COLLECTING AND DISPLAING DATA

Luigi Greco, M.D., Ph.D. M.Sc. & Francesco Giallauria, M.D., Ph.D.

Department of Translational Medical Sciences

University of Naples Federico II

WHAT WE AIM TO PROPOSE

1. **To set up a nice collection of clinical or experimental data of any kind**
2. **To distinguish the type of data (variables)**
3. To draw a simple ‘coded’ data collection form
4. To manage ‘missing’ data
5. To start to describe the collected data
6. To explore the distribution of data
7. To use few parameters to describe the data

Gulu, Uganda 2003

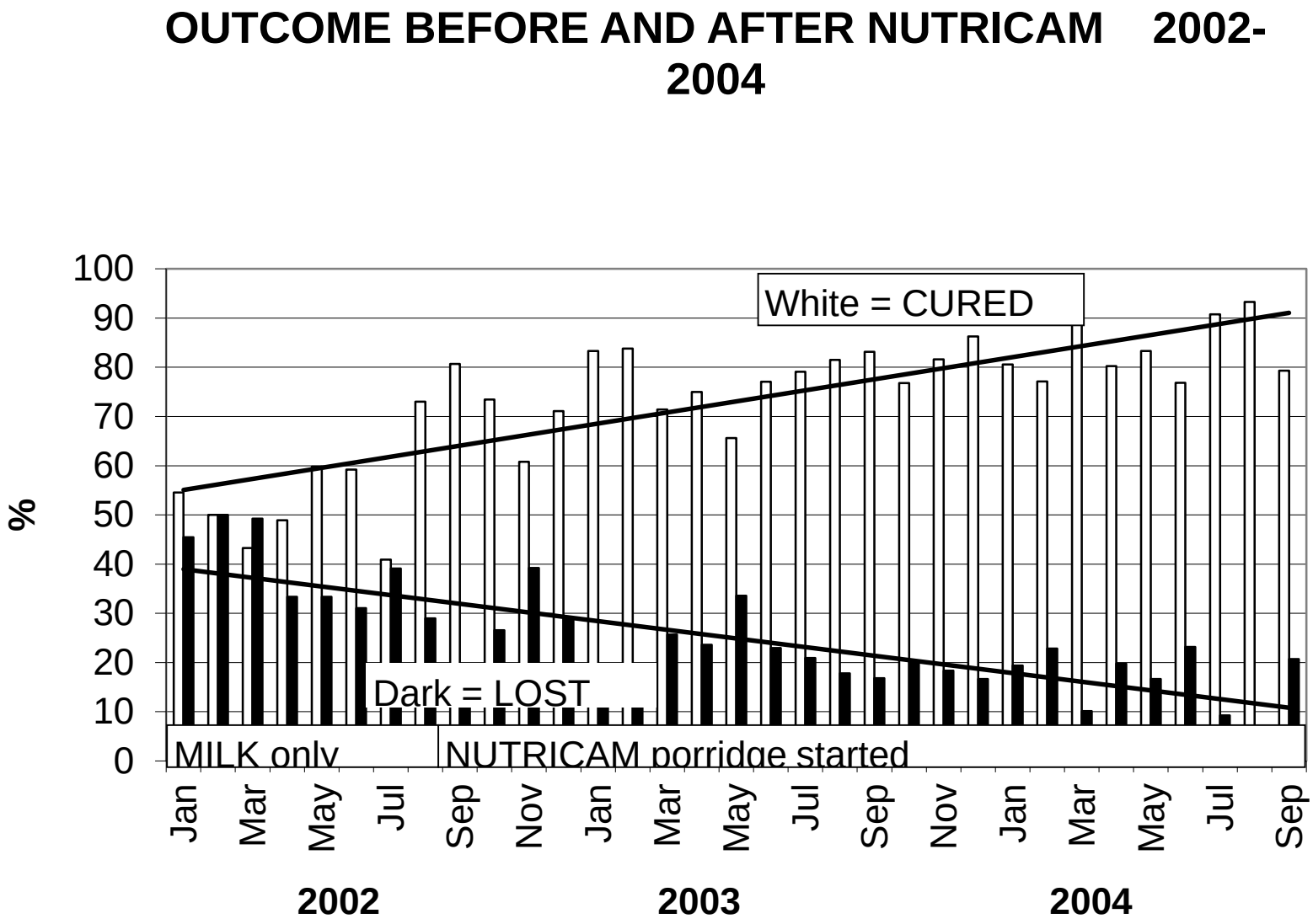
Ammessi, Curati, Morti, Persi alla NUTRITION UNIT

MONTH	CASES	CURED	DEATH	DEFAULT	LOST
Jan	66	36	14	16	30
Feb	70	35	16	19	35
Mar	67	29	13	20	33
Apr	90	44	17	13	30
May	102	61	18	16	34
Jun	103	61	14	18	32
Jul	110	45	14	29	43
Aug	100	73	13	16	29
Sep	114	92	11	9	20
Oct	64	47	13	4	17
Nov	125	76	26	23	49
Dec	128	91	18	19	37
Jan	114	95	14	5	19
Feb	74	62	6	6	12
Mar	70	50	11	7	18
Apr	72	54	12	5	17
May	125	82	14	28	42
Jun	170	131	21	18	39
Jul	129	102	10	17	27
Aug	135	110	11	13	24
Sep	95	79	9	7	16
Oct	69	53	5	9	14
Nov	98	80	7	11	18
Dec	102	88	7	10	17

Aggiungo le % di casi curati, morti, persi sul totale dei casi

MONTH	CASES	CURED	DEATH	DEFAULT	LOST	%cured	%death	%default	%lost
Jan	66	36	14	16	30	54,5	21,2	24,2	45,5
Feb	70	35	16	19	35	50,0	22,9	27,1	50,0
Mar	67	29	13	20	33	43,3	19,4	29,9	49,3
Apr	90	44	17	13	30	48,9	18,9	14,4	33,3
May	102	61	18	16	34	59,8	17,6	15,7	33,3
Jun	103	61	14	18	32	59,2	13,6	17,5	31,1
Jul	110	45	14	29	43	40,9	12,7	26,4	39,1
Aug	100	73	13	16	29	73,0	13,0	16,0	29,0
Sep	114	92	11	9	20	80,7	9,6	7,9	17,5
Oct	64	47	13	4	17	73,4	20,3	6,3	26,6
Nov	125	76	26	23	49	60,8	20,8	18,4	39,2
Dec	128	91	18	19	37	71,1	14,1	14,8	28,9
Jan	114	95	14	5	19	83,3	12,3	4,4	16,7
Feb	74	62	6	6	12	83,8	8,1	8,1	16,2
Mar	70	50	11	7	18	71,4	15,7	10,0	25,7
Apr	72	54	12	5	17	75,0	16,7	6,9	23,6
May	125	82	14	28	42	65,6	11,2	22,4	33,6
Jun	170	131	21	18	39	77,1	12,4	10,6	22,9
Jul	129	102	10	17	27	79,1	7,8	13,2	20,9
Aug	135	110	11	13	24	81,5	8,1	9,6	17,8
Sep	95	79	9	7	16	83,2	9,5	7,4	16,8
Oct	69	53	5	9	14	76,8	7,2	13,0	20,3
Nov	98	80	7	11	18	81,6	7,1	11,2	18,4
Dec	102	88	7	10	17	86,3	6,9	9,8	16,7

Un grafico che ha cambiato il destino di questi piccoli



Che cosa li ha curati? La Statistica ??

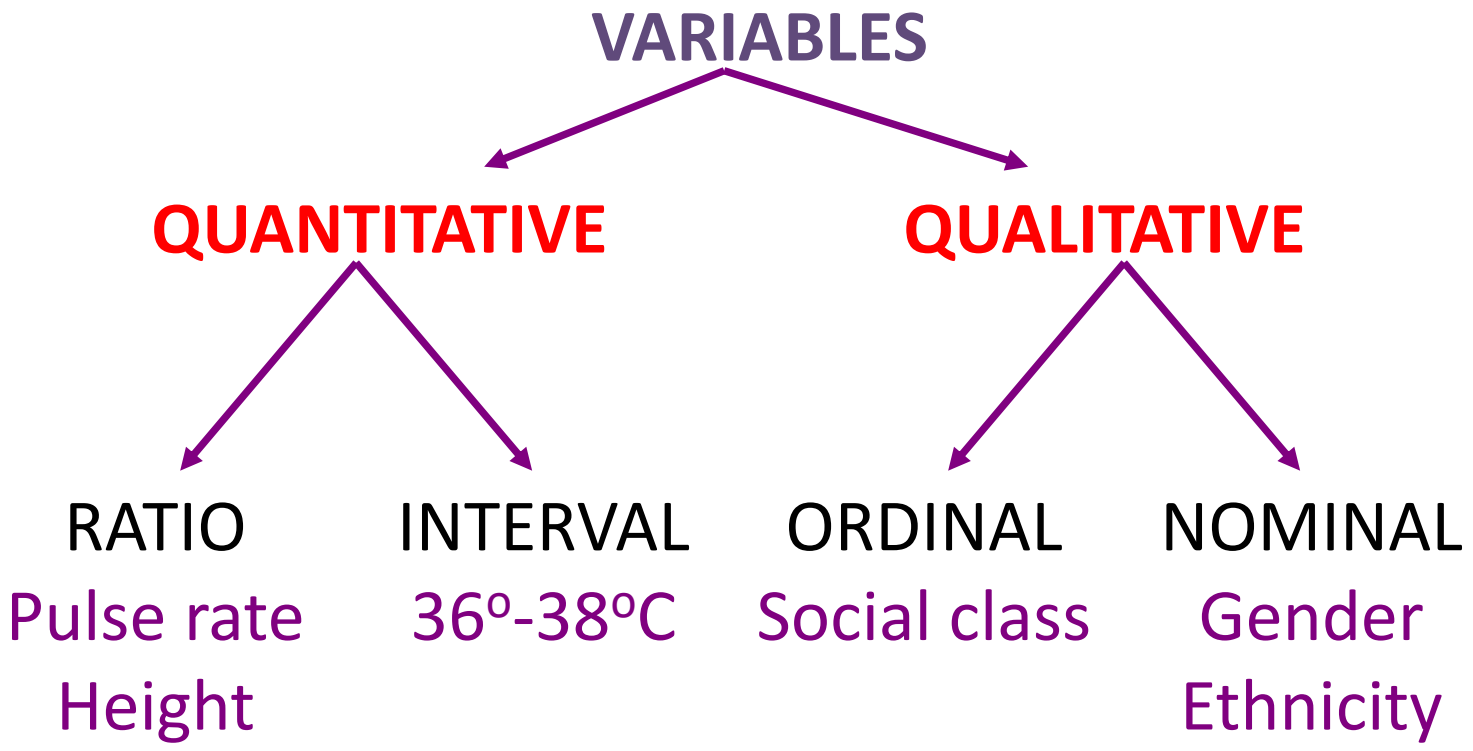
Types of data

In any statistical analysis the first step, before any calculations or plotting of data, is to decide what type of data one is dealing with. There are a number of typologies, but one that has proven useful is given in the following Table

The basic distinction is between ***quantitative*** variables (for which one asks "how much?") and ***categorical*** variables (for which one asks "what type?")

