# Module 1:

# Part 2 - Data and Descriptive Statistics

# (Full Notes)

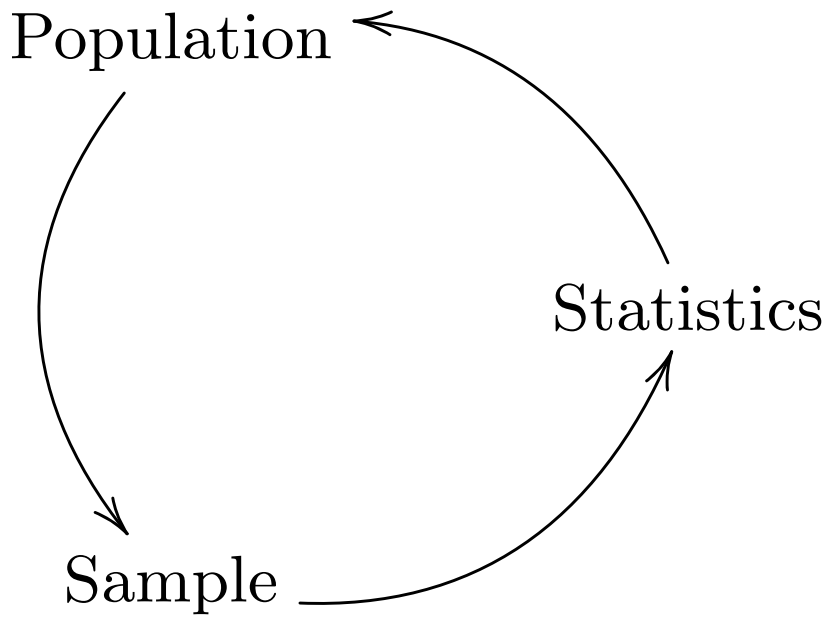
Manuela Zucknick
Oslo Center for Biostatistics and Epidemiology, UiO
manuela.zucknick@medisin.uio.no

MF9130E – Introductory Statistics
April 24, 2023

# Data and Statistics in Medicine

# The role of statistics

- **Inferential statistics** is about using information from a sample (data set) to make inference about the population it originates from.




- **Descriptive statistics** is about describing the data set (i.e. sample) itself.
- Even when a data analysis aims mostly for inferential statistics, descriptive statistics will usually be performed first.

# Data and statistics in medicine

## Populations and samples

"Except when a full census is taken, we collect data on a **sample** from a much larger group called **population**. The sample is of interest not on its own right, but for what it tells the investigator about the population. **Statistics allows us to use the sample to make inferences about the population from which it was derived**. Because of chance, different samples from the population will give different results and this must be taken into account when using a sample to make inferences about the population. This phenomenon, called **sampling variation**, lies in the heart of statistics."

Kirkwood and Sterne p. 9

## Example: Infant nutrition

- **Goal**: Describe the nutrition among infants in Norway (first 6 months)
- **Population**: Norwegian infants in 1998
- **Sample**: 3000 Norwegian infants born in 1998
  (a representative sample!)

**Fylles ut på helsestasjonen ved 6-månederskontrollen**

Dato for 6-mnd-kontrollen: ☐☐ (dag mnd)

Barnets vekt (6 mnd): ☐☐☐☐ g  Barnets lengde (6 mnd): ☐☐ cm

Fødselsvekt: ☐☐☐☐ g  Lengde ved fødsel: ☐☐ cm

14. **Har barnet begynt å få fast føde?**

☐ Ja

☐ Nei ➪ *Gå til spørsmål 21*

---

36. **Hvordan er mors familiesituasjon?**
*Sett kun ett kryss her*

☐ Gift/Samboer

☐ Bor alene med barnet/barna

☐ Annet

31. Hvor har du fått informasjon om amming/spedbarnsernæring, og hvordan vurderer du denne informasjonen?

| | Svært nyttig | Nyttig | Lite nyttig | Unyttig |
|---|---|---|---|---|
| Føde-/barselavdelingen..... | ☐ | ☐ | ☐ | ☐ |

## Variables and observations

- A **variable** is a quantity that is measured
- An **observation** is a specific value of that variable that has been measured
- In the **infant example**:

  Variable: $X =$ birth weight

  Observation: $x = 3710g$

  and

  Variable: $Y =$ solid food at 6 months

  Observation: $y = no$

## What kind of variables do we measure?

- Variables measured **on a scale**, *continuous* or *discrete*
  - Fever 39.6C
  - Blood pressure 95 mm Hg
  - Birth weight 3250 g
  - Self-reported pain (scale)


**Figur 1.1** Visuell analogskala

- ... or **in categories**
  - male/female
  - cancer/not cancer
  - solid food at 6 month/no solid food at 6 months
  - very helpful/helpful/not very helpful/useless (*ordinal*)
  - Married/cohabitant, living alone with kids, other (*nominal*)

**Data are uncertain!**

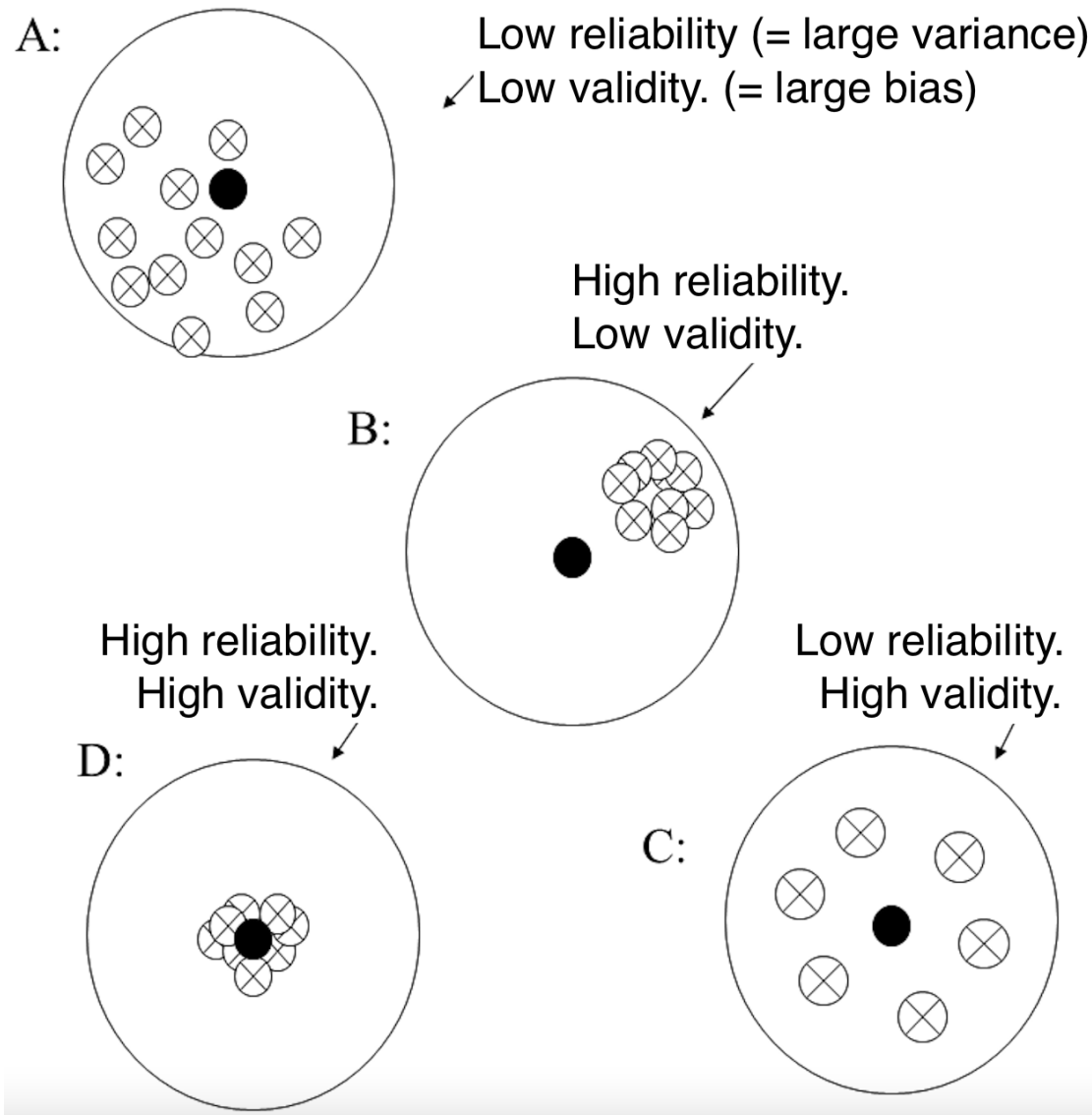
## Can we trust our measurements?

