Grundlæggende univariat analyse

Statistik E24 (15 ECTS)

ved Mikkeline Munk Nielsen

ILISIMATUSARFIK
University of Greenland

Hvad er univariat analyse?

- Uni er latin for én eller et, dvs. analyser med kun én enkelt variabel
- Univariat analyse falder inden for ***deskriptiv*** (beskrivende) statistik i modsætning til ***inferentiel*** (forklarende) statistik
- I univariat analyse beskriver vi fordelingen af én variabel.
- Vi bruger univariat analyse til at skabe overblik over et karakteristikum i en population, f.eks. at beskrive indkomstfordelingen i en population

ILISIMATUSARFIK
University of Greenland

Hvad er univariat analyse?

Typer af spørgsmål vi kan besvare med univariat analyse:

- Hvordan ser fordelingen af denne variabel ud? → tabeller eller plots
- Hvad er den typiske værdi på denne variabel? → centrummål
- Er der meget stor forskel på, hvilke værdier forskellige observationer har på denne variabel? → spredningsmål

ILISIMATUSARFIK
University of Greenland

Eksempel med datasæt

Firmadatasættet fra sidst:

Datasæt | Kode

	navn	industri	ansatte	omsaetning	tilfredshed
1	Firm 1	Finans	2766	801603.40	Meget tilfreds
2	Firm 2	Finans	962	53875.27	Meget utilfreds
3	Firm 3	Sundhed	4453	493462.09	Meget utilfreds
4	Firm 4	Sundhed	1026	765705.75	Utilfreds
5	Firm 5	Finans	2022	239915.70	Neutral
6	Firm 6	Produktion	2897	338635.58	Tilfreds
7	Firm 7	Detail	2576	254677.01	Meget tilfreds
8	Firm 8	Teknologi	1459	168516.02	Meget tilfreds
9	Firm 9	Sundhed	1799	432109.95	Neutral
10	Firm 10	Finans	4316	431312.60	Meget utilfreds
11	Firm 11	Finans	2766	801603.40	Meget tilfreds
12	Firm 12	Finans	962	53875.27	Meget utilfreds
13	Firm 13	Sundhed	4453	493462.09	Meget utilfreds
14	Firm 14	Sundhed	1026	765705.75	Utilfreds
15	Firm 15	Finans	2022	239915.70	Neutral
16	Firm 16	Produktion	2897	338635.58	Tilfreds
17	Firm 17	Detail	2576	254677.01	Meget tilfreds
18	Firm 18	Teknologi	1459	168516.02	Meget tilfreds
19	Firm 19	Sundhed	1799	432109.95	Neutral
20	Firm 20	Finans	4316	431312.60	Meget utilfreds
21	Firm 21	Finans	2766	801603.40	Meget tilfreds
22	Firm 22	Finans	962	53875.27	Meget utilfreds
23	Firm 23	Sundhed	4453	493462.09	Meget utilfreds

Frekvenser

Variablen Industri måler, hvilken industri hvert firma tilhører:

Industri
Detail
Finans
Produktion
Sundhed
Teknologi

Frekvenser

En typisk måde at opsummere frekvenser er via. frekvenstabeller (evt. procenter)

