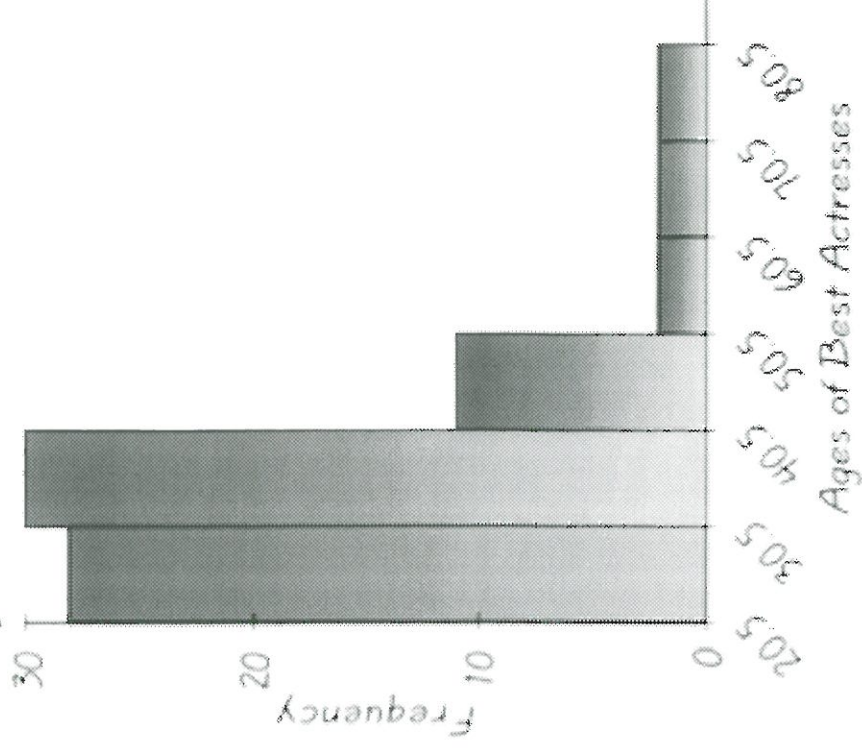


Table 2-2

**Frequency Distribution:
Ages of Best Actresses**

Age of Actress	Frequency
21-30	28
31-40	30
41-50	12
51-60	2
61-70	2
71-80	2

Relativ frekvens

Relativ frekvens histogram

På den vertikale akse kan man også bruke relativ frekvens.

Da skal enhetene
kunne være %!

