Statistik
Lektion 1

Introduktion
Grundlæggende statistiske begreber
Deskriptiv statistik

Introduktion

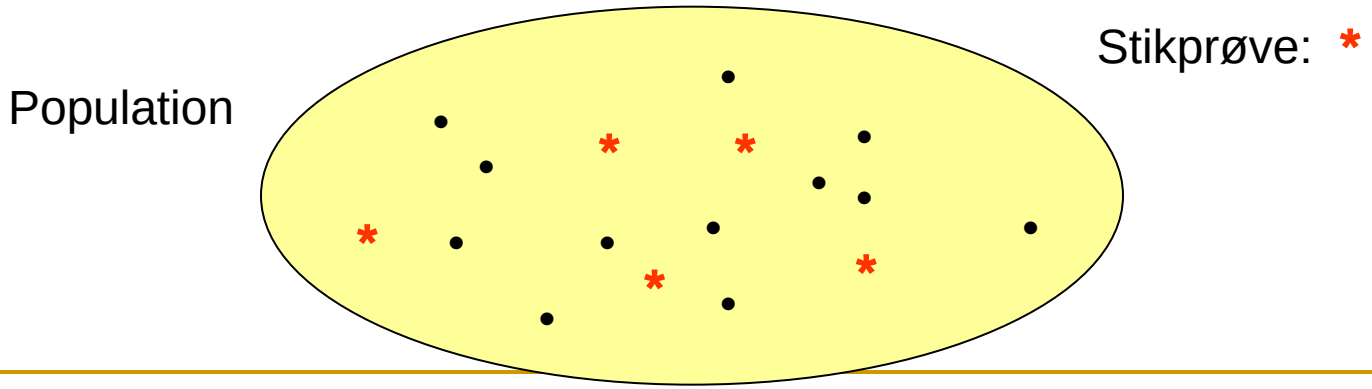
- Kursusholder: Kasper K. Berthelsen
- **Opbygning**: Kurset består af 5 blokke
  - En blok består af:
    - To ”normale” kursusgange, dvs. 2x45 minutter forelæsning efterfulgt af opgaver
    - Derefter en kursusgang uden forelæsning, hvor i regner på en eksamensopgave
- **Eksamen**: Individuel mundtlig efter 7-trins skala
- Eksamen tager udgangspunkt i de 5 opgaver.
- **Software:** SPSS

Statistik

- Disciplinen statistik består af tre dele
  - **Design** (i dag)
    - Planlægning af hvordan data skal indsamles
  - **Deskriptive** (i dag)
    - Opsummering af de indsamlede data
  - **Inferens** (resten af kurset)
    - Drage generelle konklusioner på baggrund af data

Population & Stikprøve

- **Population**
  - En population er mængden af alle individer/enheder, som er af interesse.
  - Fx. Alle danskere, nordjyske produktionsvirksomheder, alle målinger af lysets hastighed.
- **Stikprøve**
  - En stikprøve er den delmængde af populationen.



Deskriptiv og Inferentiel Statistik

- **Deskriptiv statistik**
  - Deskriptiv statistik er en opsummering af data, fx. vha. tabeller og grafer.

- **Inferentiel statistik**
  - Statistisk inferens handler om at drage konklusioner om hele populationen på baggrund af en stikprøve.

Parameter og Statistik

- **Parameter**
  - En parameter er en numerisk opsummering af en population
  - Fx. andelen af folk, der vil stemme på retsforbundet.
- **Statistik**
  - En statistik er en numerisk opsummering af en stikprøve.
  - Fx andelen af folk, der angiver at de vil stemme på retsforbundet i forbindelse med en rundringning.
- **Central problemstilling:**
  - VI vil gerne kende en parameter, men har kun en statistik.
  - Hvor pålideligt kan vi udtale os om parameteren på baggrund af statikken?

Data

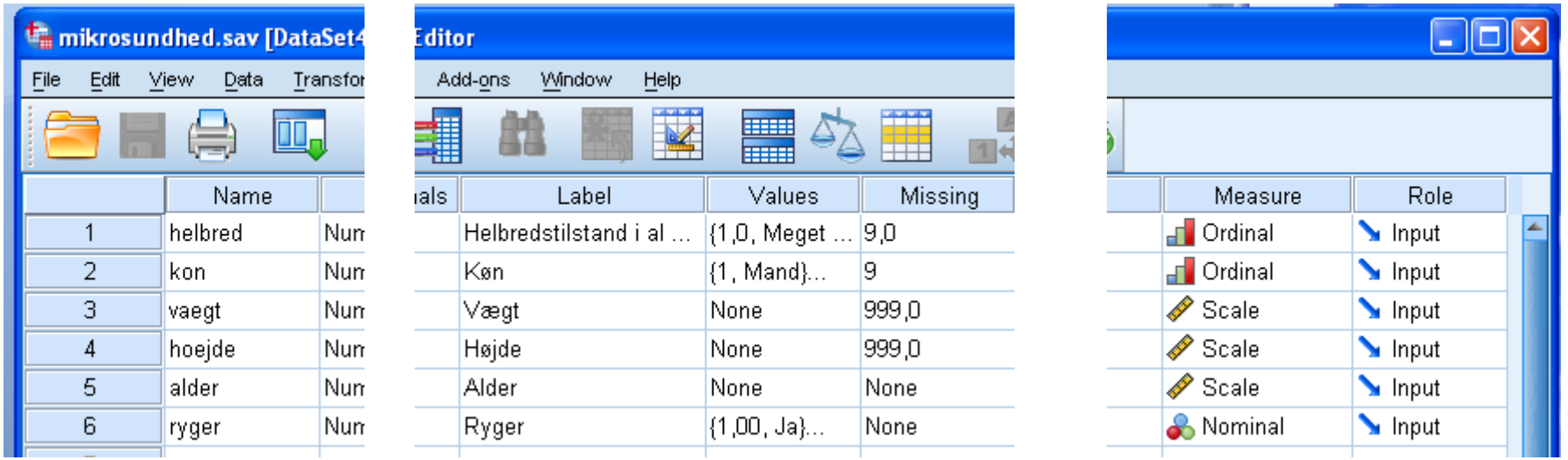
- Data består af en række **variable**.