Envejstabel | Kode

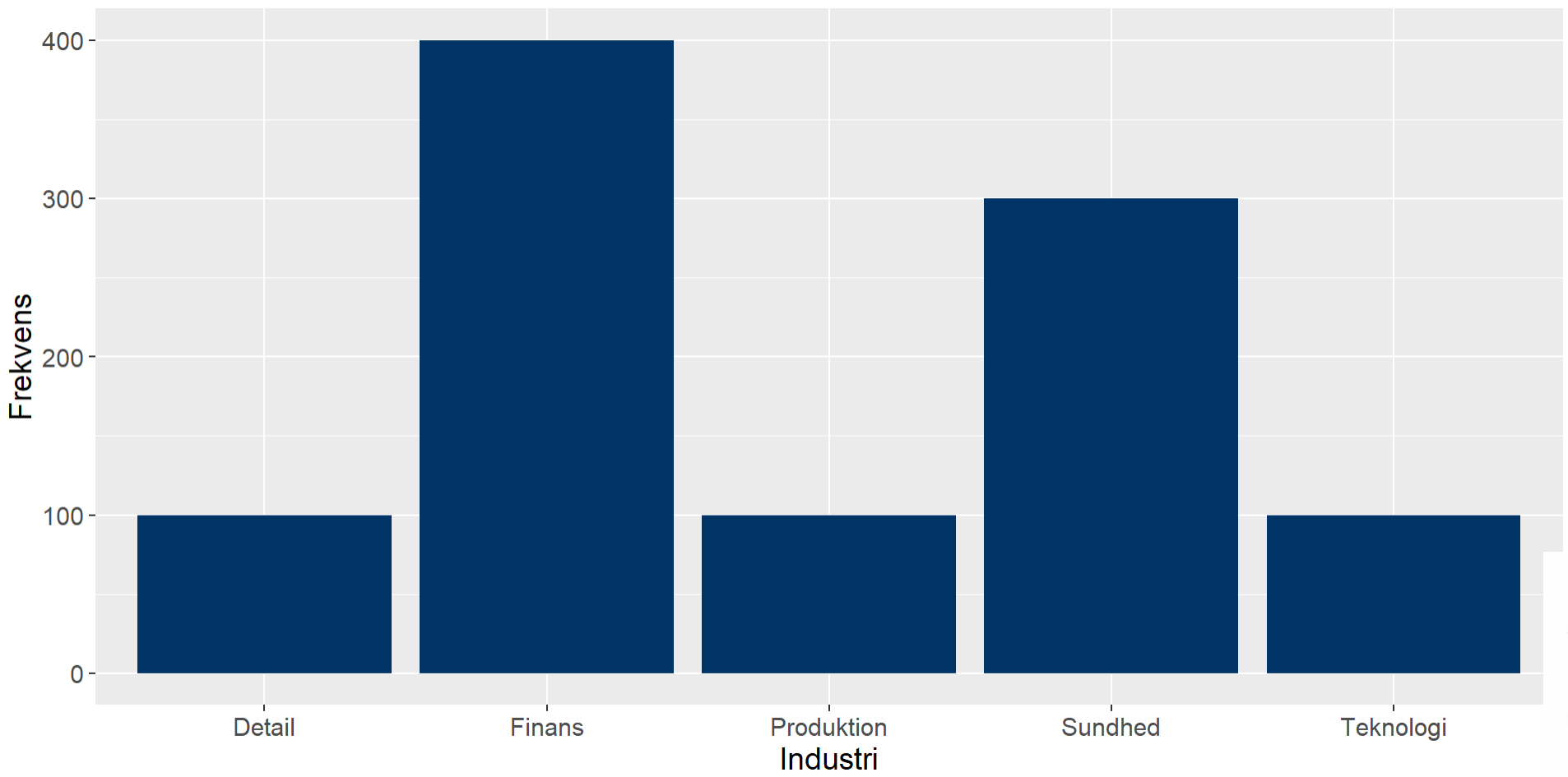
industri	n	percent
Detail	100	0.1
Finans	400	0.4
Produktion	100	0.1
Sundhed	300	0.3
Teknologi	100	0.1

ILISIMATUSARFIK
University of Greenland

Frekvenser

Frekvenser visualeres også ofte via. søjlediagrammer:

Barplot | Kode



Andele

- Andele beskriver relativ frekvens/hyppighed
- Andelsfunktion: \(g(z) = \frac{\text{antal observationer med værdien } z}{\text{Antallet af observationer}}\)