Examples of types of data	
Quantitative	
Continuous	**Discrete**
Blood pressure, height, weight, age	Number of children Number of attacks of asthma per week
Categorical	
Ordinal (Ordered categories)	**Nominal (Unordered categories)**
Grade of breast cancer Better, same, worse Disagree, neutral, agree	Sex (male/female) Alive or dead Blood group O, A, B, AB

TYPES OF DATA



Four kinds of variables (data): let us revise !

- **Continuous Variables** : from 0 to XXX
  - Age, Height, Weight, Blood Glucose, Blood Pressure

- **Discrete or Scalar Variables** :
  - School level, Pain severity, Degrees of shade

- **Qualitative Variables**:
  - Color, Work ,Type of disease, Symptoms, Residence

- **Binomial Variables** : Sex M/F, Pain yes/not, True/False

Converting Variables

Continuous Variables can be converted to discrete or scalar variables by using "cut off points”

- For example, blood pressure can be turned into a nominal variable by defining *hypertension* as a diastolic blood pressure greater than 90 mmHg, and *normotension* as blood pressure less than or equal to 90 mmHg. Height (continuous) can be converted into *short*, *average* or *tall* (ordinal)

In general, it is easier to summarise categorical variables, and so quantitative variables are often converted to categorical ones for descriptive purposes

Converting Variables

To make a clinical decision on someone, one does not need to know the exact serum potassium level (continuous) but whether it is within the normal range (nominal). It may be easier to think of the proportion of the population who are hypertensive than the distribution of blood pressure

However, categorising a continuous variable reduces the amount of information available and statistical tests will in general be more sensitive - that is they will have more power for a continuous variable than the corresponding nominal one, although more assumptions may have to be made about the data. Categorising data is therefore useful for summarising results, but not for statistical analysis. It is often not appreciated that the choice of appropriate cut off points can be difficult, and different choices can lead to different conclusions about a set of data

These definitions of types of data are not unique, nor are they mutually exclusive, and are given as an aid to help an investigator decide how to display and analyse data. One should not debate overlong the typology of a particular variable!

LET US WORK TO AN EXAMPLE OF A DATA COLLECTION FORM FOR BABIES

NAME_____ **Check of : |_|_|_|_|_____| Sex M/F ☐**

Born the |___|___|_____| Todays Age : |_____|_____|
years months

Bon of |_|_|_|_|grams at |_|_| weeks Breast fed |_|_|_| days Weaned at |_| months

Weigth |_|_|_|_|_| gr. cent |_|_| Height |_|_|_|, |_| cm cent |_|_| BMI |____| Cent |___|

Cranius |_____| c.|_____| Growth of one year |____| cm c.|_____| and |_____| grams c.|____|

MAIN PROBLEMS :

1._____**|__|__|**

2._____**|__|__|**

Referred by |_self | or :_____**|__|**

You may see that, in addition to quantitative data (weigth, heigth etc) we have qualitative data (sex), time data (date of birth), but also some 'string' to report clearly the main problems, without restricting to a limited well defined list of items. At the end of each string line a two digit code (from 00 to 99) allows you to code the 'main problem' after data collection.

BEWARE ! NO SURVEY WITHOUT SERVICE !
You asks and collect data in order to give a service

Still a boring form?

NAME_____

Check of : |_|_|_|_|_____| Sex M/F ☐

Born of |_|_|_|_|grams at |_|_| weeks

Breast fed |_|_|_| days Weaned at |_| months

Weigth |_|_|_|_|_| gr. cent |_|_| Height |_|_|_|, |_| cm cent |_|_| BMI |____| Cent |___!

Cranius |_____| c.|_____| Growth o[illegible]ne year |____| cm c.|_____| [illegible]nd |_____| grams c.|_

4 spaces for weight in grams

Height needs 3 spaces + a decimal digit

Dedicate time and attention to data collection: you will need an appropriately designed form, by which to reduce at minimum the errors, the subjectivity, the randomness of data collection. A Date needs 2 digits for day, two for month and 4 for year, no less, no more, no 'open space'

You may find gold in well collected data!

Before embarking in data collection it is mandatory, for each single question of the form, to verify:

- For what and to whom it may help this question
- How many are likely to answer ... which reliability is likely to have the answer
- How can you control for the reliability
- Which feed-back will get the responder
- How we will analyze and show this information
- How you will handle missing or impossible or not available data

WHAT WE HOPE TO SHOW

1. To set up a nice collection of clinical or experimental data of any kind
2. To distinguish the type of data (variables)
3. **To draw a simple ‘coded’ data collection form**
4. To manage ‘missing’ data
5. To start to describe the collected data
6. To explore the distribution of data
7. To use few parameters to describe the data

Let us dedicate the appropriate space, in digits, to each variables. Beware of 'open spaces'!

- A 'Coded' form is simply a form in which we assign the appropriate digits to each variable/information to be collected
- It may be a number, a letter, or a string
- EXAMPLES
- sex: male (1), female (2) or (M) and (F), missing (9);
- school: illiterate (1), primary (2), secondary (3) , not known (9)
- Height : 3 digit – comma-one digit , Weight grams: 5 digit
- Symptoms : 1°_____string 2° _____ string 3° ___ string

(give a string of 12-14 spaces, and add separately a possible list of symptoms to choose from)

SET THE LABELS: ***it helps to transform most informations into numbers, in order to facilitate the analysis of data. To start you prepare a sheet (either paper or EXCEL) on which you assign a column or more to each variable and add the labels for each value of the variable***

LABEL CODING SHEET

Variable	Type	Column	Variable Labels
Case N.	continuous	1-3	from 001 to 999
Sex	binomial	4	M =Male F = Female 9 = Unknown
Date of Birth	date	5-12	day dd/month mm/ year yyyy
Date of survey	date	13-20	day dd/month mm/ year yyyy
Age of Mom	continuous	21-22	years
Education Mom	Ordinal	23	1 'no school' 2 'primary' 3 'higher' 4 'University' 9 'unknown'
Delivery	Qualitative	24	1 'vaginal' 2 'cesarean' 9 'unknown'
Birth Wt	Continuous	25-28	Wt in grams - 9999 = missing
Apgar Score 5'	Scalar	29-30	From 01 to 10 99 = missing
at Breast	Continuous	31-33	from 001 to 500 999= 'not known' missing
Weaned	Scalar	34-35	Months when weaning started 99= missing
Symptom1	Qualitative	36-37	01 'none' 02 'diarrhoea' 03 'bronchitis' …. 99= missing
Symptom2	Qualitative	38-39	01 'none' 02 'diarrhoea' 03 'bronchitis' …. 99= missing
Symptom3	Qualitative	40-41	01 'none' 02 'diarrhoea' 03 'bronchitis' …. 99= missing

WHAT WE HOPE TO SHOW

1. To set up a nice collection of clinical or experimental data of any kind
2. To distinguish the type of data (variables)
3. To draw a simple ‘coded’ data collection form
4. **To manage ‘missing’ data**
5. To start to describe the collected data
6. To explore the distribution of data
7. To use few parameters to describe the data

Missing !!! Missing data are CRUCIAL!

- To get **'clean'** data you need a clear explanation of missing data
- The answer migth be missing because not questioned or unaswered, or a negative answer or 'not applicable' to a child

GIVE A VALUE TO MISSING DATA

- Try to avoid 0, which is often confused with blank
- You may assign 9 = dont' know or 99 : not questioned 77 = not applicable: try to choose numbers which may not have a meaning in your data
- You should prefer a number of the size of the variable which is missing
- If height is missing and height has 5 digits (114,3 cm) you may assign 999,9
- For sex, which has a value 1 or 2, you easily assign 9
- For date of birth: 99/99/9999

WHAT WE HOPE TO SHOW

1. To set up a nice collection of clinical or experimental data of any kind
2. To distinguish the type of data (variables)
3. To draw a simple ‘coded’ data collection form
4. To manage ‘missing’ data
5. **To start to describe the collected data**
6. To explore the distribution of data
7. To use few parameters to describe the data

Once data are collected, you wish to tabulate them in a Frequency Distribution

Let us prepare a Frequency Table of symptoms: You list it!

1. the symptom label, 2. the numer of occurrences, 3 the percentage of each on the total symptoms, 4 the 'valid' % once missing data are excluded (no missing here!) and 5 the Cumulative Percentage: by progressive addition of the cathegories

1. Symptoms	2. Frequency	3 Percent	4 Valid Percent	5 Cumulative Percent
Diarroea	7	3,4	3,4	3,4
Growth failure	3	1,5	1,5	4,9
vomit	2	1,0	1,0	5,9
Abdominal pain	18	8,8	8,8	14,7
constipaton	10	4,9	4,9	19,6
Anaemia	4	2,0	2,0	21,6
Skin disorders	6	2,9	2,9	24,5
Headache & neurol.	9	4,4	4,4	28,9
Hypertransaminasaemia	1	,5	,5	29,4
Gastritis	2	1,0	1,0	30,4
Gyn-OBS symptoms	2	1,0	1,0	31,4
No symptoms	140	68,6	68,6	100,0
Total	204	100,0	100,0	

Cumulative frequencies are helpful when we want to draw the 'distribution' of a certain variable

1. Symptoms	2. Frequency	3 Percent	4 Cumulative Percent
Diarroea	7	**3,4**	3,4
Growth failure	3	**1,5**	4,9
Vomit	2	**1,0**	5,9
Abdominal pain	18	**8,8**	14,7
Constipaton	10	**4,9**	**19,6**
Anaemia	4	2,0	21,6
Skin disorders	6	2,9	24,5
Headache & neurol.	9	4,4	28,9
Hypertransaminasaemia	1	,5	29,4
Gastritis	2	1,0	30,4
Gyn-OBS symptoms	2	1,0	31,4
No symptoms	140	68,6	100,0
Total	204	100,0	

You may observe that the sum of % of celiac children with G.I. symptoms (diarrhoea, growth failure, vomit and abdominal pain and constipation):

3,4 + 1,5 + 1,0 + 8,8 + 4,9 = 19,6%

shows that 19,6% of children have a GI symptoms, while 68,6 have No symptoms and 100 – (68,6 + 19,6) = 11,8% have an extra-GI symptom

WHAT WE HOPE TO SHOW

1. To set up a nice collection of clinical or experimental data of any kind
2. To distinguish the type of data (variables)
3. To draw a simple ‘coded’ data collection form
4. To manage ‘missing’ data
5. To start to describe the collected data
6. **To explore the distribution of data**
7. To use few parameters to describe the data

Stem and leaf plots

Before any statistical calculation, even the simplest, is performed the data should be tabulated or plotted. If they are quantitative and relatively few, say up to about 30, they are conveniently written down in order of size

