STK1000 Innføring i anvendt statistikk

Mandag 18. august 2008

Ingrid K. Glad

I løpet av dette kurset skal dere

- bli fortrolig med statistisk tenkemåte
- forstå teori og metoder som ligger bak knappene/menyene i vanlige statistikkpakker
- få trening i enkel analyse av data vha. dataverktøy
- lære å tolke statistiske opplysninger (spesielt i faglitteratur)

Kapittel 1: Data og fordelinger

Beskrive, forstå og utforske data

Hvordan beskrive og forstå data

Kap. 1.1 og 1.2

– Grafisk beskrivelse av data
– Sentralmål
– Spredningsmål

Kap. 1.3 om fordelinger neste uke

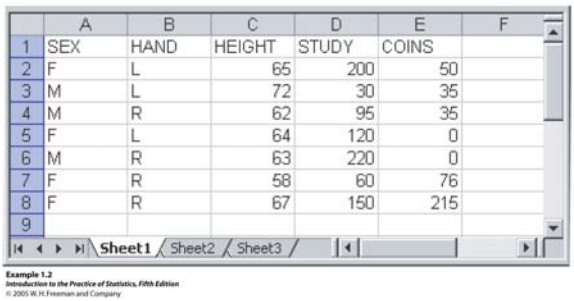
Hva er data?

- Data kommer fra et sett individer.
- Kjennetegn som kan knyttes til hvert individ organiseres i variable.
- Eksempler:

  Individer: personer, batterier, bananfluer, målestasjoner, tabletter,…

  Variable: kjønn, blodtrykk, levetid, ekspresjon av bestemt gen, lufttemperatur, vekt,...

Eksempel på data:



	A	B	C	D	E	F
1	SEX	HAND	HEIGHT	STUDY	COINS	
2	F	L	65	200	50	
3	M	L	72	30	35	
4	M	R	62	95	35	
5	F	L	64	120	0	
6	M	R	63	220	0	
7	F	R	58	60	76	
8	F	R	67	150	215	
9						

Sheet1 Sheet2 Sheet3

Example 1.2
Introduction to the Practice of Statistics, Fifth Edition

Innsamling av data

- Forsøksstudier el. observasjonsstudier
- Må planlegges (Hvilke spørsmål ønsker man å belyse, hva skal man måle/observere på hvem?)
- Statistisk forsøksplanlegging
- Her: litt i kap. 3

7

Eksplorativ dataanalyse

- Starter med å studere hver variabel for seg (Kap. 1)
- Deretter sammenhenger mellom variable (Kap. 2)
- Start med grafiske metoder (Kap. 1.1)
- Deretter numeriske oppsummeringer (Kap. 1.2)

Start alltid statistiske analyser med eksplorativ analyse!

8

To typer variable

- Kategoriske (ikke-numeriske) data
  - "god", "middels" eller "dårlig" testresultat (ordnet)
  - "6MP", "Imurel", "Prednisolon", eller "Budensonide" (ikke ordnet)
  - 'røyker' eller 'ikke-røyker' (ikke ordnet)
  - 'kvinne' eller 'mann' (ikke ordnet)
- Kvantitative (numeriske) data
  - antall fødte barn
  - antall pulsslag per minutt
  - Årsinntekt født i 1975
  - høyde, vekt
  - genekspresjon
  - temperatur

9

Fordelingen til en variabel beskriver

- Hvilke verdier variabelen kan ta
- Hvor ofte den tar disse verdiene

- Et **datasett** er et sett med observerte verdier for en eller flere variable på et antall individer. Fordelingen til en variabel kan utforskes ved hjelp av grafikk og enkle beregninger.

10

Så dette er et datasett:

	A	B	C	D	E	F
1	SEX	HAND	HEIGHT	STUDY	COINS	
2	F	L	65	200	50	
3	M	L	72	30	35	
4	M	R	62	95	35	
5	F	L	64	120	0	
6	M	R	63	220	0	
7	F	R	58	60	76	
8	F	R	67	150	215	
9						

Sheet1 Sheet2 Sheet3

Example 1.2
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

Kategoriske variable? Numeriske variable?

