Lecture 4

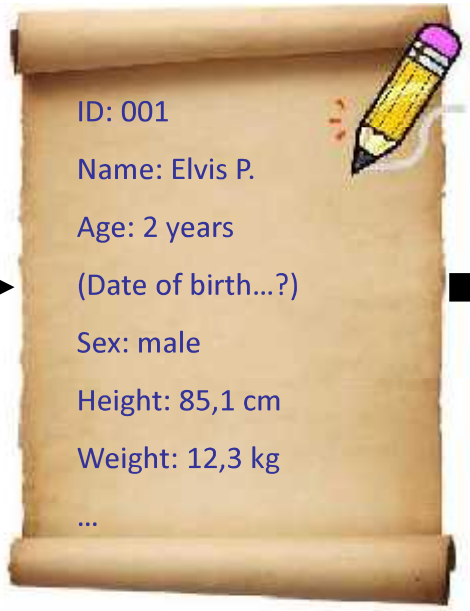
Describing data;

normal distribution;

measures of variance

Let's talk about … data

Where does the data come from?

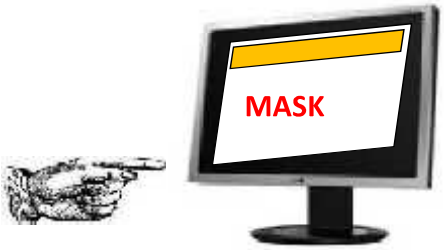


Data processing…

Example questionnaire ID _ _ _

Name(s) _____ family name _____
Date of interview: _ _ / _ _ / _ _ _ _
Date of birth: _ _ / _ _ / _ _ _ _
Sex: □ male □ female
Did you ever smoke? □ yes □ no
if yes: do you smoke currently? □ yes □ no
if yes: cigs./day _ _ _ duration _ _ (years)
if no: quitting smoking? (years) _ _ _ _

Did you ever drink alcohol? □ yes □ no
…



Data are **observations** on individuals
any aspect of the observation is called a **variable**

Data set = table = data sheet =…

ID *001*
Name(s) __*Johanna*__ family name __*Lennon*__
Date of interview: *14 / 07 / 2008*
Date of
Sex: x
Did you

ID *002*
Name(s) __*Paula*__ family name __*McCartney*__
Date of interview: *16 / 07 / 2008*
Date
Sex:
Did y

ID *003*
Name(s) __*Ringo*__ family name __*Stark*__
Date of interview: *16 / 07 / 2008*

ID	Name	Fam.name	date of birth	Sex	date of int	EverSmo	CurrSmo	CigsDay
001	John	Lenon	1980/12/31	1	2008/07/14	1	1	20
002	Paul	McCartney	1984/04/13	2	2008/07/16	0	.	.
003	Ringo	Star	1982/10/23	1	2008/07/17	1	0	…
…	…	…	…	…	…	…	…	…

1 observation = 1 record = 1 row in a data set = …

Defining the data / types of variables

Qualitative variables
(= categorical)

(special case: binary)

- nominal
- ordinal

Quantitative variables
(= numerical)

- discrete
- continuous

Qualitative variables

A variable is called

- *qualitative*, when the values represent non overlapping (distinct) categories, without any numeric order
- *qualitative ordinal*, given a natural order between the values.

Table 2.2 Categorical (qualitative) variables recorded in the study of outcome after diagnosis of tuberculosis.

Variable	Categories	Type of variable
Hospital	1, 2, 3	Categorical
Sex	Male, female	Binary
Smear result	Negative, uncertain, positive	Ordered categorical
Culture result	Negative, positive	Binary
Alive at 6 months?	No, yes	Binary

Variables based on threshold values

Table 2.3 Examples of derived variables based on threshold values.

Derived variable	Original variable
LBW (Low birthweight):	Birthweight:
Yes	< 2500 g
No	≥ 2500 g
Vitamin A status:	Serum retinol level:
Severe deficiency	< 0.35 μmol/l
Mild/moderate deficiency	0.35–0.69 μmol/l
Normal	≥ 0.70 μmol/l

Quantitative variables

A *quantitative* variable measures/counts a quantifiable characteristic, such as height, weight or the number of children you have.

- The quantitative variable value represents a quantity / count / measurement and has numerical meaning (i.e. you can add, subtract, multiply, or divide the values of a quantitative variable, and the results make sense as numbers.)

  (This characteristic isn't true of qualitative variables, which can take on numerical values only as placeholders.)

Types of quantitative variables:

- A variable like height or weight which might involve any possible value between two other values, is called *continuous*,
- the others, like number of children, are called *discrete*.

Characteristics of empirical distributions

Which do you know?

Summarizing numerical data: tables

- Start with raw data
- Identify minimum and maximum
- Form groups(5-20, according to size of dataset)
- Count events in the group

Table 3.2 Haemoglobin levels in g/100ml for 70 women.

(a) Raw data with the highest and lowest values underlined.

12.2	13.7	10.4	14.5	11.5		
13.3	12.9	12.1	9.4	13.2		
10.6	10.5	13.7	11.9	14.1		
12.1	12.9	11.4	12.7	10.6		
9.3	13.5	14.0	11.2	11.7		
12.0	12.9	11.1	8.8	10.2		
13.4	13.1	10.9	11.3	14.7		
11.9	11.4	12.5	13.0	11.6	13.1	9.7
11.2	15.1	10.7	12.9	13.4	12.3	11.0
14.6	11.1	13.5	10.9	13.1	11.8	12.2

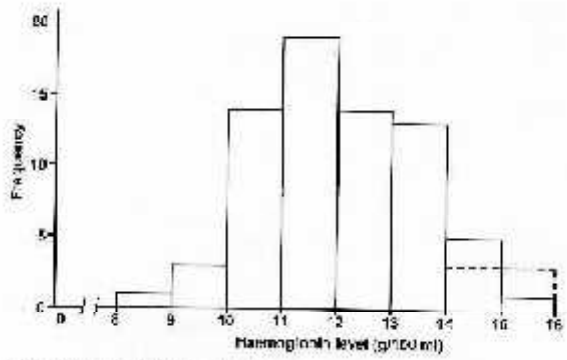
(b) Frequency distribution.

Haemoglobin (g/100 ml)	No. of women	Percentage
8–	1	1.4
9–	3	4.3
10–	14	20.0
11–	19	27.1
12–	14	20.0
13–	13	18.6
14–	5	7.1
15–15.9	1	1.4
Total	70	100.0

Summarizing numerical data: histogram

- horizontal axis: level of variable (cm, g, g/100ml)
- vertical axis (frequency per group)
- Area of the bar represents frequency
  - » if group sizes are equal: length can be taken as measure
  - » If group sizes are unequal (attention!): bar length= frequency/ bar width

see:
8-9: 1
14-15: 5
15-16: 1
as the area under the curve represents the frequency, combining 14 to 16 gives a height of 3.



