Descriptive statistics

Aggregated data

COVID-19 DEATHS IN HEALTHCARE OPERATORS

TABELLA 7. DISTRIBUZIONE DI CASI, DECESSI E LETALITÀ NEGLI OPERATORI SANITARI

Classe di età (anni)	Casi		Deceduti		Letalità (%)
	N	%	N	%	
18-29	3.800	11,8	0	NA	0%
30-39	5.744	17,8	2	1,8	0%
40-49	8.880	27,5	4	3,6	0%
50-59	10.230	31,6	23	20,5	0,2%
60-69	3.325	10,3	51	45,5	1,5%
70-79	185	0,6	16	14,3	8,6%
Età non nota	173	0,5	16	14,3	9,2%
Totale	32.337	NA	112	NA	0,3%

Prodotto dall'Istituto Superiore di Sanità (ISS), Roma, 29 settembre 2020

COVID-19 DEATHS IN HEALTHCARE OPERATORS

		FREQUENCIES	
AGE CLASS	**AGE CLASS**	f	p%
[18,30)	18-	0	0.0
[30,40)	30-	2	2.1
[40,50)	40-	4	4.2
[50,60)	50-	23	24.0
[60,70)	60-	51	53.1
[70,80)	70-	16	16.7
	Tot	96	

Prodotto dall'Istituto Superiore di Sanità (ISS), Roma, 29 settembre 2020

* Esclusi 16 casi di età non nota

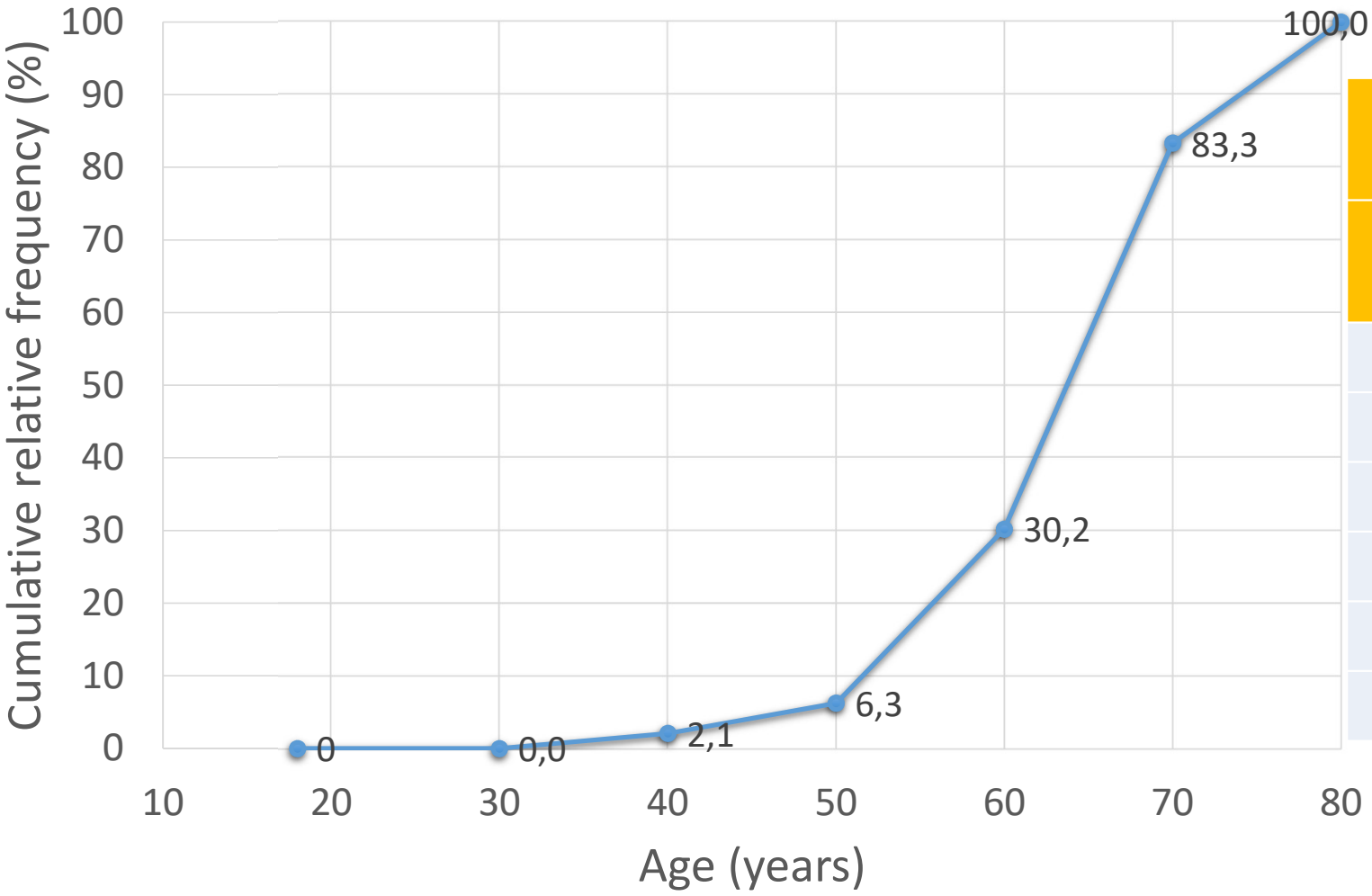
COVID-19 DEATHS IN HEALTHCARE OPERATORS

		FREQUENCIES		CUMULATIVE FREQUENCIES	
AGE CLASS	**AGE CLASS**	f	p%	F	P%
[18,30)	18-	0	0.0	0	0.0
[30,40)	30-	2	2.1	2	2.1
[40,50)	40-	4	4.2	6	6.3
[50,60)	50-	23	24.0	29	30.2
[60,70)	60-	51	53.1	80	83.3
[70,80)	70-	16	16.7	96	100.0
	Tot	96			