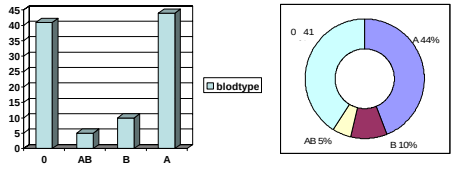
11

1.1 Fordelinger beskrevet med grafikk

Skiller mellom kategoriske og numeriske variable

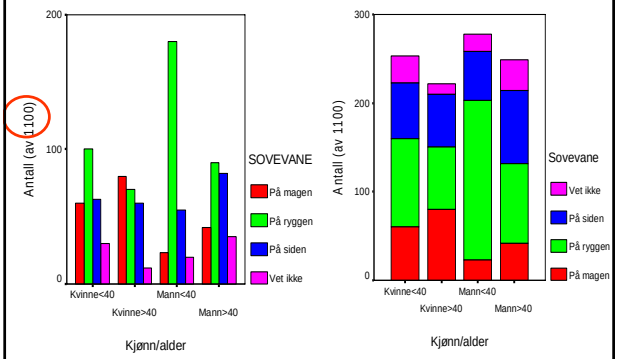
12

Diagrammer for kategoriske data



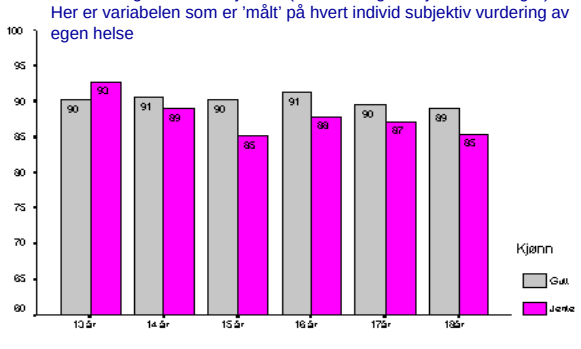
Diagrammer for kategoriske data fremstiller antall eller andel i hver kategori

Stolpe- og smultringsdiagram over blodtypefordeling



Søylediagram og stablet søylediagram for sovevaner.

Data kan også være subjektive (i motsetning til objektive målinger):
Her er variabelen som er 'målt' på hvert individ subjektiv vurdering av egen helse



Andel som betrakter sin helse som god eller svært god

Grafiske metoder for numeriske variable

Eksempel 1.4
Registrering av telefonsamtaler, kundeservice bank

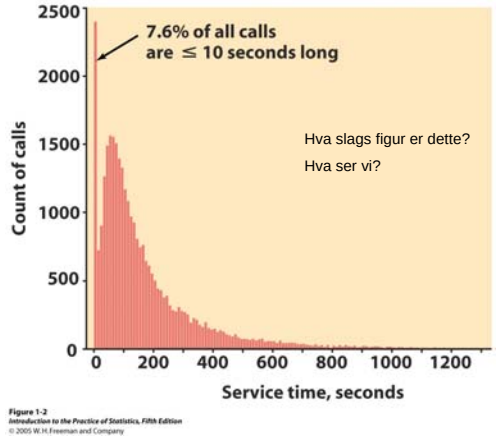
- 31492 samtaler i løpet av en måned
- Individer: hver samtale
- Variabel: lengden av samtalen (i sekunder)

De 80 første registreringene:

TABLE 1.1
Service times (seconds) for calls to a customer service center

77	289	128	59	19	148	157	203
126	118	104	141	290	48	3	2
372	140	438	56	44	274	479	211
179	1	68	386	2631	90	30	57
89	116	225	700	40	73	75	51
148	9	115	19	76	138	178	76
67	102	35	80	143	951	106	55
4	54	137	367	277	201	52	9
700	182	73	199	325	75	103	64
121	11	9	88	1148	2	465	25

Table 1-1
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company



Hva slags figur er dette?
Hva ser vi?