- **Reliability**: How precise are the data? How much can they change if the observation is repeated?
- **Validity**: Do we measure the quantity we want to measure?

- Reliability and validity is connected to the concepts of **random error** (average to zero if a large enough sample is drawn) and **systematic error** (fall in one direction; create bias)

- *The concepts of variance and bias are related:*
- *High reliability $\equiv$ low variance. High validity $\equiv$ low bias.*

# Illustration: Reliability and validity



## Example 1: Reliability of clinical investigation

- Taken from Sackett et al: Clinical Epidemiology (Little, Brown and Company, 1985). Photographs of the retina in 100 patients evaluated by two clinicians with respect to occurrence of retinopathy.

| | | 2. clinician | |
|---|---|---|---|
| | | Little/no | Moder./severe |
| 1. clinician | Little/no | 46 | 10 |
| | Moder./severe | 12 | 32 |

Observed agreement: $\frac{46+32}{100} = 78\%$

## Example 2: Validity of mammography

- From the Norwegian Medical Journal, 1990: 372 women with a lump in the breast has been referred to surgical clinic

| | | Mammography | |
|---|---|---|---|
| | | Benign | Malign |
| Final diagnosis | Benign | 331 | 16 |
| | Malign | 3 | 22 |

There are $\frac{16+3}{372} = 5\%$ wrong diagnosis

## Examples: Sources of variation

- Laboratory variation
- Observer variation
- Instrument variation
- Measurement uncertainty
- Biological variation between individuals
- Day to day variation within an individual

**Statistics is a tool for analysing uncertainty in data**

# Descriptive Statistics

# Descriptive statistics

- A **descriptive statistic** is a summary statistic that quantitatively describes or summarizes features from a collection of information, while
- **Descriptive statistics** (in the mass noun sense) is the process of using and analysing those statistics.

- Descriptive statistics is distinguished from **inferential statistics**:
- It aims to summarize a **sample**, while inferential statistics is about using the data (the sample) to learn about the underlying **population**.

# Descriptive statistics

## How do we present the data?

- Example: Ages of 150 medical students

**Tabell 2.1** Alderen til hver av de nye medisinerstudentene i Oslo i 1984. Alderen er beregnet per 31.12.84 og angitt med to desimaler (eksakt alder). Rekkefølgen nedover er alfabetisk. Antall studenter er 150.

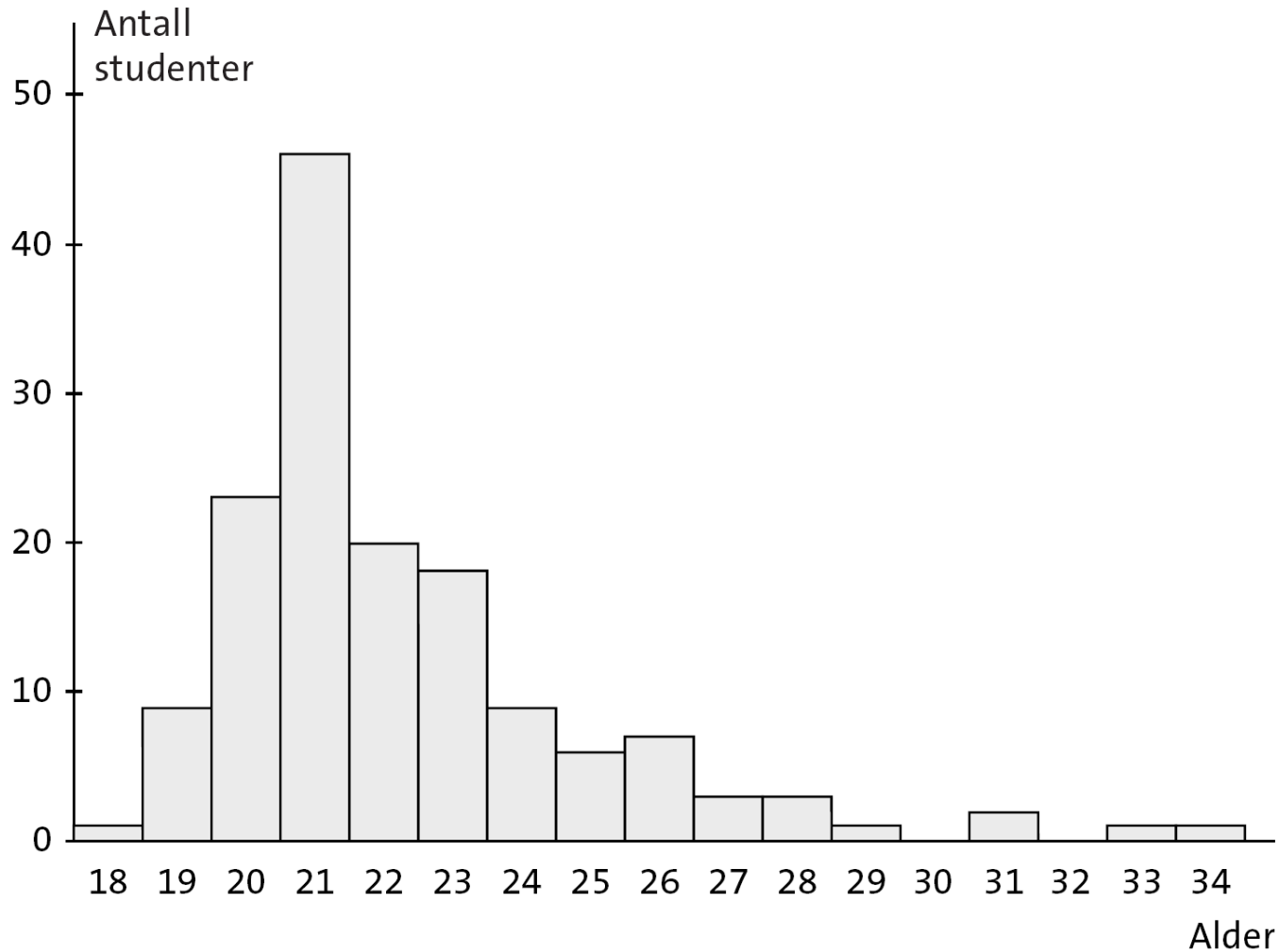
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 20.53 | 24.92 | 22.78 | 21.10 | 20.51 | 24.96 | 22.75 | 23.63 | 21.11 | 31.50 |
| 20.92 | 22.58 | 24.58 | 21.72 | 24.41 | 21.24 | 22.53 | 20.25 | 24.96 | 28.15 |
| 19.62 | 31.71 | 33.29 | 26.09 | 21.05 | 21.72 | 20.37 | 21.80 | 22.60 | 21.90 |
| 22.33 | 27.88 | 21.74 | 19.03 | 21.45 | 23.15 | 21.51 | 20.34 | 19.64 | 23.06 |
| 21.70 | 25.15 | 23.65 | 20.11 | 21.62 | 21.56 | 20.67 | 21.68 | 19.51 | 21.55 |
| 25.53 | 21.61 | 19.68 | 21.90 | 19.21 | 20.98 | 19.82 | 21.77 | 22.11 | 21.93 |
| 21.76 | 23.05 | 21.91 | 23.80 | 21.42 | 20.40 | 20.37 | 26.14 | 20.45 | 23.50 |
| 21.82 | 20.40 | 22.66 | 21.04 | 20.53 | 22.72 | 26.75 | 22.72 | 22.75 | 21.68 |
| 21.13 | 23.89 | 24.75 | 19.80 | 22.07 | 23.42 | 21.78 | 22.85 | 23.30 | 21.38 |
| 20.82 | 21.80 | 23.01 | 26.15 | 22.88 | 22.62 | 27.47 | 23.02 | 20.75 | 25.18 |
| 21.45 | 20.80 | 20.15 | 21.86 | 21.91 | 26.98 | 24.10 | 34.15 | 26.08 | 21.12 |
| 29.63 | 22.84 | 22.57 | 20.72 | 21.50 | 21.23 | 21.53 | 23.45 | 23.06 | 21.33 |
| 21.94 | 21.78 | 24.71 | 28.07 | 21.13 | 26.73 | 20.42 | 19.90 | 21.29 | 23.62 |
| 18.60 | 25.54 | 23.10 | 25.56 | 23.77 | 22.67 | 22.18 | 21.72 | 20.75 | 23.85 |
| 27.90 | 28.38 | 20.81 | 21.21 | 22.89 | 21.88 | 21.75 | 20.70 | 25.39 | 24.12 |