Prodotto dall'Istituto Superiore di Sanità (ISS), Roma, 29 settembre 2020

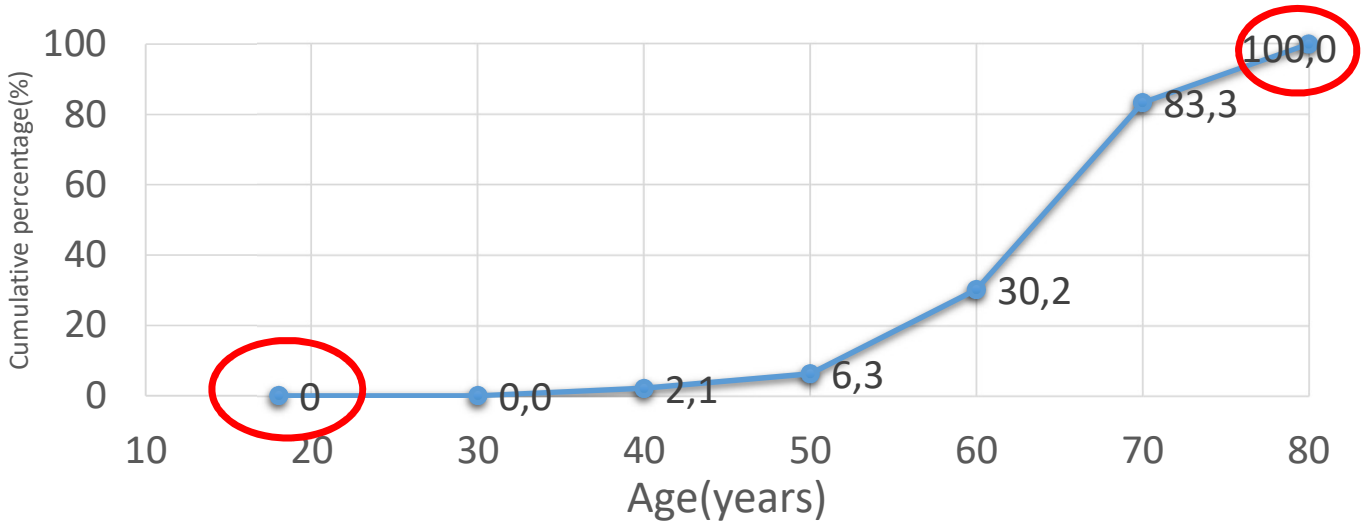
* Esclusi 16 casi di età non nota

Cumulative frequency graph



		Cumulative frequency
AGE CLASS	Upper extreme	P%
[18,30)	30	0.0
[30,40)	40	2.1
[40,50)	50	6.3
[50,60)	60	30.2
[60,70)	70	83.3
[70,80)	80	100.0