For example, a community paediatrician, concerned by the heavy traffic in the district, is investigating the amount of lead in the urine of children from a nearby housing estate, which might impair their intellectual development. In a particular street there are 15 children whose ages range from 1 year to under 16, and in a preliminary study the registrar has found the following amounts of urinary lead (µmol/24h), given in the Table, what is called an array:

Urinary concentration of lead in 15 children from housing estate (µmol/24hr)
0.6, 2.6, 0.1, 1.1, 0.4, 2.0, 0.8, 1.3, 1.2, 1.5, 3.2, 1.7, 1.9, 1.9, 2.2

A simple way to order, and also to display, the data is to use a stem and leaf plot. To do this we need to abbreviate the observations to two significant digits

In the case of the urinary concentration data, the digit to the left of the decimal point is the "stem" and the digit to the right the "leaf”

We first write the stems in order down the page. We then work along the data set, writing the leaves down "as they come". Thus, for the first data point, we write a 6 opposite the 0 stem. These data are given in the Figure:

Stem and leaf "as they come"

Stem	Leaf
0	6 1 4 8
1	1 3 2 5 7 9 9
2	6 0 2
3	2

I DATI CHE ABBIAMO RACCOLTI

0.6, 0.1, 0.4, 0.8

1.1, 1.3, 1.2, 1.5, 1.7, 1.9, 1.9

2.6, 2.0, 2.2

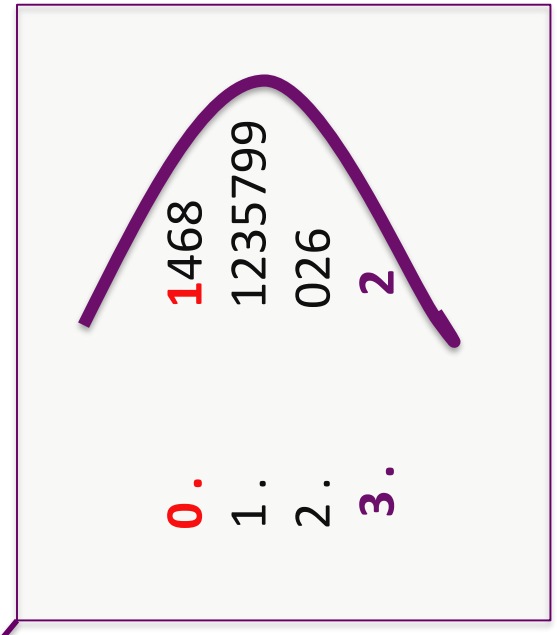
3.2

We then order the leaves as shown in this Figure:

Stem-and-Leaf Plot

Frequency	Stem &	Leaf
4,00	**0 .**	**1**468
7,00	1 .	1235799
3,00	2 .	026
1,00	**3 .**	**2**

Each leaf: 1 case(s)



The advantage of first setting the figures out in order of size and not simply feeding them straight from notes into a calculator (for example, to find their mean) is that the relation of each to the next can be looked at. Is there a steady progression, a noteworthy hump, a considerable gap? Simple inspection can disclose irregularities. **Furthermore, a glance at the figures gives information on their range.** The smallest value is **0.1** and the largest is **3.2** µmol/24h

WHAT WE HOPE TO SHOW

1. To set up a nice collection of clinical or experimental data of any kind
2. To distinguish the type of data (variables)
3. To draw a simple ‘coded’ data collection form
4. To manage ‘missing’ data
5. To start to describe the collected data
6. To explore the distribution of data
7. **To use few parameters to describe the data**

DESCRIBING DATA

MEAN	**Average or arithmetic mean of the data**
MEDIAN	**The value which comes half way when the data are ranked in order**
MODE	**Most common value observed**

- In a normal distribution, mean and median are the same
- If median and mean are different, indicates that the data are not normally distributed
- The mode is of little if any practical use

MEAN OR AVERAGE

To calculate the mean we add up the observed values and divide by the number of them

This familiar process is conveniently expressed by the following symbols:

$$\bar{x} = \frac{(\text{å}\ x)}{n}$$

(pronounced "x bar") signifies the mean; x is each of the values; n is the number of these values; and , the Greek capital sigma (our "S") denotes "sum of". A major disadvantage of the mean is that it is sensitive to outlying points

MEDIAN

The median is known as a measure of location; that is, it tells us where the data are

You do not need to know all the exact values to calculate the median; if you made the smallest value even smaller or the largest value even larger, it would not change the value of the median

Thus the median does not use all the information in the data and so it can be shown to be less efficient than the mean or average, which does use all values of the data

Median

To find the median (or mid point) we need to identify the point which has the property that half the data are greater than it, and half the data are less than it. For 15 points, the mid point is clearly the eighth largest, so that seven points are less than the median, and seven points are greater than it. This is easily obtained from the previous by counting the eighth leaf, which is 1.5 µmol/24h

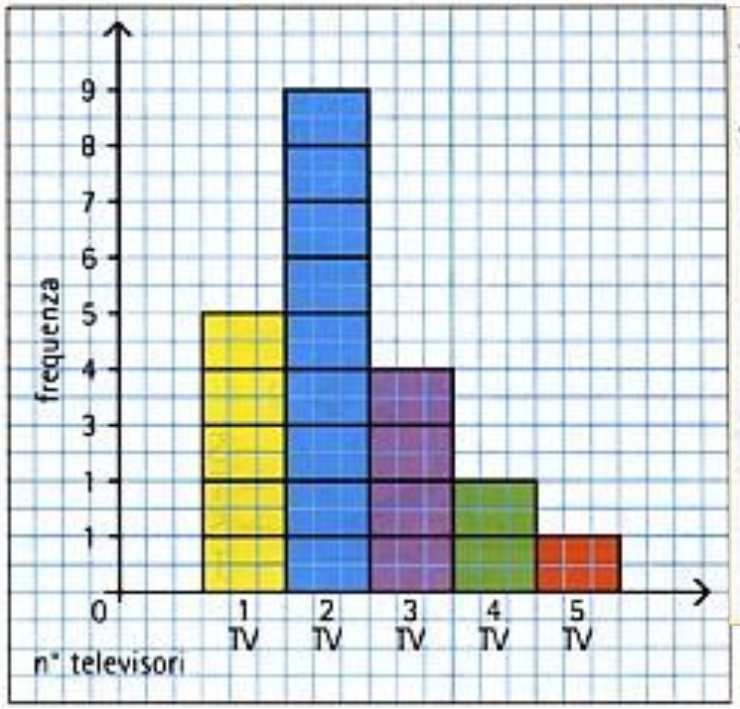
To find the median for an even number of points, the procedure is as follows. Suppose the paediatric registrar obtained a further set of 16 urinary lead concentrations from children living in the countryside in the same county as the hospital

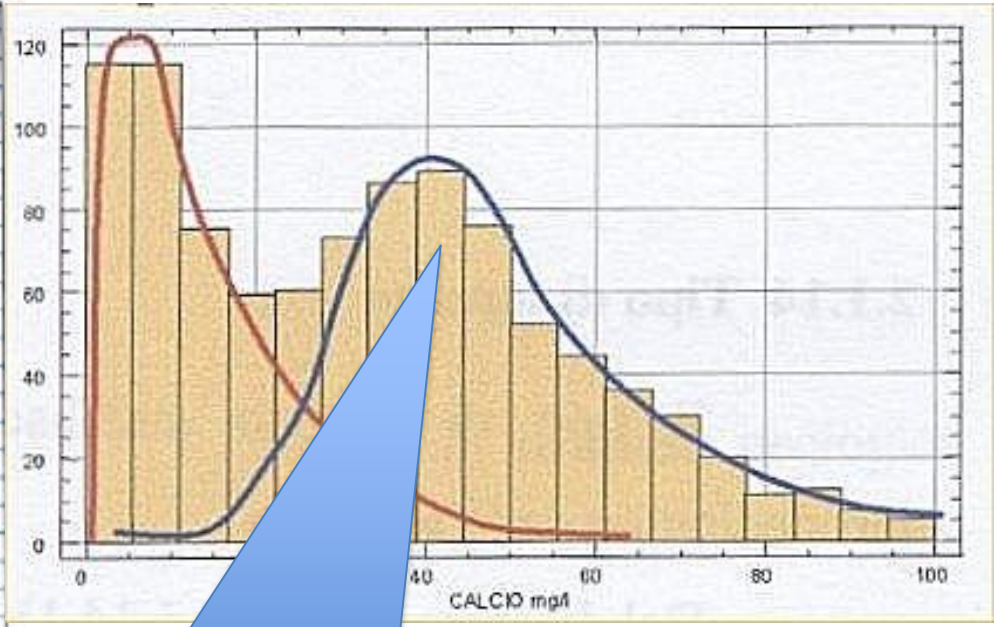
Urinary concentration of lead in 16 rural children ($\mu mol/24hr$)
0.2, 0.3, 0.6, 0.7, 0.8, 1.5, 1.7, 1.8, 1.9, 1.9, 2.0, 2.0, 2.1, 2.8, 3.1, 3.4

To obtain the median we average the eighth and ninth points (1.8 and 1.9) to get 1.85 µmol/24h. In general, if n is even, we average the n/2th largest and the n/2 + 1th largest observations

The main advantage of using the median as a measure of location is that it is "robust" to outliers. One disadvantage is that it is tedious to order a large number of observations by hand (there is usually no "median" button on a calculator)