Summarizing numerical data: histogram

Table 3.3 Road accident casualties in the London Borough of Harrow in 1985 (excluding 65 with unknown age)

Age	Frequency
0– 4	28
5– 9	46
10–15	58
16	20
17	31
18–19	64
20–24	149
25–59	316
60	103
Total	815

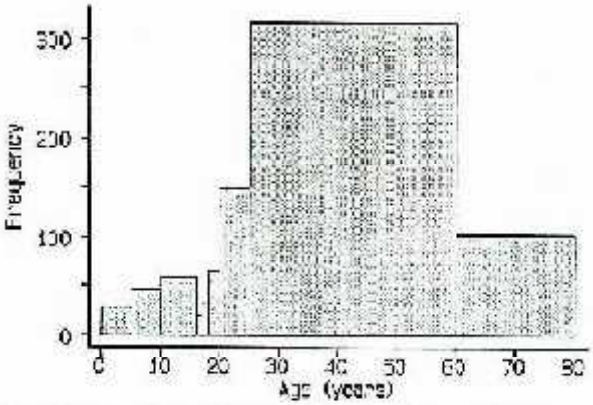


Figure 3.5 Incorrect histogram of road accident data of Table 3.3

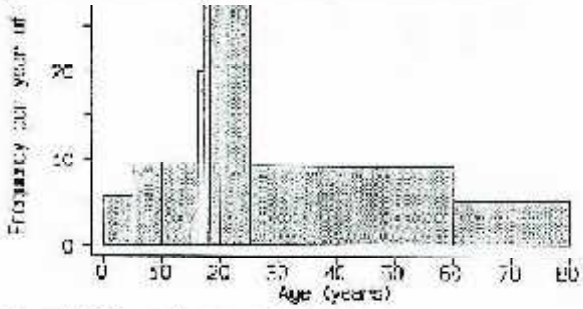
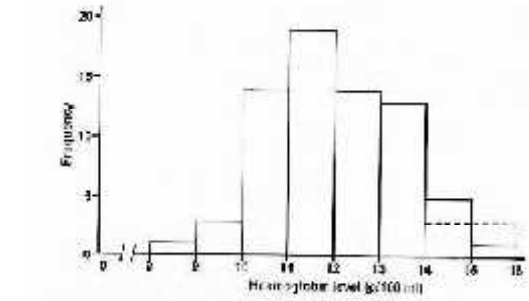
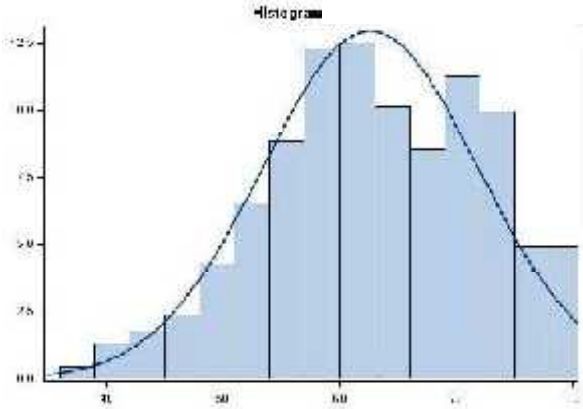
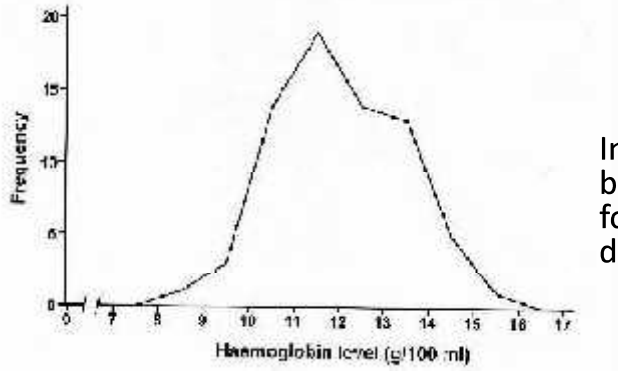


Figure 3.6 Correct histogram of road accident data

Summarizing numerical data: frequency polygon

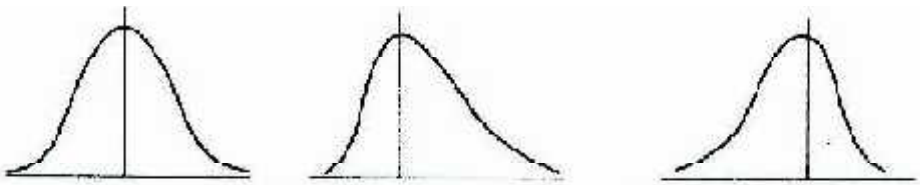






In case the dataset is big and the size of bars is small for continuous variables, the form of the histogram is similar to the density function of continuous variables.

Shapes of frequency distributions

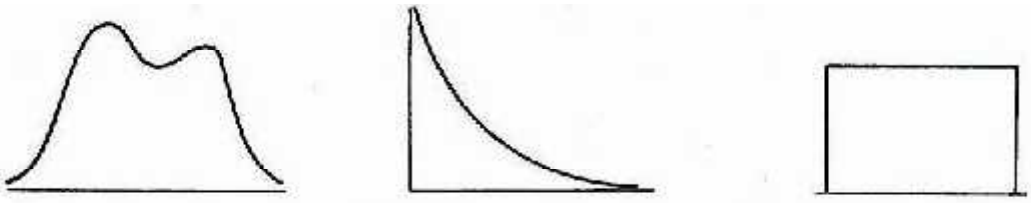


(a) Symmetrical and bell-shaped, e.g. height

(b) Positively skewed or skewed to the right, e.g. triceps skinfold measurement

(c) Negatively skewed or skewed to the left, e.g. period of gestation

Fig. 3.5 Three common shapes of frequency distributions with an example of each.



(a) Bimodal, e.g. hormone levels of males and females

(b) Reverse J-shaped, e.g. survival time after diagnosis of lung cancer

(c) Uniform, e.g. month of occurrence of disease with no seasonal pattern

Fig. 3.6 Three less-common shapes of frequency distributions with an example of each.

Summarizing numerical data: measures of location

"Where does the sample lie?"