# Frequency table of ages

**Tabell 2.2** Hyppighetstabell over medisinerstudentenes alder. De kumulative relative hyppighetene er beregnet som summen av de relative hyppighetene opp til og inklusive den aktuelle klassen. (At totalsummen blir 100.1 istedet for 100.0, skyldes avrundingsfeil.)

| Klasse-intervaller | Alder i år | Hyppighet | Relativ hyppighet | Kumulativ rel. hypp. |
|---|---|---|---|---|
| 18.00–18.99 | 18 | 1 | 0.7 % | 0.7 % |
| 19.00–19.99 | 19 | 9 | 6.0 % | 6.7 % |
| 20.00–20.99 | 20 | 23 | 15.3 % | 22.0 % |
| 21.00–21.99 | 21 | 46 | 30.7 % | 52.7 % |
| 22.00–22.99 | 22 | 20 | 13.3 % | 66.0 % |
| 23.00–23.99 | 23 | 18 | 12.0 % | 78.0 % |
| 24.00–24.99 | 24 | 9 | 6.0 % | 84.0 % |
| 25.00–25.99 | 25 | 6 | 4.0 % | 88.0 % |
| 26.00–26.99 | 26 | 7 | 4.7 % | 92.7 % |
| 27.00–27.99 | 27 | 3 | 2.0 % | 94.7 % |
| 28.00–28.99 | 28 | 3 | 2.0 % | 96.7 % |
| 29.00–29.99 | 29 | 1 | 0.7 % | 97.4 % |
| 30.00–30.99 | 30 | 0 | 0.0 % | 97.4 % |
| 31.00–31.99 | 31 | 2 | 1.3 % | 98.7 % |
| 32.00–32.99 | 32 | 0 | 0.0 % | 98.7 % |
| 33.00–33.99 | 33 | 1 | 0.7 % | 99.4 % |
| 34.00–34.99 | 34 | 1 | 0.7 % | 100.1 % |
| Totalt | | 150 | 100.1 % | |

# Histogram of age for the 150 medical students



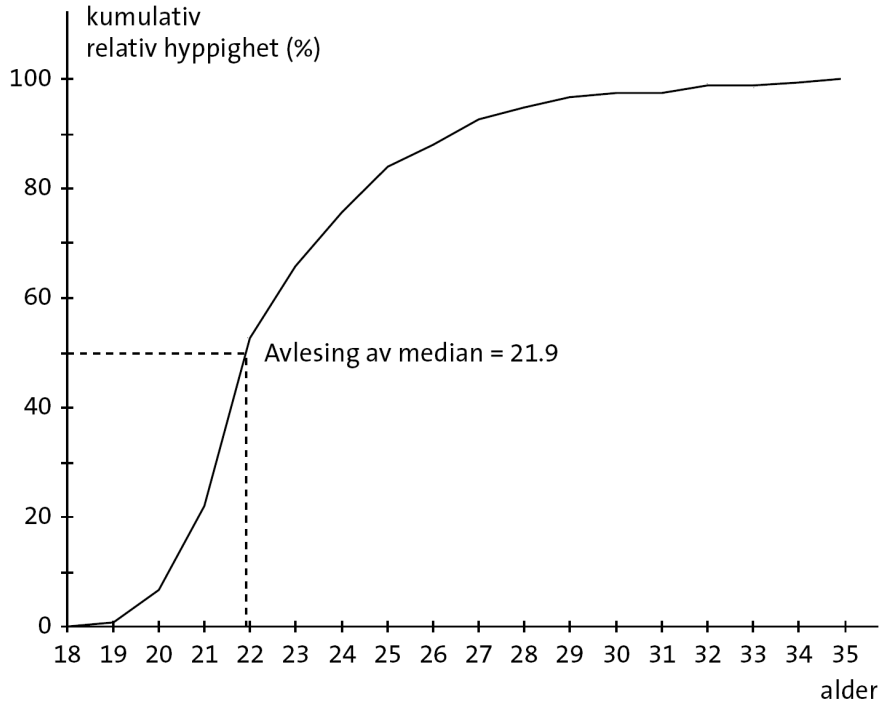

**Figur 2.1** Aldersfordeling for medisinerstudentene 1984

## Central measures

- Consider $n$ measurements $x_1$, $x_2$, ..., $x_n$
- **Mean:** $\bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n) = \frac{1}{n}\sum_i x_i$
- **Median:** First put all measurements in increasing order, then take the one in the middle (or the average of the two in the middle) to be the median
- The mean is vulnerable to **skewness** in the distribution, while the median is robust

