Kyiv
School of
Economics

Descriptive statistics

Ass. Professor Andriy Stavytskyy

Outline

- Sample
- Descriptive statistics

What is statistics?

Statistics is the study of the collection, analysis, interpretation, presentation, and organization of data.

Assumptions:

- The observations are the values of a random variable
- The sample represents the population from which it is selected

Basic concepts

- Population: Collection of objects for which a conclusion shall be made (can be human beings but also a collection of atoms when applied in physics)

- Sample: a representative part/sub-set of the population

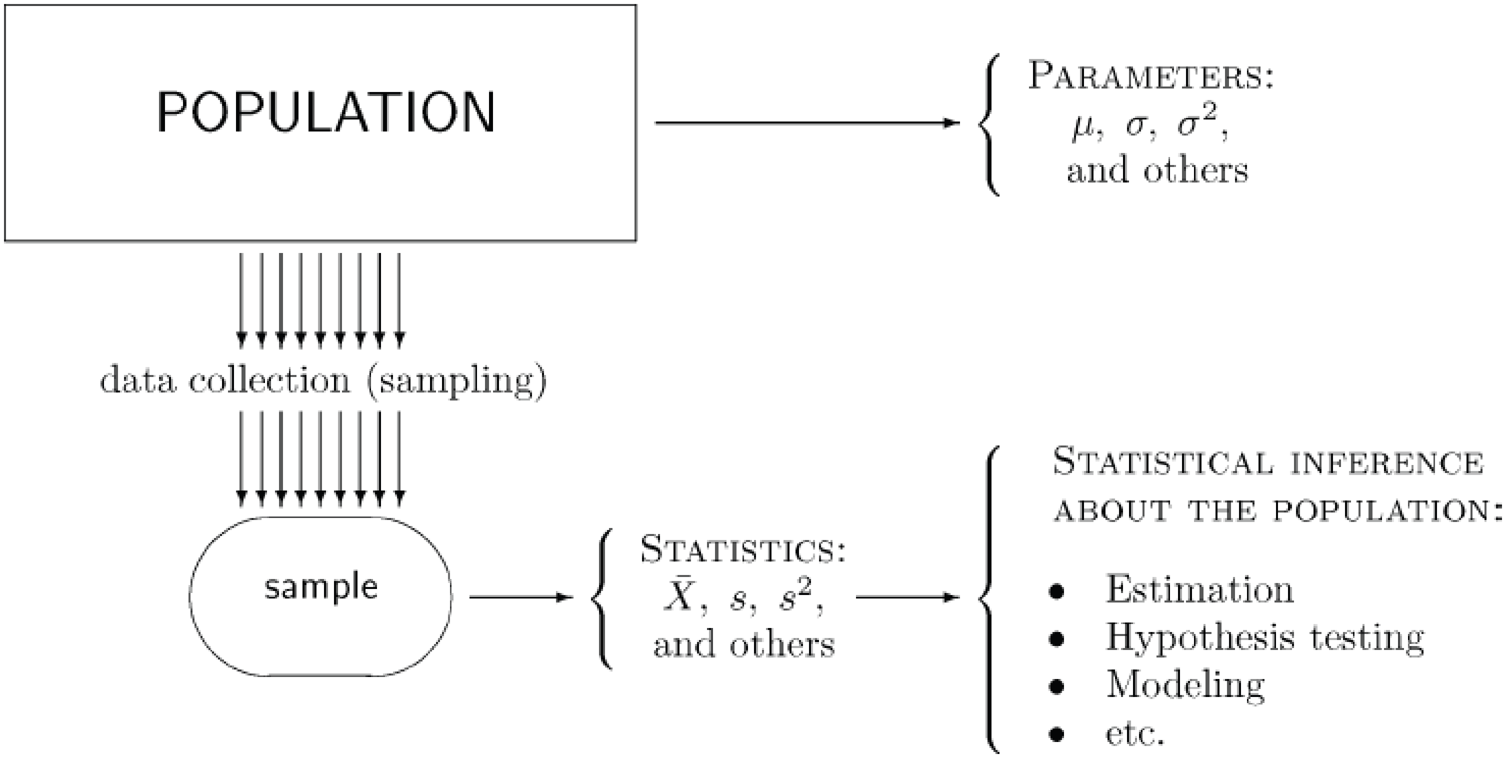
- Random sample: elements of the population drawn randomly and independently of each other

Example: „Mietspiegel“ (= statistics of rents) for the city of Bonn

  - Population: all rooms, flats etc. for rent in Bonn (← too many to investigate all)
  - Sample: selected part; all flats from Poppelsdorf
  - Random sample: Investigation of $n = 100, 200, \ldots$ random objects from Bonn

Population and sample

Population parameters and sample statistics.



Frequencies – 1

- Absolute frequency n_i:
  - Number of obersvations with attribute value i (counts)

- Relative frequency h_i:
  - Portion of elements with attribute value i
  - To be computed as absolute frequency devided by total number of objects N: n_i / N
  - Relative frequencies lie between 0 and 1
  - Relative frequencies have to add up to 1 (<- can be used to check computation)

Example

AB0 blood group

	value	tally sheet	absolute frequency n_i		relative frequency h_i	
			Kyiv	Lviv	Kyiv	Lviv
1	0	𝍸 𝍸 𝍸 ‖	17	78	0.34	0.39
2	A1	𝍸 𝍸 𝍸 ‖‖	19	76	0.38	0.38
3	A2	𝍸 |	6	20	0.12	0.10
4	B	𝍸	5	18	0.10	0.09
5	A1B	‖	2	6	0.04	0.03
6	A2B	|	1	2	0.02	0.01
7	other		0	0	0.00	0.00
			N = 50	200	1.00	1.00

$$\sum_{i=1}^{k} n_i = N$$

$$h_i = \frac{n_i}{N}$$

$$0 \leq h_i \leq 1$$

$$\sum_{i=1}^{k} h_i = 1$$

Frequencies - 2

- Cumulative frequency:
  - Sum of all frequencies up to a given value i.
  - Denoted as N_i for absolute frequencies and denoted as H_i for relative frequencies
  - Often used when values are subdivided into classes

- Classification:
  - Arrangement of attribute values into disjoint groups, so called „classes“
  - Classes are disjoint, i.e. non-overlapping, and neighbouring intervals of attribute values, which are defined by a lower and an upper bound. Neighbouring values implies that each value belongs to a class and does not lie outiside (completeness of the classification).

