

# Kapitel 3

## Central tendens og spredning

Peter Tibert Stoltze  
stat@peterstoltze.dk

Elementær statistik  
F2011

1 / 25

## Indledning

- ▶ I kapitel 2 omsatte vi de rå data til en tabel, der bedre viste materialets fordeling
- ▶ Fordelingen illustrerede vi med forskellige former for grafik
- ▶ Nu vil vi gerne karakterisere fordelingerne kvantitativt gennem deres **beliggenhed** og **variation**

2 / 25

# Centraltendens

- ▶ Typetal eller modus (eng: mode)
- ▶ Aritmetisk middelværdi eller stikprøvegennemsnit (eng: mean or sample mean)
- ▶ Median

3 / 25

## Modus

- ▶ **Modus** eller typetallet er den hyppigst forekommende værdi
- ▶ Eneste anvendelige mål for data målt på nominalskala (m/k, ja/nej) men vel nok mest anvendt ifm. data målt på ordinalskala (karakterer, scores)
- ▶ Læsescores: 71 for pigerne mod 55 for drengene
- ▶ Beregnes i Excel med funktionen `hyppigst`

4 / 25

## Aritmetisk middelværdi

- ▶ Betegnes også tit som stikprøvegennemsnit
- ▶ Den **aritmetiske middelværdi** er summen af observationerne i forhold til antallet af observationer

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \dots + x_n}{n}$$

- ▶ Læsescores: 59,7 for pigerne mod 51,9 for drengene
- ▶ Beregnes i Excel med funktionen `midde1`

5 / 25

## Median

- ▶ **Medianen** er den midterste observation (gennemsnittet af de to midterste hvis  $n$  er lige)
- ▶ Specialtilfælde af generelt fraktilbegreb, som vi gennemgår senere
- ▶ Læsescores: 62,5 for pigerne mod 50,5 for drengene
- ▶ Beregnes i Excel med funktionen `median`

6 / 25

## Centraltendenser for læsescores

Score	Drenge	Piger
$n$	30	30
Modus	55	71
$\bar{x}$ (stikprøvegennemsnit)	51,9	59,7
Median	50,5	62,5

7 / 25

## Spredning og varians

- ▶ Spredning kaldes også for standardafvigelse
- ▶ Spredning på læsescores er 16,6 for pigerne mod 19,2 for drengene — der er større spredning på drengenes score end på pigernes, der i gennemsnit ligner hinanden mere
- ▶ Med Excel benyttes funktionerne `stdafv` og `varians` til beregning af spredning og varians

8 / 25

# Spredning og varians

- ▶ **Spredningen**  $s$  er kvadratroden af **variansen**  $s^2$
- ▶ Variansen  $s^2$  er kvadratet på spredningen  $s$
- ▶ Variansen beregnes efter følgende formel:

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

hvor

$$\text{SAK}_x = \sum_{i=1}^n x_i^2 - \frac{1}{n} \left( \sum_{i=1}^n x_i \right)^2$$

- ▶ Hvis der *ikke* er tale om en stikprøve *kan* man benytte  $n$  i stedet for  $n - 1$  i nævneren, men ...

9 / 25

## $n - 1$ giver centrale estimater

						$\frac{\text{SAK}}{n}$	$\frac{\text{SAK}}{n-1}$
Population	7	9	13	14	27	48,80	61,00
Stikprøve 1	7	9	13	14		8,19	10,92
Stikprøve 2	7	9	13		27	61,00	81,33
Stikprøve 3	7	9		14	27	60,69	80,92
Stikprøve 4	7		13	14	27	53,19	70,92
Stikprøve 5		9	13	14	27	45,69	60,92
Gennemsnit						45,75	61,00

# Standardafvigelse og standardfejl

- ▶ Spredning kaldes også for **standardafvigelse** (eng: standard deviation)
- ▶ Må ikke forveksles med **standardfejl** (eng: standard error), der er spredningen på middelværdien:

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Dette vender vi tilbage til i kapitel 4

11 / 25

## Først lidt opsamling på nomenklatur

- ▶ Vi regner på en simpelt tilfældigt udtaget **stikprøve** fra en **population**.
- ▶ Vi antager at observationerne er målt på en intervalskala eller en ratioskala.
- ▶ Gennemsnittet i stikprøven  $\bar{x}$  er et **estimat** for middelværdien  $\mu$  i populationen:

$$\hat{\mu} = \bar{x}$$

- ▶ Spredningen i stikprøven  $s$  er estimat for spredningen  $\sigma$  i populationen:

$$\hat{\sigma} = s$$

- ▶ Spredningen er målt på samme skala som observationerne — det er variansen ikke.