Arithmetic mean: "Average"

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \cdots + x_n\right) = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example: ages of students

21, 18, 19, 22, 22, 23, 22, 20, 19, 24

$\bar{x} =$

Summarizing numerical data: measures of location

Median: "Middle observation" divides the sample into halves
≥50% of values ≤Median
≥50% of values ≥Median

1. order the observations
2. take the middle observation
   odd number of events: one middle observation
   even number of events: average two 'middle' values

Median is outlier-robust

<u>Example1:</u> 21, 18, 19, 22, 22, 23, 22, 20, 19, 24
ordered observations: 18, 19, 19, 20, 21, 22, 22, 22, 23, 24
median:

<u>Example2:</u> 21, 18, 19, 22, 22, 23, 22, 20, 19, 24, 25
ordered observations: 18, 19, 19, 20, 21, 22, 22, 22, 23, 24, 25
median:

Summarizing numerical data: measures of location

- **Mode**: the value which occurs most often

(not often used in publications)

- **Range**: report minimum and maximum

How much does the sample spread around the measure of location?

(often used in publications)

Example: 21, 18, 19, 22, 22, 23, 22, 20, 19, 24

Mode: 22

Range: 18-24

Summarizing numerical data: quartiles

- divide your sample into 4 equally sized groups

➢ Lower quartile = 1st quartile = Q_1 = 25th percentile

➢ 2nd quartile = 50th percentile = Q_2 = Median

➢ Upper quartile = 3rd quartile = Q_3 = 75th percentile

➢ Interquartile-range = 3rd quartile -1st quartile

➢ Quartiles are outlier-robust

Summarizing numerical data: quartiles

Table 3.3 Cumulative percentages for different ranges of haemoglobin levels of 70 women.

Observation	Cumulative percentage	Haemoglobin level (g/100 ml)		Quartile
1	1.4	8.8	Minimum = 8.8	1
2	2.9	9.3		1
3	4.3	9.4		1
4	5.7	9.7		1
5	7.1	10.2		
⋮	⋮	⋮		
15	21.4	10.8		1
16	22.9	10.9		1
17	24.3	10.9	Lower quartile = 10.9	1
18	25.7	10.9		1
19	27.1	11.0		2
20	28.6	11.0		2
	⋮	⋮		
33	47.1	11.7		2
34	48.6	11.8		

Summarizing numerical data: measures of location

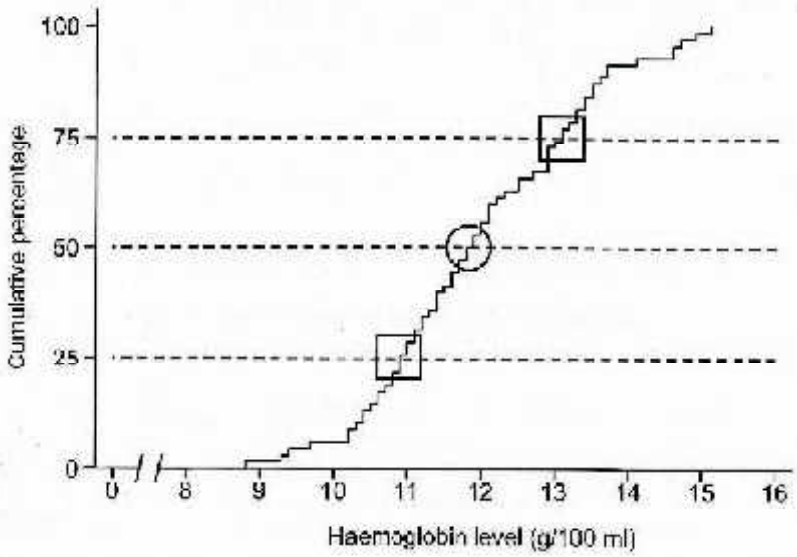
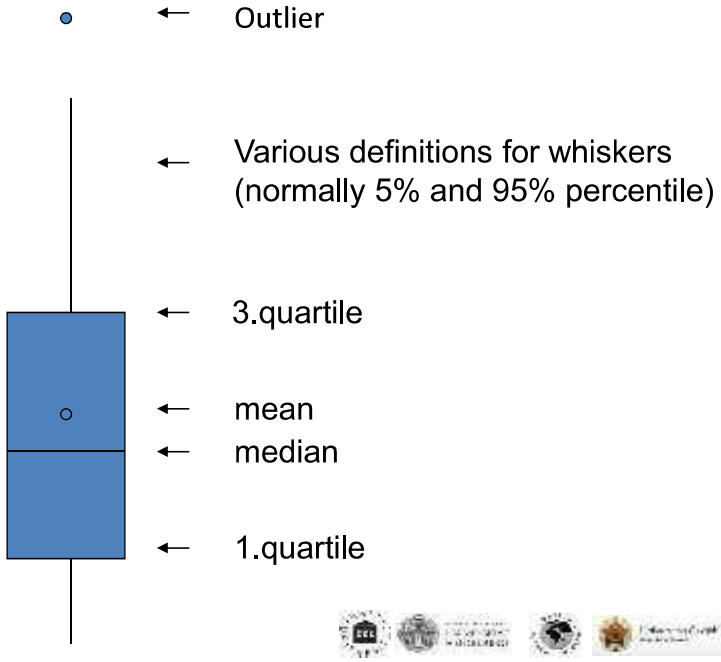


Fig. 3.7 Cumulative frequency distribution of haemoglobin levels of 70 women, with the median marked by a circle, and lower and upper quartiles marked by squares.

Summarizing numerical data: boxplot

Box-and Whiskers-Plot



Summarizing numerical data: boxplot

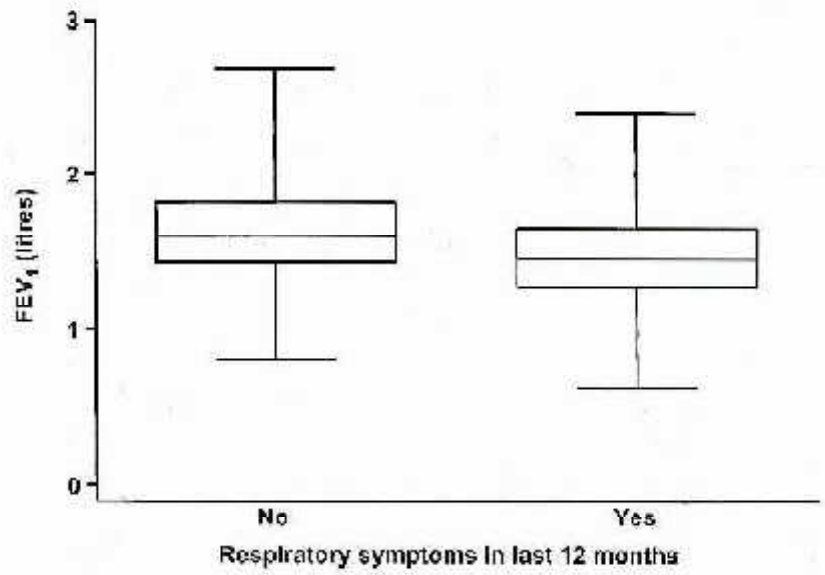


Fig. 3.12 Box and whiskers plots of the distribution of FEV_1 in 636 children living in a deprived suburb of Lima, Peru, according to whether they reported respiratory symptoms in the previous 12 months.