Envejstabel | Kode

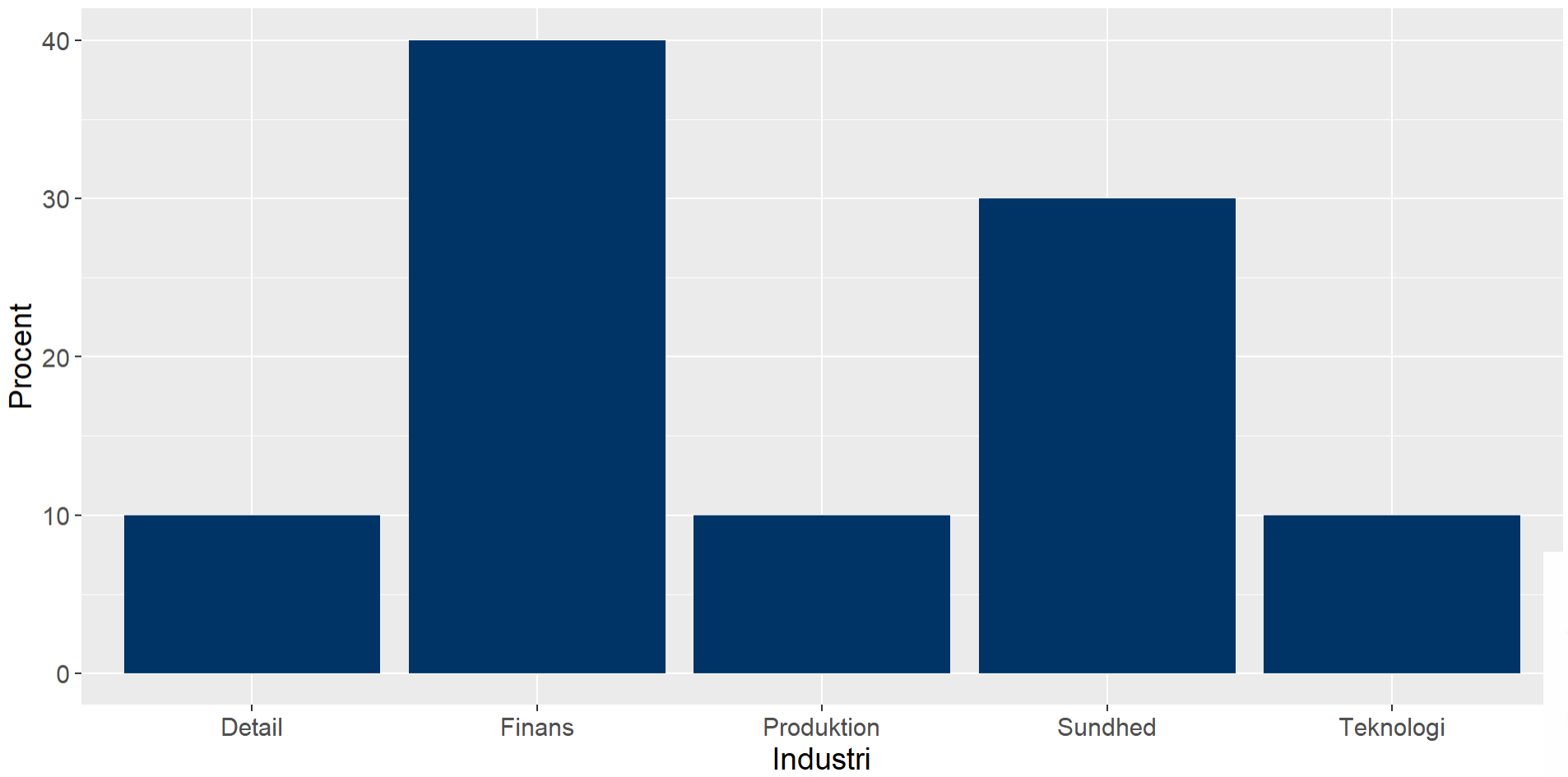
industri	n	percent
Detail	100	0.1
Finans	400	0.4
Produktion	100	0.1
Sundhed	300	0.3
Teknologi	100	0.1