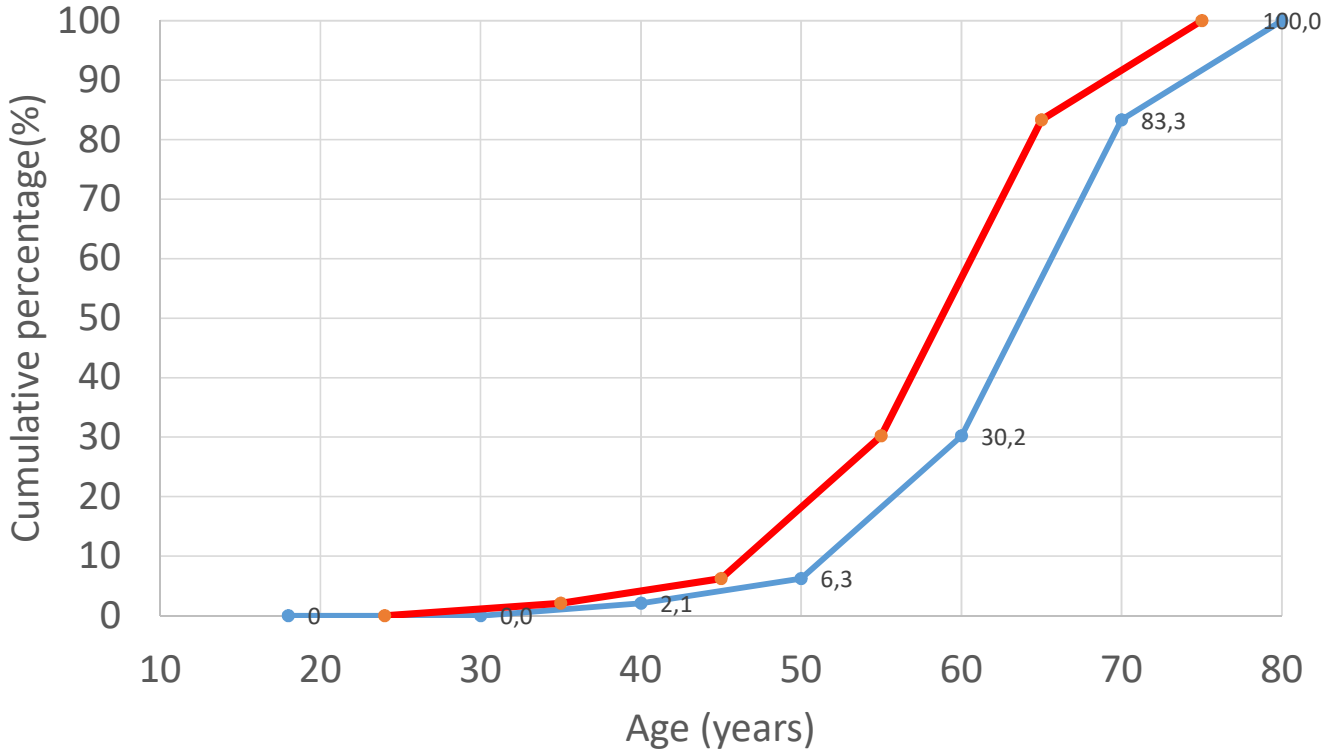
Cumulative frequency graph



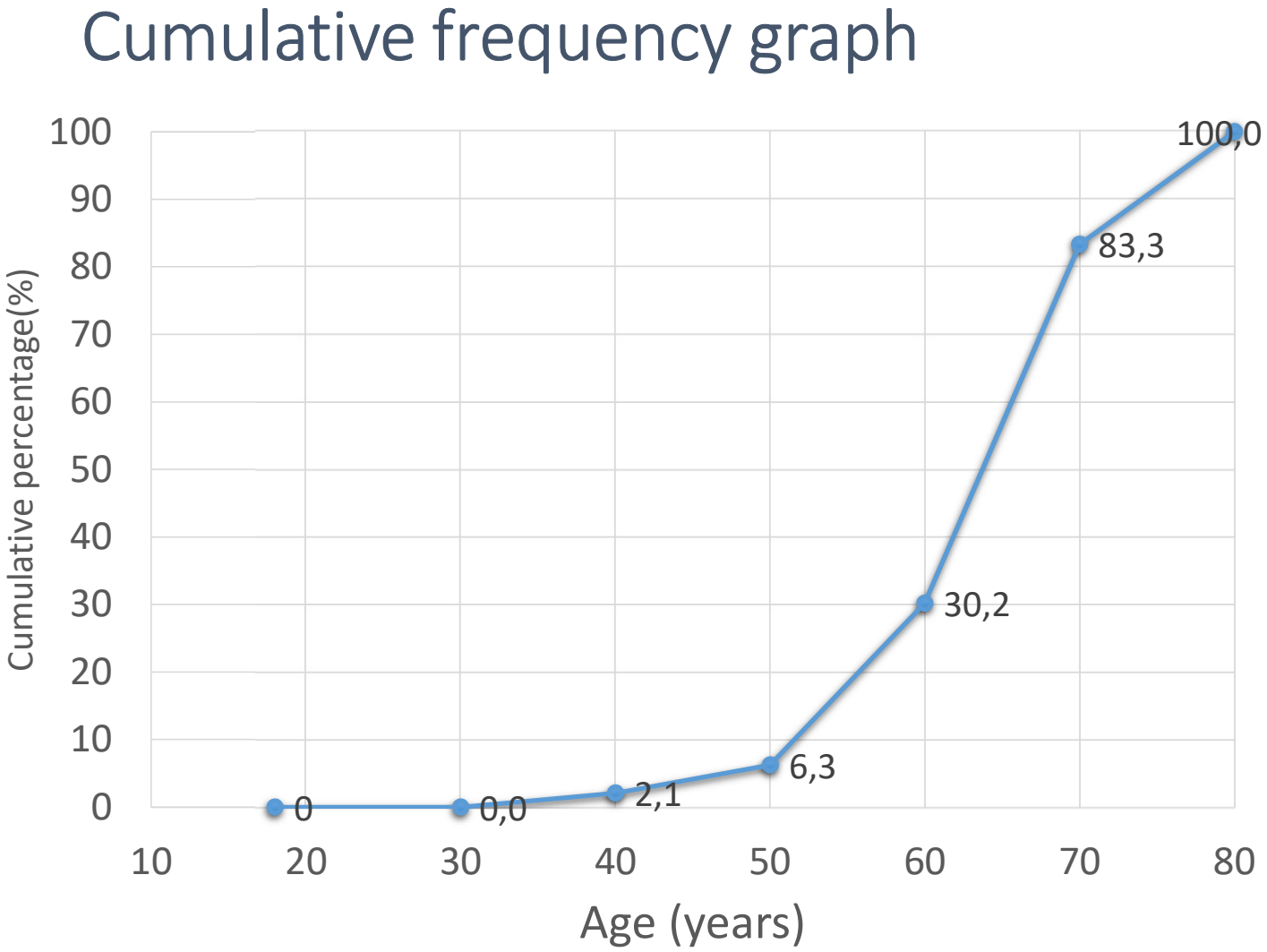
- ✓ The broken line starts at 0 and ends at 1 or 100%.
- ✓ The broken line is obtained by joining with segments the two points whose coordinates are :

  [lower bound, previous cumulative frequency] [upper bound, cumulative frequency]
- ✓ The distribution of data in the classes is assumed to be uniform (linear interpolation)

Cumulative frequency graph

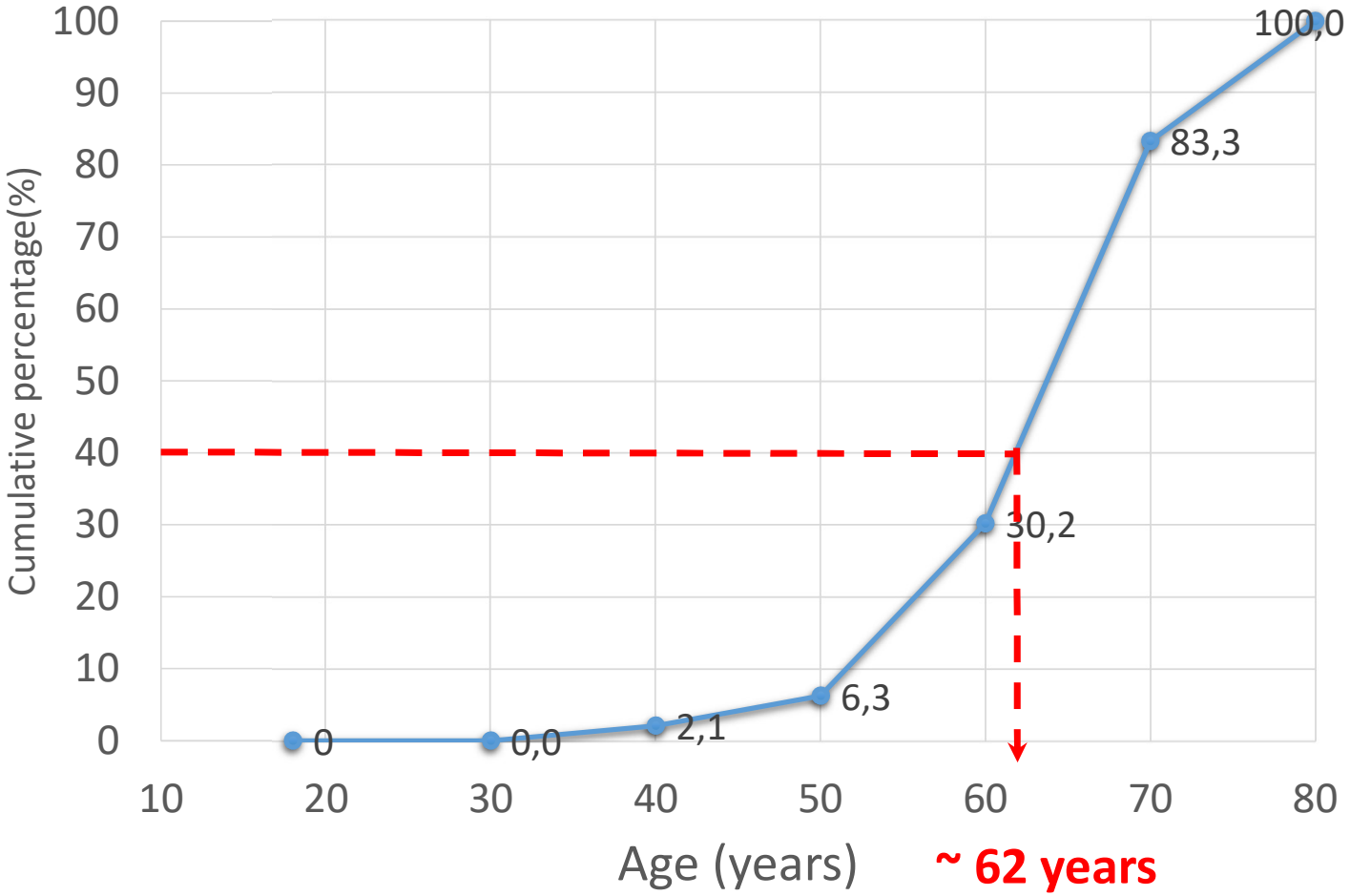


If the central values were joined together, an incorrect representation would be obtained.



What is the age value below which I find 40% of deaths?