Summarizing numerical data: dot plot

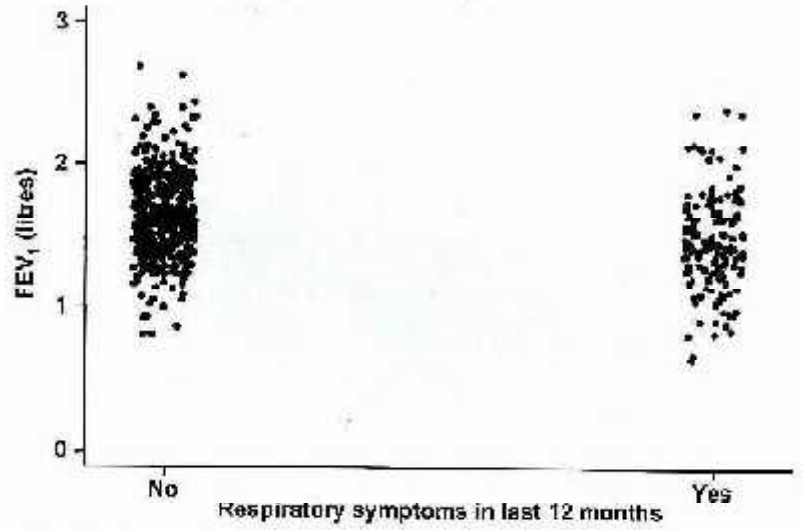


Fig. 3.11 Scatter plot showing the relationship between FEV_1 and respiratory symptoms in 606 children living in a deprived suburb of Lima, Peru. The position of the points on the horizontal axis was moved randomly ('jittered') in order to separate them.

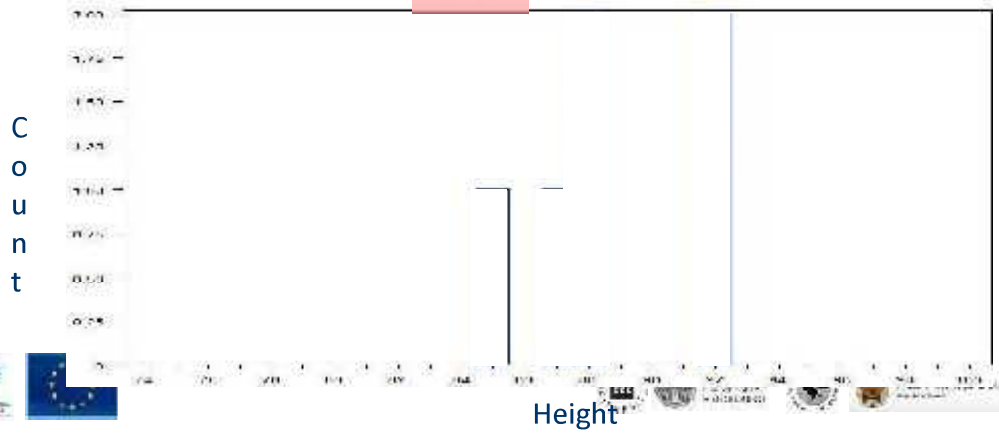
Standard notation in statistics

	Population parameter	Sample estimate
Size	N	n
Mean	μ	$\bar{X}$
Variance	σ^2	s^2
Standard deviation	σ	s

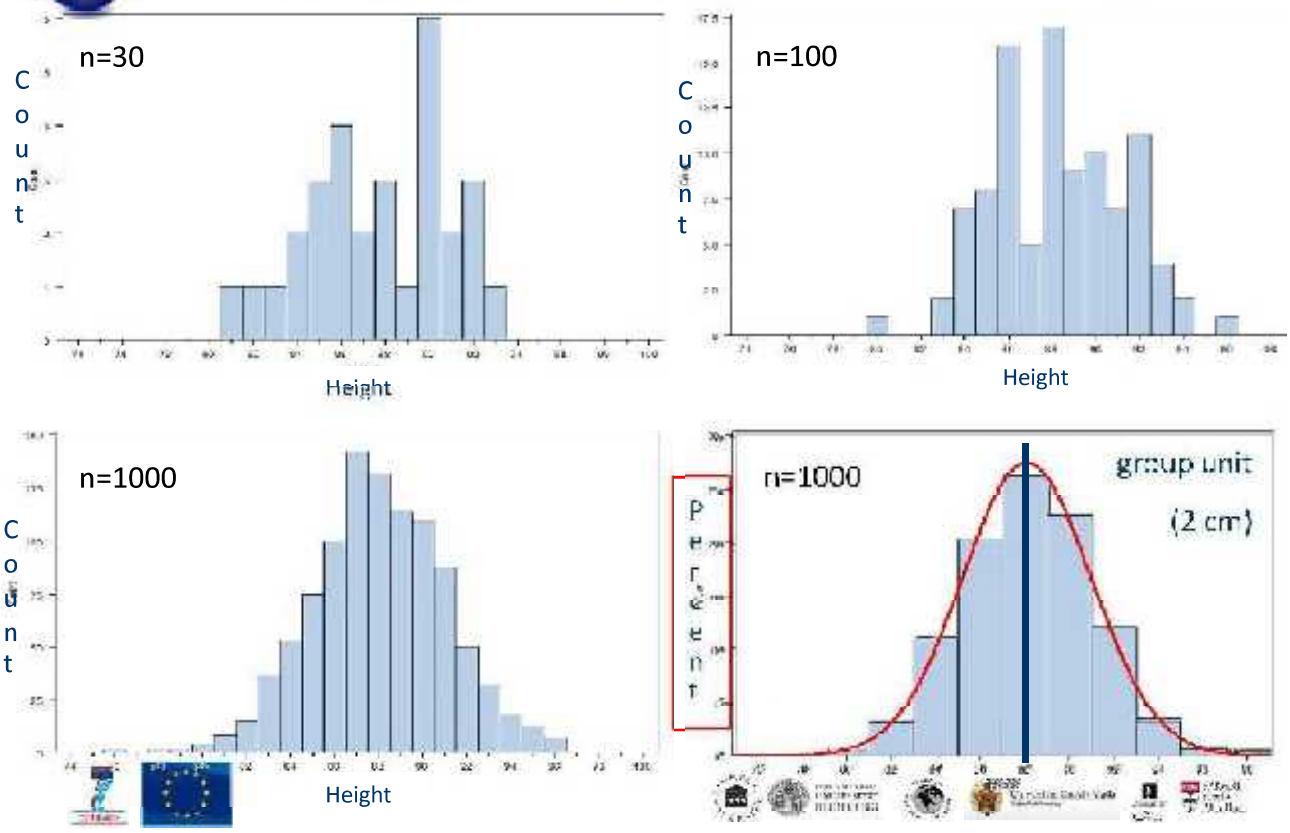
Histogram

Table 1: 6 Observations from a pediatrician

ID	Age (years)	Sex	Height	Weight (kg)	Number of siblings	Musicality
001	2	M	85.1	12.3	1	high
002	2	M	91.5	13.1	2	low
003	2	F	87.8	12.2	0	normal
004	2	F	92.3	13.6	1	low
005	2	M	86.5	11.8	3	Normal
006	2	M	88.2	12.7	0	Normal