ILISIMATUSARFIK
University of Greenland

Andele

Andele kan også visualiseres med søjlediagrammer (barplots)…

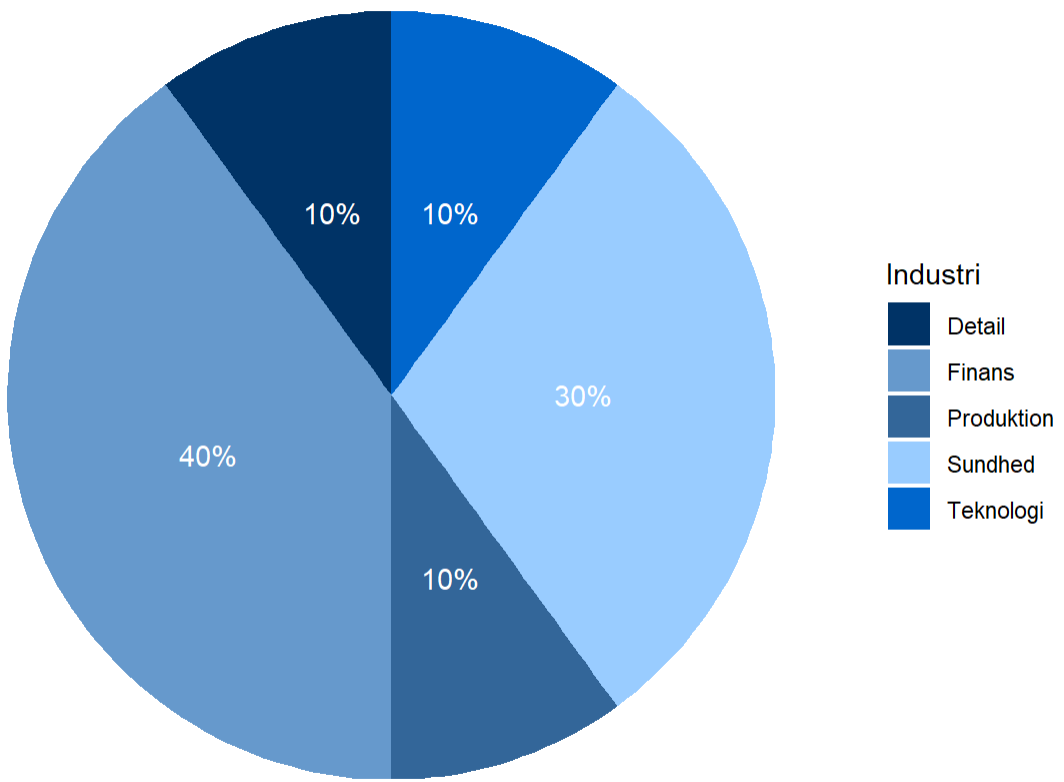
Barplot (procent) | Kode



Andele

… eller f.eks. cirkeldiagrammer

Pie chart (procent) | Kode



ILISIMATUSARFIK
University of Greenland

Andele

- Kummulativ andelsfunktion: $g(z) = \frac{\text{antal observationer} \leq z}{\text{antallet af observationer}}$
- Først rigtig meningsfuld, når værdier kan rangordnes (fra ordinalt måleniveau)