Cumulative frequency graph



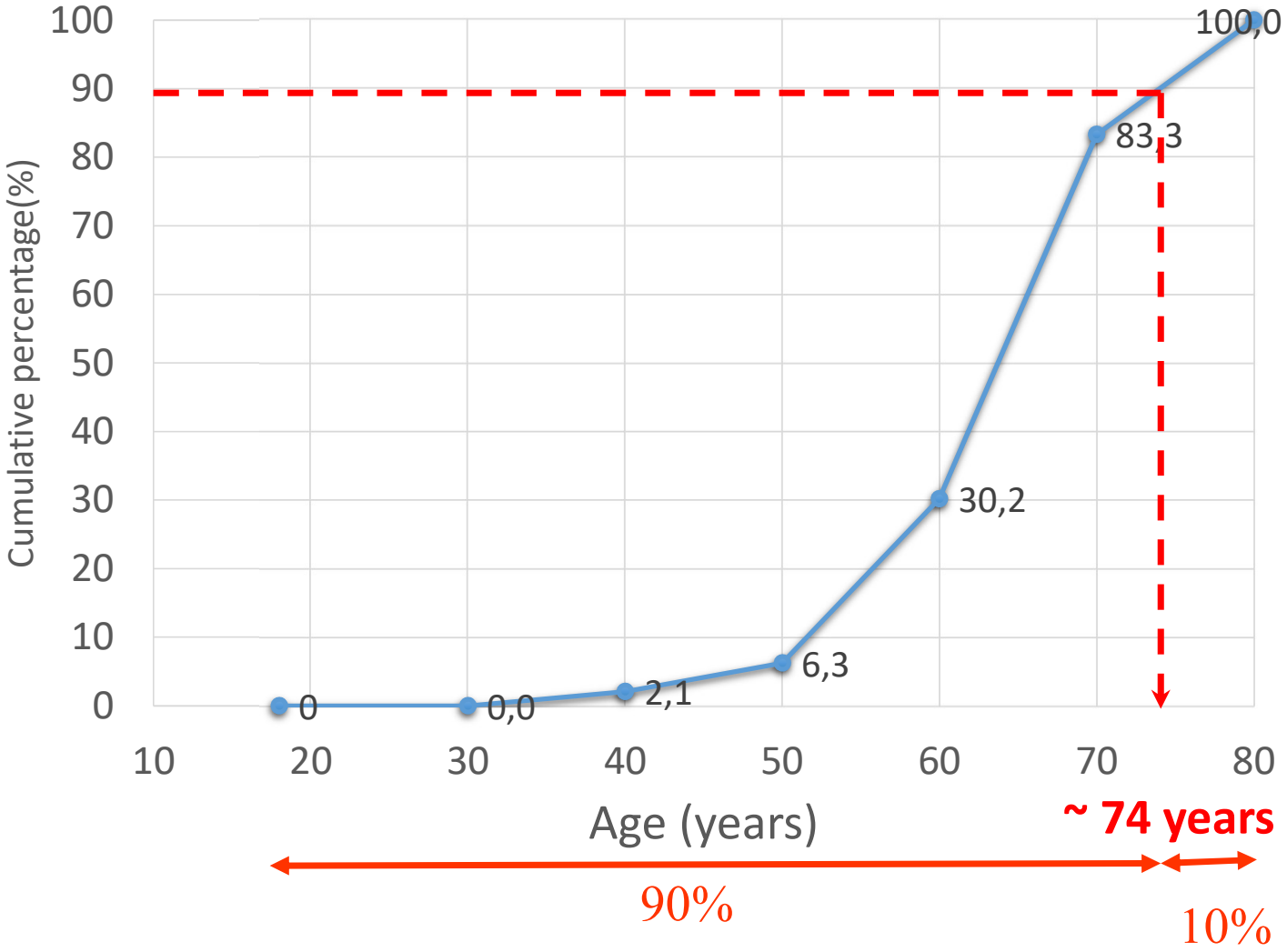
What is the age value below which I find 40% of deaths?

Percentiles from cumulative frequencies

Es. p=0.90

90th percentile

$x_{0.9}$~ 74 years



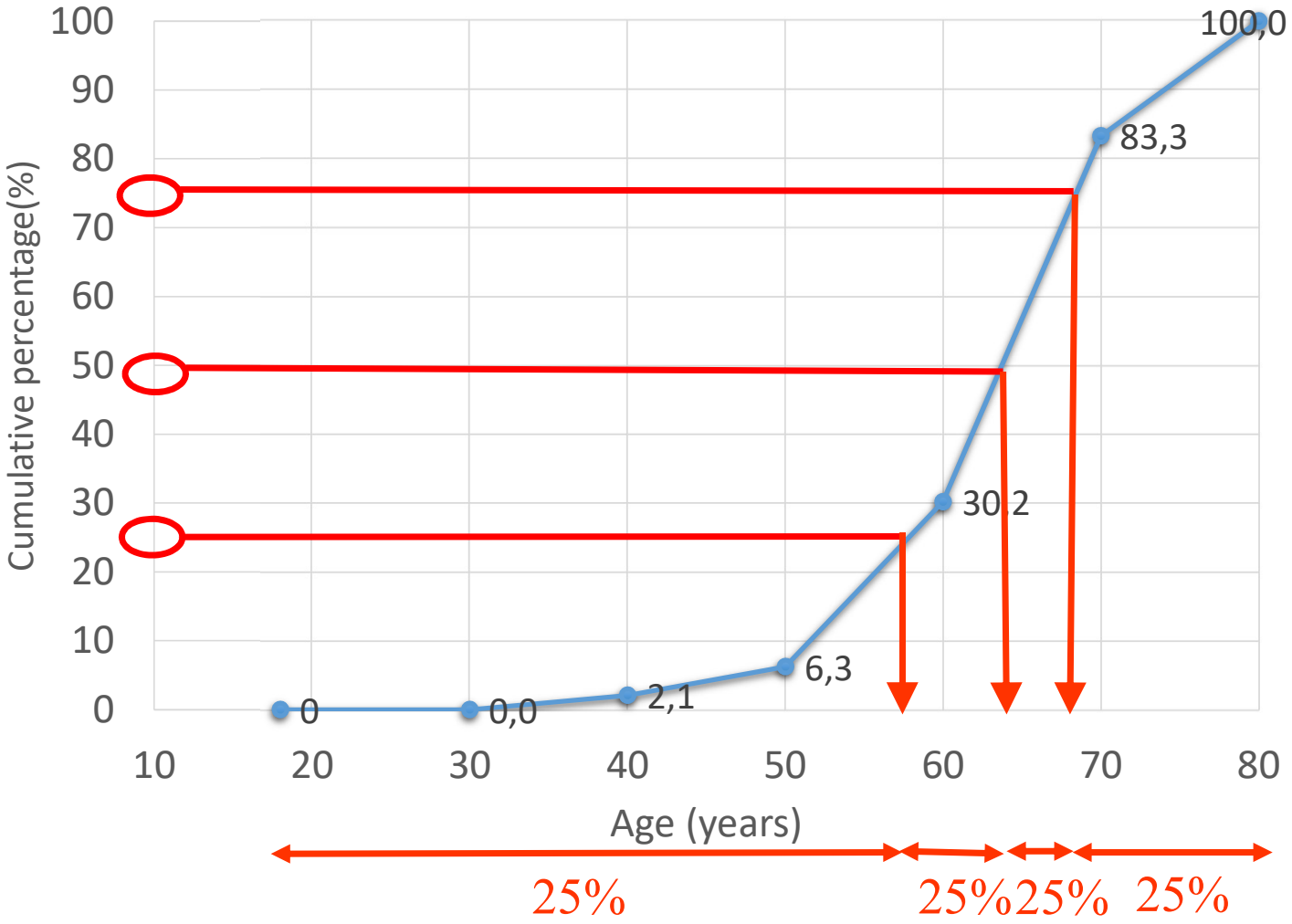
Particular percentiles : quartiles

Quartiles:

p=0.25, $x_{0.25}$~ 58

p=0.50 , $x_{0.5}$~ 64

p=0.75, $x_{0.75}$~ 68

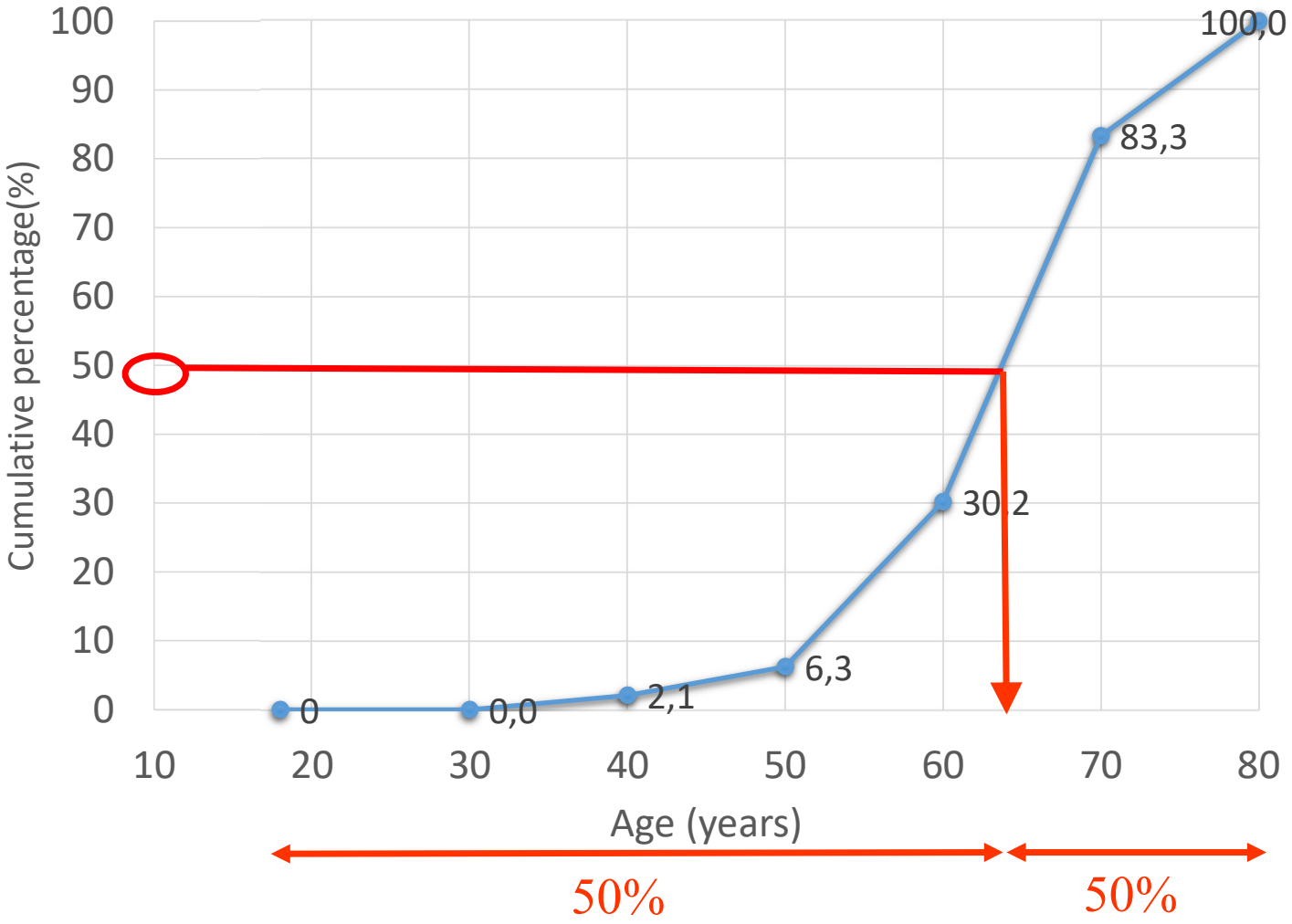


Quartiles: divide the data into four equal parts (25%)

Particular percentiles : median

Median

p=0.50 , $x_{0.5}$~ 64



Median from frequency table

		Simple frequencies		Cumulative frequencies	
Age class	Age class	f	p%	F	P%
[18,30)	18-	0	0.0	0	0.0
[30,40)	30-	2	2.1	2	2.1
[40,50)	40-	4	4.2	6	6.3
[50,60)	50-	23	24.0	29	30.2
[60,70)	60-	51	53.1	80	83.3
[70,80)	70-	16	16.7	96	100.0
	Tot	96			

Median class [60, 70)

Prodotto dall'Istituto Superiore di Sanità (ISS), Roma, 29 settembre 2020

* Esclusi 16 casi di età non nota

Arithmetic mean

		Simple frequencies		Cumulative frequencies	
Age class	Central value ($_cx_i$)	f	p%	F	P%
[18,30)	**24**	0	0.0	0	0.0
[30,40)	**35**	2	2.1	2	2.1
[40,50)	**45**	4	4.2	6	6.3
[50,60)	**55**	23	24.0	29	30.2
[60,70)	**65**	51	53.1	80	83.3
[70,80)	**75**	16	16.7	96	100.0
Tot		96			

To calculate the average it is necessary to consider its central value as the representative value of each class $_cx_i$

Prodotto dall'Istituto Superiore di Sanità (ISS), Roma, 29 settembre 2020

* Esclusi 16 casi di età non nota

Mean from aggregated data

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{n} = \frac{24 \cdot 0 + 35 \cdot 2 + \cdots . + 75 \cdot 16}{96} = 62.8$$

$$\bar{x} = \sum_{i=1}^{k} x_i \cdot p_i = \ldots$$

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot p_i \%}{100} = \ldots$$

The average calculated on the data grouped into classes represents an approximation of the one determined from the single data.