Example

height ***[cm]***

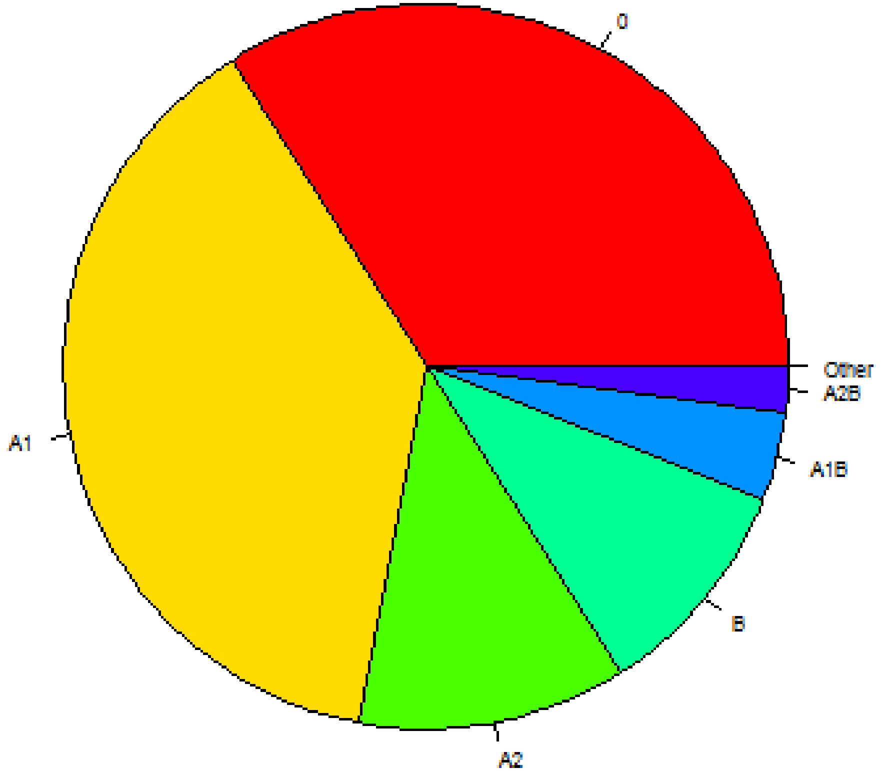
Class number i	Class limits $(a_{i-1}; a_i]$	Tally sheet	frequency		Cumulative frequency	
			absolute n_i	relative h_i	absolute N_i	relative H_i
1	≤ 150		0	0.00	0	0.00
2	(150; 160]		5	0.05	5	0.05
3	(160; 170]		30	0.30	35	0.35
4	(170; 180]		35	0.35	70	0.70
5	(180; 190]		25	0.25	95	0.95
6	(190; 200]		5	0.05	100	1.00
7	> 200		0	0.00	100	1.00
			N=100	1,00		

$$\boxed{N_i = \sum_{k=1}^{i} n_k}$$

$$\boxed{H_i = \frac{N_i}{N}}$$

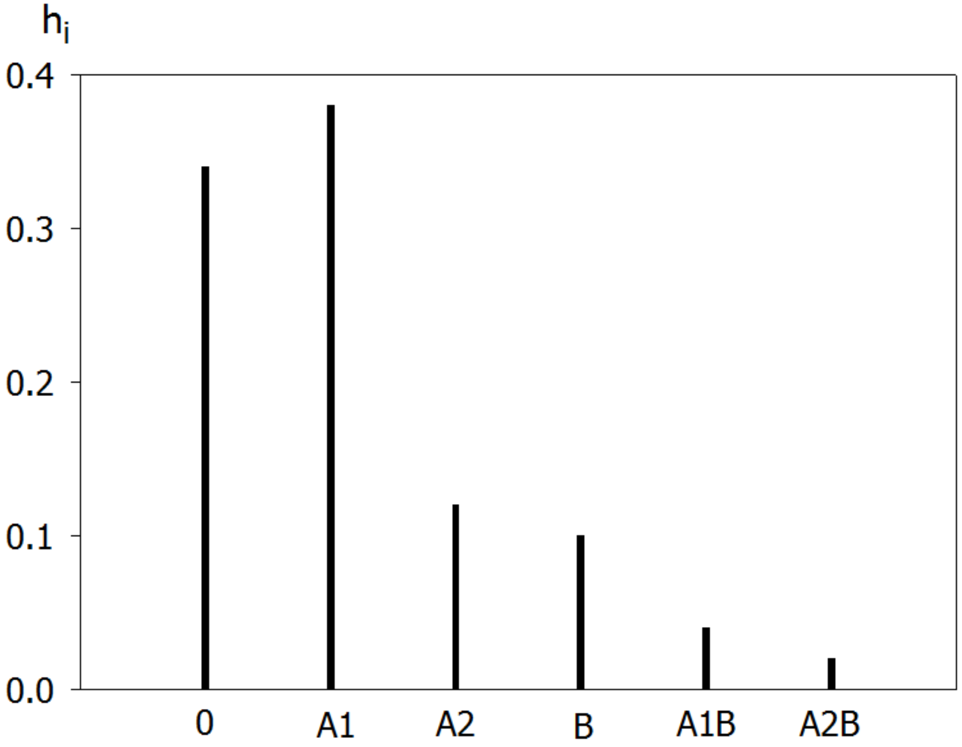
Graphical representation – 1

- Pie chart
  - Shows absolute frequencies
  - Example: blood groups



Graphical representation – 2

- Bar chart
  - Shows relative frequencies
  - Example: blood groups



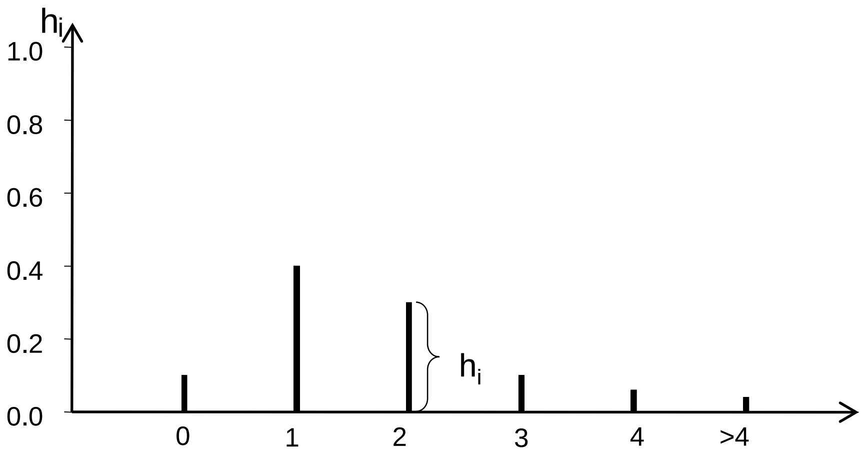
Empirical distribution function – 1

- Representation of cumulative frequencies with empirical distribution function F
  - Discrete trait: Number of Children