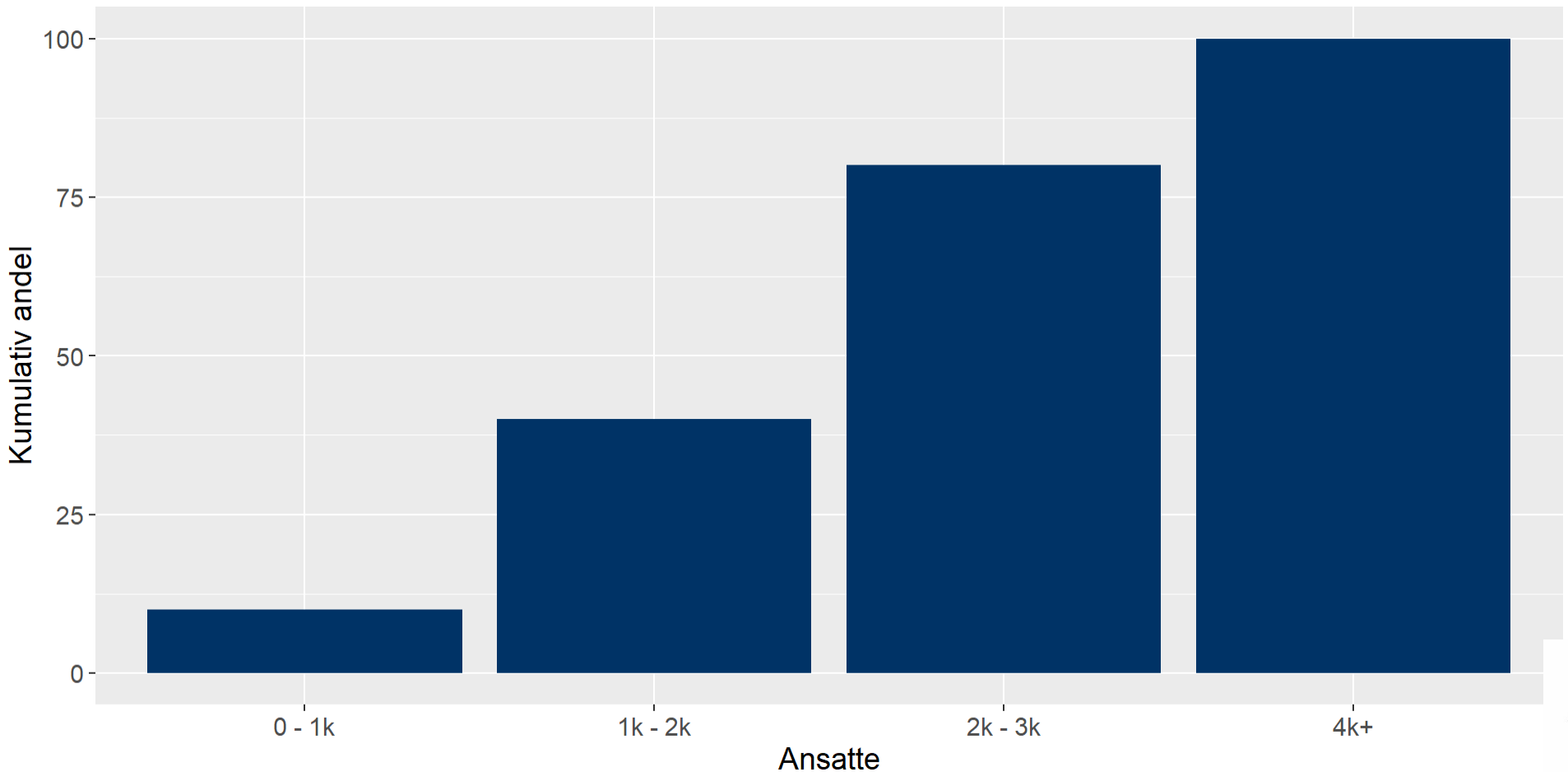
Ny variabel: **Antal ansatte (grupperet)**:

Værdi
0-1000
1001-2000
2001-3000
3001-4000
4000+

ILISIMATUSARFIK
University of Greenland

Andele

Søjlediagram | Kode



ILISIMATUSARFIK
University of Greenland

Mål for den centrale tendens (centrummål)

Mål for en variabels centrale tendens siger noget om en ”typiske” værdi på variablen. Man taler oftest om tre centrummål:

- **Typetal**: den værdi, der optræder flest gange (kaldes også modus eller modalværdi)
- **Middelværdi/gennemsnit**: $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- **Median**: værdien af den midterste observation eller gennemsnittet mellem de to midterste (den værdi, der deler enhederne i to lige store dele).

ILISIMATUSARFIK
University of Greenland

Typetal

- Den værdi, der optræder flest gange (kaldes også modus eller modalværdi
- Vi kan bruge `Mode()` funktionen fra pakken `DescTools` til at finde typetallet

```
1  library(DescTools)
2  (mode <- Mode(df$industri, na.rm = TRUE))
```

```
[1] Finans
attr(,"freq")
[1] 400
Levels: Detail Finans Produktion Sundhed Teknologi
```

- `Mode()` fortæller os, at typetallet/modus er “Finans”, og at det optræder 400 gange i datasættet