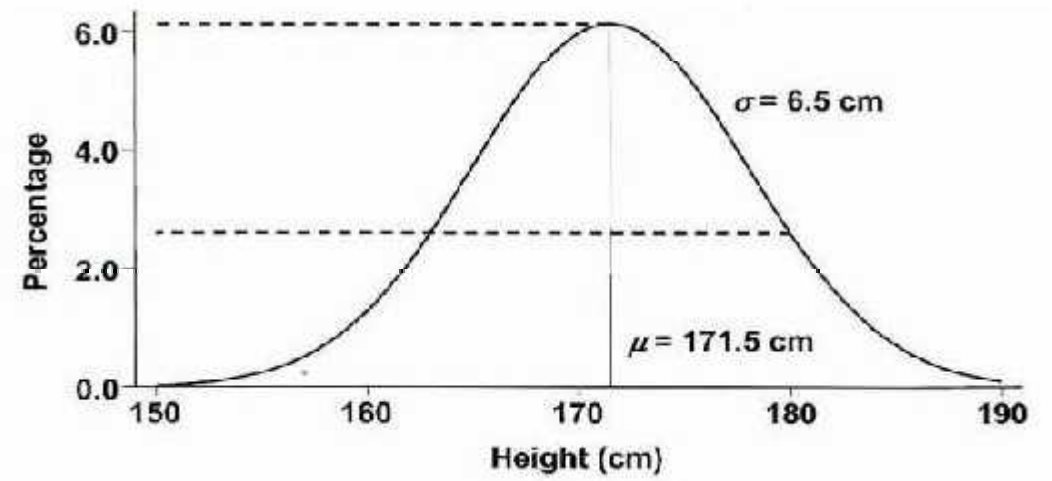


Histogram (children at their 2nd birthday)



The normal distribution

Diagram showing the approximate normal curve describing the distribution of heights of adult men



source: Kirkwood, Essential Medical Statistics

The normal distribution

Many variables are distributed according to the normal distribution e.g. height, weight, blood pressure, temperature, …

<u>Normal Distribution = Gaussian Distribution</u>

- symmetrical around the mean
- mean = median
- bell-shaped
- location given by the *mean* $\overline{X}$
- shape given by its *variance* $\mathbf{s}^2$
- Notation: $X \sim N(\overline{X}, \mathbf{s}^2)$

(Arithmetic) mean

Table 1: 6 Observations from a pediatrician

ID	Age (years)	Sex	Height	Weight (kg)	Number of siblings	Musicality
001	2	M	85.1	12.3	1	high
002	2	M	91.5	13.1	2	low
003	2	F	87.8	12.2	0	normal
004	2	F	92.3	13.6	1	low
005	2	M	86.5	11.8	3	normal
006	2	M	88.2	12.7	0	normal

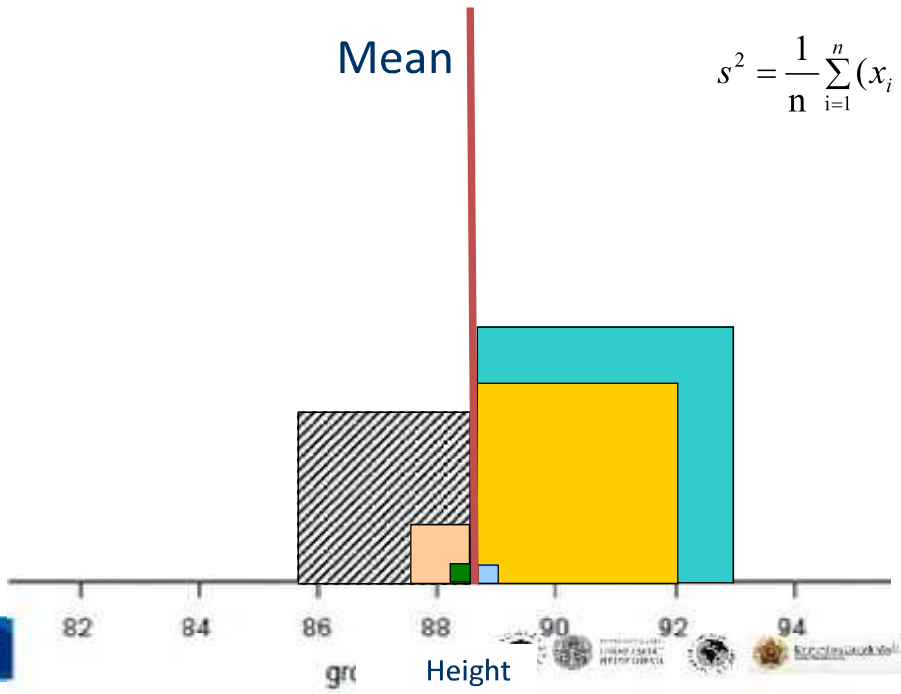
$$\bar{x} = \frac{x_1 + x_2 + .. + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$x =$

Variance σ

= calculating the mean of squared distances

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



Definitions

- Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

  Mathematically better: division by (n-1) instead of n

- Standard deviation: $s = \sqrt{s^2}$

  Note: s^2 and s are estimates for the unknown population parameters σ^2 and σ

- Standard error of a mean:

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

mean, variance and standard deviation

Table 1: 6 Observations from a pediatrician

ID	Height	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
001	85.1		
002	91.5		
003	87.8		
004	92.3		
005	86.5		
006	88.2		