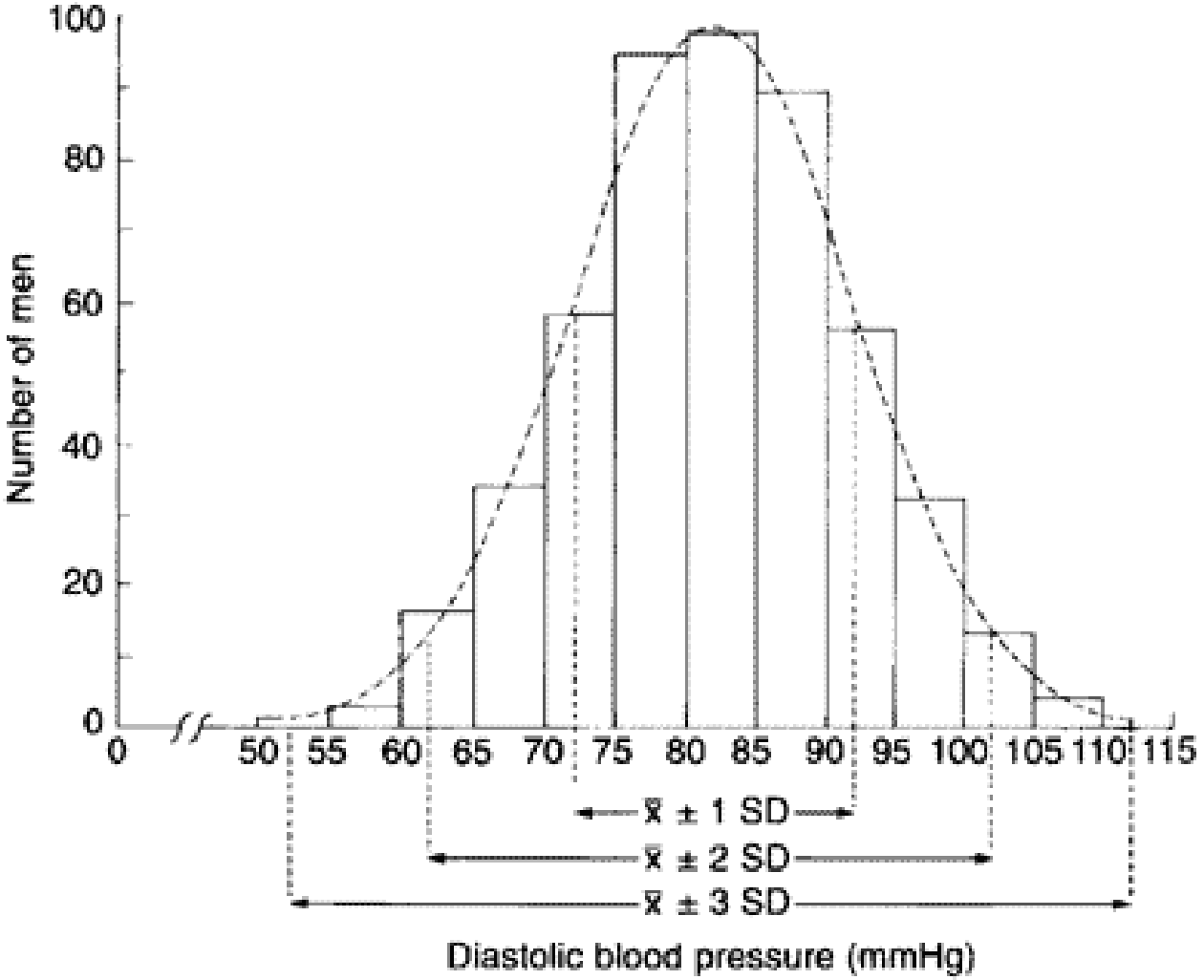
It is easy to diplay data in a graphic: bars for small series, lines for large series





Some continuous variables show a simmetrical distribution around a central point with dispersion of data at right and at the left of a central point

Data with normal distribution are scattered simmetrically around a central point, which is the mean of the numbers (Mean), the most frequent value (Mode) and the point of the distribution with equal distance from the extreme (Median)



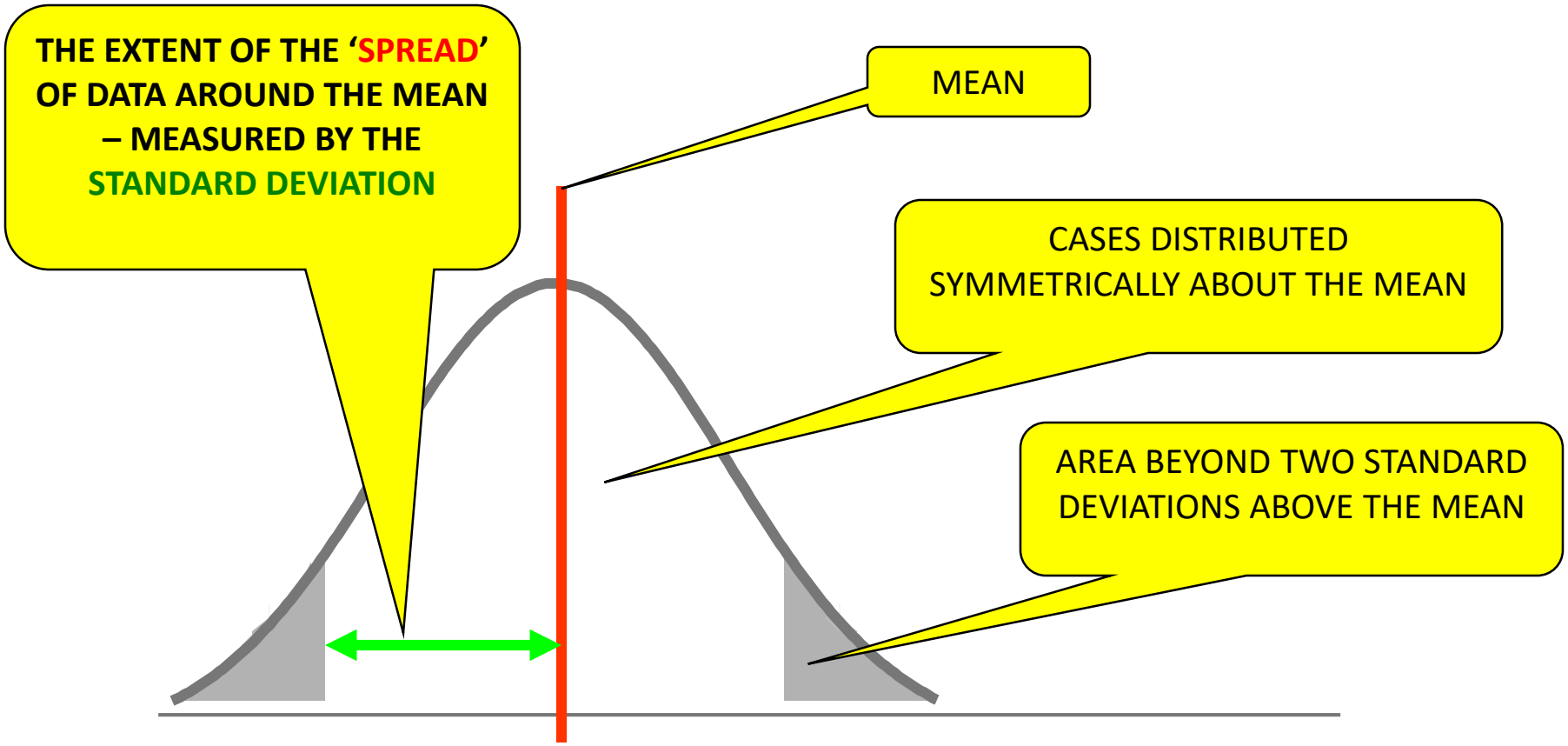
The Normal Distribution

The Normal distribution is represented by a family of curves defined uniquely by two parameters, which are the mean and the standard deviation of the population

The curves are always symmetrically bell shaped, but the extent to which the bell is compressed or flattened out depends on the standard deviation of the population

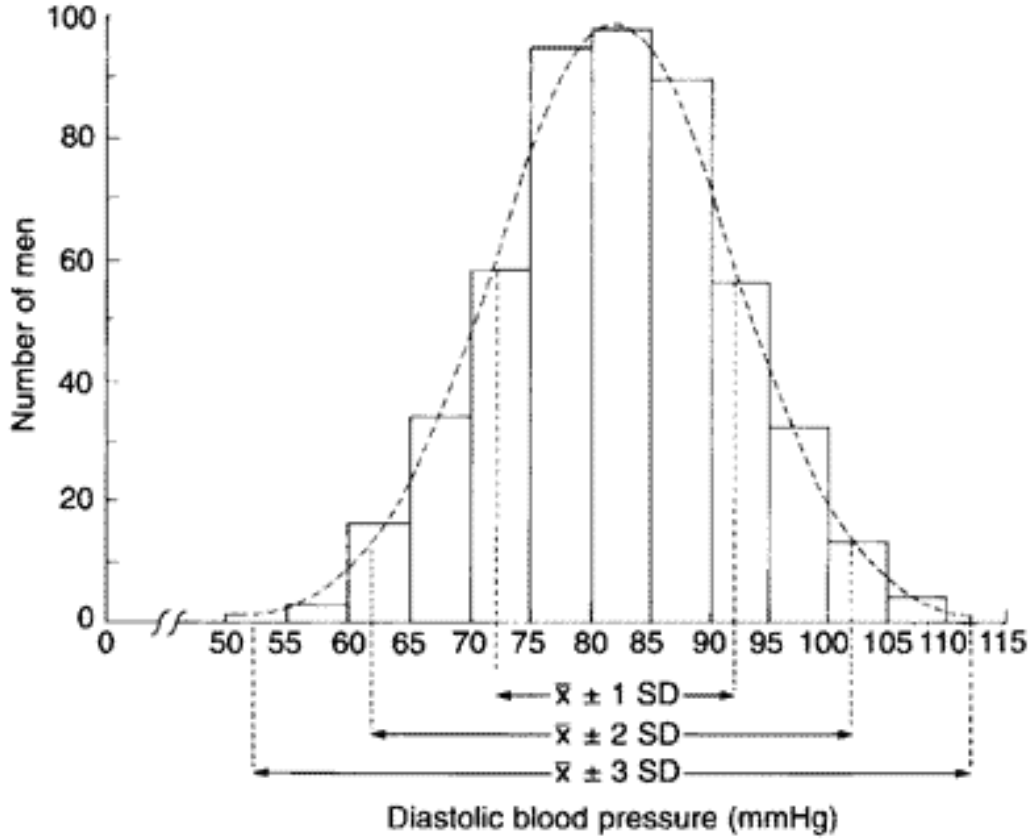
However, the mere fact that a curve is bell shaped does not mean that it represents a Normal distribution, because other distributions may have a similar sort of shape

The NORMAL DISTRIBUTION has very interesting features : it shows how the data are spread around a central point, symmetrically



Many biological characteristics conform to a Normal distribution closely enough for it to be commonly used - for example, heights of adult men and women, blood pressures in a healthy population, random errors in many types of laboratory measurements and biochemical data

Figure shows a Normal curve calculated from the diastolic blood pressures of 500 men, mean 82 mmHg, standard deviation 10 mmHg. The ranges representing ±1SD, ±2SD, and ±3SD and about the mean are marked



Measures of variation

It is informative to have some measure of the variation of observations about the median. The range is very susceptible to what are known as outliers, points well outside the main body of the data (For example, if we had made the mistake of writing 34 instead 3.4 in the Table shown before, then the range would be written as 0.1 to 34 which is clearly misleading)

A more robust approach is to divide the distribution of the data into four groups, and find the points below which are 25%, 50% and 75% of the distribution. These are known as **quartiles**, and the median is the second quartile. The variation of the data can be summarised in the interquartile range, the distance between the first and third quartile. With small data sets and if the sample size is not divisible by four, it may not be possible to divide the data set into exact quarters, and there are a variety of proposed methods to estimate the quartiles. A simple, consistent method is to find the points midway between each end of the range and the median

Statistics Derived from Sample Percentiles

-
- The ***50th*** percentile of a sample of *n* observations is referred to as the ***sample median***