- **Variabel**
  - En variabel er en ”egenskab” der kan variere blandt de individer/enheder vi studerer.
  - Fx. højde, antal søskende, omsætning, hastighed, farvoritparti osv.

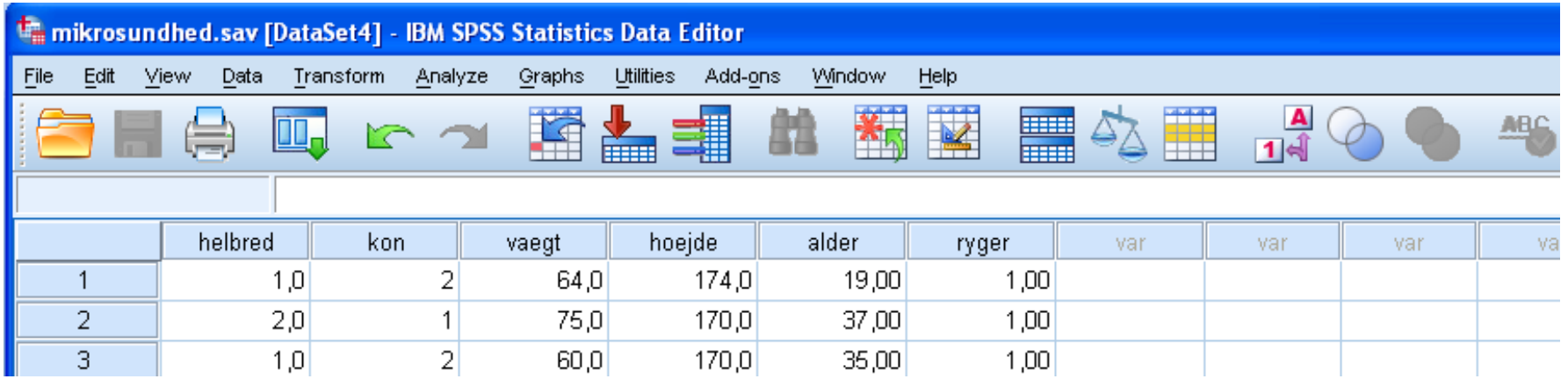
- **Variabeltyper**
  - Vi håndterer variable forskelligt alt efter hvilken type de er.

Data i SPSS

Variable view

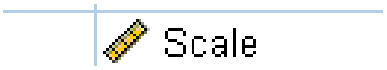


Data view



Kvantitative vs Kvalitative variable

- **Kvantitativ variabel**



  - En kvantitativ variabel er en variabel, der kan måles.
  - Fx. højde, hastighed, omsætning, antal søskende

- **Kvalitativ / kategorisk variabe**
  - En variabel der tilhører en af flere kategorier
  - Fx. Hjemkommune, farvoritfarve, indkomstgruppe

- **Ordinal kategorisk**

  - Kategorierne kan ordnes efter rækkefølge
- **Nominal kategorisk**

  - Kategorierne har *ikke* en naturlig rækkefølge.

Diskret vs Kontinuert Variabel

- **Diskret variabel**
  - En variabel, der kan tage en antal separate værdier.
  - Fx Antal biler = 0,1,2,3,…
- **Kontinuert variable**
  - Variabel, der kan tage alle værdier i et interval.
  - Fx. højden $\in [0,\infty)$
- **Spørgsmål**:
  - Hvad med indtægt?

Tilfældige Stikprøver

- Vi skal bruge en stikprøve, men hvordan skal vi udtage vores stikprøve?

- **Stikprøvestørrelse**
  - Stikprøvestørrelsen er antallet af individer/enheder i stikprøven

- **En simpel tilfældig stikprøve**
  - I en (simpel) tilfældig stikprøve har alle individer lige stor sandsynlighed for at blive udvalgt.