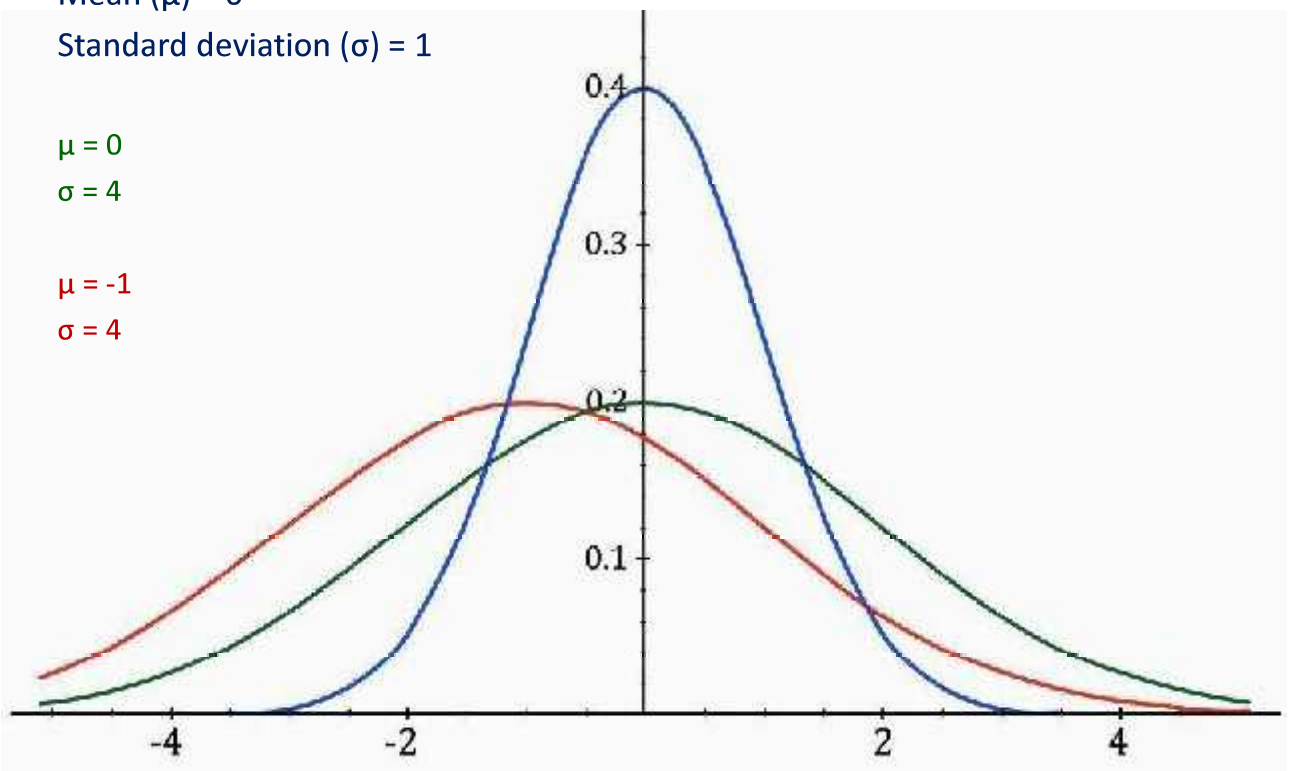
$\bar{x} = 531.4/6 = 88.6$

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

We can shift and shrink the normal distribution

Mean (μ) = 0

Standard deviation (σ) = 1



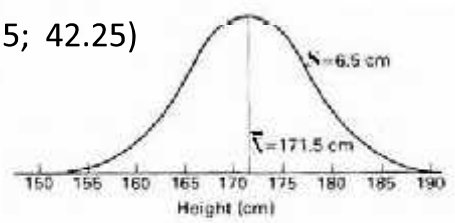
Standard normal distribution/z-transformation

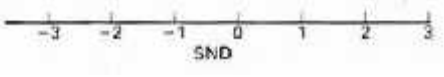
If a normally distributed variable X has mean $\bar{x}$ and standard deviation s, i.e. X ~ N($\bar{x}$, s^2) we can transform it into Z ~ *N*(0,1)

- x- $\bar{x}$ has a mean of 0. This holds for $\frac{x-\bar{x}}{s}$ as well.
- $\frac{x-\bar{x}}{s}$ has a standard deviation of 1.

The distribution Z ~ *N*(0,1) is *the* **Standard normal distribution**

X ~ *N*(171.5; 42.25)





Which variable Z is standard normal distributed?
Z ~ *N*(0; 1)

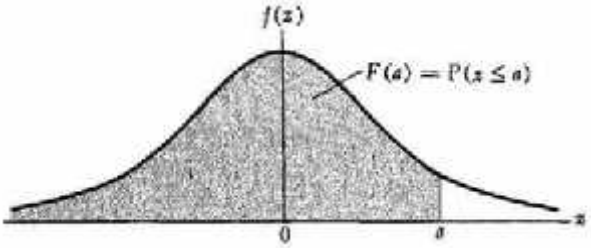
Z- scores

- Area under *standard normal distribution* = 1
- Area less than z = probability for this value or lower

  i.e., the table gives the cumulative probability up to the standardised normal value z

- Z-score tables are easily available:

  https://www.boundless.com/image/z-score-table--2/

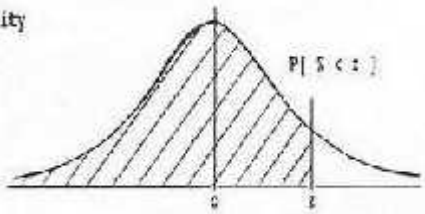


Z Score Table

The table gives the cumulative probability up to the standardized normal value z

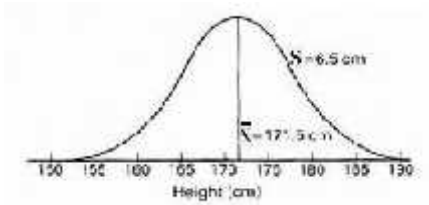
i.e.

$$P[\,Z < z\,] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$$



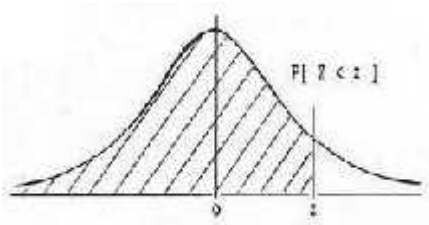
z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633

z-core: an example



Find the proportion for men being taller than 185 cm?

Area between 2 values:



What is the proportion of men with height between 165 and 175 cm?

Brief outlook: confidence interval

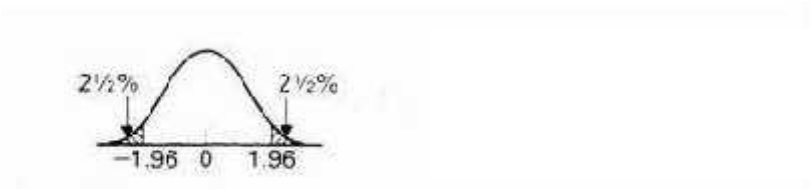
What is the limit z such that the area between -z and z is 95%?