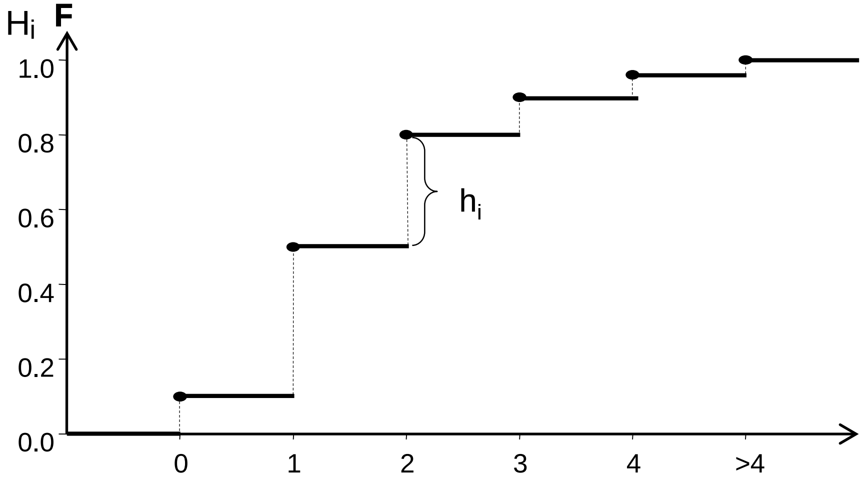
	Number of children	Tally sheet	Frequencies		Cumulative frequencies	
			absolute n_i	relative h_i	absolute N_i	relative H_i
1	0	𝍸	5	0.10	5	0.10
2	1	𝍸 𝍸 𝍸 𝍸	20	0.40	25	0.50
3	2	𝍸 𝍸 𝍸	15	0.30	40	0.80
4	3	𝍸	5	0.10	45	0.90
5	4	𝍷𝍷𝍷	3	0.06	48	0.96
6	>4	𝍷𝍷	2	0.04	50	1.00
			N = 50	1.00		

Empirical distribution function – 2

Number of children



<u>Bar chart</u>



F: <u>Empirical distribution function</u>

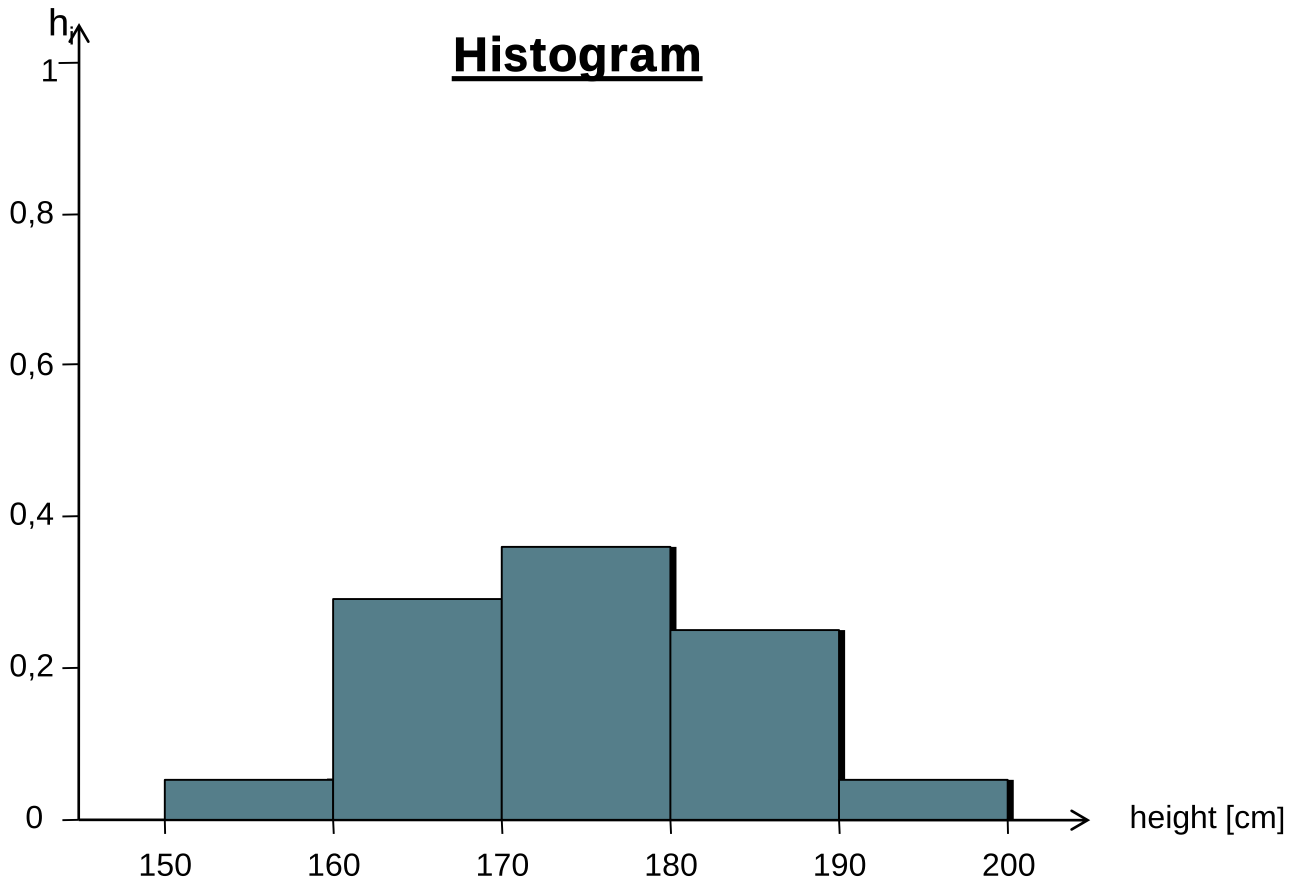
Since the attribute is quantitative discrete, we obtain a step function