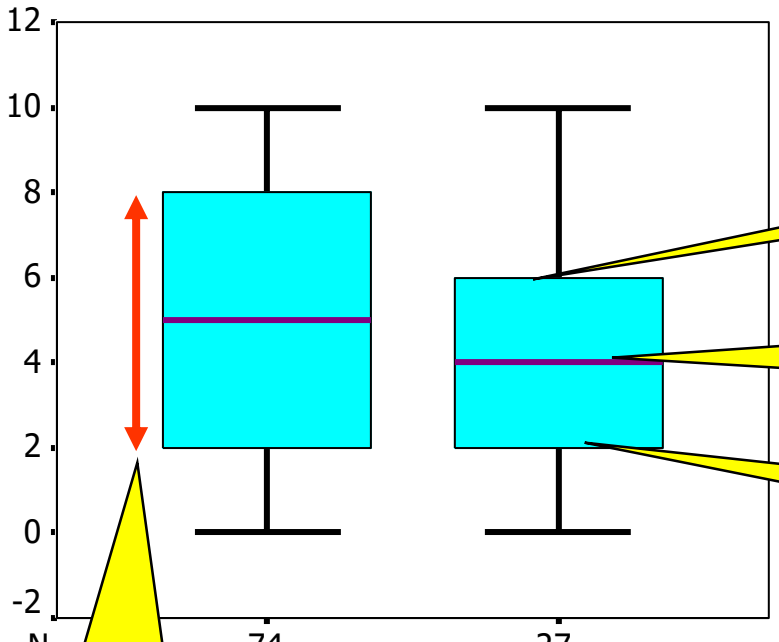
- The ***sample median*** estimates the central location of the distribution of the values of the variable within the study population

- The ***25th*** percentile of a sample of n observations is referred to as the ***lower quartile***, and the ***75th*** percentile of a sample of n observations is referred to as the ***upper quartile***

- The difference between the upper quartile value and the lower quartile value is referred to as the ***interquartile range*** of the frequency distribution

- The ***interquartile range*** estimates the degree of the dispersion of the observations within the middle 50% of the distribution of the values of the variable within the study population

Inter-Quartile Range BOXPLOT



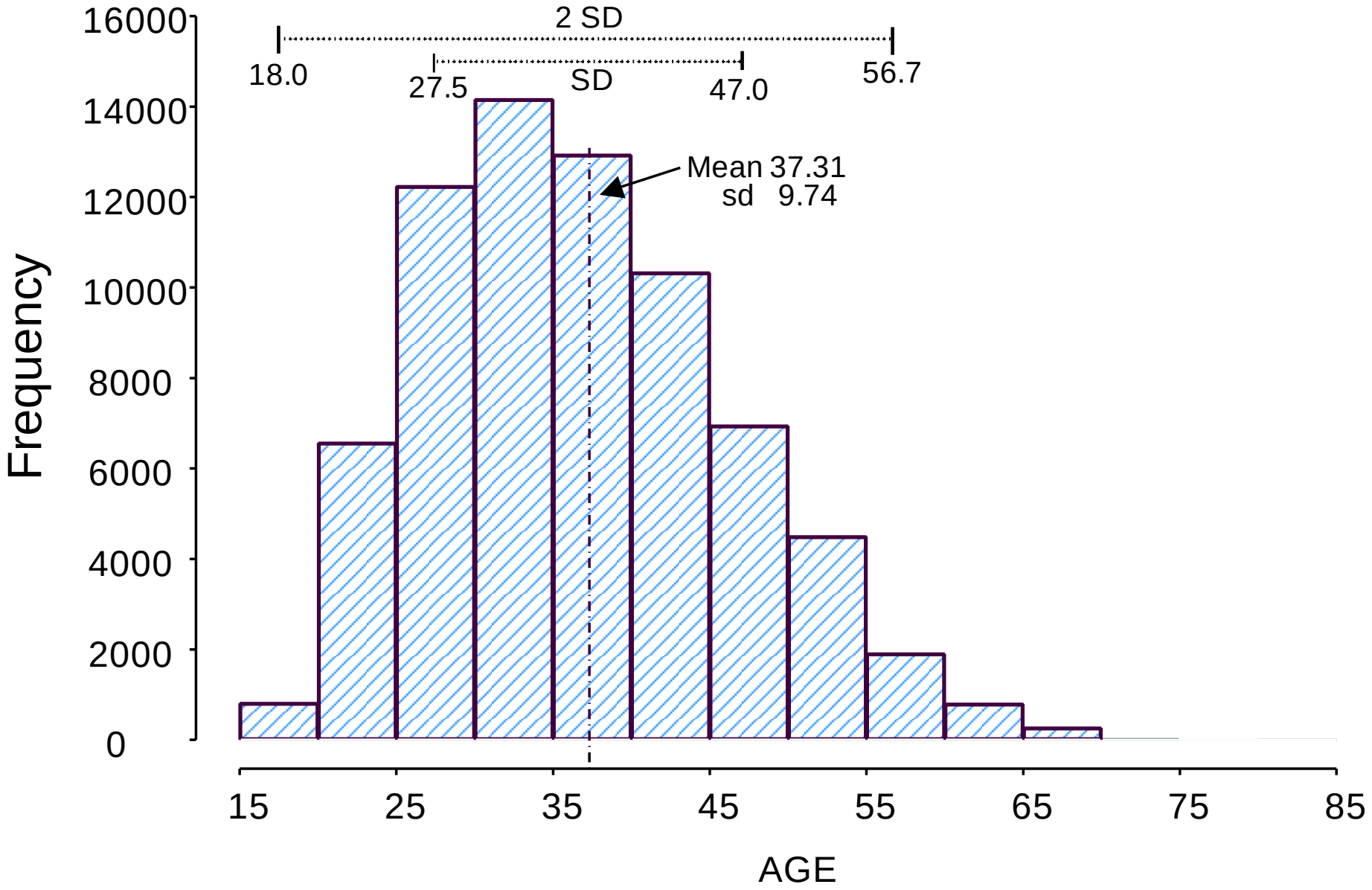
75th Centile = 2nd Quart

MEDIAN
(50th centile)