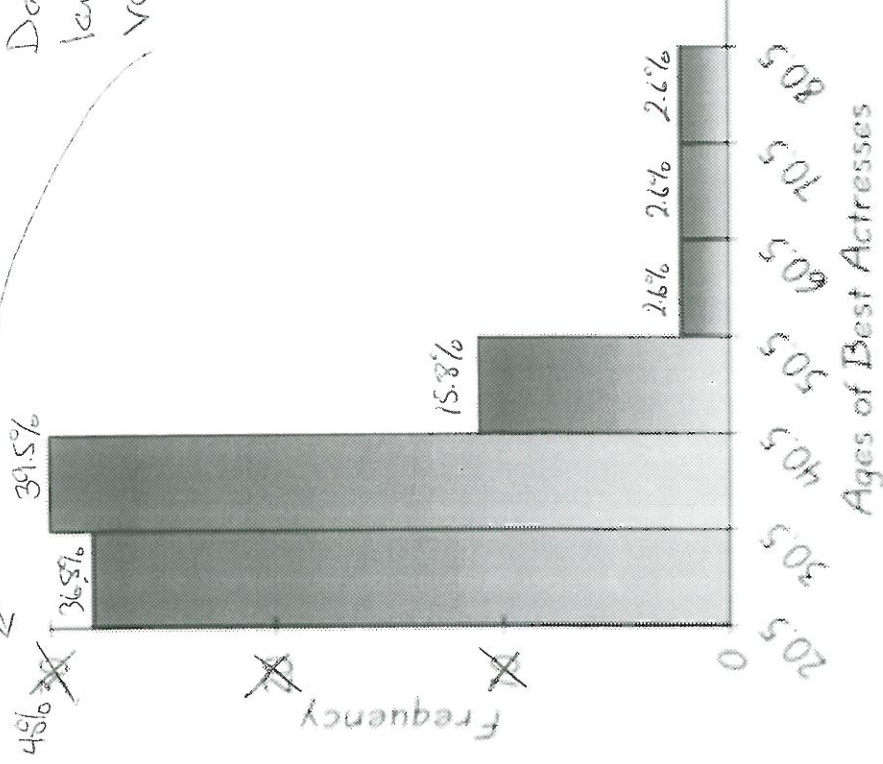


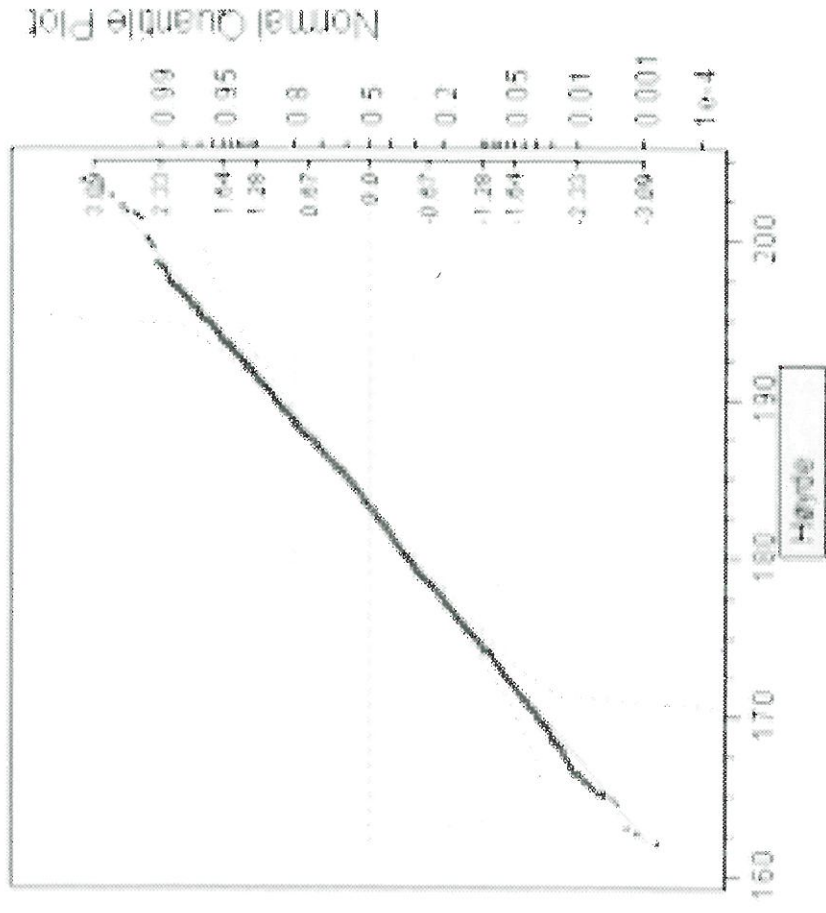
Table 2-2
Frequency Distribution:
Ages of Best Actresses