Stikprøve Fejl og Bias

- **Stikprøve fejl**
  - Stikprøvefejlen er den fejl vi begår når vi bruger en statistik baseret på stikprøven til at udtale os om populationen
  - Fx forudsige valgresultat på baggrund af tilfældig stikprøve
- **Stikprøve bias**
  - Stikprøve bias er en systematisk fejl i statistikken pga. den måde stikprøven bliver udtaget.
  - Pga. **ukendt sandsynlighed**: Fx. vores stikprøve stammer fra en webpoll på retsforbundets hjemmeside…
  - Pga. **manglede svar**: Fx. er det kun brokhoveder, der udfylder spørgeskemaet.
  - Pga. **ledende spørgsmål**.

Andre Stikprøvestrategier

- **Systematisk stikprøve**
  - Fx udtage systematisk hver 4. individ.
- **Stratificeret stikprøve**
  - Inddel populationen i delpopulationer, og udtag (lige store) stikprøver fra hver.
  - Fx. sammenligning af hjemløs og ”resten”.
- **Klynge stikprøve**
  - Fx udvælg tilfældige gader i Aalborg og spørg så alle der.

Deskriptiv statistik

- Deskriptiv statistik handler om at præsentere data vha.
  - Diverse **tabeller, grafer og plot**
    - Barplot, histogrammer, boxplot, krydstabeller, scatterplot
  - **Numeriske opsummeringer**, dvs. opsummere data ved få talværdier. De primære
    - **Centralitet** – ”Hvor ligger data?”
      - Typetal, middelværdi, median
    - **Variation** - ”Hvor meget varierer data?”
      - Standardafvigelse, varians, spænd, IQR

Relative Frekvenser

- **Relative frekvenser**
  - Relative frekvenser for en kategori, er andelen af observationerne, der falder i den kategori.
- Kan opsummeres vha. en tabel.
- **SPSS**: Analyze → Descriptive statistics → Frequencies
- **Eksempel**: Typer af husholdninger i USA

type

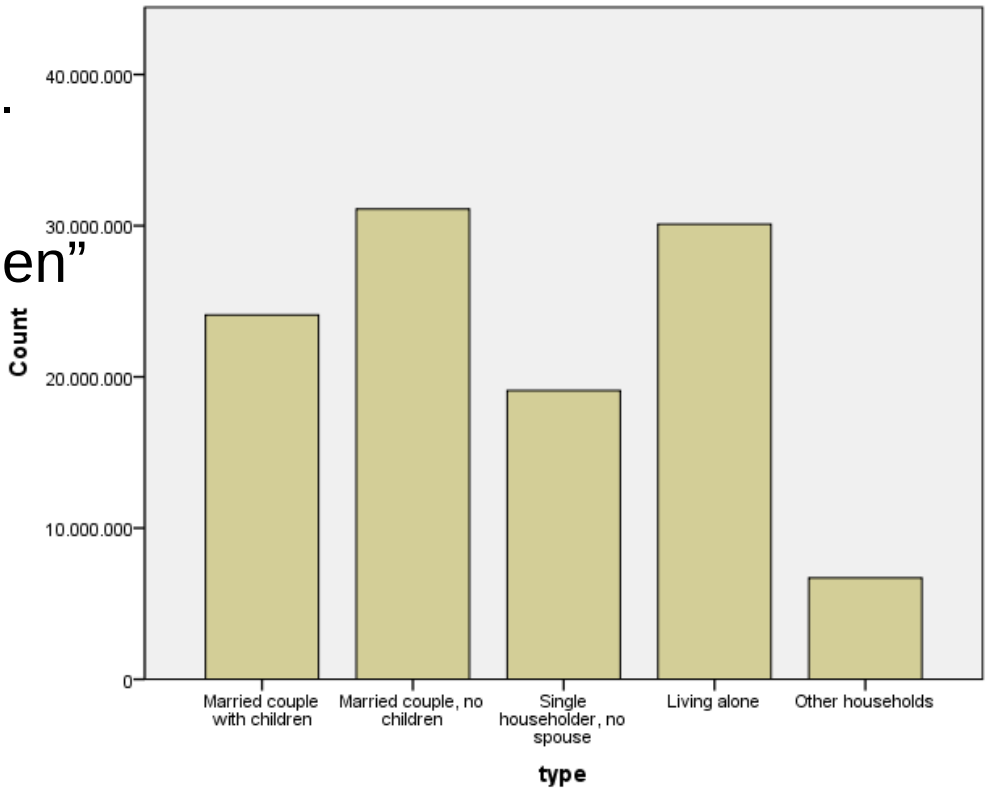
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Married couple with children	24100000	21,7	21,7	21,7
	Married couple, no children	31100000	28,0	28,0	49,7
	Single householder, no spouse	19100000	17,2	17,2	66,9
	Living alone	30100000	27,1	27,1	94,0
	Other households	6700000	6,0	6,0	100,0
	Total	111100000	100,0	100,0	

Bar-plot

- De relative frekvenser, kan også opsummeres grafisk med et bar-plot
- **SPSS**:
- Graphs → Chart builder...