Histograms – 1

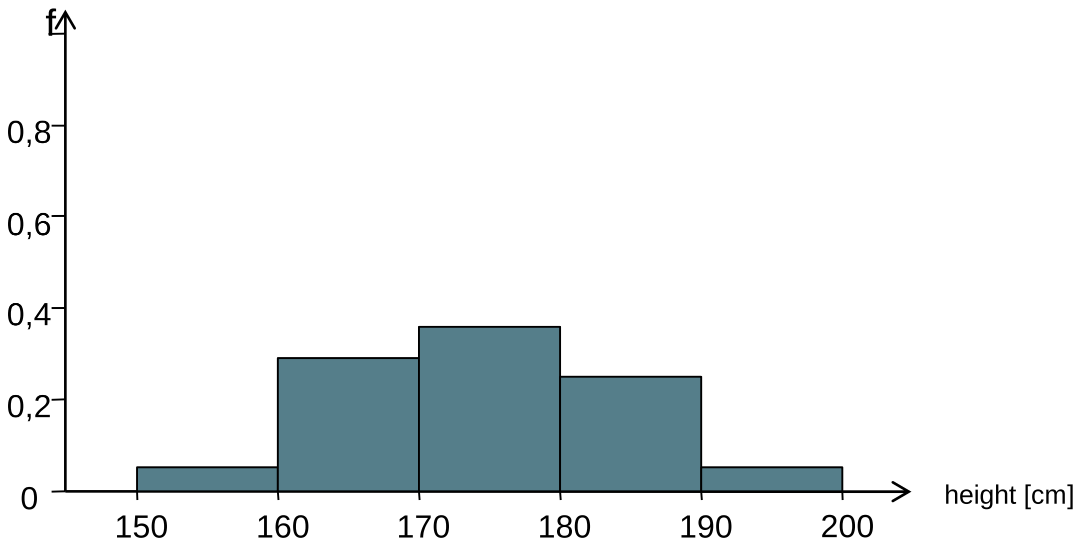
- Construction:
  - Data is subdivided into classes
  - Surface area of columns is proportional to the respective frequencies
  - Columns are neighbouring since classes are neighbouring