Age of Actress	Frequency
21-30	28
31-40	30
41-50	12
51-60	2
61-70	2
71-80	2

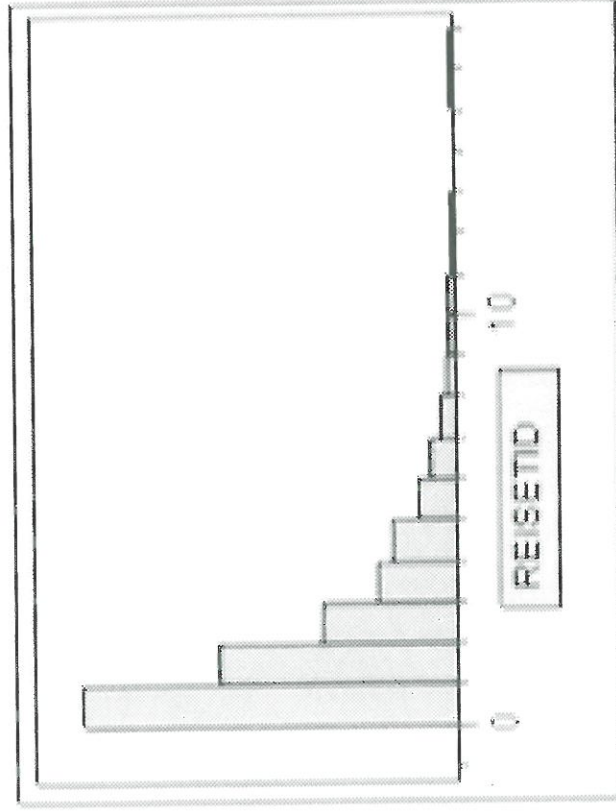
Er dataene normalfordelte? JA



Dette ser (ca)
normal fordelt ut.

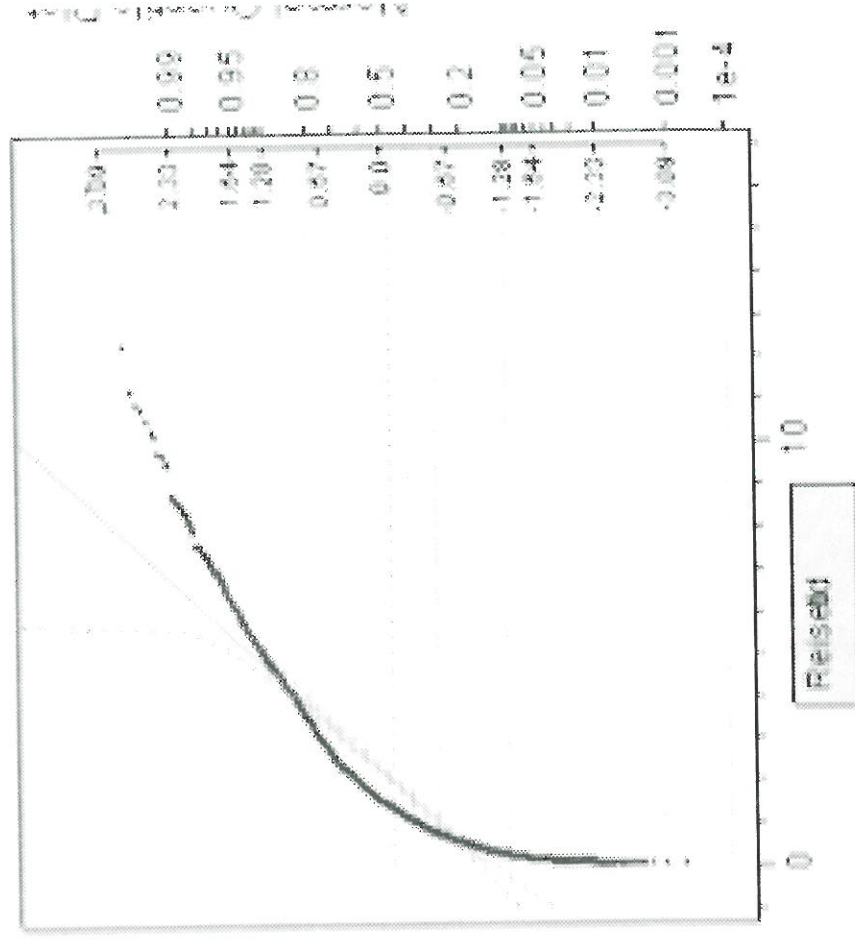


Er dataene normalfordelte? NEI



↑

dette ser ikke normalfordelt
ut - skjev fordeling
(ikke symmetrisk)