Mean from aggregated data

		Simple frequencies			
Age class	**Central value ($_cx_i$)**	f	p%	$_cx_i$*f	
[18,30)	24	0	0.0	0	
[30,40)	35	2	2.1	70	
[40,50)	45	4	4.2	180	
[50,60)	55	23	24.0	1265	
[60,70)	65	51	53.1	3315	
[70,80)	75	16	16.7	1200	
Sum		96		6030	

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot f_i}{n} = \frac{6030}{96} = 62.8 \text{ years}$$

Standard deviation from aggregated data

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

$$=\sqrt{\frac{(24-62.8)^2\cdot 0+(35-62.8)^2\cdot 2+\cdots.+(75-62.8)^2\cdot 16}{96-1}}=$$

=8.5 years

Standard deviation from aggregated data

		Simple frequencies			
Age class	**Central value ($_cx_i$)**	f	p%	$(_cx_i-\bar{x})^2$	$f*(_cx_i-\bar{x})^2$
[18,30)	24	0	0.0	1506	0
[30,40)	35	2	2.1	773.5	1547
[40,50)	45	4	4.2	317.3	1269
[50,60)	55	23	24.0	61.04	1404
[60,70)	65	51	53.1	4.785	244
[70,80)	75	16	16.7	148.5	2377
Somma		96			6841

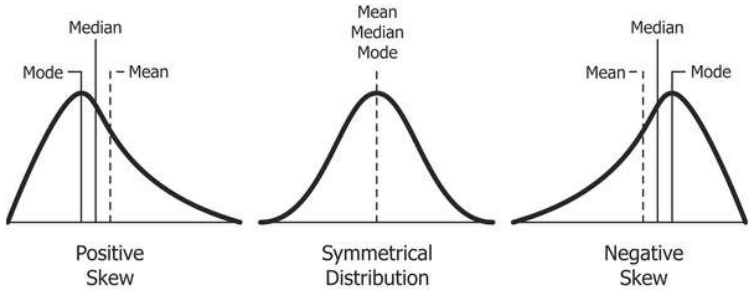
$$s=\sqrt{\frac{(24-62.8)^2\cdot 0+(35-62.8)^2\cdot 2+\cdots.+(75-62.8)^2\cdot 16}{96-1}}=\sqrt{\frac{6841}{95}}=8.5$$

Exercise: represent the data through a histogram

Shape of data is measured by

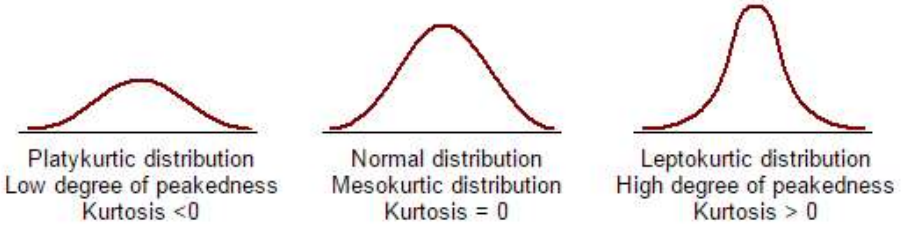
- ✓ Skewness

  Positive or right skewed: Longer right tail
  Negative or left skewed: Longer left tail



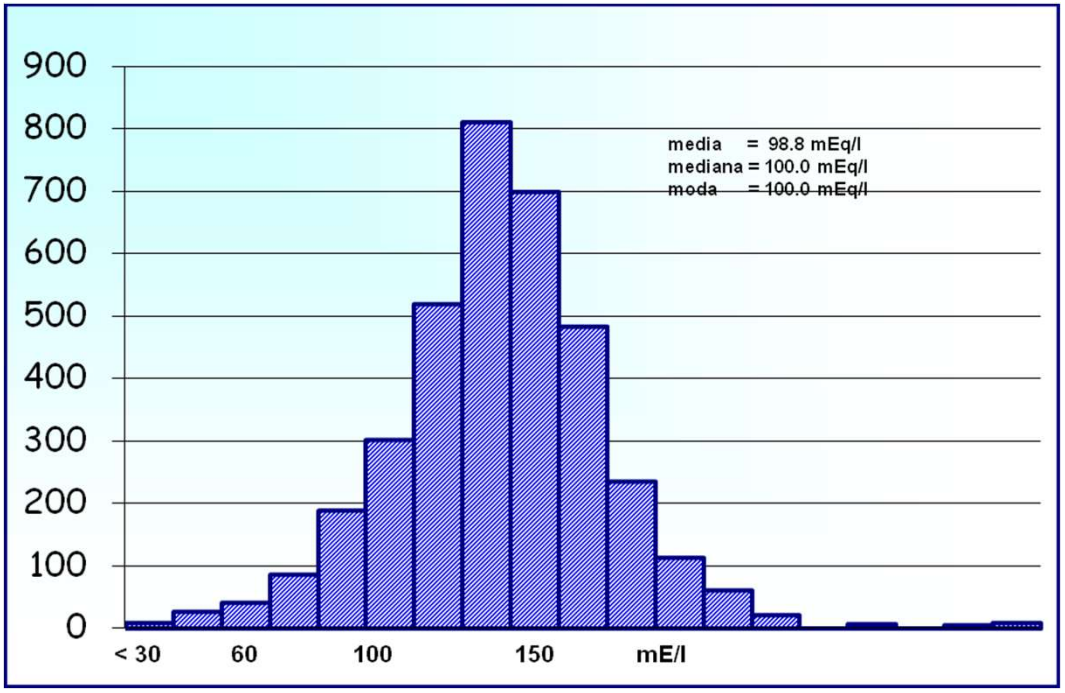
- ✓ Kurtosis

  Measures peakedness of the distribution of data.
  The kurtosis of normal distribution is 0



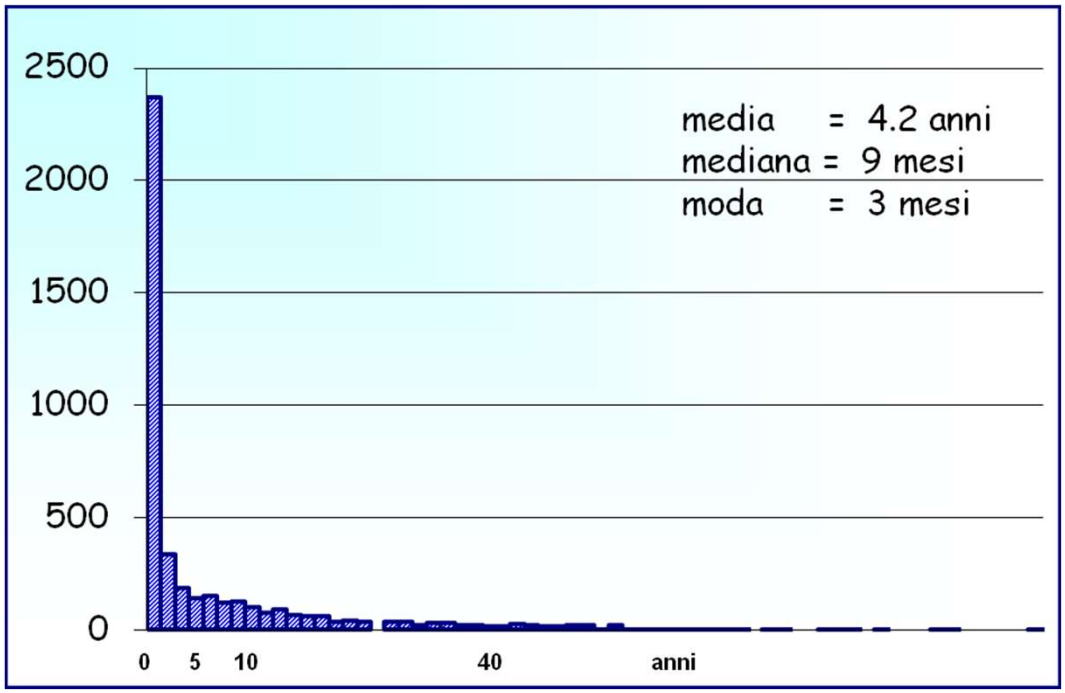
Example:

Chlorine concentration in sweat (symmetrical)



Example:

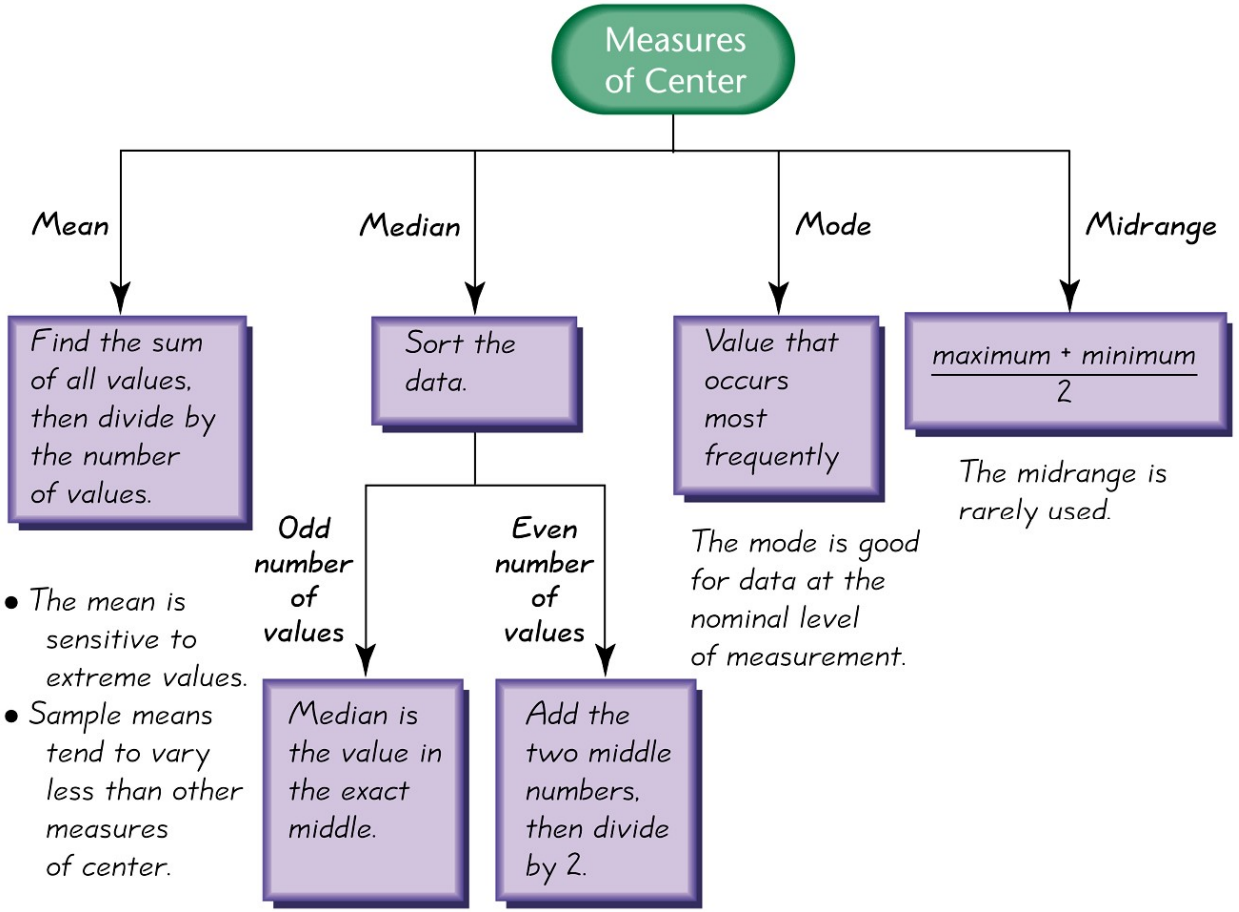
Age at diagnosis in cystic fibrosis (positive asimmetry)



Summary indicators (position and variability)

	Nominal	Ordinal	Discrete	Continuous
Modal value	■	■	■	■
Mean			■	■
Median		■	■	■
Standard Deviation			■	■
Interquartile range		■	■	■
Range		■	■	■

Best Measure of Center



Example

Effect of n-3 polyunsaturated fatty acids in patients with chronic heart failure (the GISSI-HF trial): a randomised, double-blind, placebo-controlled trial

*GISSI-HF investigators**

Summary

Background Several epidemiological and experimental studies suggest that n-3 polyunsaturated fatty acids (PUFA) can exert favourable effects on atherothrombotic cardiovascular disease, including arrhythmias. We investigated whether n-3 PUFA could improve morbidity and mortality in a large population of patients with symptomatic heart failure of any cause.

	n-3 PUFA (n=3494)	Placebo (n=3481)
Patients' characteristics		
Age (years)	67 (11)	67 (11)
Age >70 years	1465 (41·9%)	1482 (42·6%)
Women	777 (22·2%)	739 (21·2%)
Heart disease risk factors		
BMI (kg/m²)	27 (5)	27 (5)
SBP (mm Hg)	126 (18)	126 (18)
DBP (mm Hg)	77 (10)	77 (10)
Heart rate (beats per min)	72 (13)	73 (14)
Current smoking	502 (14·4%)	485 (13·9%)
History of hypertension	1886 (54·0%)	1923 (55·2%)
NYHA class		
II	2226 (63·7%)	2199 (63·2%)
III	1178 (33·7%)	1187 (34·1%)
IV	90 (2·6%)	95 (2·7%)
LVEF (%)	33·0% (8·5)	33·2% (8·5)
LVEF >40%	333 (9·5%)	320 (9·2%)
Medical history		
Admission for HF in previous year	1746 (50·0%)	1638 (47·1%)
Previous AMI	1461 (41·8%)	1448 (41·6%)

Data are mean (SD) or number (%). PUFA=polyunsaturated fatty acids. BMI=body-mass index. SBP=systolic blood pressure. DBP=diastolic blood pressure. NYHA=New York Heart Association. LVEF=left ventricular ejection fraction. HF=heart failure. AMI=acute myocardial infarction. CABG=coronary artery bypass graft. PCI=percutaneous coronary intervention. ICD=implantable cardioverter defi brillator. COPD=chronic obstructive pulmonary disease. ACE=angiotensin-converting enzyme. ARBs=angiotensin receptor blockers. *Available for 6899 patients (3455 n-3 PUFA, 3444 placebo).