→ What is the limit z such that the area above z is 2.5%

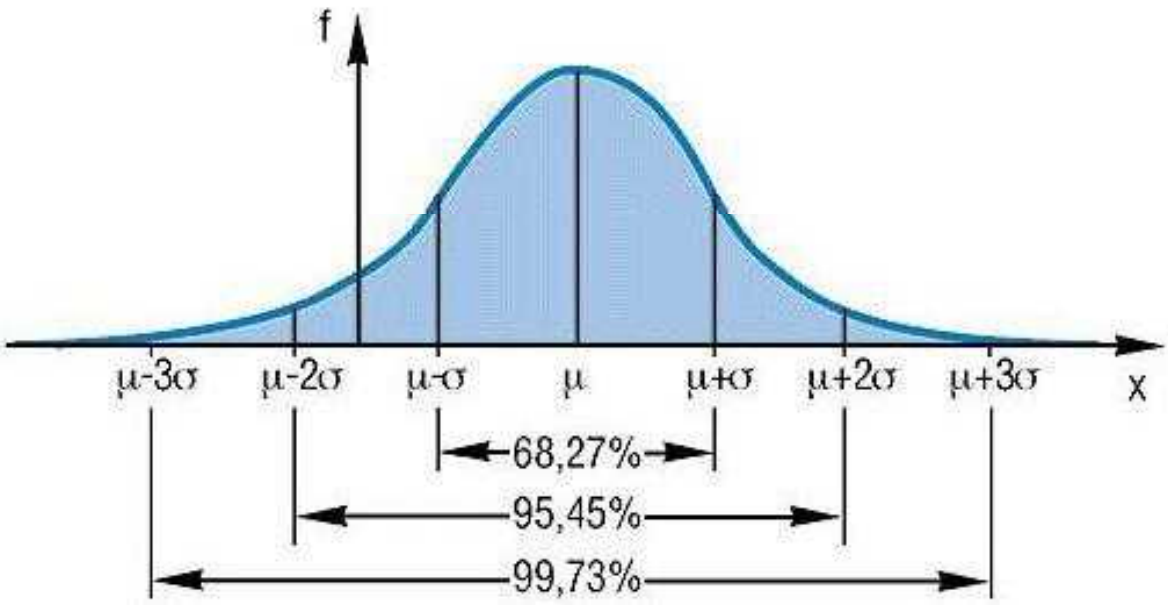
z = 1.96

1.96 is called the **two sided 5% percentage point** of the SND

i.e. on each side (positive axis and negative axis, we have 2.5%)



More generally: a and z



So far…

…we talked about 1 sample and the characteristics of a distribution.

…now we'll talk about standard deviation and standard error,
in other words, more samples, not only 1.

Sampling variation and standard error

What is a good estimation of the mean?

- If we take the mean of a sample, we will get different means for different samples.
- How can we describe the ***variation of the mean of the sample*****?**

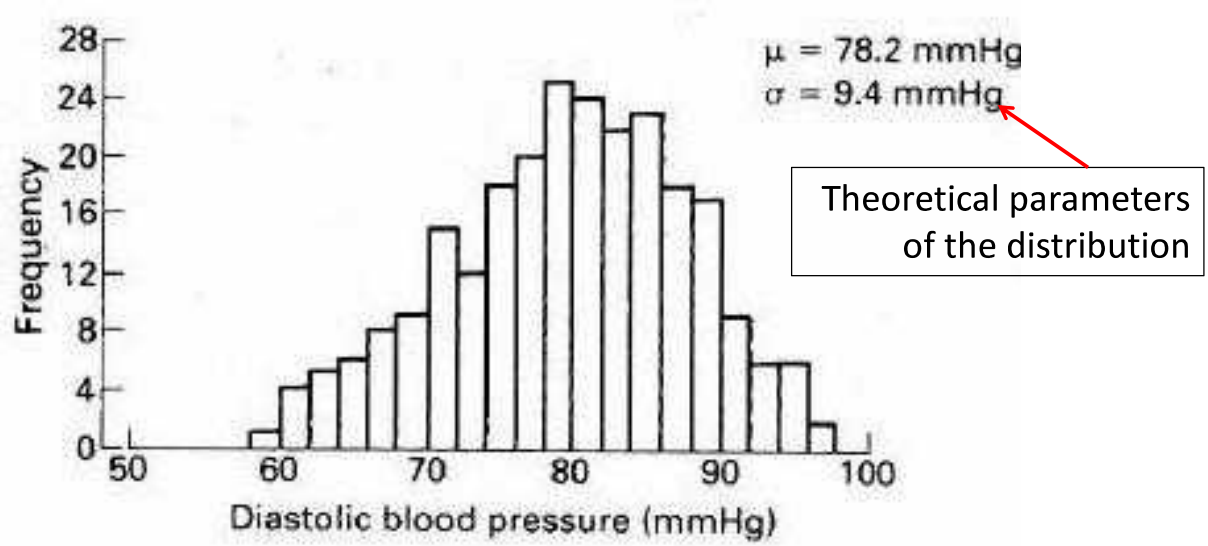
Standard error (of the sample mean)

- The variation of the mean of the sample depends:
  - on the sample size (the larger the better!)
  - on the variation of the original values (measured by σ or s)

  It is calculated by: $\text{s.e.} = \frac{s}{\sqrt{n}}$

Sampling variation and standard error

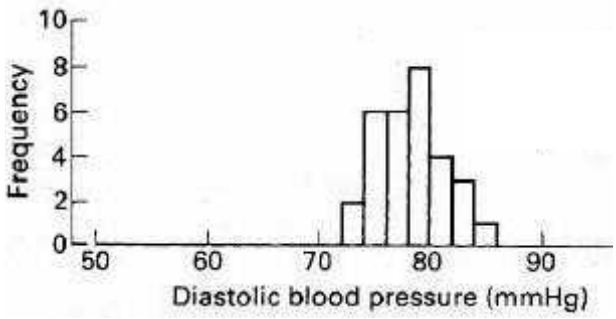
a) Distribution of diastolic blood pressure for a population of 250 airline pilots



Sampling variation and standard error

Sampling distribution for 30 sample means

b) sample size =10



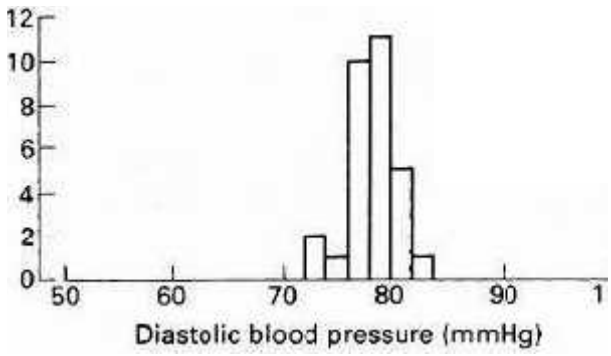
Mean (of sample means) = 78.2 mmHg

s.d. (sample means) = 3.01 mmHg

s.e. (theoretical) = $9.4/\sqrt{10}$

= 2.97 mmHg

c) sample size =20



Mean (of sample means) =78.1mmHg

s.d. (sample means) = 2.07 mmHg

s.e. (theoretical) = $9.4/\sqrt{20}$

= 2.1 mmHg

Standard deviation and standard error

- The **standard deviation** measures the amount of variability in the population

- The **standard error** measures the amount of variability in the sample mean; it indicates how closely the population mean is likely to be estimated by the sample mean.