# Central measures for the age of the students

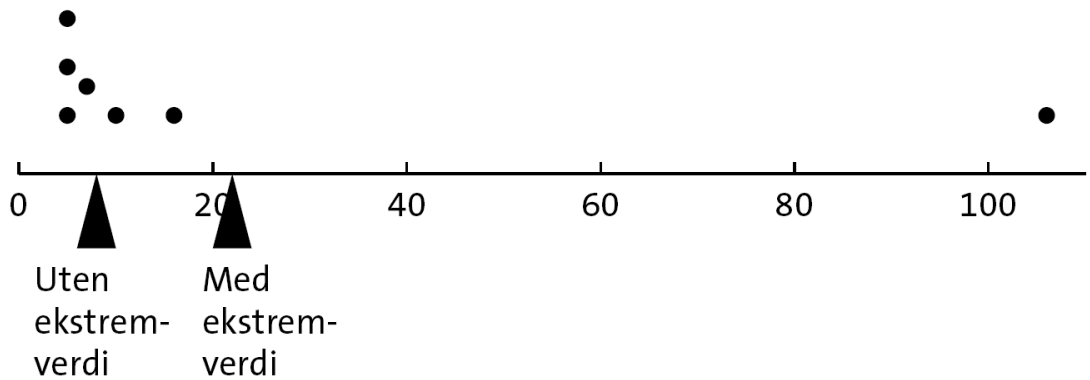
- **Mean**: 22.8 years
- **Median**: 21.9 years




**Figur 2.2** Kumulativ fordelingsfunksjon for alder blant medisinerstudentene. Som eksempel er vist hvorledes kurven kan brukes til å avlese medianalder (se definisjon lenger ute i kapitlet)

## Example: Extreme values

- Data on length of stay in hospital for patients (days): 5, 5, 5, 7, 10, 16, 106
- Mean: 22 days
- Median: 7 days
- Mean without extreme value: 8 days




**Figur 2.4** Data for liggetider i sykehus. Gjennomsnittlig liggetid med og uten ekstremverdi er vist

## Measures of variation

- **Range**
- **Interquartile range** (percentiles, quartiles)
- **Variance**
- **Standard deviation**

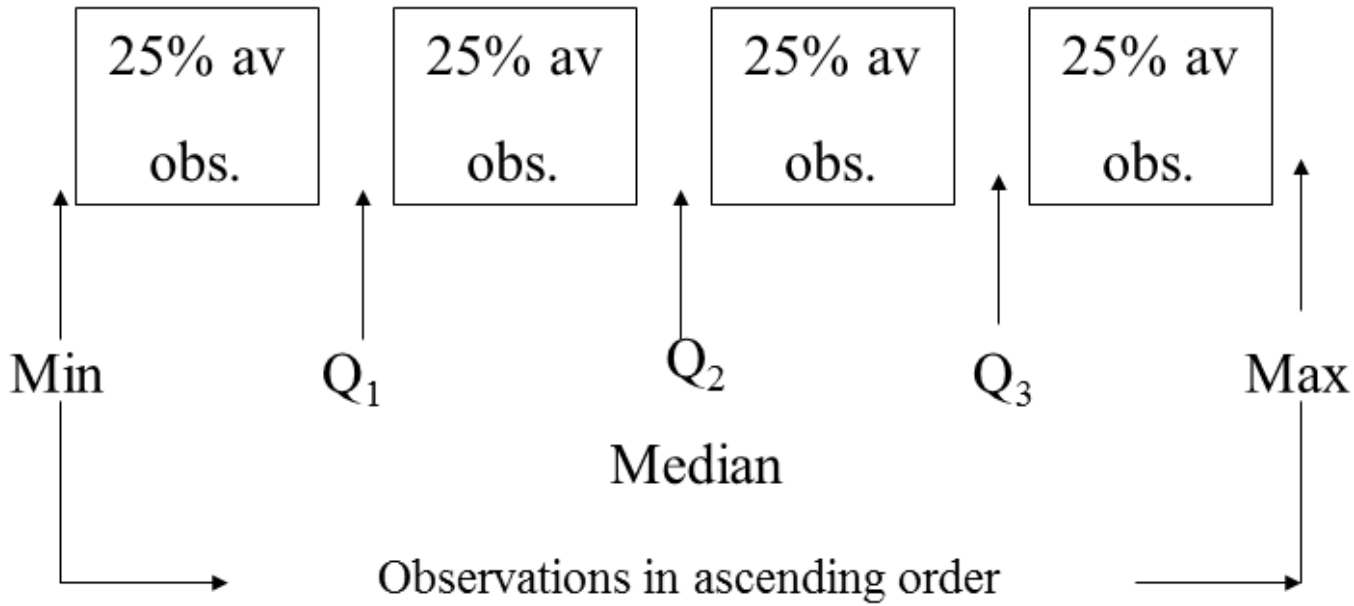
## The range of the observations

- **Range** = maximum value - minimum value
- **Student age example**:
  - Minimum: 18.6 years
  - Maximum: 34.15 years
  - Range: 34.15 years - 18.6 years = 15.55 years

## Percentiles and quartiles

- For example: The 25th **percentile** is a value which has 25% of the data below and 75% above
- Corresponding definition of the ***i*'th percentile**
- **Quartiles**: 25th, 50th and 75th percentile
- **Interquartile range**: 75th percentile - 25 percentile
- Example: **Age-data**
  - 25th percentile: 21.1
  - 75th percentile: 23.8
  - Interquartile range: 2.7

## Quartiles illustrated




- **25th percentile** = 1st quartile ($Q_1$)
  **50th percentile** = 2nd quartile ($Q_2$)
  **75th percentile** = 3rd quartile ($Q_3$)

## Standard deviation

Consider $n$ measurements of the variable $X$: $x_1$, $x_2$, ..., $x_n$

Mean: $\bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n) = \frac{1}{n}\sum x_i$

- **Empiric variance**: $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$
- **Standard deviation**: $s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$

# Example: Calculation of standard deviation

**Tabell 2.4** Eksempel på beregning av standardavvik

| Enkeltdata I | Avstand til gjennomsnittet II | Kvadratavstand III |
|---|---|---|
| $x_i$ | $x_i - \overline{x}$ | $(x_i - \overline{x})^2$ |
| 4 | $-1$ | 1 |
| 2 | $-3$ | 9 |
| 5 | 0 | 0 |
| 9 | 4 | 16 |
| | | 26 |

- $s = \sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2} = \sqrt{\frac{26}{3}} = 2.94$

# Example: Summary measures in scientific papers

*Public Health Nutrition: 7(4), 495–503* DOI: 10.1079/PHN2003550

Breast-feeding at 12 months of age and dietary habits among breast-fed and non-breast-fed infants

Britt Lande[1,2,*], Lene Frost Andersen[2], Marit B Veierød[3], Anne Bærug[4], Lars Johansson[1], Kerstin U Trygg[2] and Gunn-Elin Aa Bjørneboe[1]