  Vælg: Bar → Simple Bar

  Træk Type ned på ”x-aksen”



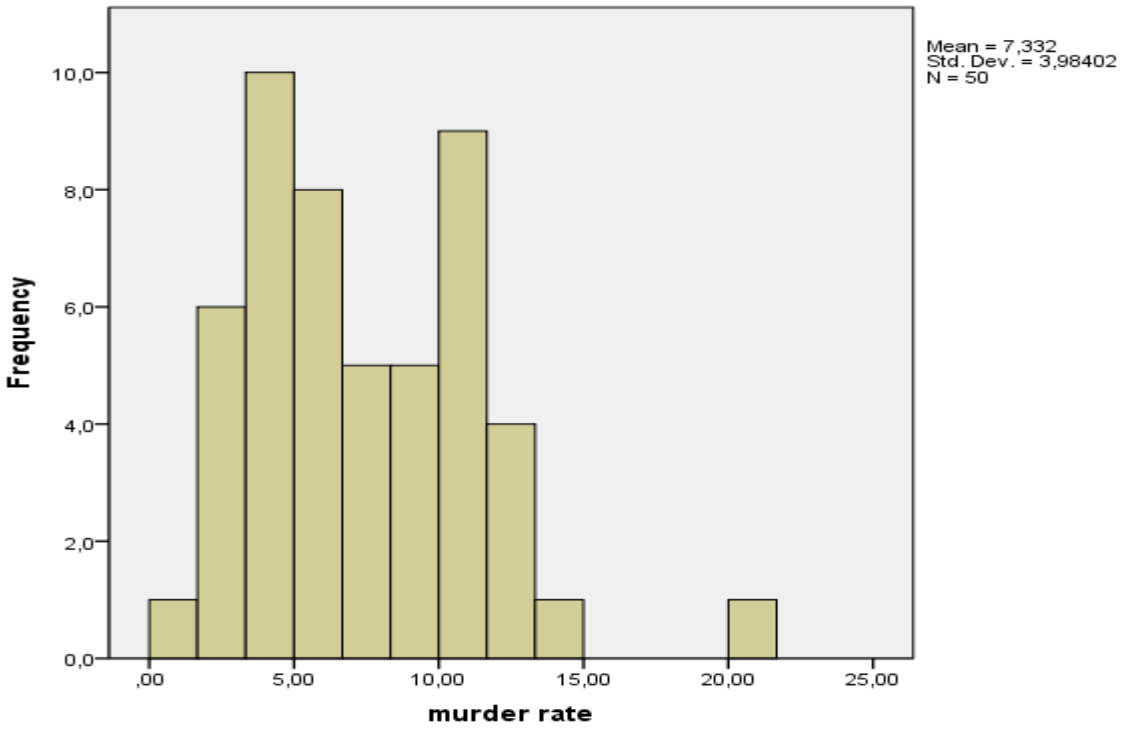
Frekvens-fordeling: Kvantitative data

- For kvantitative data inddeler vi observationerne i intervaller.
- Derefter opsummere vi, hvor mange observationer, der falder i hvert interval.
- **Eksempel**: Mord pr. 100,000 inddelt efter stater i USA
- **SPSS**:

  Graphs → Chart builder…

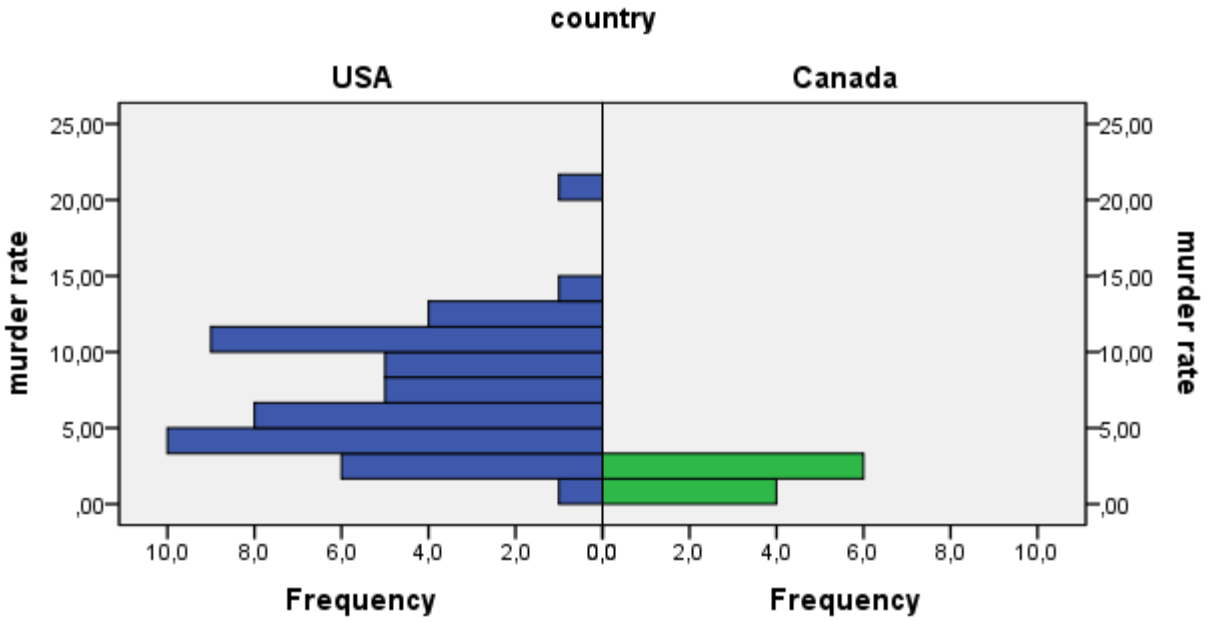
  Vælg: Histogram → Simple histogram

  Flyt murder rate over på x-aksen



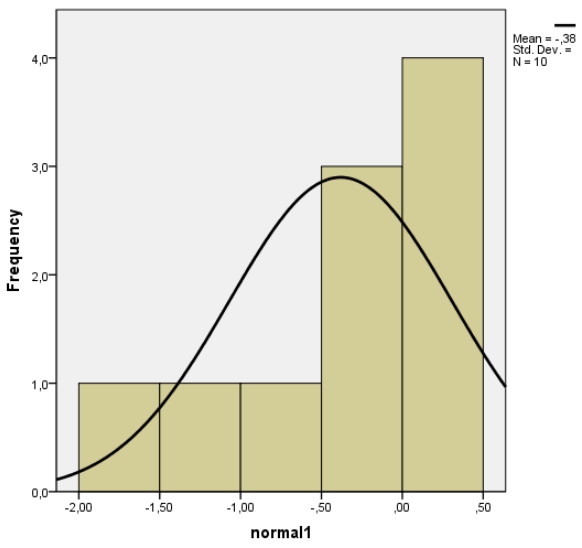
Histogram for to grupper

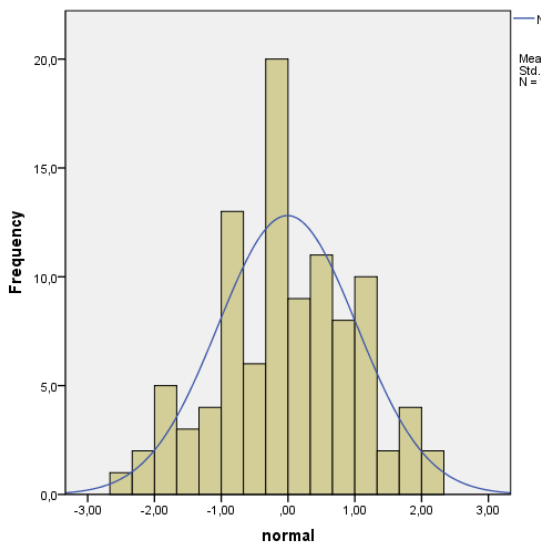
- Histogram af antal mord pr. 100,000 indbyggere fordelt på stater grupperet efter land (USA og Canada)

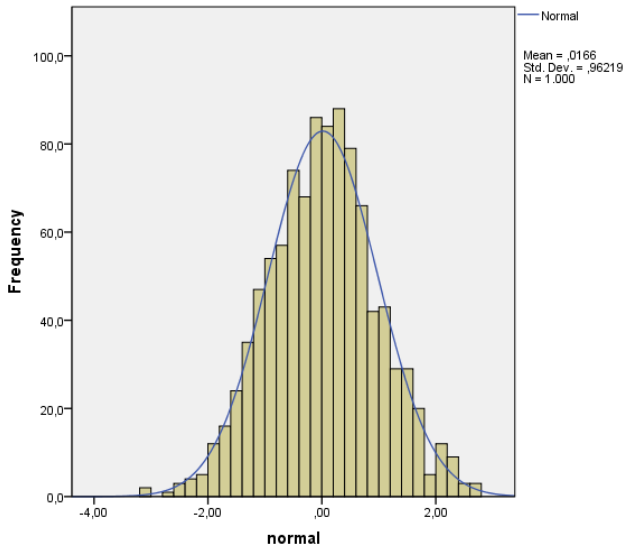


Fordelingen i data og population

- Efterhånden som stikprøven vokser, vil histogrammet ligene den sande populationsfordeling mere og mere