Histograms – 2

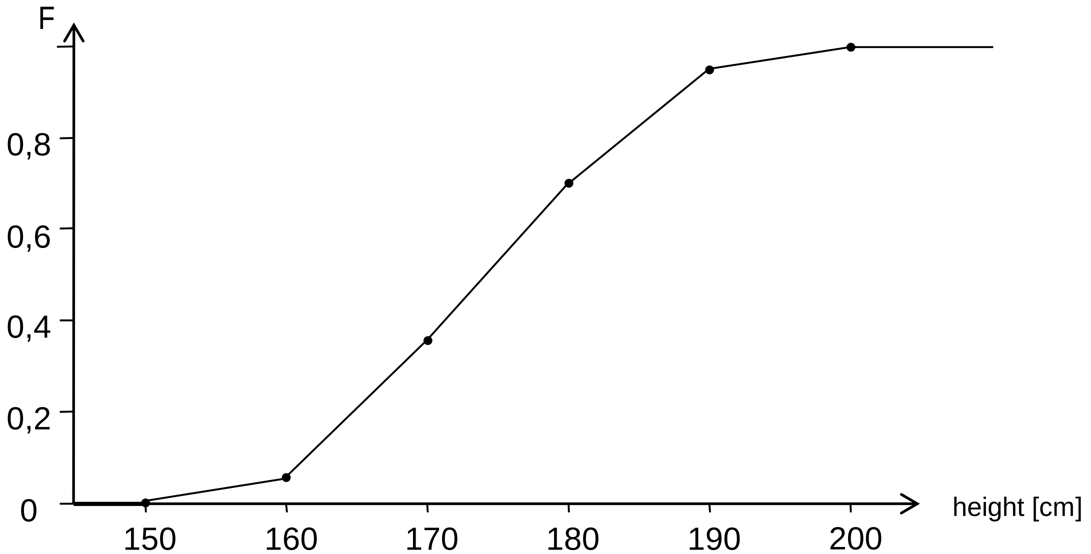
Example: ***Height [cm]***



Histograms – 3



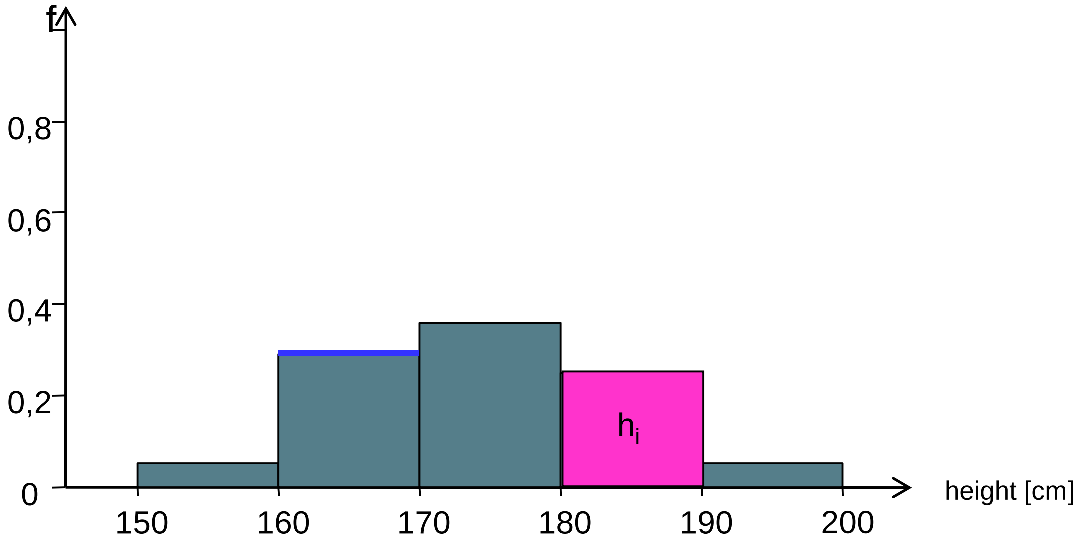
empirical density function f



empirical distribution function F

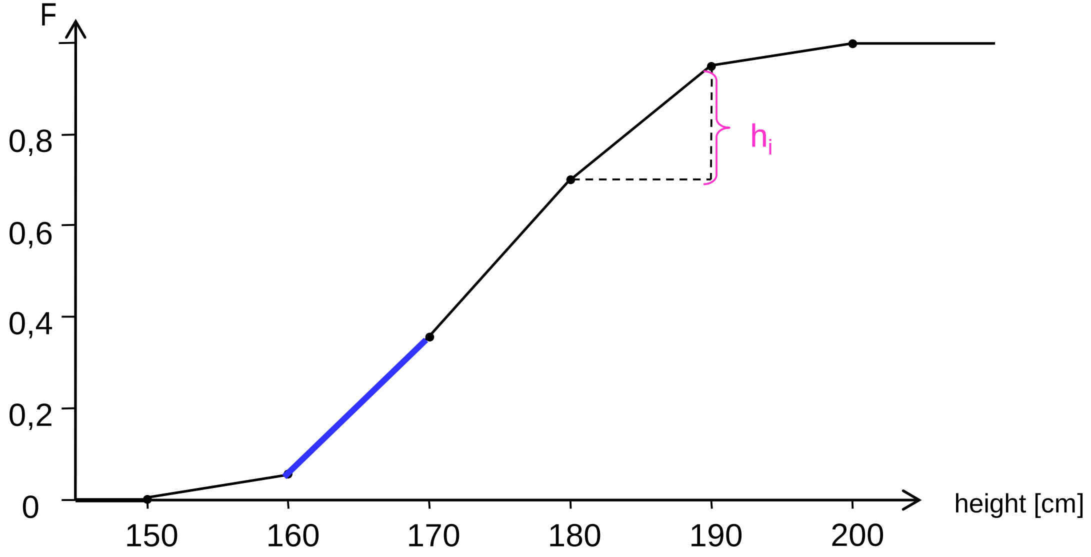
(for continuous trait)

Histograms – 4



empirical density function f

$$f = F'$$



empirical distribution function F

$$F = \int f$$

Measures of central tendency

A number to characterize the „center“ of the data

- Most important:
  - Mean
  - Median

Median

- Sample: $x_1, x_2, \dots, x_n \rightarrow$ Order according to: $x_1 \leq x_2 \leq \cdots \leq x_n$
  $\rightarrow$Ordered sample: $x_{(1)}, x_{(2)}, \dots, x_{(n)}$
  - Median $\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \textit{in case } n \textit{ is odd (value in the "middle")} \\ \frac{1}{2}\left[x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right], & \textit{in case } n \textit{ is even} \end{cases}$

sample	ranks
$x_1=5$	$x_{(1)}=3$
$x_2=9$	$x_{(2)}=4$
$x_3=3$	$x_{(3)}=5$
$x_4=8$	$x_{(4)}=6$
$x_5=19$	$x_{(5)}=8$
$x_6=4$	$x_{(6)}=9$
$x_7=6$	$x_{(7)}=19$

n = 7 odd:

$$\tilde{x} = x_{\left(\frac{7+1}{2}\right)} = x_{(4)} = 6$$

sample	ranks
$x_1=5$	$x_{(1)}=3$
$x_2=9$	$x_{(2)}=4$
$x_3=3$	$x_{(3)}=5$
$x_4=8$	$x_{(4)}=6$
$x_5=19$	$x_{(5)}=7$
$x_6=4$	$x_{(6)}=8$
$x_7=6$	$x_{(7)}=9$
$x_8=7$	$x_{(8)}=19$

n = 8 even:

$$\tilde{x} = \frac{1}{2}\left[x_{\left(\frac{8}{2}\right)} + x_{\left(\frac{8}{2}+1\right)}\right] = \frac{1}{2}\left[x_{(4)} + x_{(5)}\right] = \frac{1}{2}[6+7] = 6.5$$

Median for interval sample

$$Me = y_i + h_i \frac{\frac{n}{2} - \sum_{k=1}^{i-1} m_k}{m_i}$$

Mean

- Mean
  - Sample: $x_1, x_2, \ldots, x_n$
  - Sample size: n

  - Mean $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$

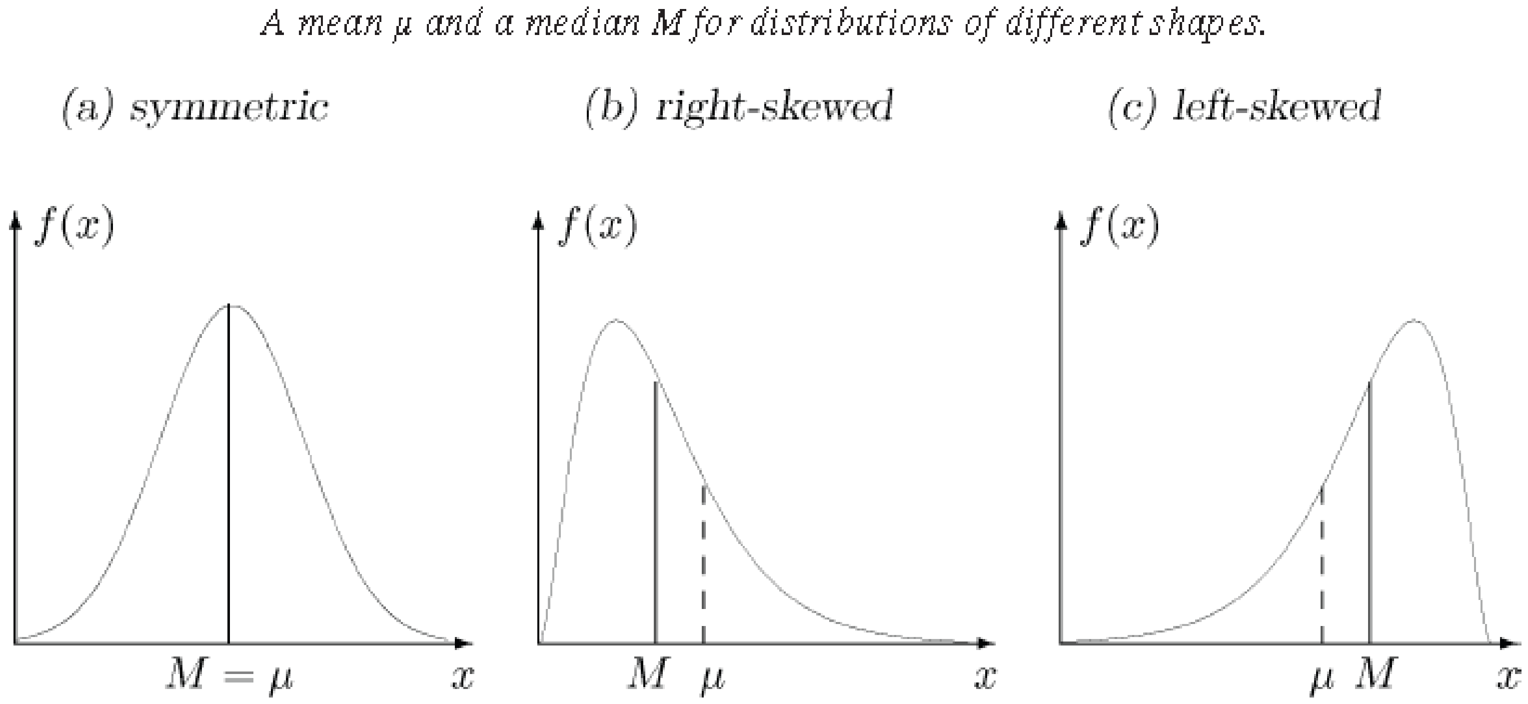
Comparison of median and mean

- Both samples have median 2500
- $\bar{x} = 3000$ and $\overline{x}' = 5000$ are the mean values
- Mean can strongly be influenced by a single value

- Median is more robust against extreme values („outliers“)
- Nevertheless, the mean is more often used in practice since it has other desirable properties.