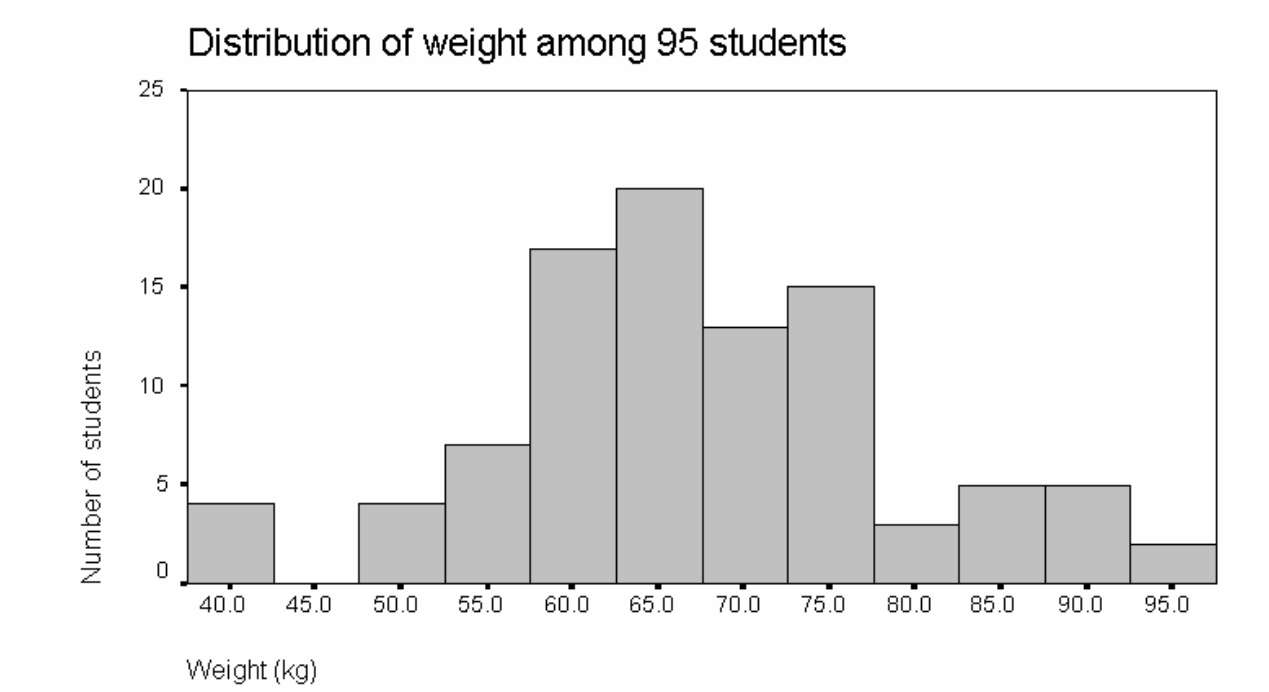
**Table 3** Intakes of foods and drinks

| | Intake* (g day$^{-1}$) | | | |
|---|---|---|---|---|
| | Breast-fed ($n = 701$) | | Non-breast-fed ($n = 1231$) | |
| | Mean | Median ($P_{25}$, $P_{75}$) | Mean | Median ($P_{25}$, $P_{75}$) |
| *Foods* | | | | |
| Porridge | 260 | 200 (100, 400) | 257 | 200 (71, 400) |
| Commercial infant porridge | 224 | 200 (43, 400) | 226 | 200 (43, 400) |
| Home-made porridge | 35 | 0 (0, 14) | 31 | 0 (0, 0) |
| Bread | 55 | 51 (27, 77) | 64 | 57 (30, 90) |
| Vegetables and potatoes‡ | 93 | 75 (38, 125) | 97 | 77 (39, 126) |
| Meat/meat products§ | 28 | 24 (13, 37) | 33 | 28 (17, 43) |
| Commercial infant dinner with meat§¶ | 90 | 56 (0, 112) | 101 | 84 (0, 139) |
| Fish/fish products | 8 | 5 (0, 12) | 11 | 8 (2, 16) |
| Commercial infant dinner with fish¶ | 9 | 0 (0, 0) | 14 | 0 (0, 0) |
| Fruit and berries‖ | 94 | 75 (41, 124) | 90 | 75 (38, 117) |
| Yoghurt | 52 | 31 (0, 86) | 66 | 36 (13, 94) |
| Cheese | 10 | 7 (2, 14) | 11 | 7 (2, 14) |
| Margarine and butter (as spreads) | 11 | 9 (5, 15) | 13 | 10 (5, 17) |
| *Drinks* | | | | |
| Infant formula** | 30 | 0 (0, 0) | 153 | 0 (0, 240) |
| Cow's milk** | 99 | 51 (0, 120) | 238 | 180 (17, 360) |
| Whole milk (3.8% fat) | 62 | 0 (0, 60) | 163 | 34 (0, 300) |
| Semi-skimmed milk (1.5% fat) | 35 | 0 (0, 17) | 72 | 0 (0, 34) |
| Skimmed milk (0.1% fat) | 2 | 0 (0, 0) | 2 | 0 (0, 0) |
| Juice | 22 | 0 (0, 17) | 21 | 0 (0, 17) |
| Commercial baby drinks | 13 | 0 (0, 0) | 14 | 0 (0, 0) |
| Sugar-sweetened drinks | 59 | 17 (0, 77) | 93 | 34 (0, 120) |
| Nectar | 13 | 0 (0, 0) | 18 | 0 (0, 17) |
| Squash ('saft') | 44 | 0 (0, 43) | 72 | 17 (0, 120) |
| Carbonated soft drinks | 3 | 0 (0, 0) | 3 | 0 (0, 0) |
| Artificial sweetened squash | 7 | 0 (0, 0) | 20 | 0 (0, 0) |
| Water | 206 | 180 (120, 300) | 191 | 120 (60, 240) |

## Simple graphical presentations: How do plot data?

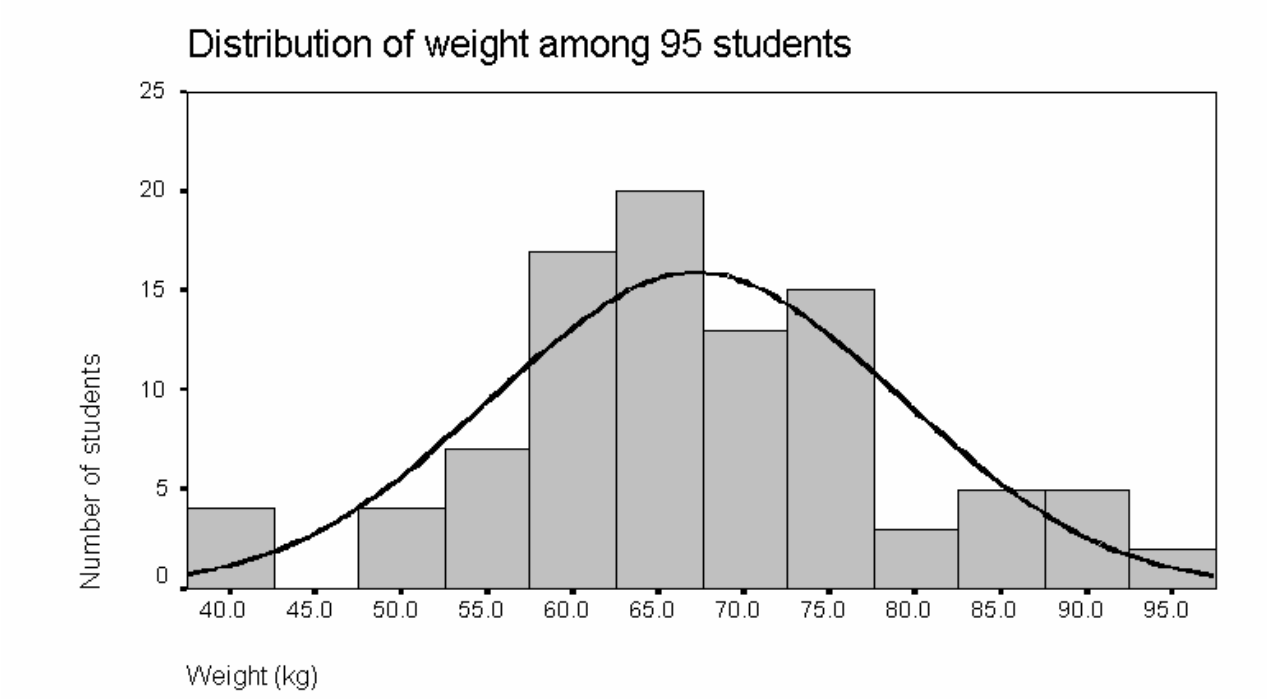
- **Histogram** and **box plot**: Describing the distribution of continuous data (scale variables)
- **Bar graphs:** Describing categorical data
- **Scatter plot**: Association between variables
- **Time series plot**: Showing trend over time