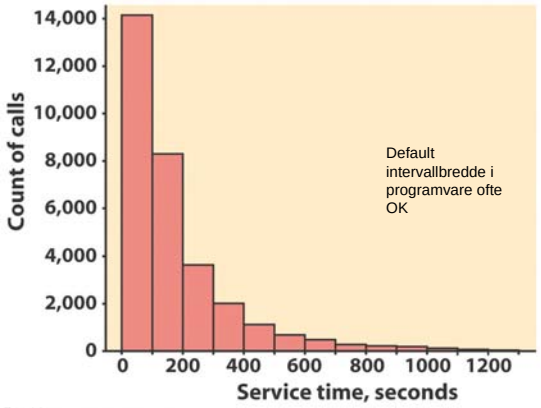
Figure 1-2
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W. H. Freeman and Company

Histogrammer

Enkleste metode:

1. Del verdiområdet til variabelen opp i intervaller
2. Tell opp antall individer i hvert intervall
3. Tegn søyler som tilsvarer antall i intervallet

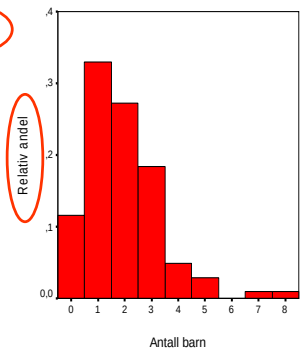
Problem: antall intervaller



Histogram med relative andeler (normert)

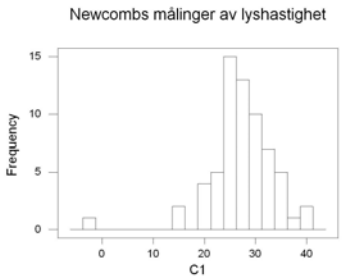
Antall fødte barn	Antall kvinner med så mange fødte barn	Relativ frekvens
0	12	0,116
1	34	0,330
2	28	0,272
3	19	0,184
4	5	0,049
5	3	0,029
6	0	0,000
7	1	0,010
8	1	0,010
Totalt	103	1.000

Frekvenstabell for kvinners fruktbarhet.



Arealet av hver søyle tilsvarer relativ andel obs. i intervallet søylen dekker

Historisk datasett



Gjentatte målinger av samme størrelse

Papir-og-blyant-alternativ til histogram:

Stilk-og-blad-plott: første siffer stilk, siste siffer blad

```
-0 2
 0
 0
 1
 1 669
 2 0112233344444
 2 55555666667777778888888899999
 3 0001122222334
 3 666679
 4 0
```

(Newcombs målinger)

Stem-and-Leaf Display: Call Length

```
Stem-and-leaf of C1   N   = 80
Leaf Unit = 10

     0   00000000111233444
     0   555555566677777778889
     1   00001112223344444
     1   57789
     2   0012
     2   7789
     3   2
     3   678

HI 43, 46, 47, 70, 70, 95, 114, 263
```

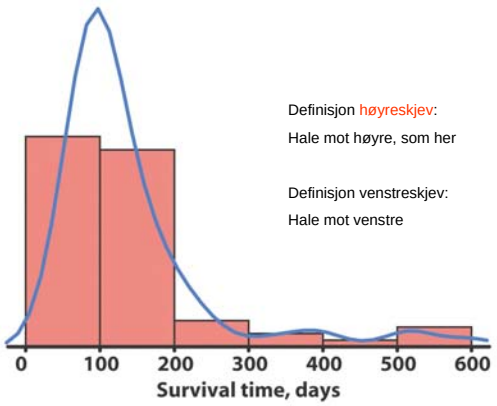
Hva ser vi etter?