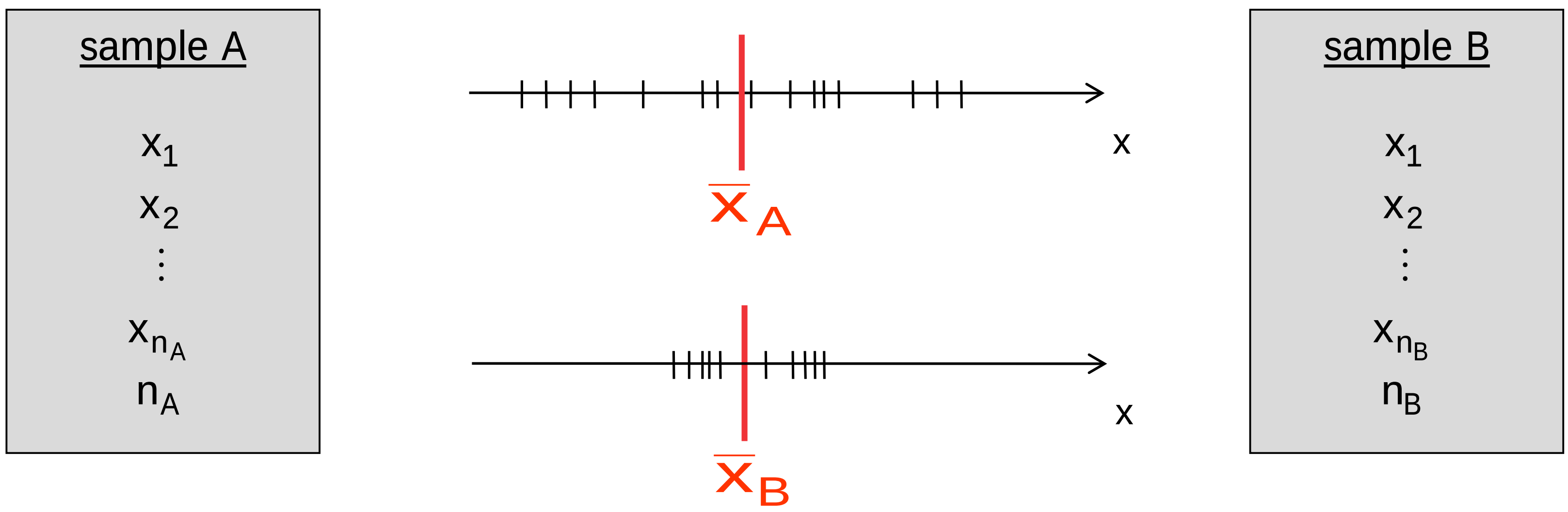
i	x_i	x_i'	ordered x_i and x_i'
1	2000	2000	1500
2	5000	15000	2000
3	4000	4000	2500
4	1500	1500	4000
5	2500	2500	5000 / 15000

Mean vs. median



Center of gravity vs. half of the area

Amount of variation of the data



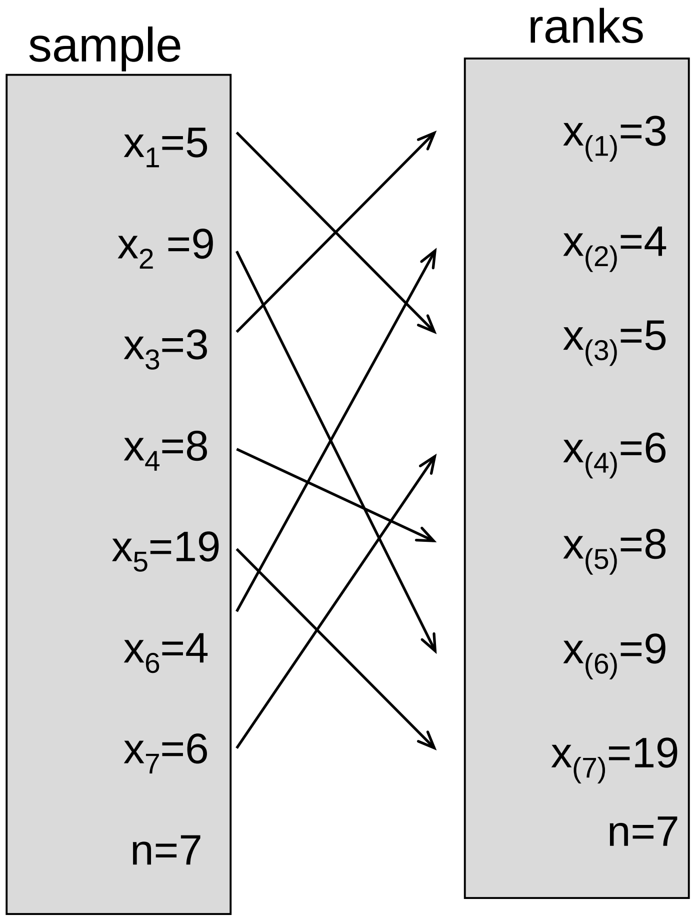
$\rightarrow$ The mean (or median) is not sufficent to describe a sample

Measures of dispersion and spread

- Measures of dispersion and spread:
  - Numbers to characterize the amount variation around the center (= mean)

- Most important:
  - Minimum, maximum, range (dispersion)
  - Empirical variance (spread)
  - Empirical standard deviation (spread)

Range



minimum: $\min = x_{(1)}$

maximum: $\max = x_{(n)}$

range: $R = x_{(n)} - x_{(1)} = 16$

Variance and standard deviation

•A measure to express the spread around the center (mean) by a single value

•The squared deviation $(x_i - \bar{x})^2$ of each attribute value x_i from the mean is considered.

•Formula for the empirical variance from a sample of n elements:

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

•The empircal standard deviation s is just the square root of the variance,

$$s = \sqrt{s^2}$$

Why divide by $n - 1$ instead of n?