Statistiske grafer

Forskjellige typer grafer

- Histogrammet er en viktig graf
- Det viser hvordan dataene er fordelt på de ulike mulige verdiene
- Men det finnes mange flere typer grafer
- En god graf gjør at man forstår noe nytt om dataene

Løv og blad diagram

Stem (tens)	Leaves (units)	Date Svarer til observervationsnr.
2	12445556667778888999999	21,22, 24,24,...
3	001112233334445555677888899	30,30, 31,31,...
4	011111223569	...
5	04	60,61,63
6	013	74
7	4	80
8	0	

tiere 30 enere

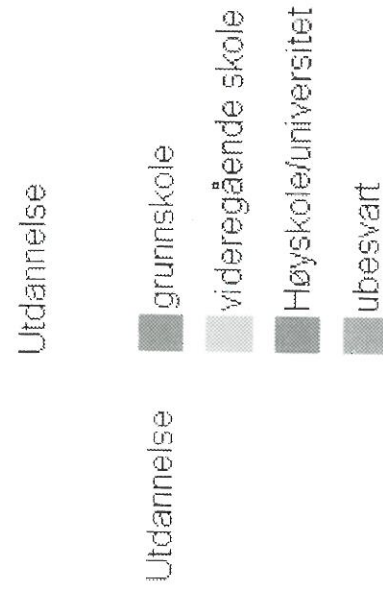
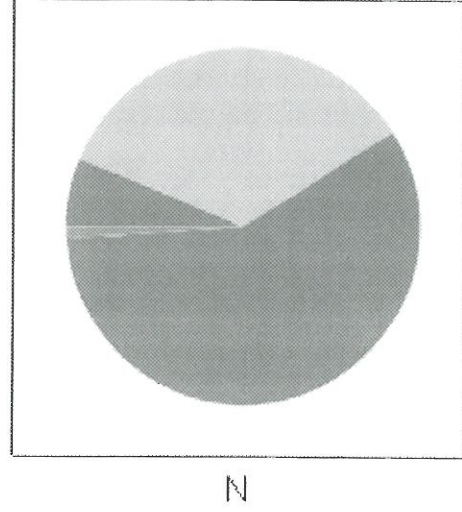
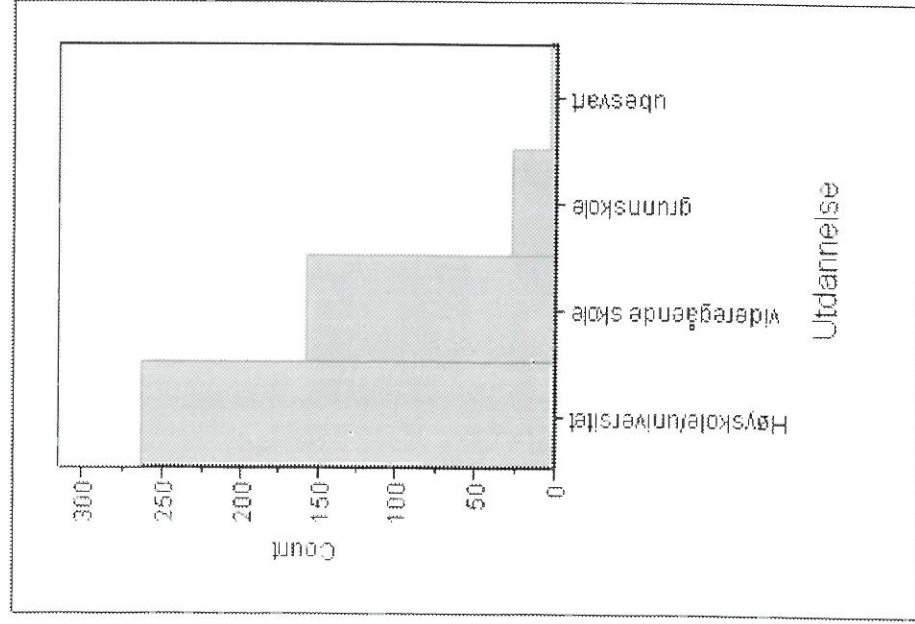
←Values are 50 and 54
 ←Value is 80.