Når vi vurderer fordelingen til datasettet ser vi spesielt etter

-Form, senter og spredning (en eller flere topper, symmetrisk eller skjev, midtpunkt, minste og største verdi...)

-Uteliggere (outliers) (typisk målinger der noe gikk galt, men kan også være reelle, dvs. tegn på skjeve fordelinger el. tunge haler. Forsøk alltid å finne en forklaring!)

25



Definisjon høyreskjev:

Hale mot høyre, som her

Definisjon venstreskjev:

Hale mot venstre

Figure 1-22b
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H.Freeman and Company

26

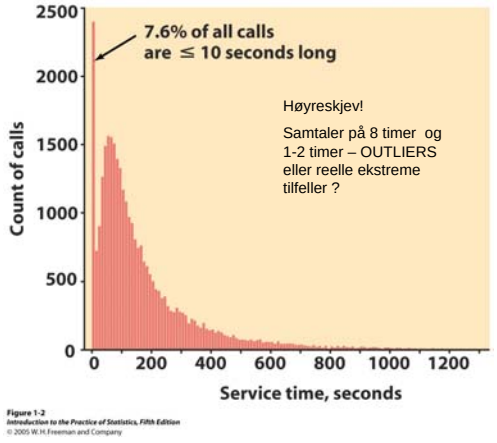
TABLE 1.8

Survival times (days) of guinea pigs in a medical experiment

43	45	53	56	56	57	58	66	67	73
74	79	80	80	81	81	81	82	83	83
84	88	89	91	91	92	92	97	99	99
100	100	101	102	102	102	103	104	107	108
109	113	114	118	121	123	126	128	137	138
139	144	145	147	156	162	174	178	179	184
191	198	211	214	243	249	329	380	403	511
522	598								

Table 1-8
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H.Freeman and Company

27



Høyreskjev!

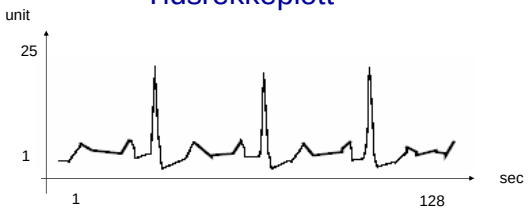
Samtaler på 8 timer og 1-2 timer – OUTLIERS eller reelle ekstreme tilfeller ?

Figure 1-2
Introduction to the Practice of Statistics, Fifth Edition
© 2005 W.H.Freeman and Company

28

Andre typer plott for eksplorativ dataanalyse

Tidsrekkeplott

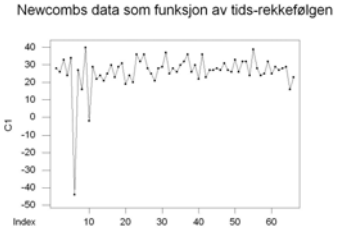


Observasjoner gjort over tid

29

Rekkefølgen forsøkene ble foretatt i

Ta hensyn til læring!



1.2 Fordelinger beskrevet med tall

Grafisk fremstilling suppleres med numeriske mål (tall!) som beskriver fordelingen ytterligere

- Sentralmål (beliggenhet)
- Spredningsmål

31

	Kjønn	Høyde (cm)
1	G	178,0
2	G	177,0
3	J	164,0
4	G	185,0
5	J	156,0
6	J	176,0
7	G	152,0
8	G	186,0
9	G	192,0
10	J	181,0
11	J	145,0
12	J	168,0
13	J	174,0
14	G	178,0
15	J	171,0
16	G	184,0
17	G	182,0
18	J	155,0
19	J	167,0
20	J	169,0

Oversikt over kjønn og høyde for 20 studenter, 9 gutter og 11 jenter

Gjennomsnitt - sentralmål

Gjennomsnittlig høyde for alle studentene:

$$\bar{x} = \frac{178+177+164+\cdots+169}{20} = 171{,}95$$

Gjennomsnittlig høyde for guttene:

$$\overline{x_g} = \frac{178 + 177 + 185 + \cdots + 182}{9} = 179{,}22$$

Gjennomsnittlig høyde for jentene:

$$\overline{x_j} = \frac{164 + 156 + 176 + \cdots + 169}{11} = 166{,}00$$

32

Def. gjennomsnitt (mean)