12 / 25

## Praktisk beregning af middelværdi og spredning

- ▶ Data (stikprøven) præsenteres på en liste på følgende måde

$$x = \{9; 2; 8; 11; 16; 16; 6; 5; 4; 6\}$$

- ▶ Start med at bestemme antallet af observationer

$$n = 10$$

- ▶ og dernæst summen af observationerne (indeks på sumtegnene er udeladt, da der summeres over alle observationer)

$$\sum x = 9 + 2 + 8 + 11 + 16 + 16 + 6 + 5 + 4 + 6 = 83$$

- ▶ Nu kan du bestemme gennemsnittet i stikprøven

$$\bar{x} = \frac{\sum x}{n} = \frac{83}{10} = 8,3$$

13 / 25

## Praktisk beregning af middelværdi og spredning

- ▶ Så beregner du summen af de kvadrerede observationer (kvadratsummen)

$$\sum x^2 = 9^2 + 2^2 + 8^2 + 11^2 + 16^2 + 16^2 + 6^2 + 5^2 + 4^2 + 6^2 = 895$$

- ▶ så du kan beregne summen af de kvadrerede afvigelser fra gennemsnittet (summen af afvigelsesernes kvadrater)

$$SAK_x = \sum (x_i - \bar{x})^2 = \sum x^2 - \frac{1}{n} (\sum x)^2 = 895 - \frac{83^2}{10} = 206,1$$

- ▶ Så kan du beregne variansen for xerne i stikprøven

$$\hat{v}\text{ar}(x) = \frac{SAK_x}{n-1} = \frac{206,1}{10-1} = 22,9$$

- ▶ og endelig spredningen (stikprøvestandardafvigelsen)

$$s = \sqrt{\hat{v}\text{ar}(x)} = \sqrt{22,9} = 4,8$$

14 / 25

# Fraktiler

- ▶ Om  $P\%$ -fraktilen gælder, at  $P$  procent af observationerne er mindre end eller lig denne værdi
- ▶ Der er nogle fraktilværdier, man ofte er specielt interesseret i:
  - ▶ Medianen (50)
  - ▶ Kvartiler (25, 50, 75)
  - ▶ Deciler (10, 20, ..., 90)
  - ▶ Percentiler (1, 2, ..., 99)
- ▶ Specielt er 25% fraktilen den nedre kvartil og 75% fraktilen den øvre kvartil, og forskellen mellem øvre og nedre kvartil kaldes for interkvartilafstanden (IQR)

15 / 25

## Direkte beregning af fraktiler

- ▶ Lad en stikprøve med  $n$  elementer være opstillet i rækkefølge, således at  $x_1$  er den mindste observation og  $x_n$  er den største, da er den  $i$ 'te observation  $P$ -fraktilen i stikprøven, hvor

$$P = \frac{i - 0,5}{n}$$

- ▶ For store stikprøver er således  $P \approx 0$  for  $i = 1$  og  $P \approx 1$  for  $i = n$
- ▶ Ønsker man at kende en bestemt fraktil, da kan man regne baglæns i ovenstående udtryk, hvor resultatet dog kun sjældent vil være heltalligt  $i$  og dermed en bestemt observation. Dette kan løses ved **lineær interpolation**...

16 / 25

## Beregning af fraktiler for grupperet data

$$P\% = L + \frac{k(\frac{Pn}{100} - F)}{f}$$

hvor

- ▶  $P$  er den ønskede fraktil
- ▶  $L$  nedre grænse i klassen, hvor den ønskede fraktil befinder sig
- ▶  $k$  er klassebredden
- ▶  $n$  er antal observationer
- ▶  $F$  er antal observationer op til nedre grænse i den klasse, hvor fraktilen befinder sig
- ▶  $f$  er antal observationer i den klasse, hvor fraktilen befinder sig