ILISIMATUSARFIK
University of Greenland

Middelværdi

- Det mest brugte centrummål er middelværdien, bedre kendt som gennemsnittet
- Gennemsnittet er defineret som $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$
- Det beregnes altså ved at lægge alle værdierne på en variabel sammen og dividere med antallet af værdier/observationer. Det er derfor følsomt over for meget store eller meget små værdier, som kan trække gennemsnittet op eller ned.

```
1  Mean(df$omsaetning, na.rm = T)
```

```
[1] 397981.3
```

ILISIMATUSARFIK
University of Greenland

Median

- Medianen er et mål for den “midterste værdi” på en variabel, når værdierne er rangeret i stigende rækkefølge.
- De enkelte empiriske observationer af en variabel \((Z)\) noteres som \(a\) tildeles et nummer svarende til den enhed (f.eks. individ), der er observeret: \(a_1, a_2, a_3, ... a_n\). Dernæst rangordnes observationerne efter værdi:

18 26 32 34 41 48 55 62 74 78

- Intuitionen bag medianen er, at halvdelen af observationerne vil være større end medianen og halvdelen af observationerne vil være mindre end medianen.

ILISIMATUSARFIK
University of Greenland

Median

- Hvis der er et lige antal observationer er medianen gennemsnittet af de to midterste tal.

  $$ \text{Median} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2} + 1\right)}}{2} $$

- Hvis der er et ulige antal observationer er medianen det midterste tal i rækken

  $$ \text{Median} = x_{\left(\frac{n+1}{2}\right)} $$

ILISIMATUSARFIK
University of Greenland

Fraktiler

- Medianen er et tilfælde af typen af statistiske mål, der hedder ***fraktiler***
- En p-fraktil er en værdi, hvor andelen p af elementerne i en population har en værdi mindre end p fraktilen. Medianen er altså 0,5 fraktilen
- Andre nyttige fraktiler er $0,25$ og $0,75$ fraktiler. Tilsammen kaldes $0,25$, $0,50$, og $0,75$ fraktilerne for ***kvartiler***, da de inddeler enhederne i fire lige store grupper

```
1  Median(df$omsaetning, na.rm=T)
```

```
[1] 384974.1
```

```
1  quantile(df$omsaetning, probs = c(0.25, 0.5, 0.75), na.rm = TRUE)
```

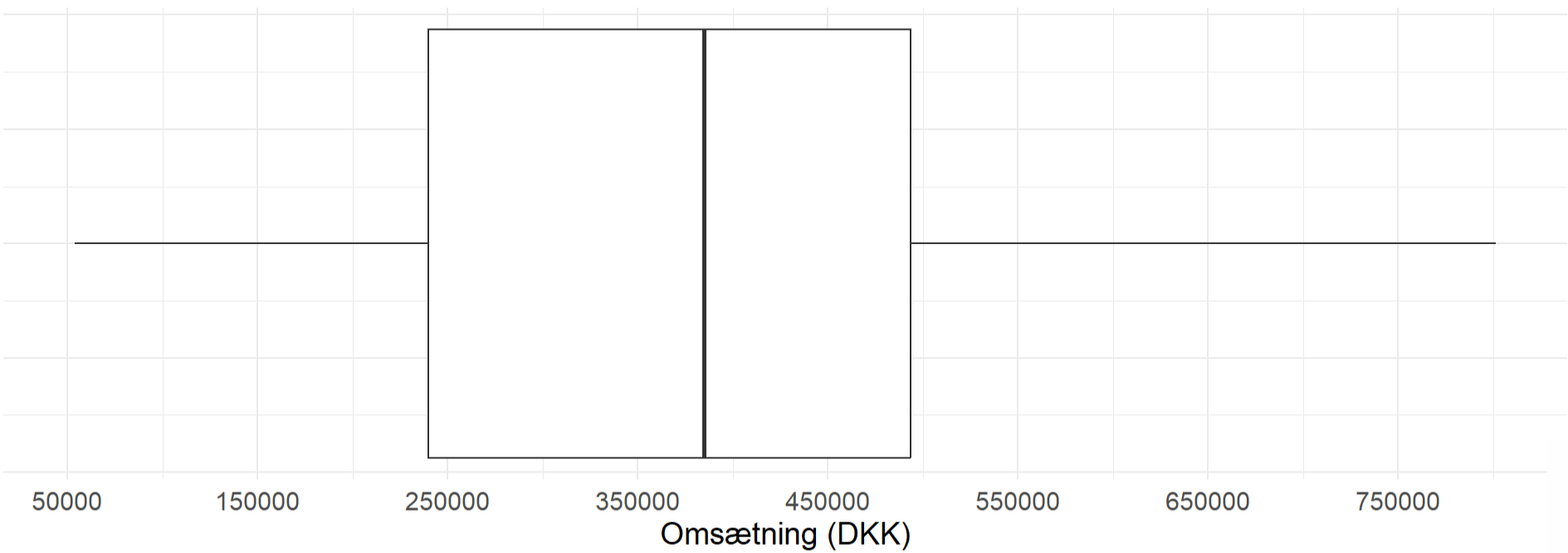
```
     25%      50%      75%
239915.7 384974.1 493462.1
```