33

	Kjønn	Høyde (cm)
1	J	145
2	G	152
3	J	155
4	J	156
5	J	164
6	J	167
7	J	168
8	J	169
9	J	171
10	J	174
11	J	176
12	G	177
13	G	177
14	G	178
15	J	181
16	G	182
17	G	184
18	G	185
19	G	186
20	G	192

ordne data!

Median M - sentralmål: halvparten av observasjonene er mindre enn M, halvparten er større.

Median høyde for alle studentene blir

$$\tilde{x} = \frac{x_{(10)} + x_{(11)}}{2} = \frac{174+176}{2}$$

Median høyde for (9) guttene:

$$\tilde{x}_g = x_{g(5)} = 182$$

Median høyde for (11) jentene:

$$\tilde{x}_j = x_{j(6)} = 168$$

Tabell: Oversikt over kjønn og høyde for 20 studenter (sortert etter høyde)

34

Def. median M

- Medianen M i et datasett med n observasjoner er et tall slik at halvparten av observasjonene er mindre enn tallet og den andre halvparten er større
- n oddetall:
  M = midterste observasjon
- n partall:
  M = gjennomsnitt av de to midterste observasjonene

35

Gjennomsnitt vs. Median

Forskjellen mellom gjennomsnitt og median, eksempler :

(a) 1, 2, 9 ..... median 2.0 .... gjennomsnitt 4.0
(b) 1, 8, 9 ..... median 8.0 .... gjennomsnitt 6.0
(c) 1, 2, 8, 9 ... median 5.0 .... gjennomsnitt 5.0
(d) 1, 2, 3, 9 ... median 2.5 .... gjennomsnitt 3.75

I (c) er de to sentralmålene like. Dette er kun tilfelle når fordelingen er symmetrisk. I skjeve fordelinger ligger gjennomsnittet lenger ut i halen (d). Beregning av begge er nyttig for å vurdere skjevhet.

Gjennomsnittet er svært følsomt for ekstreme observasjoner. Medianen er mer robust i forhold til disse.

(e) 1, 2, 35 ..... median 2 ..... mean 12.7
(f) 1, 2, 350 ..... median 2 ..... mean 119.3

36

•Range: (minimum, maksimum) - spredningsmål

evt. maksimum - minimum

•IQR, interkvartil avstand: - spredningsmål

p% persentil (fraktil): p % av obs. er mindre enn dette tallet.

M = median = 50% persentil

Q1 = 1. kvartil = 25% persentil

Q3 = 3. kvartil = 75% persentil

37

Q1 og Q3 beregner vi lettest som medianen i de obs. som er hhv. mindre og større enn medianen M

n partall: 1 2 3 4 5 | 6 7 8 9 10

Q1 Q3

n oddetall: 1 2 | 3 4 5 6 7 | 8 9

IQR = Q3-Q1 = 'Inter Quartile Range'

= det intervallet de midterste 50% av observasjonene ligger i

38

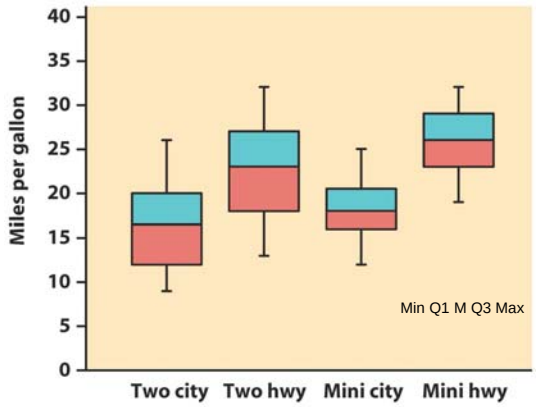
Fem-talls-oppsummering

- Et datasett oppsummeres ofte med fem størrelser:

Min Q1 M Q3 Max

Et boxplott er en grafisk fremstilling av disse!