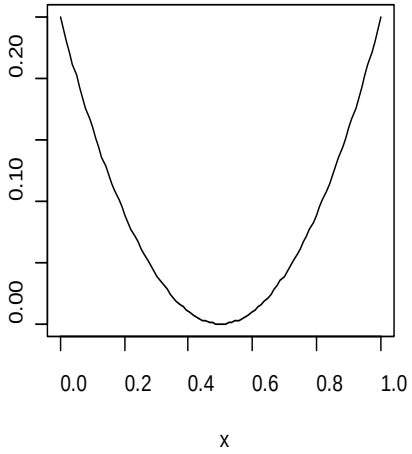






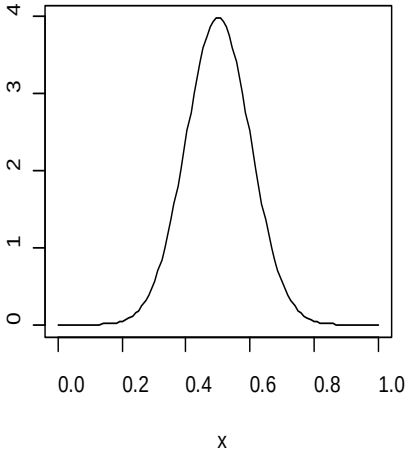
Faconer

U-formet

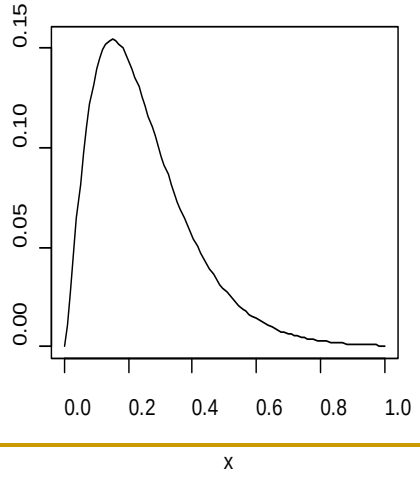


Klokkeformet

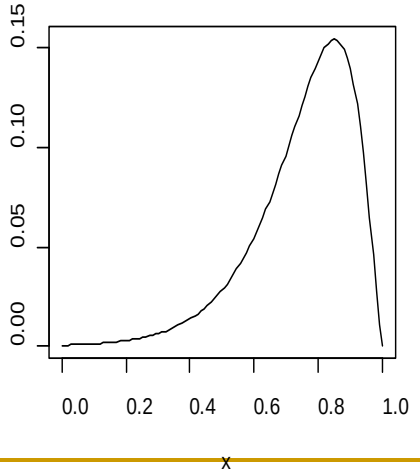
Aka Normalfordelt
Aka Gauss



Højreskæv



Venstreskæv



Hvor ligger data?

- Der flere mål for, hvor data ligger:
  - **Middelværdien** - tyngdepunktet
  - **Medianen** - midten
  - **Typetallet**
    - Det tal, der oftest forekommer i data.

Middelværdi / Gennemsnit

- **Gennemsnit**
  - Gennemsnittet er summen af observationer divideret med antallet af observationer
- Notation:
  - *n* betegner antallet af observationer (stikprøvestørrelsen)
  - $y_1, y_2, y_3, \ldots, y_n$ betegner de *n* observationer
  - $\bar{y}$ betegner gennemsnittet, og er givet som:

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n} = \frac{\sum_i y_i}{n}$$

- Gennemsnittet er følsomt overfor ekstreme observationer.
- Gennemsnittet er ”tyngdepunktet” for data.

Medianen

- **Medianen**
  - Medianen er den midterste observation i en sorteret stikprøve. Hvis der et lige antal observationer, er medianen gennemsnittet af de to midterste observationer.

- Medianen kan bruges for kvantitative data og ordinale kategoriske data.
- I symmetriske fordelinger er gennemsnit og median ens.
- Medianen er ikke følsom overfor ekstreme observationer.

Eksempel

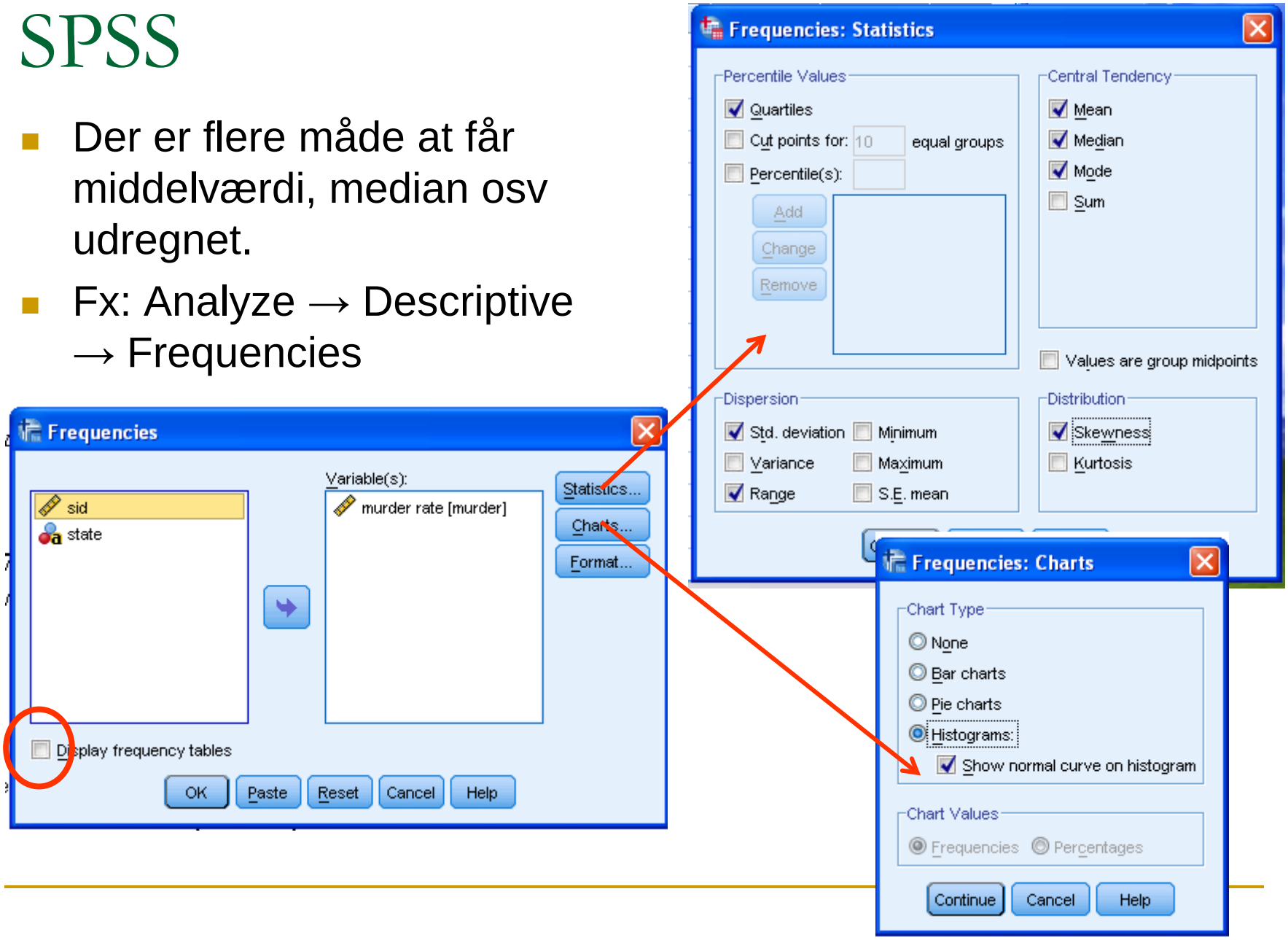
- Antag vi følgende data: 7, 9, 11,12,13,15,17