ILISIMATUSARFIK
University of Greenland

Boksplot

Kvartiler kan visualiseres ved hjælp af boksplots, der viser: `min`, `p25`, `p50`, `p75`, og `max`

Boksplot | Kode



Centrummål

Øvelse: find eksempler på, hvornår hvert af disse centrummål er interessante at anvende

- Typetal
- Middelværdi
- Median

Centrummål

	Nominel	Ordinal	Interval
Typetal	✓	✓	✓*
Median		✓	✓
Gennemsnit			✓

*Typisk kun meningsfuld for variable med diskrete værdier

Shortcut i R

En god funktion er `summary()` som viser: minimum (`Min.`), første kvartil (`1st Qu.`), median (`Median`), gennemsnit (`Mean`), tredje kvartil (`3rd Qu.`), maximum (`Max.`), og manglende værdier (`NA`):

```
1  summary(df$omsaetning)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  53875  239916  384974  397981  493462  801603 
```

ILISIMATUSARFIK
University of Greenland

Spredningsmål

- Udover den ”typiske værdi” er det informativt at undersøge spredningen på en variabel, dvs. hvor langt de forskellige enheders værdier ligger fra hinanden.
- Spredningsmål siger altså noget om variationen i data, og beskæftiger sig med, hvor meget observationerne afviger fra middelværdien

- Der er to hovedbegreber i statistik for spredningsmål*:
  - Varians: \(Var(x)\) eller \(\sigma^2{(x)}\)
  - Standardafvigelse: \(sd(x)\) eller \(\sigma{(x)}\)

*Grundbogen *Metoder i Statskundskab* nævner flere for andre måleniveauer, kap. 14

Varians

- Variansen siger noget om, hvor stor spredning, der er på en variabel. Med andre ord, ligger observationerne kort eller langt fra middelværdien?
- Variansen er givet ved gennemsnittet af de kvardrerede afstande mellem hver observation og middelværdien:

$$ \text{Var}(x) = \sigma^2(x) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 $$