# Histograms

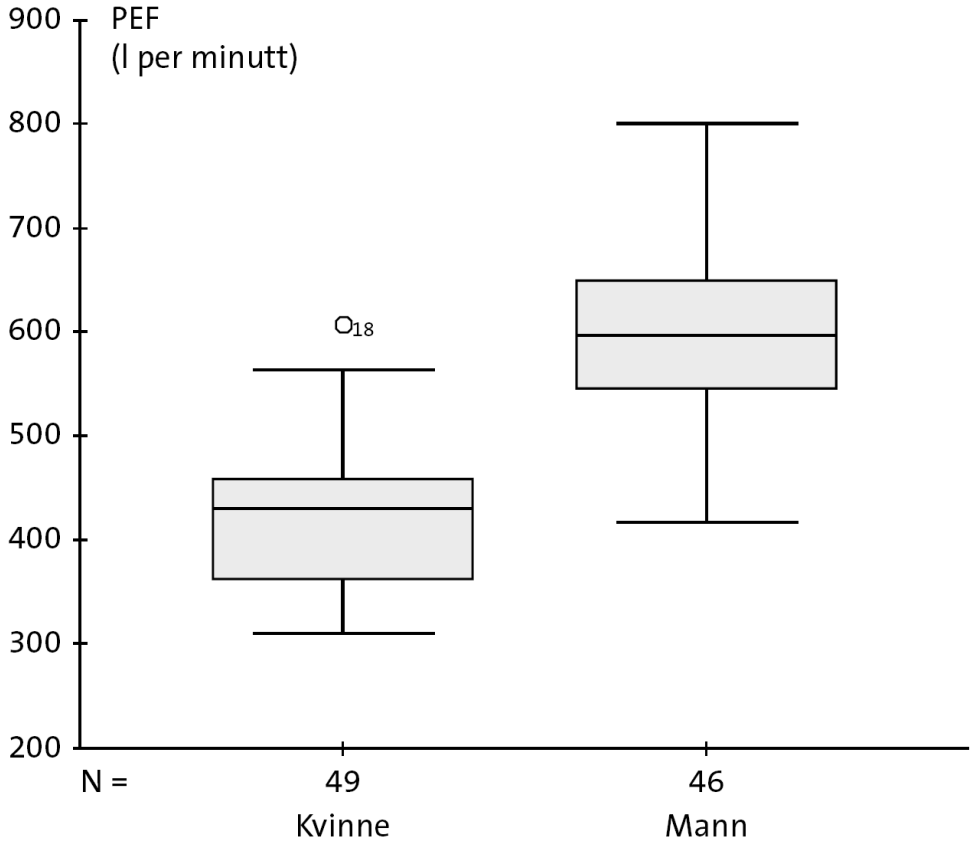



- Describes the distribution of **continuous data**

# Fitting a normal distribution to a histogram



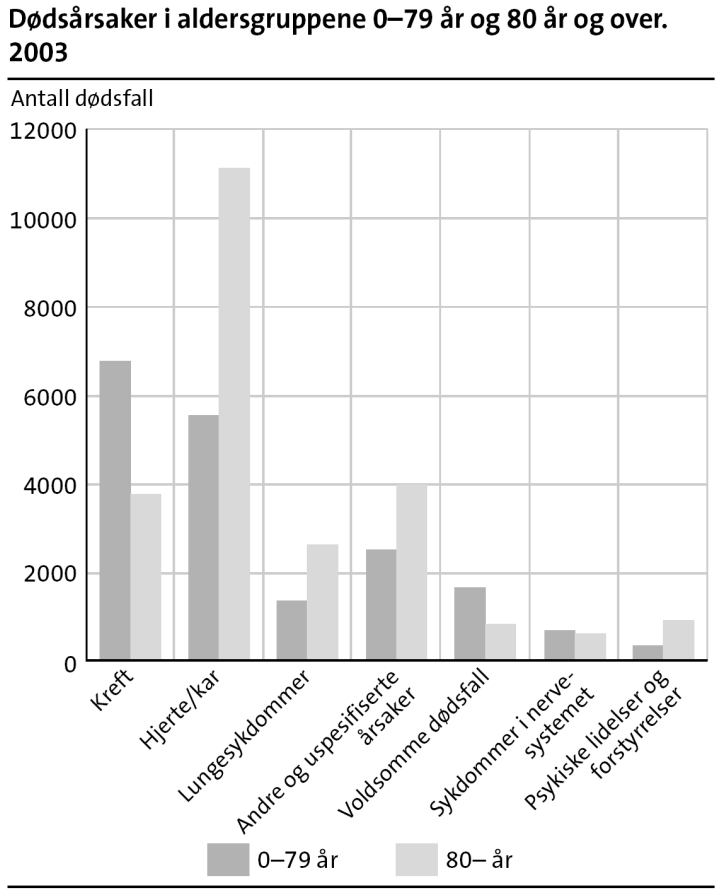
# Comparing groups with box plots




**Figur 2.9** Sammenlikning av lungefunksjon, målt ved PEF, for kvinner og menn. Data for 95 studenter i medisin og odontologi

- Gives the **median, quartiles, max and min value, and possible outliers (extremes)**