- Hvad er gennemsnittet?
- Hvad er medianen?

- Hvad sker der med medianen og gennemsnittet, hvis vi erstatter 17 med 27?

SPSS

- Der er flere måde at får middelværdi, median osv udregnet.
- Fx: Analyze → Descriptive → Frequencies



SPSS: Resultat

Statistics

murder rate

N	Valid	50
	Missing	0
Mean		7,3320
Median		6,7000
Mode		3,40[a]
Std. Deviation		3,98402
Skewness		,727
Std. Error of Skewness		,337
Range		18,70
Percentiles	25	3,8750
	50	6,7000
	75	10,3250

a. Multiple modes exist. The smallest value is shown

Standardafvigelsen

- **Afvigelse**
  - Forskellen mellem observation y_i og gennemsnittet $\bar{y}$ betegnes afvigelsen.

- **Standardafvigelsen** (for en stikprøve)
  - Standardafvigelsen *s* for en stikprøve med *n* observationer er:

$$s = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\text{summen af kvaderede afvigelser}}{\text{stikprøvestørrelse} - 1}}$$

  - Variansen s^2 for en stikprøve af størrelse *n* er

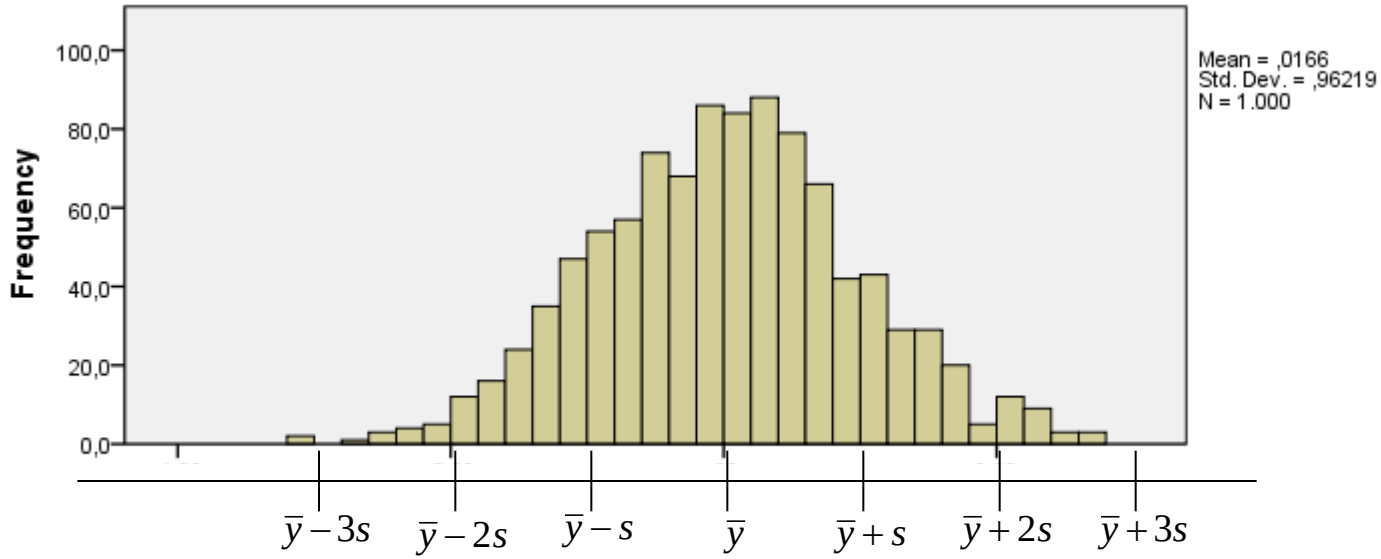
$$s^2 = \frac{\sum (y_i - \bar{y})^2}{n-1} = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2}{n-1}$$

Eksempel

- Antag vi følgende data: 7, 9, 11,12,13,15,17
- Hvad er variansen og standardafvigelsen?
- Hvad sker der med standardafvigelsen og variansen hvis vi lægger 5 til alle observationer?
- Hvad sker der med standardafvigelsen og variansen hvis vi ganger alle observationer med 10?