- N er antallet af observationer

- x_i er den enkelte observation

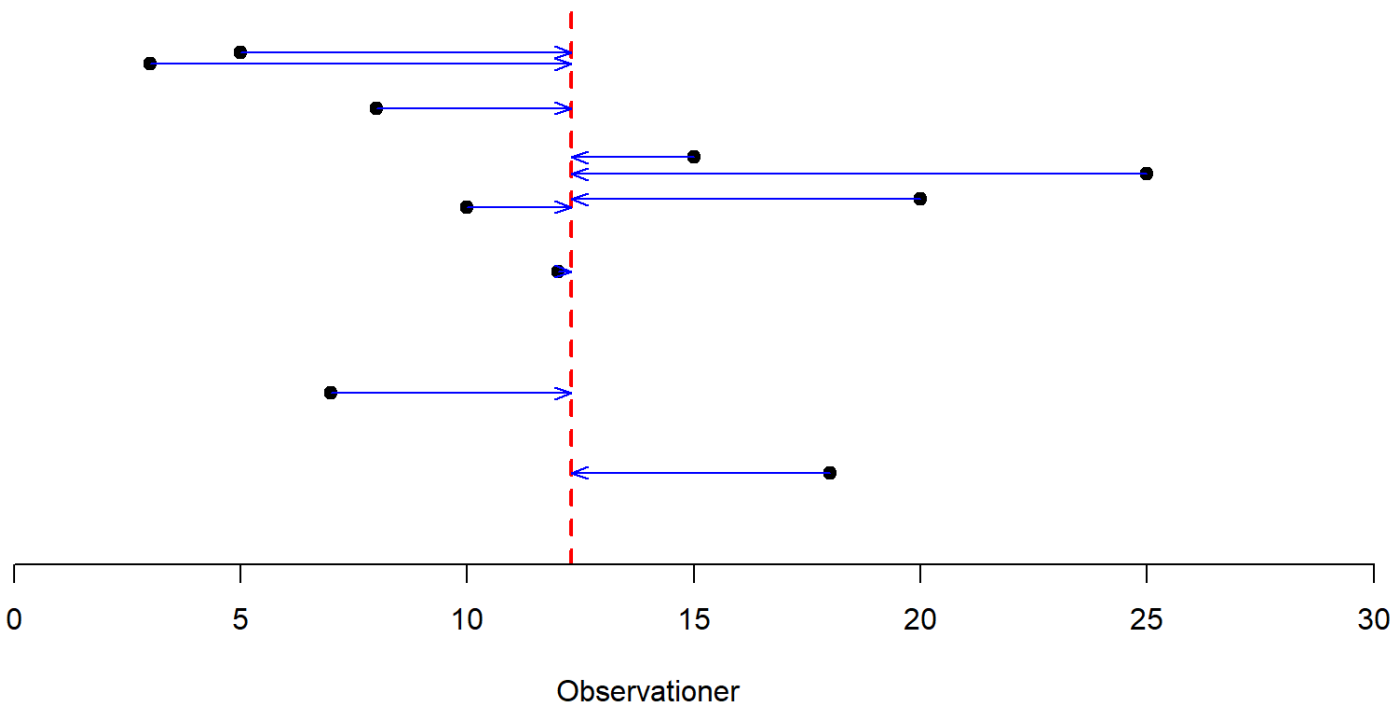
- $\bar{x}$ er gennemsnittet

ILISIMATUSARFIK
University of Greenland

Varians

```
1  data <- c(3, 5, 7, 8, 10, 12, 15, 18, 20, 25)
2  mean(data)
```

```
[1] 12.3
```



ILISIMATUSARFIK
University of Greenland

Standardafvigelse

- Tager du kvadratroden af variansen, får du standardafvigelsen (sd), som er angivet i samme enhed som variablen (f.eks. år eller kroner)

$$\text{Var}(x) = \sigma^2(x) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

$$\text{sd}(x)= \sigma(x)=\sqrt{\sigma(x)^2}=\sqrt{\sigma^2(x) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2}$$

ILISIMATUSARFIK
University of Greenland

Standardafvigelse

- Standardafvigelsen udtrykker kvadratroden af den gennemsnitlige kvadrerede afvigelse fra gennemsnittet på en variabel
- Den er med andre ord tæt på at være et mål for den gennemsnitlige afvigelse fra gennemsnittet, på variablens oprindelige skala - derfor bruges den ofte deskriptivt
- Find f.eks. variansen og standardafvigelsen på variablen ***omsaetning*** i firmadatasættet:

```
1  var(df$omsaetning, na.rm=T)
```

```
[1] 53016059543
```

```
1  sd(df$omsaetning, na.rm=T)
```

```
[1] 230252.2
```

ILISIMATUSARFIK
University of Greenland

Opsamling

	Nominal	Ordinal	Interval
Typetal	✓	✓	✓
Median		✓	✓
Middelværdi		(✓)	✓
Standardafvigelse			✓

Opsamling

Centrale funktioner:

- `tabyl()`
- `ggplot()`
- `Mode()`
- `Mean()`
- `Median()`
- `Quantile()`
- `var()`
- `sd()`