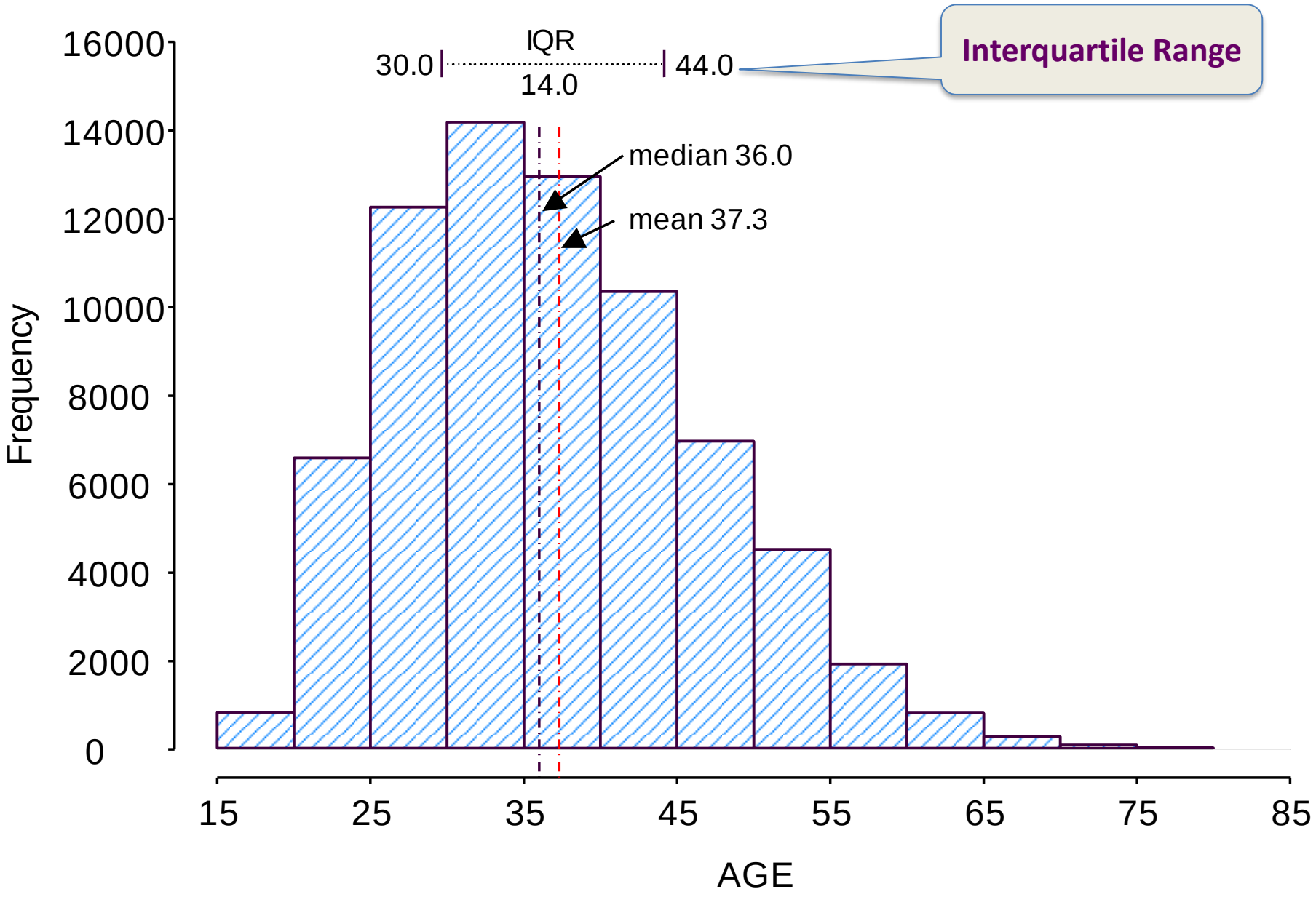
25th Centile= 1st Quart

Inter-quartile range

Distribution of the Age of Hospital Doctors in France

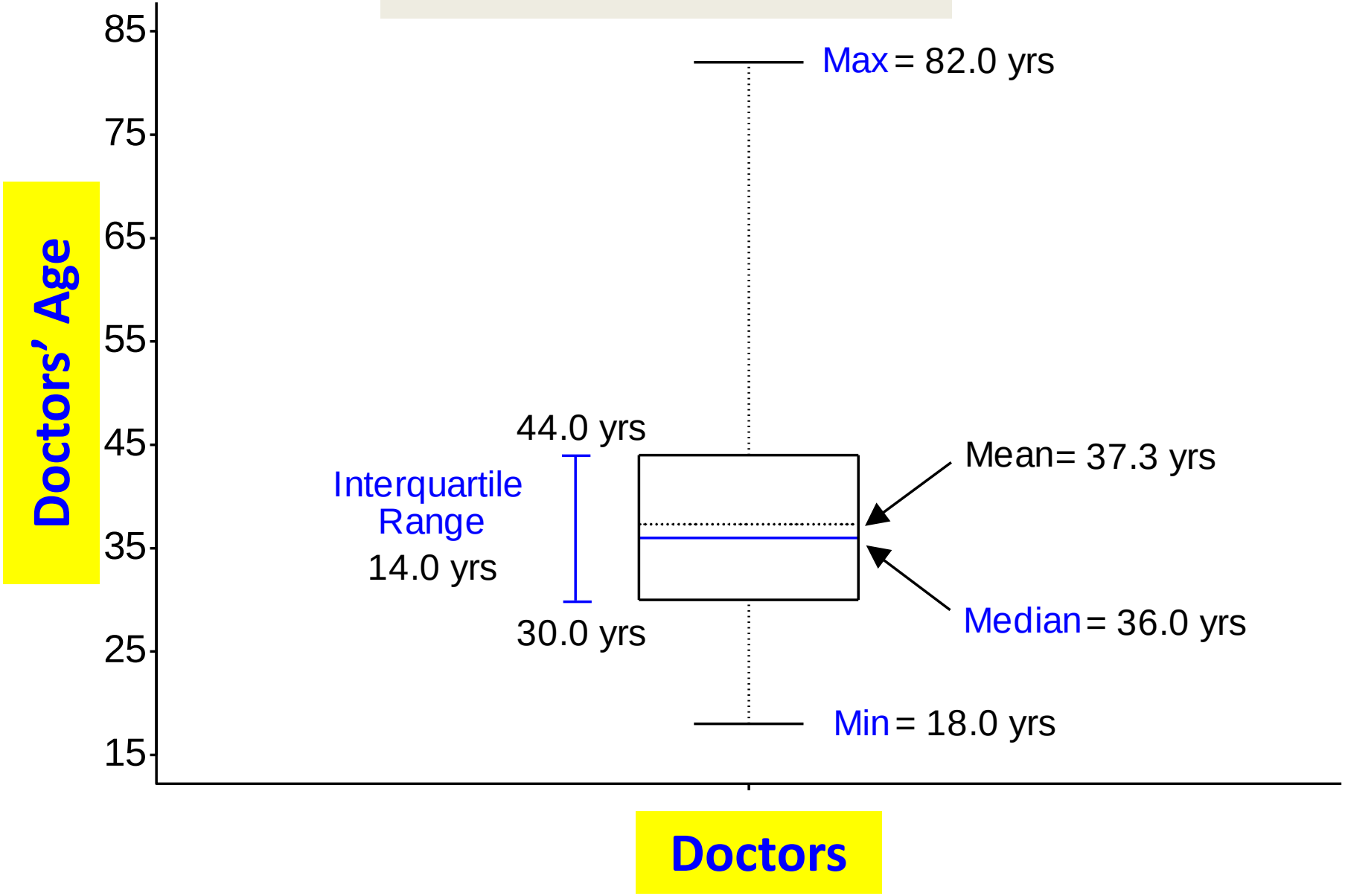


Distribution of the Age of Hospital Doctors in France

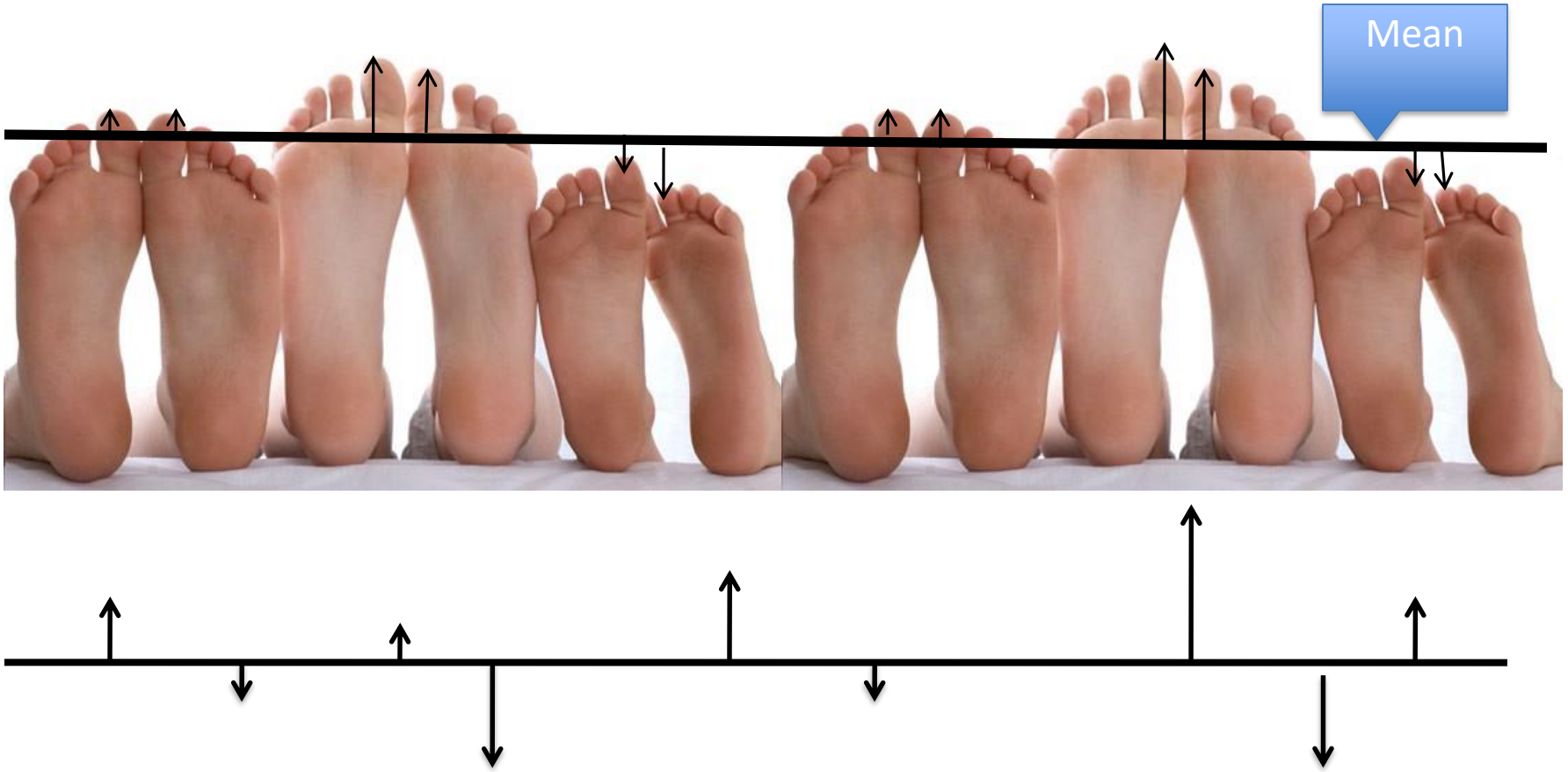


Distribution of the Age of Hospital Doctors in France

Same Data in a Box Plot!



Not all feets are equal!!!



Differences from the mean are positive as well as negative: the algebraic sum is 0
To get an estimate of the total differences from the mean we just take the square of the observed differences

Let’us get a sum of deviations from the mean!

- DEVIANCE = $\sum (m - x)^2$

Sum of squared deviation from the mean

- VARIANCE= $\frac{\sum (m - x)^2}{n - 1}$

(Note that if you have the sum of values, you may deduct the last number, by subracting from the sum all the other data, so you have n-1 ‘degrees of freedom’)

- Sum of squared deviation from the mean/by degrees of freedom (n of cases – 1)

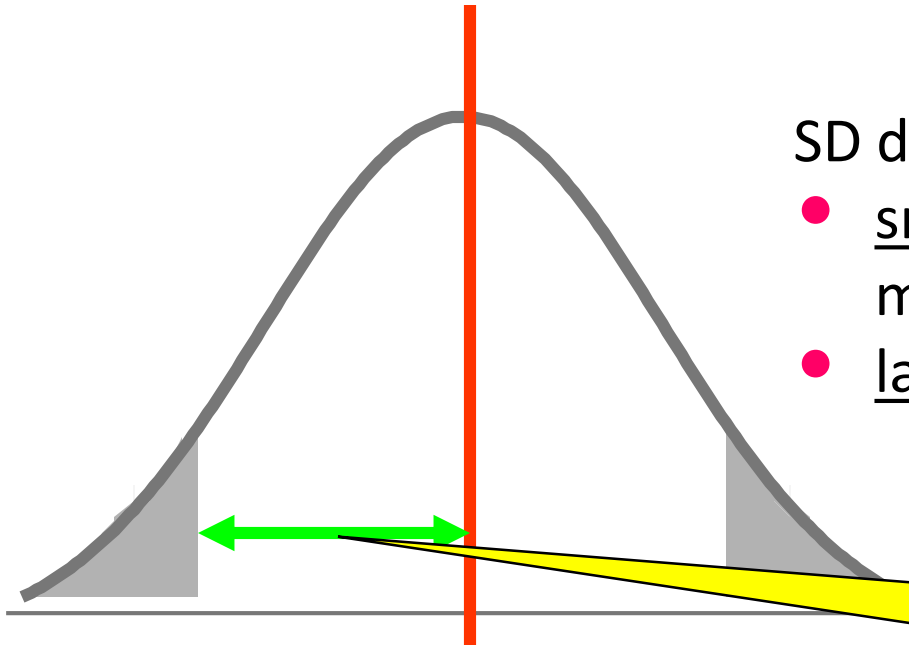
- Standard Deviation = $\sqrt{\frac{\sum (m - x)^2}{n - 1}}$

The square root of variance

STANDARD DEVIATION – MEASURE OF THE SPREAD OF VALUES OF A SAMPLE AROUND THE MEAN

THE SQUARE OF THE SD IS KNOWN AS THE **"VARIANCE"**

$$SD = \sqrt{\frac{\text{Sum}(\text{Value} - \text{Mean})^2}{\text{Number of values}}}$$



SD decreases as a function of:

- <u>smaller spread</u> of values about the mean
- <u>larger number</u> of values