Mean and standard deviation are summary indexes of location and variability

The 68% of observation lie within 1 standard deviation (s) $[\bar{x} - s \ , \ \bar{x} + s]$

The 95% of observation lie within 2 s $[\bar{x} - 2 \cdot s \ , \ \bar{x} + 2 \cdot s]$

The 99% of observation lie within 1 s $[\bar{x} - 3 \cdot s \ , \ \bar{x} + 3 \cdot s]$

They are suitable only for representing symmetric distributions (with approximately normal shape)

They allow comparison between phenomena in the same unit of measurement and with the same order of magnitude

Exercise:

Compare variability between BMI and blood pressure (SBP and DBP) in the GISSI-prevention trial in the n-3PUFA arm

	n-3 PUFA (n=3494)	Placebo (n=3481)
Patients' characteristics		
Age (years)	67 (11)	67 (11)
Age >70 years	1465 (41·9%)	1482 (42·6%)
Women	777 (22·2%)	739 (21·2%)
Heart disease risk factors		
BMI (kg/m²)	27 (5)	27 (5)
SBP (mm Hg)	126 (18)	126 (18)
DBP (mm Hg)	77 (10)	77 (10)
Heart rate (beats per min)	72 (13)	73 (14)
Current smoking	502 (14·4%)	485 (13·9%)
History of hypertension	1886 (54·0%)	1923 (55·2%)
NYHA class		
II	2226 (63·7%)	2199 (63·2%)
III	1178 (33·7%)	1187 (34·1%)
IV	90 (2·6%)	95 (2·7%)
LVEF (%)	33·0% (8·5)	33·2% (8·5)
LVEF >40%	333 (9·5%)	320 (9·2%)
Medical history		

Coefficient of variation (CV)

We have seen some empirical methods for:

- get an idea of the trend and distribution of a phenomenon
- compare phenomena in the same unit of measurement

How to compare the variability of phenomena expressed in the same unit of measurement but with different orders of magnitude?

How to compare the variability of different phenomena?

Coefficient of variation (CV)
or relative standard deviation

The coefficient of variation (CV) is the ratio of the standard deviation to the arithmetic mean:

$$CV = \frac{s}{\overline{x}}$$

It is a pure number that can assume positive or negative values depending on the sign of the mean.

- the unit of measurement is eliminated
- the variability is standardized for the order of magnitude of the phenomenon

CV : example

Same phenomenon (income) in groups with different order of magnitude (blue-collar workers and billionaires)

Income (thousand of €)			
A		B	
1	20	**1**	800020
2	60	**2**	800060

$\bar{x}_{bc} = 40\ mila\ \ euro$

$\bar{x}_{B} = 800040\ mila\ \ euro$

$s_{bc} = s_{b} = 28.3\ mila\ euro$

$$CV_{A} = \frac{28.3}{40} = 0.71 \qquad CV_{B} = \frac{28.3}{800040} = 0.00004$$

The variability of the income of B is negligible compared to the order of magnitude of the phenomenon.

Exercise:

Compare variability between BMI and blood pressure (SBP and DBP) in the GISSI-prevention trial in the n-3PUFA arm

	n-3 PUFA (n=3494)	Placebo (n=3481)
Patients' characteristics		
Age (years)	67 (11)	67 (11)
Age >70 years	1465 (41·9%)	1482 (42·6%)
Women	777 (22·2%)	739 (21·2%)
Heart disease risk factors		
BMI (kg/m²)	27 (5)	27 (5)
SBP (mm Hg)	126 (18)	126 (18)
DBP (mm Hg)	77 (10)	77 (10)
Heart rate (beats per min)	72 (13)	73 (14)
Current smoking	502 (14·4%)	485 (13·9%)
History of hypertension	1886 (54·0%)	1923 (55·2%)
NYHA class		
II	2226 (63·7%)	2199 (63·2%)
III	1178 (33·7%)	1187 (34·1%)
IV	90 (2·6%)	95 (2·7%)
LVEF (%)	33·0% (8·5)	33·2% (8·5)
LVEF >40%	333 (9·5%)	320 (9·2%)
Medical history		

$$CV(BMI) = \frac{5}{27} = 0.1852$$

$$CV(SBP) = \frac{18}{126} = 0.1429$$

$$CV(DBP) = \frac{10}{77} = 0.1299$$

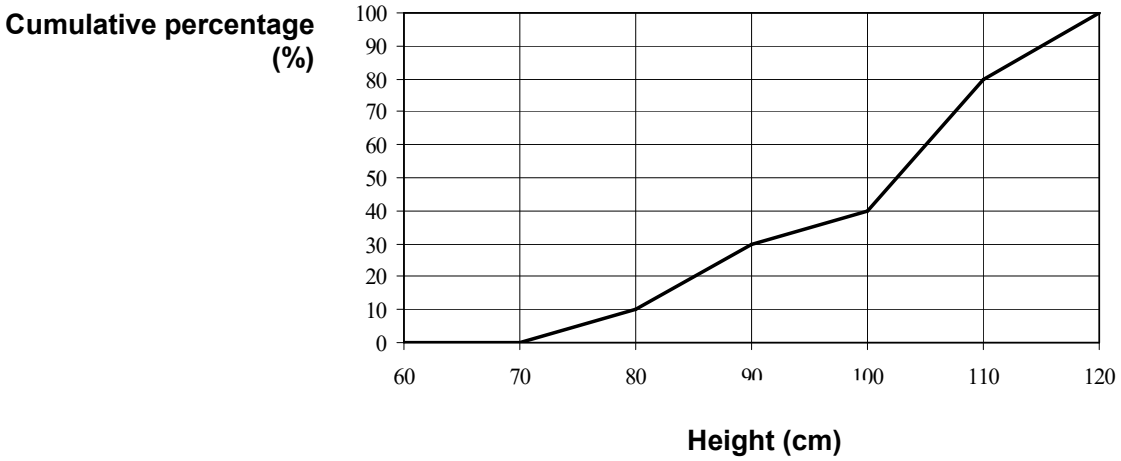
Exercise 1 and 2 from the course web-page for next time

Exercise 1 (given the size, we suggest using Excel or other software to calculate the mean and variance)
Use data on covid-19 cases in health care professionals from the ISS.

Age class	Frequency (f)
[18,30)	3800
[30,40)	5744
[40,50)	8880
[50,60)	10230
[60,70)	3325
[70,80)	185
Sum	32164

a) Calculate the cumulative relative frequencies
b) Represent the data with cumulative relative frequencies
c) Identify the quartiles
d) Calculate the mean and standard deviation
e) Identify the modal class

Exercise 2 (by hand with calculator)
The cumulative percentage distribution of height of preschool children (cm), measured in a sample of 200 children, is reported below. On the horizontal axis are reported 5 classes of height on which data have been aggregated:



On data of this sample, identify:
a) the modal class or modal classes;
b) the mean;
c) the variance;
d) report here quartile values of this distribution (identified by the graph)
e) compute the coefficient of variation (CV)