Fortolkning af s

- **Tommelfingerregler**
- Hvis histogrammet er ca. klokkeformet, så
  - Ca 68% af observationerne ligger mellem $\bar{y} - s$ og $\bar{y} + s$
  - Ca 95% af observationerne ligger mellem $\bar{y} - 2s$ og $\bar{y} + 2s$
  - Alle eller næsten alle observationer ligger mellem $\bar{y} - 3s$ og $\bar{y} + 3s$



Kvartiler og fraktiler

- **Fraktiler**
  - *p*% fraktilen er den observation, hvor *p*% af data falder under.
- Bemærk at medianen svarer til 50% fraktilen er
- **Kvartiler**
  - 25% fraktilen kaldes **den nedre kvartil**
  - 75% fraktilen kaldes **den øvre kvartil**
- Afstanden fra nedre kvartil til øvre kvartil kaldes Inter Quatile Range (IQR)
- IQR er (endnu) et mål for variationen i data.

SPSS: Resultat

Statistics

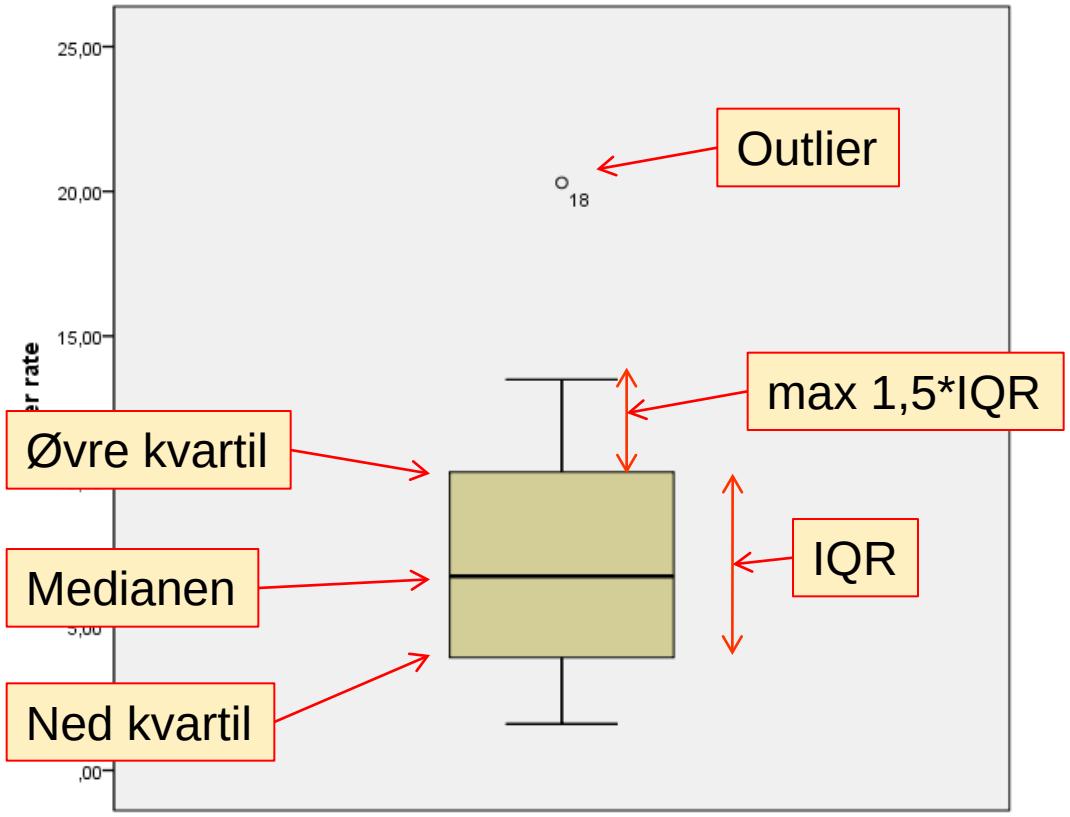
murder rate

N	Valid	50
	Missing	0
Mean		7,3320
Median		6,7000
Mode		3,40[a]
Std. Deviation		3,98402
Skewness		,727
Std. Error of Skewness		,337
Range		18,70
Percentiles	25	3,8750
	50	6,7000
	75	10,3250

a. Multiple modes exist. The smallest value is shown

Boxplot

- Et boxplot er en grafisk præsentation af bla. kvartiler:
- SPSS: Chart Builder… → Boxplot → 1-D boxplot
- Den grå kasse, angiver, hvor de midterste 50% af data ligger.
- Knurhårene strækker til observationer der ligger højst 1.5 gange kassens højde (IGQ) fra kassen.
- En observation mere end 1.5 IQR fra kassen kaldes en **outlier**.



Mord i USA og Canada

- Vi kan sammenlige grupper vha. boxplot
- SPSS: Chart Builder… → Boxplot → Simple boxplot