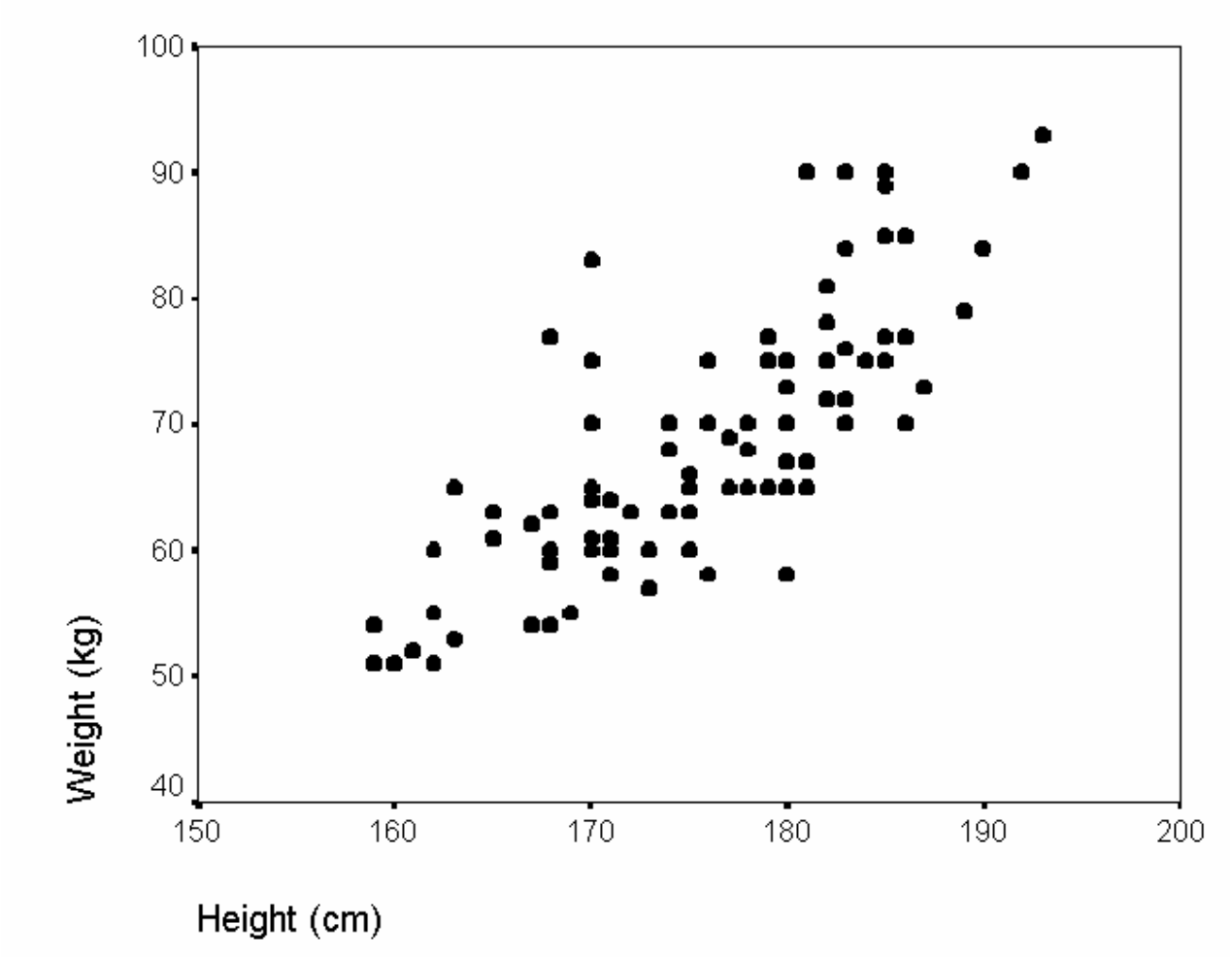
# Bar graphs




**Figur 2.5** Søylediagram som viser hyppighet av forskjellige dødsårsaker. Figur fra Statistisk sentralbyrå

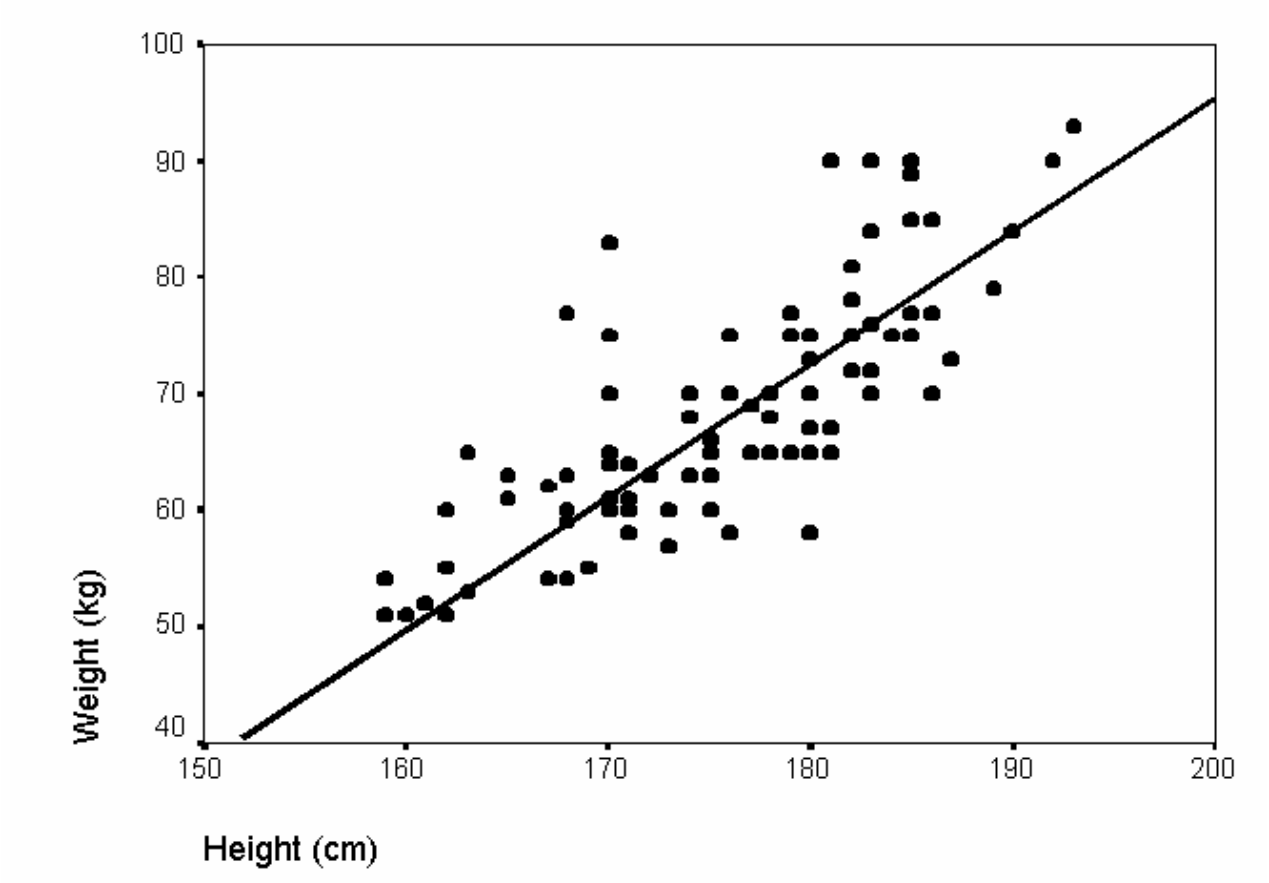
- Used to describe **categorical data**

# Scatter plots



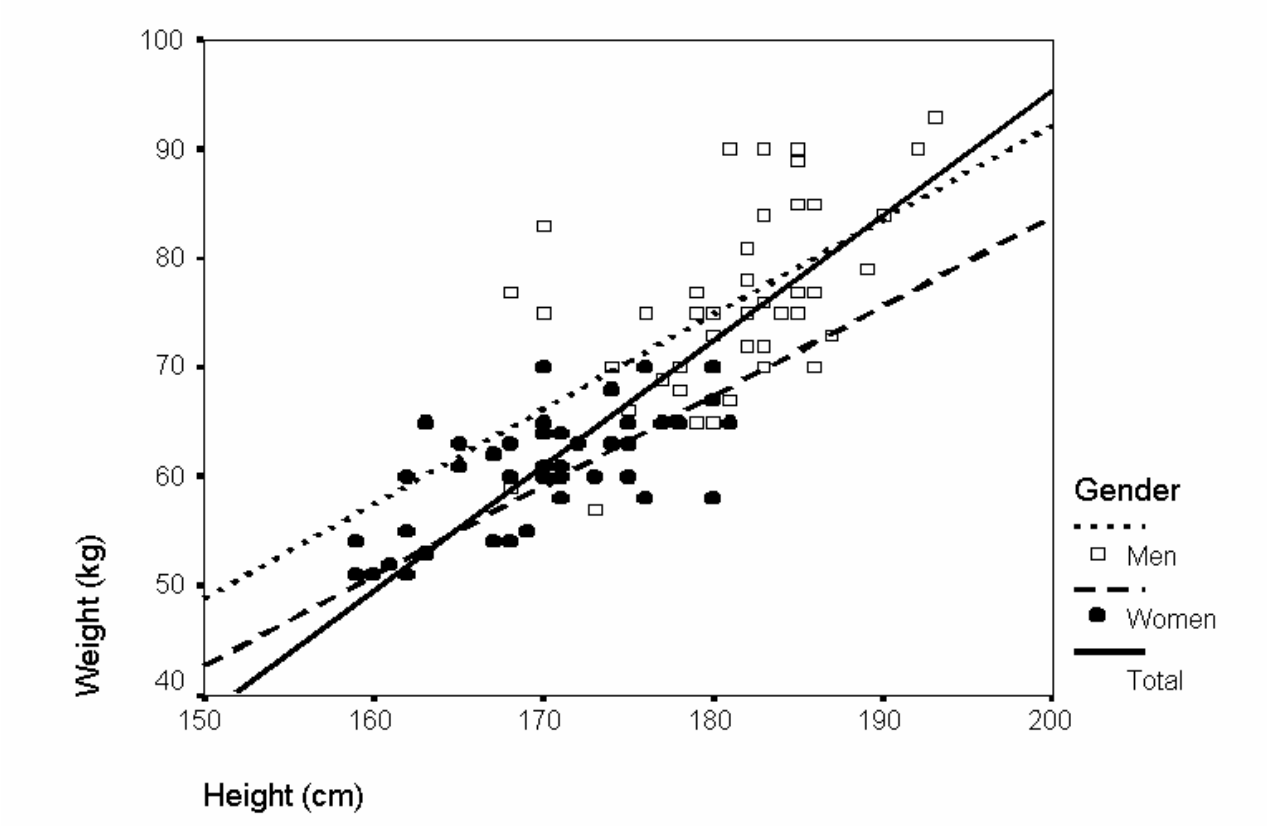

- Describes the **association between variables** (and variance in association)

# Scatter plot with regression line

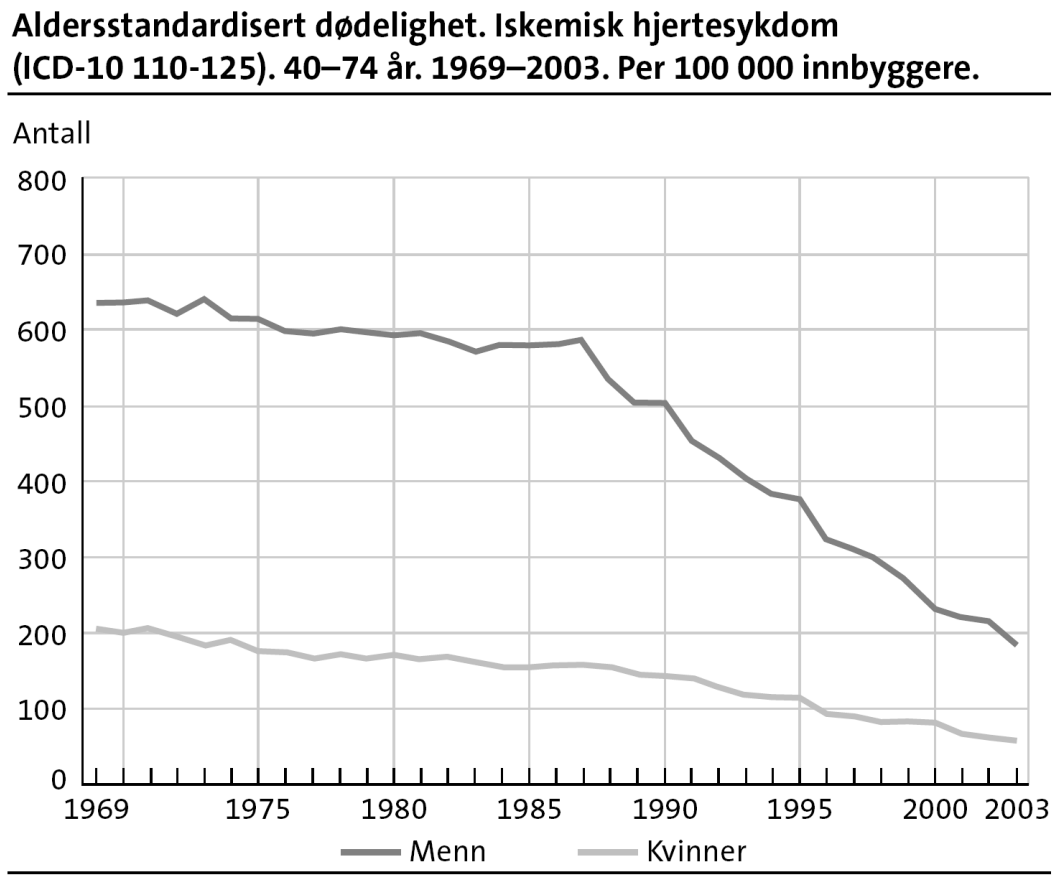



- **Inference** from scatter plots are based on regression

# Scatter plot for men and woman separate



# Time series plot




**Figur 2.7** Tidsserieplott over dødelighet av hjertesykdom i Norge. Figur fra Statistisk sentralbyrå

- When interested in trends. Do we have an increase or decrease?

# Summary

## Key terms and concepts:

- **Descriptive statistics** (as opposed to inferential statistics)
- **Population** and **sample**
- **Variables** and **observations**
- **Scale variables** and **categorical variables**
- **Reliability** and **validity**; Sources of variation
- Frequency table
- Graphics: **histogram**, **boxplot**, **bar graph**, **scatter plot**
- Central measures: **Mean**, **median**, mode
- **Skewness:** right-skewness and left-skewness
- Measures of variation: **Variance**, **standard deviation**, **range**, **interquartile range**
- **Percentiles** and quartiles, minimum (min), maximum (max)

## Mathematical notations and formulas:

- Variable $X$ with corresponding observations $x_i$ (with $i = 1, ..., n$)
- Mean: $\bar{x} = \frac{1}{n} \sum_i x_i$
- Variance: $s^2 = \frac{1}{n-1} \sum_i (x_i - \bar{(}x))^2$
- Standard deviation: $s = \sqrt{s^2}$