39

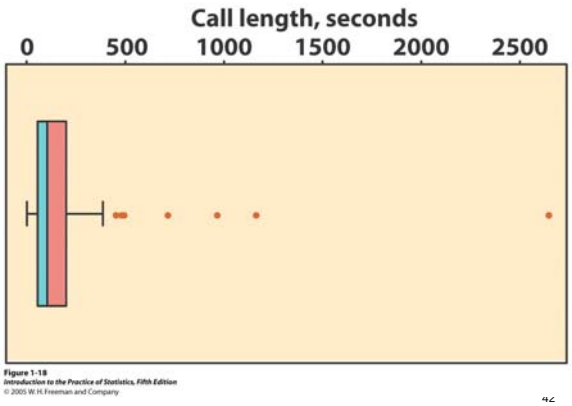


Et boksplott er mindre informativt enn et histogram, men egner seg godt til å sammenligne to eller flere datasett!

OUTLIERS: 1.5xIQR-kriteriet

Hvis en observasjon er **større** enn Q3+1.5xIQR eller **mindre** enn Q1-1.5xIQR, så er observasjonen en **potensiell** uteligger.

41

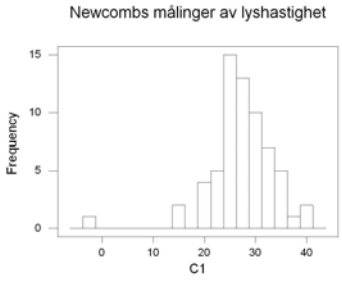


Boksplott detaljer

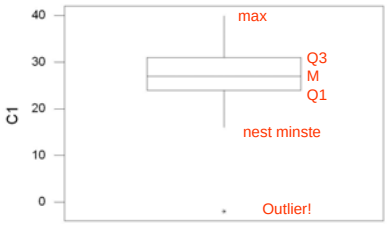
Default Minitab:

- Boks med horisontale linjer korresponderende til Q_1, median og Q_3
- Linje oppover fra boks opp til største observasjon som er mindre enn $Q_3 + 1.5IQR$ og linje nedover fra boks ned til minste observasjon som er større enn $Q_1 - 1.5IQR$.
- Punktvise plott av potensielle outliere

Alternativ: Horisontale streker til max og min (boken)



Boxplott Newcombs data, fra Minitab



Spredning: Varians og standardavvik

Data $x_1, x_2, ..., x_n$

Kvadratavvik fra gjennomsnitt:

$$(x_1 - \bar{x})^2, (x_2 - \bar{x})^2, ..., (x_n - \bar{x})^2$$

Varians s^2: Gjennomsnitt av kvadratavvik

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + ... + (x_n - \bar{x})^2}{n-1}$$

$$= \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

Standard avvik $s = \sqrt{s^2}$

Standardavvik eksempel

Stoffskifteraten måler hvor raskt kroppen forbruker energi (i kalorier per døgn). Viktig for studier av dietter og aktivitet

Data fra 7 menn i en studie

1792 1666 1362 1614 1460 1867 1439

Gjennomsnitt 1600

Avvik 192 66 -238 14 -140 267 -161

Kvadratisk avvik

36864 4356 56644 196 19600 71289 25921

$$\text{Varians} = s^2 = \frac{214870}{6} = 35811.67$$

$$\text{Standard avvik} = s = \sqrt{35811.67} = 189.24$$

Standardavvik – eksempel

Kvinne nr.	Varighet i dager av menstruasjonsperioder
1	25 25 26 29 25 24 24 25 29 24 23 22 27 24 27 25
2	26 24 27 29 26 29 28 26 23 29 27 27 23 22 25 23
3	30 35 32 37 34 29 35 35 28 29 30 27 32 27

$$\bar{x}_1 = \frac{1}{16}(25 + 25 + 26 + \cdots + 25) = 25{,}25$$

$$s_1 = \sqrt{\frac{1}{15}[(25 - 25{,}25)^2 + (25 - 25{,}25)^2 + (26 - 25{,}25)^2 + \cdots + (25 - 25{,}25)^2]} = 1{,}95$$