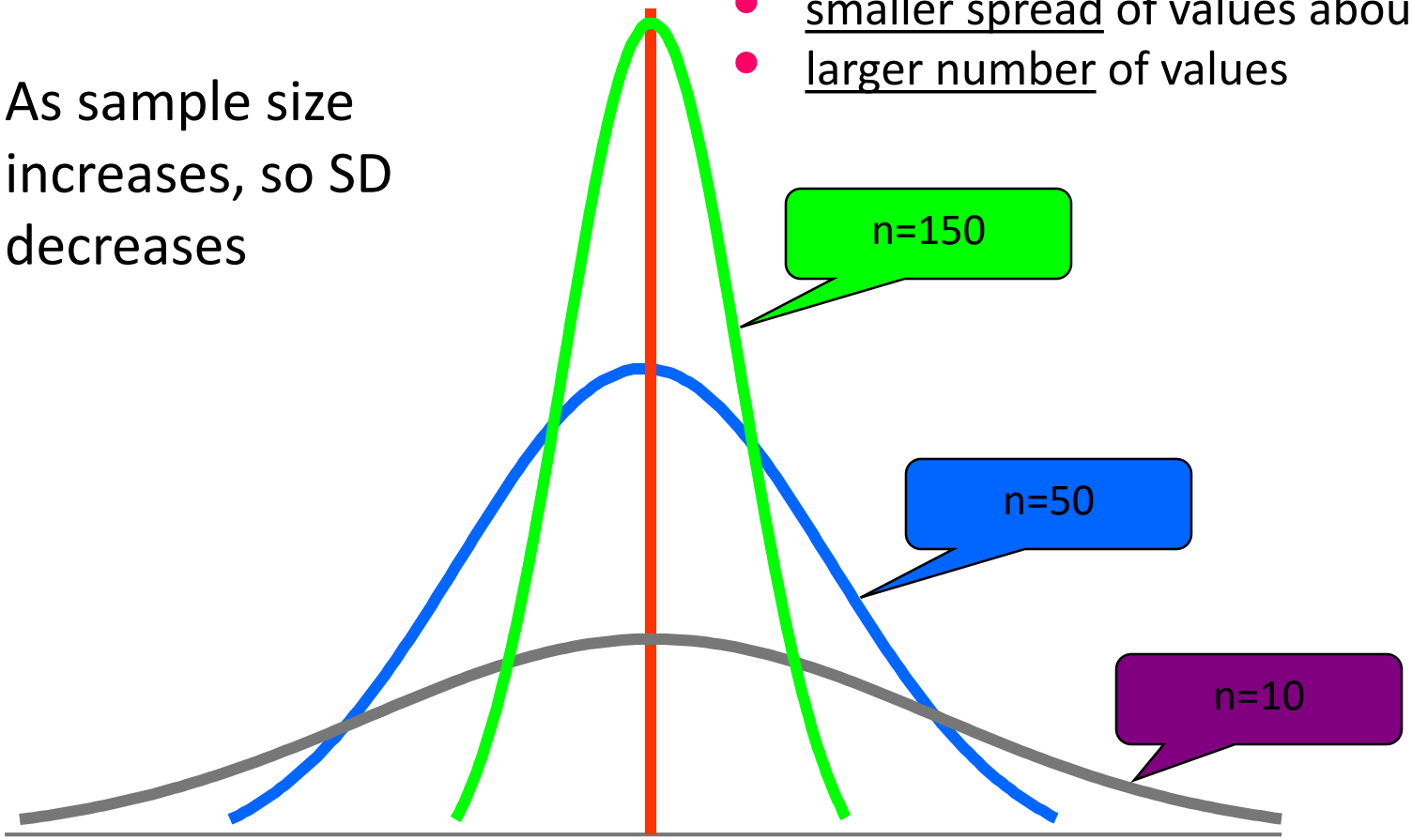
IN A NORMAL DISTRIBUTION, 95% OF THE VALUES WILL LIE WITHIN 2 SDs OF THE MEAN

STANDARD DEVIATION AND SAMPLE SIZE

SD decreases as a function of:

- smaller spread of values about the mean
- larger number of values

As sample size increases, so SD decreases



The Standard Deviation

As well as measures of location we need measures of how variable the data are

The range is an important measurement, for figures at the top and bottom of it denote the findings furthest removed from the generality. However, they do not give much indication of the spread of observations about the mean. This is where the **standard deviation (SD)** comes in

A practical point to note here is that, when the population from which the data arise have a distribution that is approximately "Normal" (or Gaussian), then the standard deviation provides a useful basis for interpreting the data in terms of probability

The reason why the standard deviation is such a useful measure of the scatter of the observations is this: if the observations follow a Normal distribution, a range covered by one standard deviation above the mean and one standard deviation below it ($\bar{x} \pm 1SD$) includes about 68% of the observations; a range of two standard deviations above and two below ($\bar{x} \pm 2SD$) about 95% of the observations; and of three standard deviations above and three below ($\bar{x} \pm 3SD$) about 99.7% of the observations

Consequently, if we know the mean and standard deviation of a set of observations, we can obtain some useful information by simple arithmetic. By putting one, two, or three standard deviations above and below the mean we can estimate the ranges that would be expected to include about 68%, 95%, and 99.7% of the observations

Standard deviation from ungrouped data

The standard deviation is a summary measure of the differences of each observation from the mean. If the differences themselves were added up, the positive would exactly balance the negative and so their sum would be zero

Consequently the squares of the differences are added. The sum of the squares is then divided by the number of observations *minus one* to give the mean of the squares, and the square root is taken to bring the measurements back to the units we started with

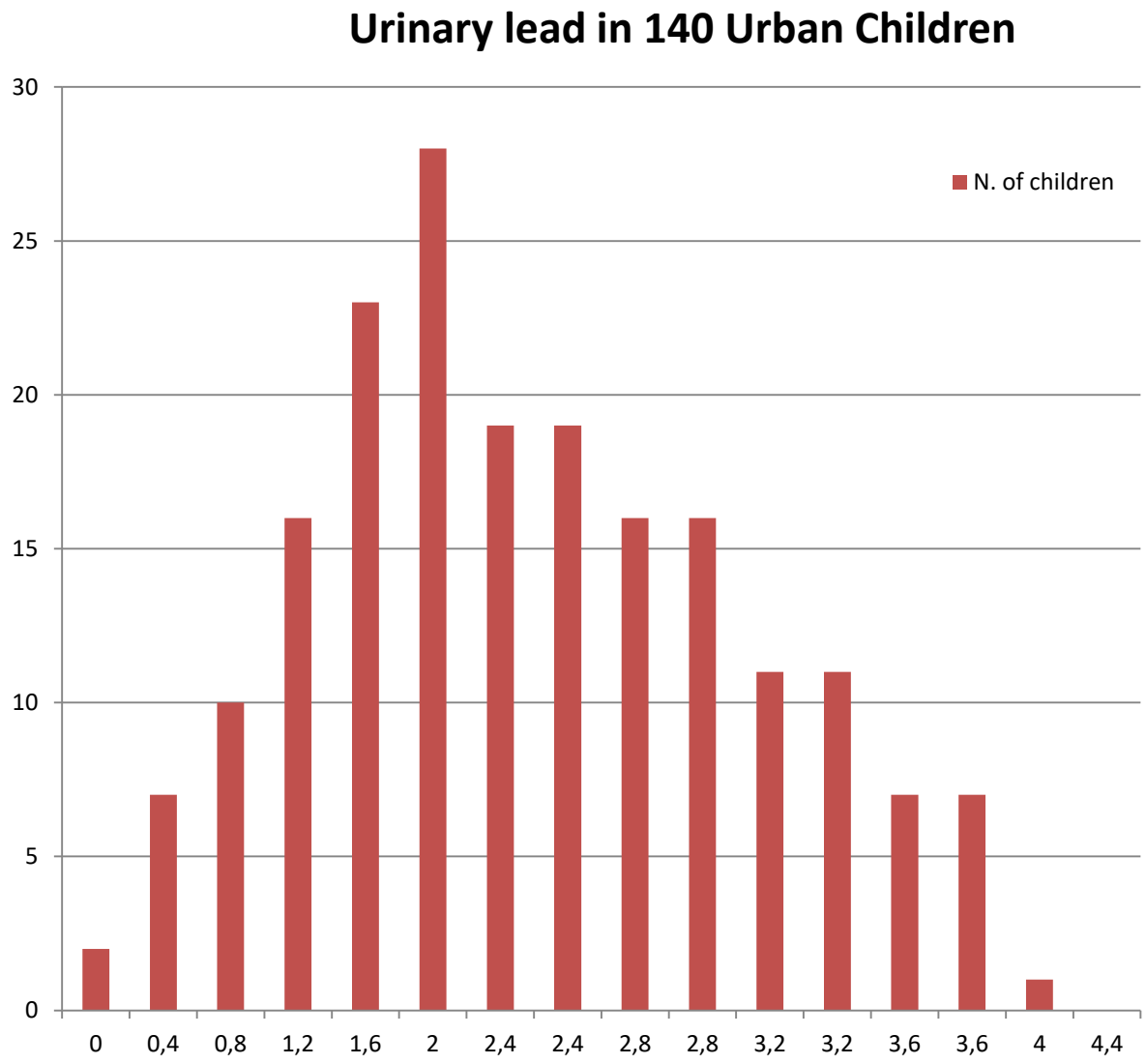
NB: The division by the number of observations *minus one* instead of the number of observations itself to obtain the mean square is because "degrees of freedom" must be used. In these circumstances they are one less than the total. The theoretical justification for this need not trouble the user in practice

Histograms

Suppose the paediatric registrar referred to earlier extends the urban study to the entire estate in which the children live. He obtains figures for the urinary lead concentration in 140 children aged over 1 year and under 16. We can display these data as a grouped frequency table

Lead Concentration in 140 Urban Children	
Lead conc	**N. of children**
0	2
0,4	7
0,8	10
1,2	16
1,6	23
2	28
2,4	19
2,4	19
2,8	16
2,8	16
3,2	11
3,2	11
3,6	7
3,6	7
4	1
4,4	0