17 / 25

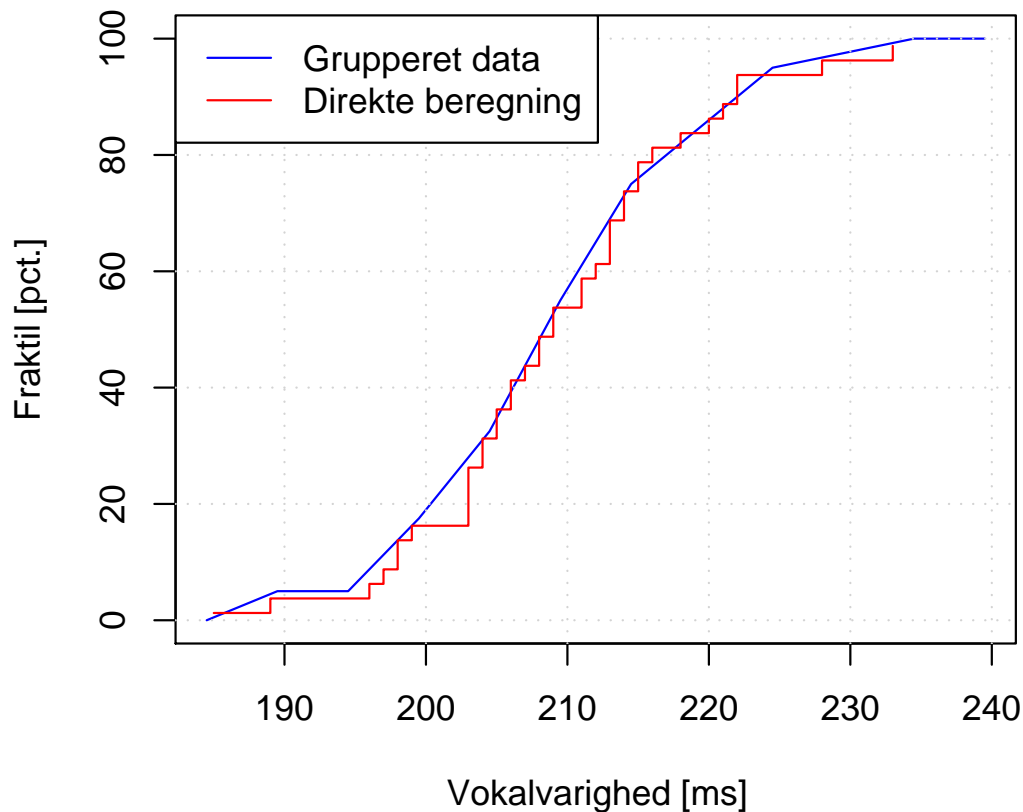
## Eksempel: Vokalvarighed

Frekvensfordeling for vokalvarighed i ms. Klassebredde er 5ms.

Nedre	Øvre	Frekvens	Kumulativ frekvens	Relativt
	184,5	0	0	0,0
184,5	189,5	2	2	5,0
189,5	194,5	0	2	5,0
194,5	199,5	5	7	17,5
199,5	204,5	6	13	32,5
204,5	209,5	9	22	55,0
209,5	214,5	8	30	75,0
214,5	219,5	4	34	85,0
219,5	224,5	4	38	95,0
224,5	229,5	1	39	97,5
229,5	234,5	1	40	100,0
234,5		0	40	100,0
		40		

18 / 25

# Kumulativ fordeling af vokalvarighed



19 / 25

## Eksempel på beregning

- ▶ Vi vil beregne 50% fraktilen (medianen) for datasættet med vokalvarighed:

$$Median = L + \frac{k(\frac{Pn}{100} - F)}{f} = 204,5 + \frac{5(\frac{50 \cdot 40}{100} - 13)}{9} = 208,39$$

- ▶ Vi kunne også lave interpolation i tabellen:

204,5	32,5
$x^*$	50,0
209,5	55,0

dvs. bestemme  $x^*$

20 / 25

## Lineær interpolation — generelt

- ▶ Vi kender punkterne  $(x_1, y_1)$  og  $(x_2, y_2)$  og ønsker at bestemme punktet  $(x^*, y^*)$  idet vi kender den ene af koordinaterne.
- ▶ Vi antager at der gælder

$$\frac{y_2 - y_1}{x_2 - x_1} = \frac{y^* - y_1}{x^* - x_1}$$

- ▶ Kender vi  $x^*$  kan vi bestemme  $y^*$  som

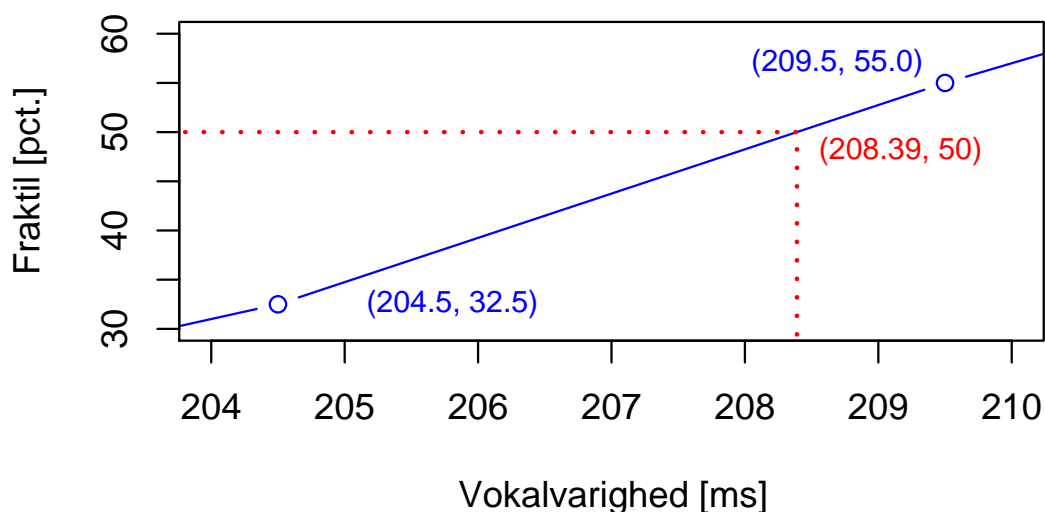
$$y^* = y_1 + \frac{y_2 - y_1}{x_2 - x_1} \cdot (x^* - x_1)$$

- ▶ Kender vi omvendt  $y^*$  kan vi bestemme  $x^*$  som

$$x^* = x_1 + \frac{x_2 - x_1}{y_2 - y_1} \cdot (y^* - y_1)$$

21 / 25

## Lineær interpolation — grafisk



$$\begin{aligned} \frac{55,0 - 32,5}{209,5 - 204,5} &= \frac{50,0 - 32,5}{x^* - 204,5} \\ \Rightarrow x^* &= 204,5 + \frac{209,5 - 204,5}{55,0 - 32,5} \cdot (50,0 - 32,5) = 208,39 \end{aligned}$$

22 / 25

## Beregning med Excel

- ▶ Beregnes i Excel med funktionen `fraktil`
- ▶ Der benyttes her en lidt anden definition end den her anvendte, men resultaterne minder en del om hinanden (specielt for store  $n$ )
- ▶ Beregning med Excel af de tre kvartiler samt interkvartilafstand ( $IQR$ ) og spredning ( $s$ ) for læsescores:

Fraktil	Drenge	Piger
25% fraktil	39,3	55,0
50% fraktil	49,0	62,7
75% fraktil	65,0	67,6
$IQR$	25,8	12,6
$s$	19,2	16,6

23 / 25

## Omsamling omkring fraktiler

- ▶ Om  $P\%$ -fraktilen for et datasæt gælder, at  $P$  procent af observationerne i datasættet er mindre end eller lig denne værdi
- ▶ Specielle fraktiler har navne som kvartiler, deciler og percentiler, men det er altså alle fraktiler
- ▶ Beregning kan foretages direkte på stikprøven, typisk vha en regnearksfunktion
- ▶ Der kan også laves beregning for grupperede data — enten med en lidt kryptisk formel eller ved lineær interpolation

24 / 25

# Opsamling

- ▶ Vi regner på en simpelt tilfældigt udvalgt **stikprøve** fra en **population**
- ▶ **Centraltendens** kan beskrives ved modus, stikprøvegennemsnit og median
- ▶ **Spredning** kan beskrives ved variationsområde, stikprøvestandardafvigelse og interkvartilafstand
- ▶ Den samlede fordeling kan beskrives med den **empiriske fordelingsfunktion**, dvs. fraktilværdien afsat som funktion af observationsværdien.