$\bar{x}_2 = 25{,}88$ $\quad s_2 = 2{,}33$

$\bar{x}_3 = 31{,}43$ $\quad s_3 = 3{,}32$

Kommentarer

- Hvorfor kvadrerer vi avvikene? (alternativer?)
  1. Naturlig for **normalfordeling** (Kap.1.3), matematisk bekvemt
  2. Gjennomsnitt er naturlig senter mhp kvadratsum:

$$\sum(x_i - a)^2 \geq \sum(x_i - \bar{x})^2$$

- Hvorfor standard avvik og ikke varians?
  1. s på samme skala som data (samme benevning)

- Hvorfor dele på $n - 1$?
  Har $\sum(x_i - \bar{x}) = 0$. Dvs. hvis vi kjenner

$$(x_1 - \bar{x}), (x_2 - \bar{x}), \ldots, (x_{n-1} - \bar{x})$$

så kan vi regne ut $x_n - \bar{x}$.

Vi sier at vi kun har $n - 1$ **frie** avvik, og deler på $n - 1$.

Terminologi: s^2 og s har $n - 1$ **frihetsgrader**.

Egenskaper standardavvik

- Måler spredning rundt **gjennomsnitt** og bør kun brukes sammen med denne.
- $s = 0$ betyr **ingen** spredning, dvs alle observasjoner er like.

  Ellers er $s > 0$.
- s, som $\bar{x}$, er **ikke** robust.
  Noen få outliere kan gjøre s veldig stor.

Oppsummering – hva skal vi velge?

To hovedalternativer

- fem siffer oppsummering
  min Q_1 Median Q_3 max
- $\bar{x}$ og s

Generelt prinsipp:
Fem siffers oppsummering er vanligvis bedre enn kun $\bar{x}$ og s for å beskrive en **skjev** fordeling eller fordeling med outliers.

$\bar{x}$ og s er ideelt for symmetriske fordelinger uten outliers.

Endring av skala

Enkelt å transformere oppsummerende mål til annen skala.

Fordi: Endring av skala er en **lineær** transformasjon:

$$x_{\text{ny}} = a + bx$$

a svarer til endring av null-punkt.

b svarer til endring av skala.

Kilometer til miles
Hvis x er i kilometer, så er

$$x_{\text{ny}} = 0.62 * x$$

i miles.
Endring av skala, men ikke av null-punkt.

Fahrenheit til Celsius
Hvis x er temperatur i Fahrenheit, så er

$$x_{\text{ny}} = \frac{5}{9}(x - 32) = -\frac{160}{9} + \frac{5}{9}x$$

temperatur i celsius.

Endring både av null-punkt og skala

Lineær transformasjon

- Endrer ikke **form** på fordeling:
  - x høyre-skjev medfører $x_{\text{ny}} = a + bx, b > 0$ høyreskjev
  - x symmetrisk og entoppet medfører x_{ny} symmetrisk, entoppet
- Senter og spredning **vil** endre seg i henhold til enkle regneregler.

55

Regler for effekten av lineær transformasjon

- Multiplisering av observasjoner med $b > 0$ multipliserer både
  - mål for senter (gjennomsnitt, median) og
  - mål for spredning (interkvartil avstand, standard avvik)

  med b.
- Ved å legge et tall a til alle observasjoner, så vil det også føre til en endring på a for
  - mål for senter
  - kvartiler og persentiler

  men medfører ingen endring i spredning.

56