Confidence interval (CI)

- Mean of the sample, $\bar{x}$, is an estimate for the "true mean", μ, of the population

- How reliable is this value $\bar{x}$?
- How much confidence do we have in a single value?

- Mean is a "point – estimator"
- Confidence interval is an "interval – estimator"

There is a 95% probability that the interval between:

($\bar{x}$ -1.96 * s.e.) and ($\bar{x}$ + 1.96 * s.e.)

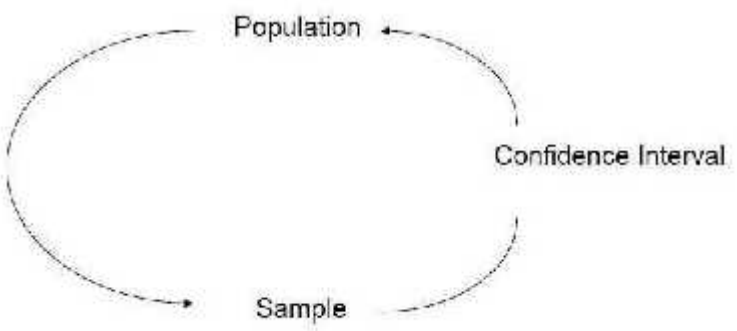
contains the population mean (which remains unknown).

Confidence interval (CI)

- 95%-CI: $$\left[\bar{x} - 1.96 * \frac{s}{\sqrt{n}},\ \bar{x} + 1.96 * \frac{s}{\sqrt{n}}\right]$$
- Determine an interval in which the population parameter lies with probability 1-α, e.g. 1-α= 95%
- The higher 1-α, the larger the CI:
- If you want more confidence (99%) the interval gets broader (replace 1.96 by 2.58)
- If you need less confidence (90%) it gets narrower (replace 1.96 by 1.64)

Confidence interval (CI)

- Use the confidence interval to make inferences about the population from which the sample is drawn



- CI depends on the actual sample

Confidence interval for a mean

Example: Indoor residual spraying, i.e. spraying the inside of dwellings with an insecticide to kill mosquitos that spread malaria. How much insecticide is needed for 10,000 houses?

- A sample of 100 houses was chosen.
- Mean sprayable surface of 100 houses 24.2 m^2, with a standard deviation of 5.9 m^2.

$$\left[24.2 - 1.96 * \frac{5.9}{\sqrt{100}},\ 24.2 + 1.96 * \frac{5.9}{\sqrt{100}}\right] = [23.04,\ 25.36]$$

- Which means: with probability of 95% the true (but unknown) population mean lies between 23.04 and 25.36.

$$[23.04 \le \mu \le 25.36]$$

Variability of confidence intervals

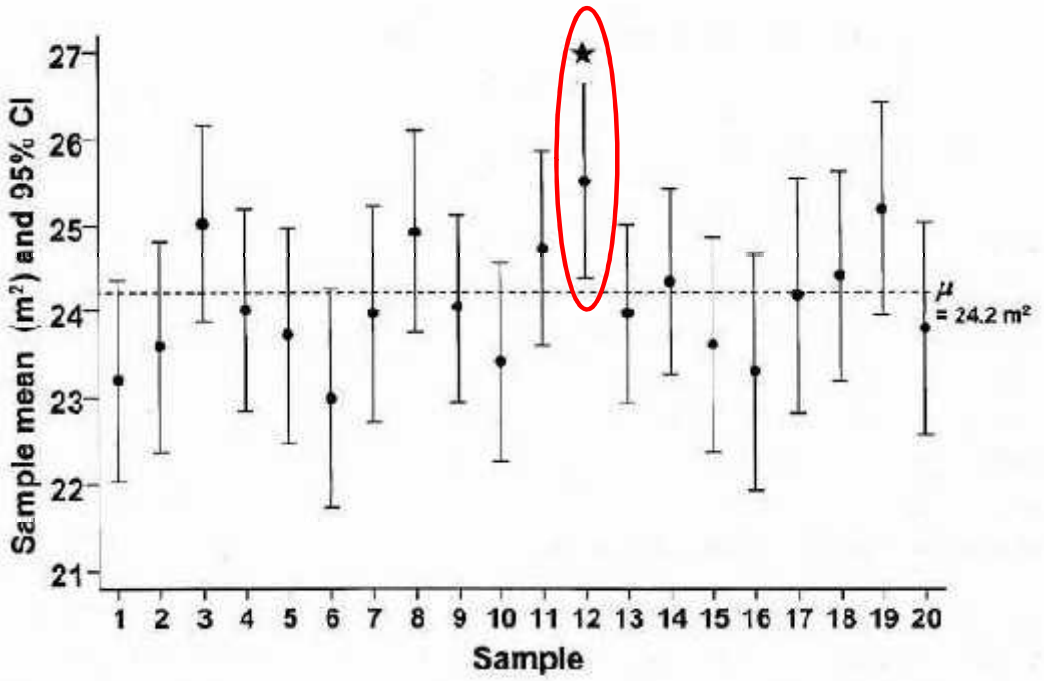


Fig. 6.2 Mean sprayable areas, with 95% confidence intervals, from 20 samples of 100 houses in a rural area. The star indicates that the CI does not contain the population mean.

Confidence interval for a small sample

- For large sample size $\left[\bar{x} - 1.96 * \frac{s}{\sqrt{n}},\ \bar{x} + 1.96 * \frac{s}{\sqrt{n}}\right]$

  as an approximation for $\left[\bar{x} - 1.96 * \frac{\sigma}{\sqrt{n}},\ \bar{x} + 1.96 * \frac{\sigma}{\sqrt{n}}\right]$

- When the sample size is small, s may not be a reliable estimate of σ

  → use t-distribution $\left[\bar{x} - t * \frac{s}{\sqrt{n}},\ \bar{x} + t * \frac{s}{\sqrt{n}}\right]$

- "Student's t-distribution":

  5% point of the t distribution with (n-1) degrees of freedom

Interpretation

Strictly speaking, what is the best interpretation of a 95% confidence interval for the mean?

a) If repeated samples were taken and the 95% confidence interval was computed for each sample, 95% of the intervals would contain the population mean.

b) A 95% confidence interval has a 0.95 probability of containing the population mean.

c) 95% of the population distribution is contained in the confidence interval.