Histogram of data from previous Table



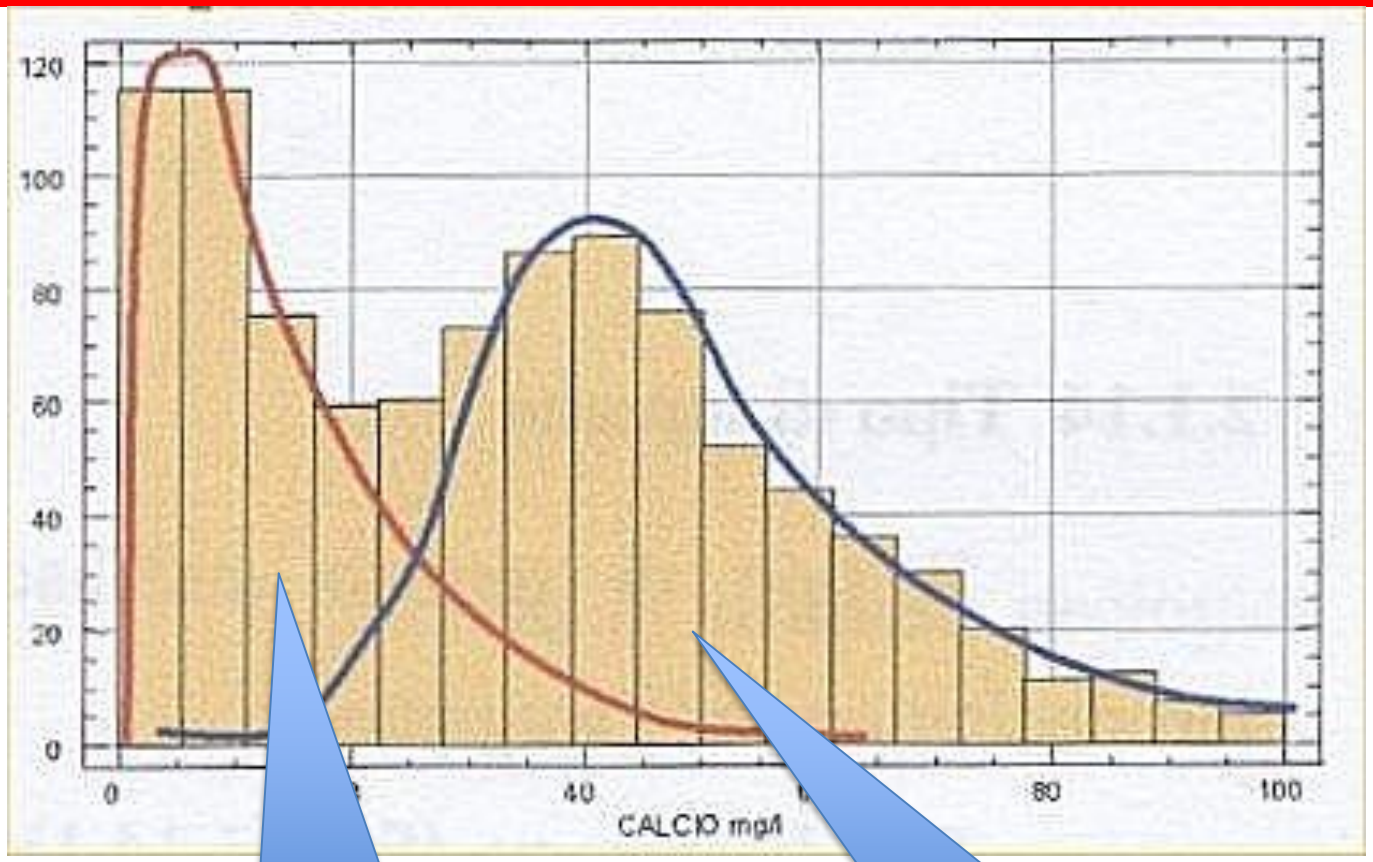
The calculation of the variance is illustrated in the following table with the 15 readings in the preliminary study of urinary lead concentrations (as an example). The readings are set out in column (1). In column (2) the difference between each reading and the mean is recorded. The sum of the differences is 0. In column (3) the differences are squared, and the sum of those squares is given at the bottom of the column

Calculation of standard deviation

	(1) Lead concentration µmol/24hr	(2) Differences from mean $x - \bar{x}$	(3) Differences squared $(x - \bar{x})^2$
	0.1	-1.4	1.96
	0.4	-1.1	1.21
	0.6	-0.9	0.81
	0.8	-0.7	0.49
	1.1	-0.4	0.16
	1.2	-0.3	0.09
	1.3	-0.2	0.04
	1.5	0	0
	1.7	0.2	0.04
	1.9	0.4	0.16
	1.9	0.4	0.16
	2.0	0.5	0.25
	2.2	0.7	0.49
	2.6	1.1	1.21
	3.2	1.7	2.89
Total	**22.5**	**0**	**9.96**

n= 15, $\bar{x}$ = 1.5

Not alla data show a symmetrical ‘normal’ distribution : Antibodies ? All skewed



VERY SKEWED AND NOT SYMMETRIC

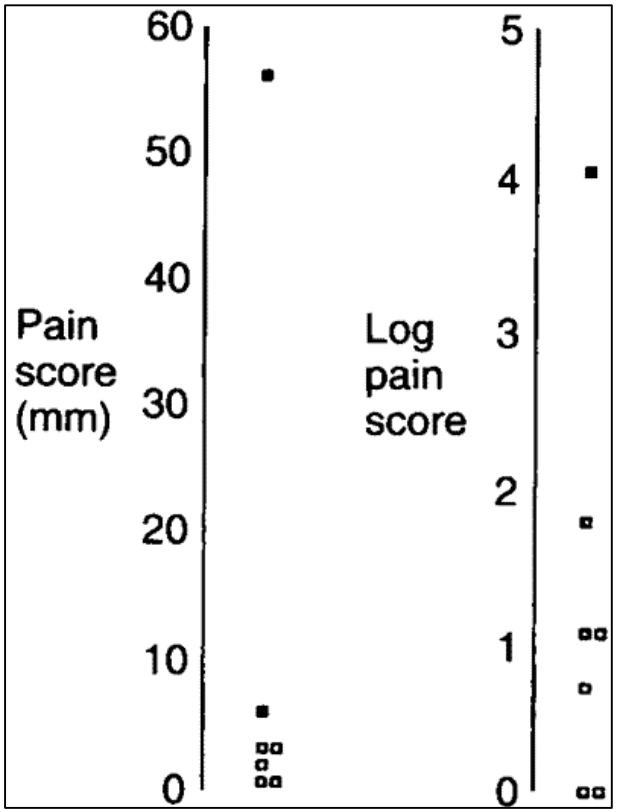
QUITE NORMAL AND SYMMETRIC

Data transformation

An anaesthetist measures the pain of a procedure using a 100 mm visual analogue scale on seven patients. The results are given in the Table, together with the **$\log_e$ transformation**

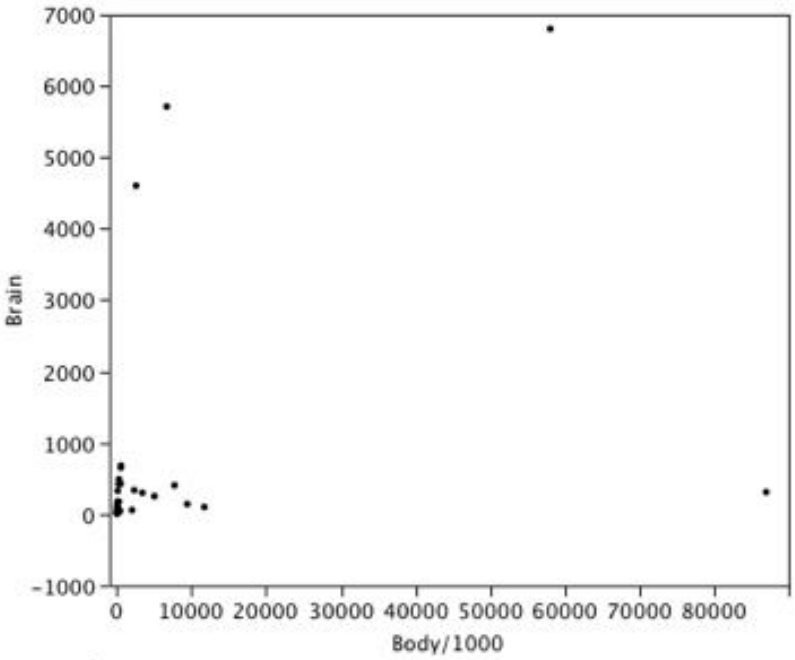
Results fom pain score on seven patients (mm)	
Original scale:	1, 1, 2, 3, 3, 6, 56
$\log_e$ scale:	0, 0, 0.69, 1.10, 1.10, 1.79, 4.03

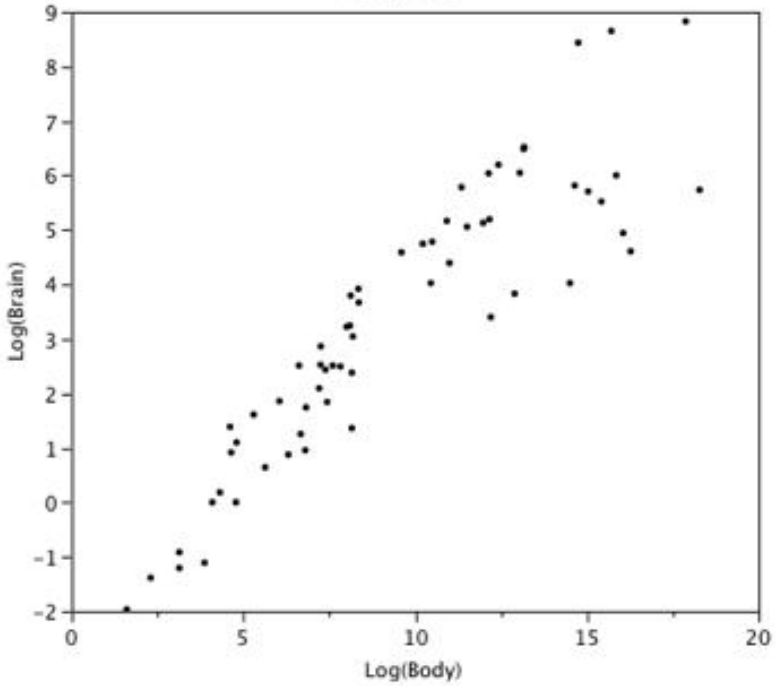
The data are plotted in the Figure, which shows that the **outlier does not appear so extreme in the logged data.** The mean and median are 10.29 and 2, respectively, for the original data, with a standard deviation of 20.22. Where the mean is bigger than the median, the distribution is positively skewed



For the logged data the mean and median are 1.24 and 1.10 respectively, indicating that the **logged data have a more symmetrical distribution.** Thus it would be better to analyse the logged transformed data in statistical tests than using the original scale

log $_e$ transformation - Example





The log transformation can be used to make highly skewed distributions less skewed. This can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics

Figure shows an example of how a log transformation can make patterns more visible. **Both graphs plot the brain weight of animals as a function of their body weight**. The **raw** weights are shown in the upper panel; the **log-transformed** weights are plotted in the lower panel

It is hard to discern a pattern in the upper panel whereas the strong relationship is clearly shown in the lower panel!

In reporting these results, **the median of the raw data would be given**, but it should be explained that the statistical test was carried out on the transformed data

Note that the median of the logged data is the same as the log of the median of the raw data - however, this is not true for the mean

The mean of the logged data is not necessarily equal to the log of the mean of the raw data. The antilog (**exp** or e^X on a calculator) of the mean of the logged data **is known as the *geometric mean,*** and is often a better summary statistic than the mean for data from positively skewed distributions. For these data the geometric mean in 3.45 mm

Between subjects and within subjects standard deviation

If repeated measurements are made of the height of a child, these measurements are likely to vary. This is **within subject, or intrasubject, variability** and we can calculate a standard deviation of these observations : $\sqrt{\frac{\sum (m - x)^2}{n - 1}}$

SQRT (SUM((each measure-mean of all measures)^2) /(n of measures-1))

this is the ***measurement error***

Measurements made on different subjects vary according to between subject, or intersubject, variability. If many observations were made on each individual, and the average taken, then we can assume that the intrasubject variability has been averaged out and the variation in the average values is due solely to the intersubject variability. Single observations on individuals clearly contain a mixture of intersubject and intrasubject variation

The ***coefficient of variation*** **(CV%)** is the **intrasubject standard deviation divided by the mean**, expressed as a percentage. It is often quoted as a measure of repeatability for biochemical assays, when an assay is carried out on several occasions on the same sample.

It has the advantage of being independent of the units of measurement
Obviously it is easy to read: 10%, 25%, 50% and estimate the ‘reliability’ of the measure