# Self-study and Group Work Sessions:

# Tasks and guiding Questions

# Self-study session

- *Review the two summary slides ("Key terms and concepts", "Mathematical notations and formulas").*
- *Use the provided reading material for this course module to read up more details and learn about any terms and concepts from the summary slides that you are not familiar with.*

1. Make sure you understand all terms highlighted bold (slide 39)
2. Prepare for the group work session by keeping in mind the "Guiding questions for the group work session" (next slide) when reviewing the material.

- **Course material:**
  - Full version of the slides: `Module1-Part2-Data_and_descriptive_statistics_full.pdf`
  - Course textbook: Kirkwood and Sterne (2003), chapters 2-4
  - Alternative textbook: Aalen et al. (2006), chapters 1-2

# Group work session

- *In your group (which should include 4-6 participants), jointly choose one of the provided papers and read the paper.*
- *Together, prepare answers for the following questions. PS: If you are finished early, choose another paper and repeat the exercise.*

1. What do you consider the key research question in the article?
2. Which is the main outcome variable, that the authors are most interested in studying? Find out as much as you can about the characteristics and distribution of this variable.
3. Which descriptive statistics are used in this article, that is central measures, measures of variation, and/or graphics? Only looking at statistics used that we covered in this class: are in your opinion the variables in the study described sufficiently well by these statistics? How so (or why not)?