Example:

x_1
x_2
$\vdots$
x_n
n
$\bar{x}$

$$
\begin{aligned}
x_1 &= 75 \\
x_2 &= 2 \\
x_3 &= 270 \\
x_4 &= 4 \cdot 100 - 75 - 2 - 270 = 53 \\
n &= 4 \\
\bar{x} &= 100
\end{aligned}
$$

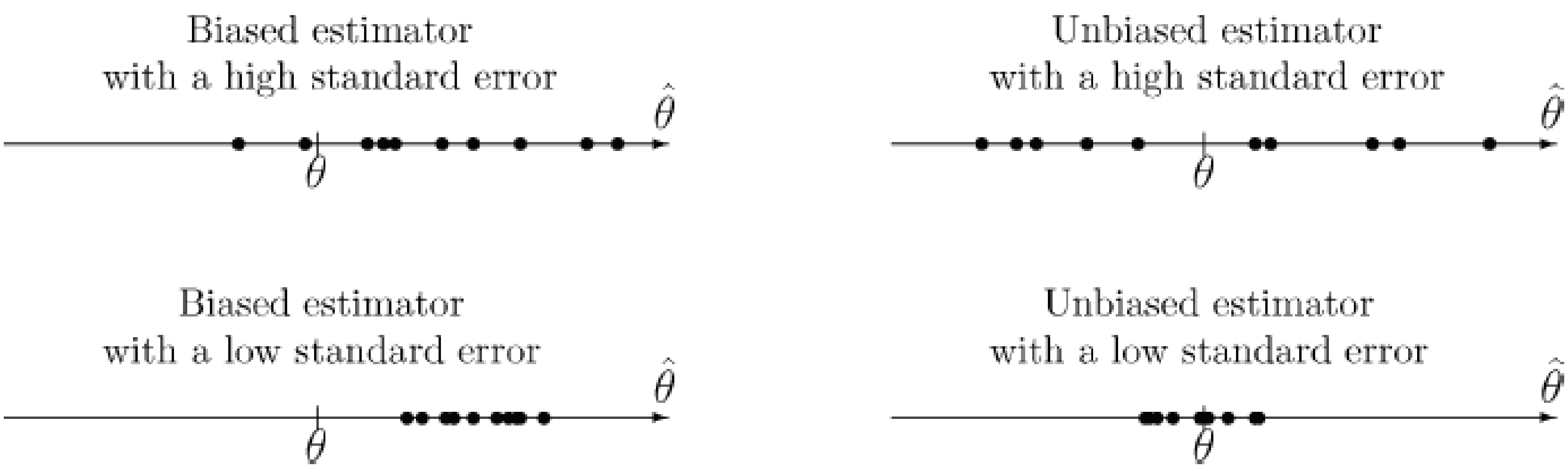
x_4 =53 is not free,
but given by other values when the mean is known.

s^2 has (n-1) degrees of freedom (f)

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

Standard errors of estimates

- For an estimator T for parameter θ, its standard error is Std(T), and it indicates the precision and reliability of T



Example: interval sample

- Calculate mean, variance, median for the sample:

Interval	[–2; 0)	[0; 4)	[4; 6)	[6; 10]
m_i	5	10	20	15

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{d} y_i^* m_i = \frac{1}{50}\left(-1\cdot 5 + 2\cdot 10 + 5\cdot 20 + 8\cdot 15\right) = 4,7.$$

$$M_2 = \frac{1}{n}\sum_{i=1}^{d}\left(y_i^*\right)^2 m_i = \frac{1}{50}\left(\left(-1\right)^2\cdot 5 + 2^2\cdot 10 + 5^2\cdot 20 + 8^2\cdot 15\right) = 30,1.$$

$$S^2 = M_2 - \left(\overline{x}\right)^2 = 30,1 - \left(4,7\right)^2 = 8,01.$$

$$\hat{S}^2 = \frac{n}{n-1}S^2 = \frac{50}{49}\cdot 8,01 \approx 8,17.$$

$$Me = 4 + 2\cdot\frac{25-15}{20} = 5.$$

Example: Discrete case – 1

Calculate mean, variance, median, histogram, empirical distribution function for the sample:

y_i	0	1	2	3	4	5	7
m_i	8	17	16	10	6	2	1

$$\overline{x} = \frac{1}{8+17+16+10+6+2+1} \cdot (0\cdot 8 + 1\cdot 17 + 2\cdot 16 + 3\cdot 10 + 4\cdot 6 + 5\cdot 2 + 7\cdot 1) = \frac{1}{60}\cdot 120 = 2$$

$$M_2 = \frac{1}{60}\cdot (0^2\cdot 8 + 1^2\cdot 17 + 2^2\cdot 16 + 3^2\cdot 10 + 4^2\cdot 6 + 5^2\cdot 2 + 7^2\cdot 1) = \frac{1}{60}\cdot 366 = 6.1$$

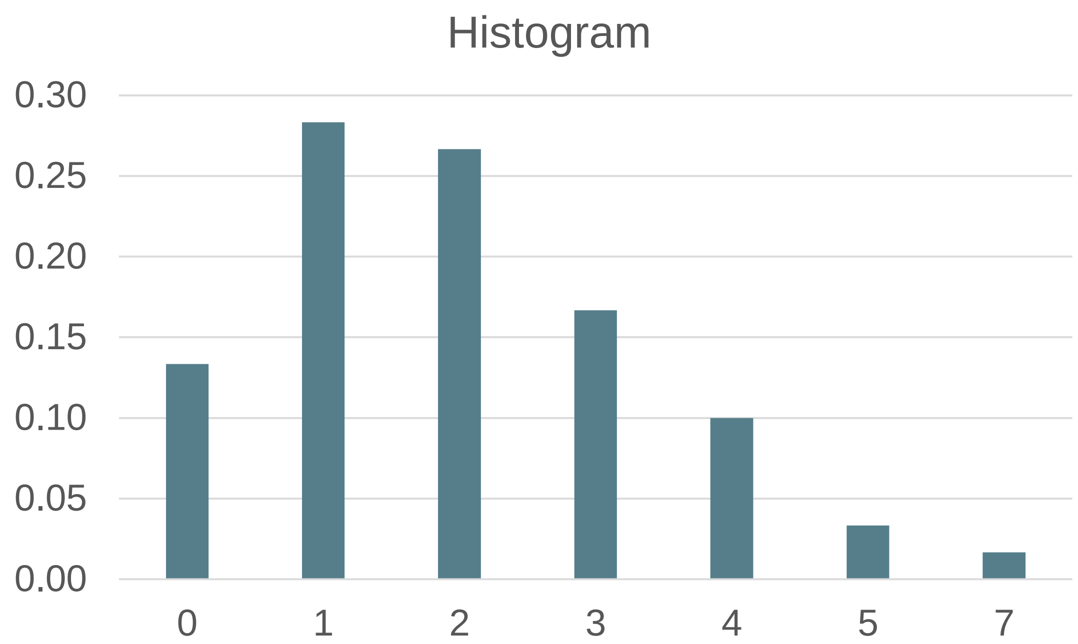
$$S^2 = M_2 - \overline{x}^2 = 6.1 - 2^2 = 2.1$$

$$\hat{S}^2 = \frac{n}{n-1} S^2 = \frac{60}{59}\cdot 2.1 = 2.14$$

Example: Discrete case – 2

Calculate mean, variance, median, histogram, empirical distribution function for the sample:

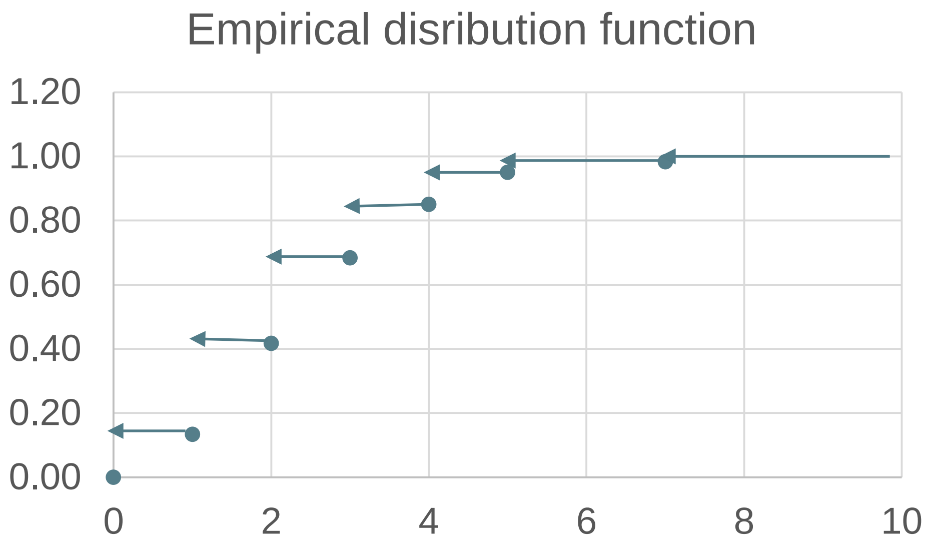
y_i	0	1	2	3	4	5	7
m_i	8	17	16	10	6	2	1



Example: Discrete case – 3

Calculate mean, variance, median, histogram, empirical distribution function for the sample:

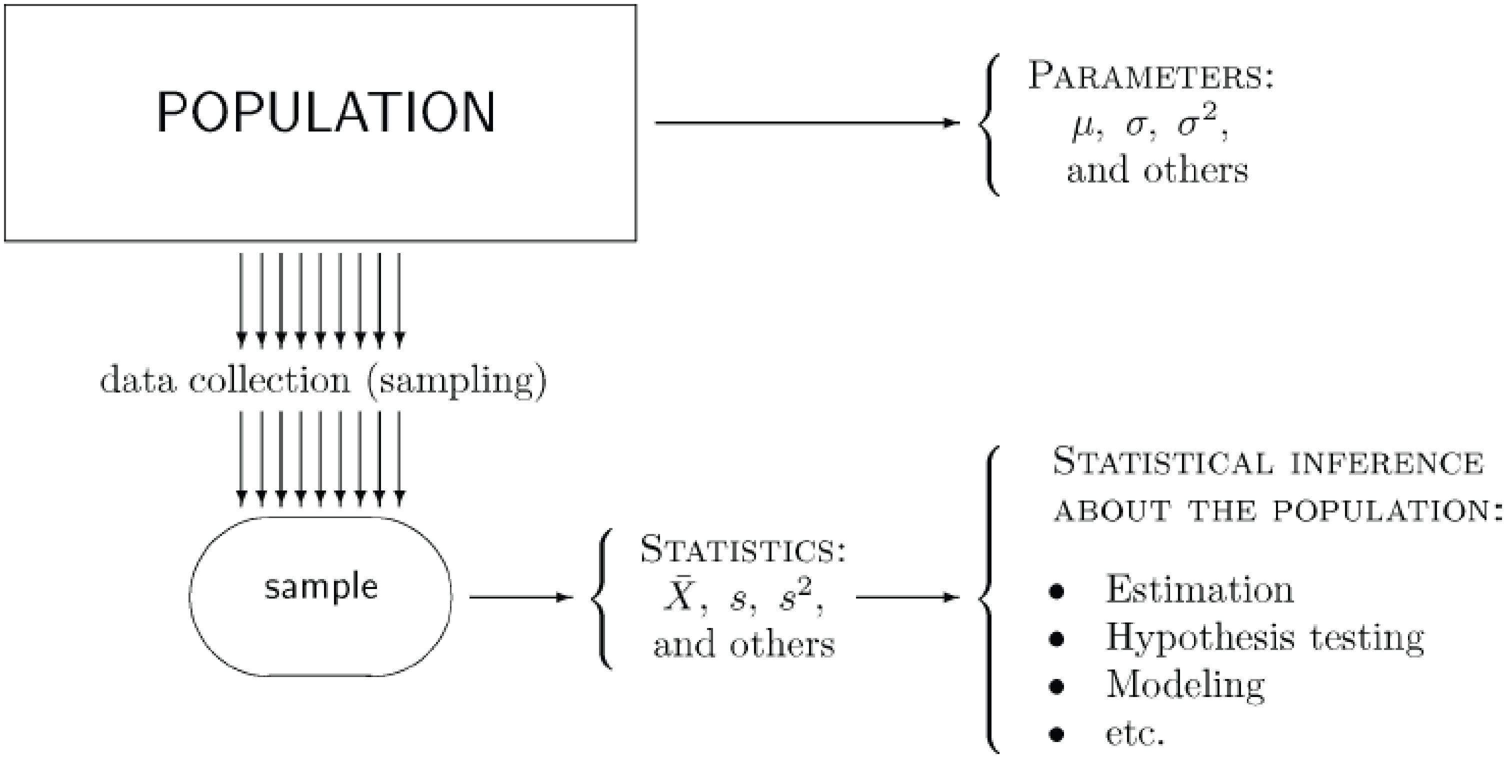
y_i	0	1	2	3	4	5	7
m_i	8	17	16	10	6	2	1



REVIEW

Population and sample

Population parameters and sample statistics.



Descriptive statistics

- Most important:
  - Mean
  - Median
  - Variance
  - Standard Deviation
  - Range
  - Minimum
  - Maximum

Thank you for attention!

KSE