Løv og blad diagram - eng: Stem and Leaf diagram

Tallene i stikkprøven deles i to

- 1 Stammen
- 2 Bladet

Kategoriske data: Pareto og Kake

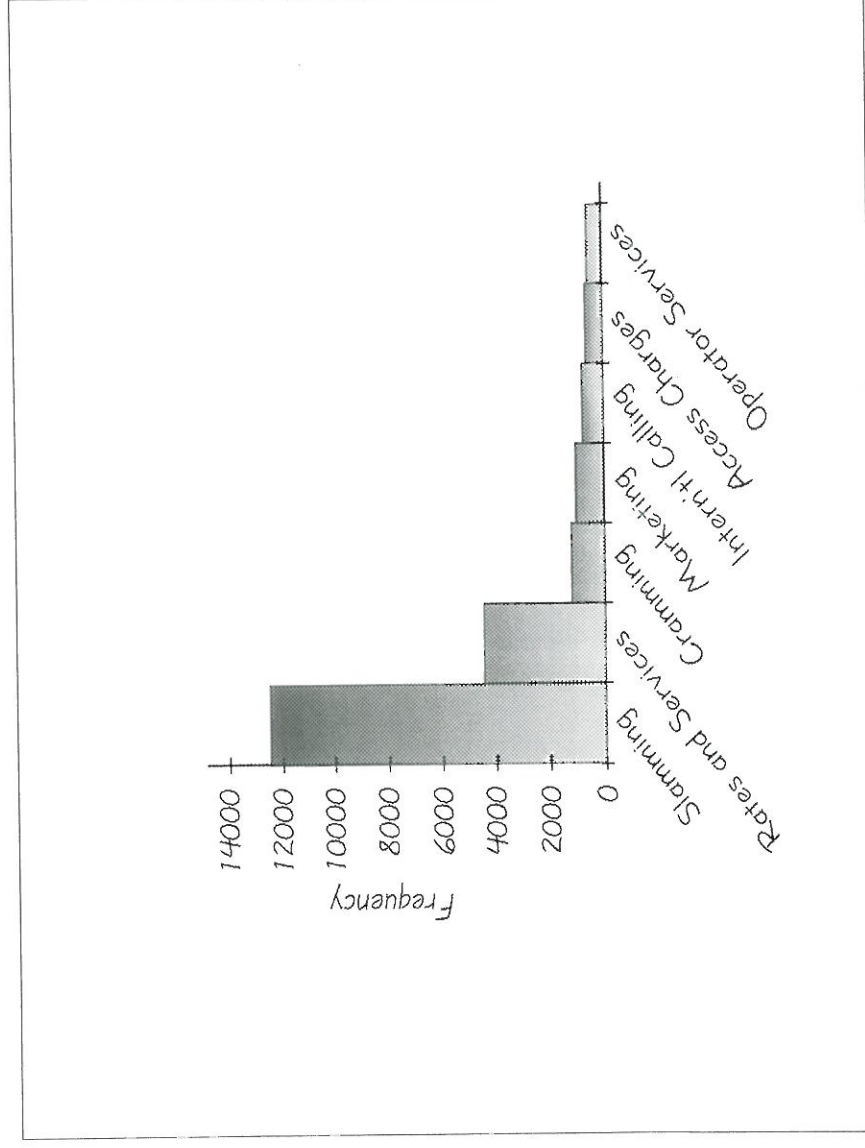


Pareto

- Kategoriske data
- Ordnet slik at høyeste frekvens er lengst til venstre

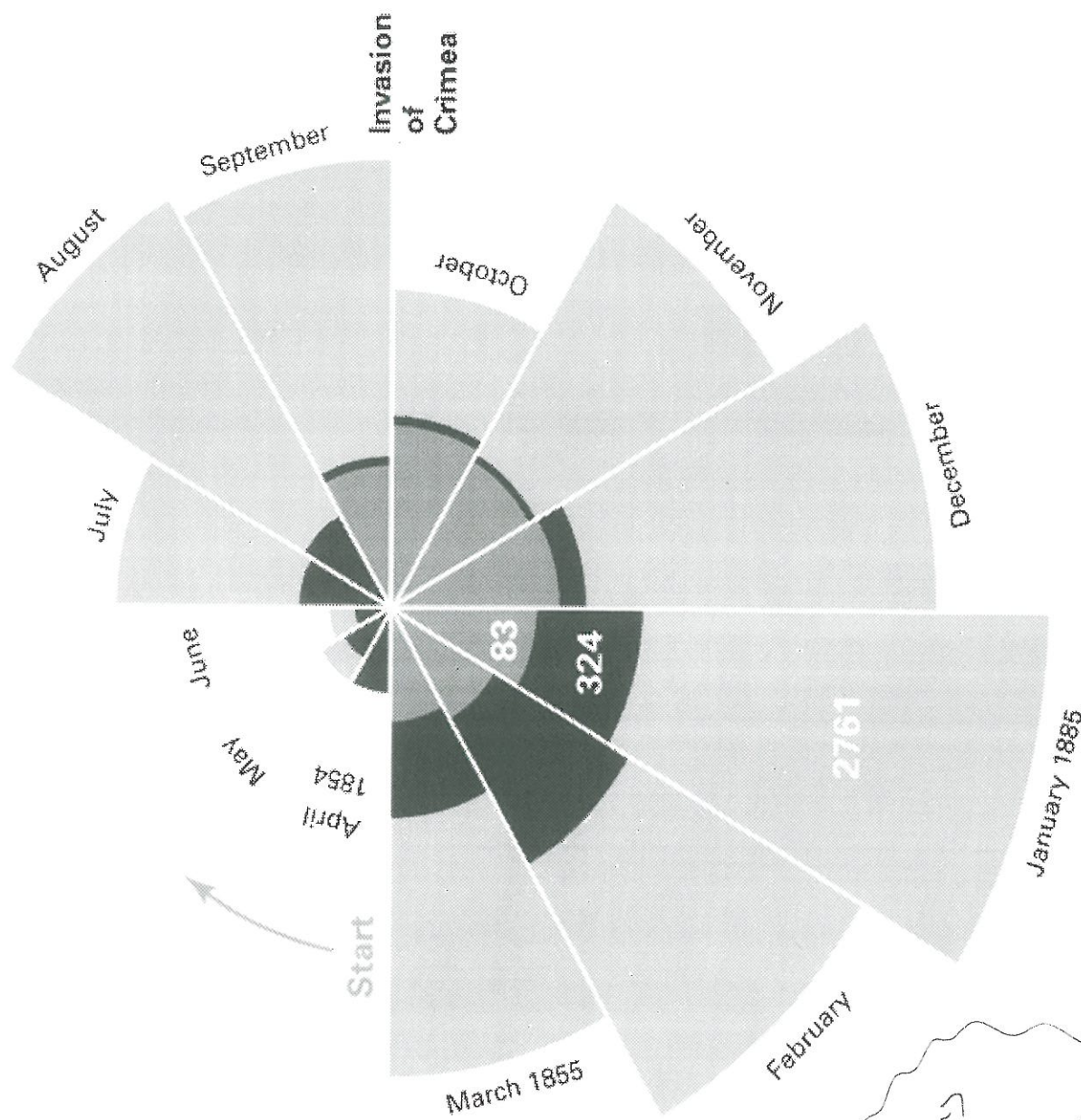
Kake

Paretodiagram



For kvalitative data. Kategorier ordnes i aftagende rekkefølge.

Mange andre typer grafer



Viste også en annen type framstilling i Forelesning 3 - Se YouTube-lenke i H's Learning

Oppsummering

Vi har sett på måter å sammenfatte store datasett på. Grafer er bilder av hvordan dataene fordeles seg. De lar oss beskrive, utforske og sammenligne data.