Practical Biostatistics

Department of Clinical Epidemiology, Biostatistics and Bioinformatics

AMC

Introduction & descriptive statistics

Aim

- ☒ Analyse your own data with the appropriate statistical tests
- ☒ Learn how to interpret the obtained results

- ☒ SPSS, version 11 or 12

Program

- ☒ Introduction of the subject
  9.30 to 10.15/10.30
- ☒ Break
- ☒ SPSS exercise
  10.30/10.45 to 12.30

- ☒ Books: Petrie & Sabin
  Altman

Program per Lecture

1. Introduction; What are data?
2. Principle of statistical tests (distr)
3. Comparing continuous variables (t-test & ANOVA)
4. Comparing 2 proportions: RR, OR, Chi^2
5. Correlation & linear regression
6. Multivariate linear regression

Program per Lecture

7. Logistic regression
8. Survival analysis – Kaplan Meier Curves
9. Survival analysis – Cox regression
10. Repeated measurements
11. Update – integration of college 1 to 10

Assumptions & Goals

- ☒ The database is clean: ready to use (course clinical data management)
- ☒ Analyse your own data
- ☒ Adequate ability to interpret results

A ‘cleaned’ database

☒ Missing values are marked (88, 99)
☒ Variables are labelled
☒ Values are labelled

For example:

- Variable: sex
- Label: gender of the patient
- Value label: 1 = male, 0 = female

Petrie chapter 2: Data entry

Data Checking

- ☒ Missing checks
- ☒ Range checks
- ☒ Outlier checks
- ☒ Distribution checks
- ☒ Logical checks

Petrie chapter 3: Data checking

Types of data (1)

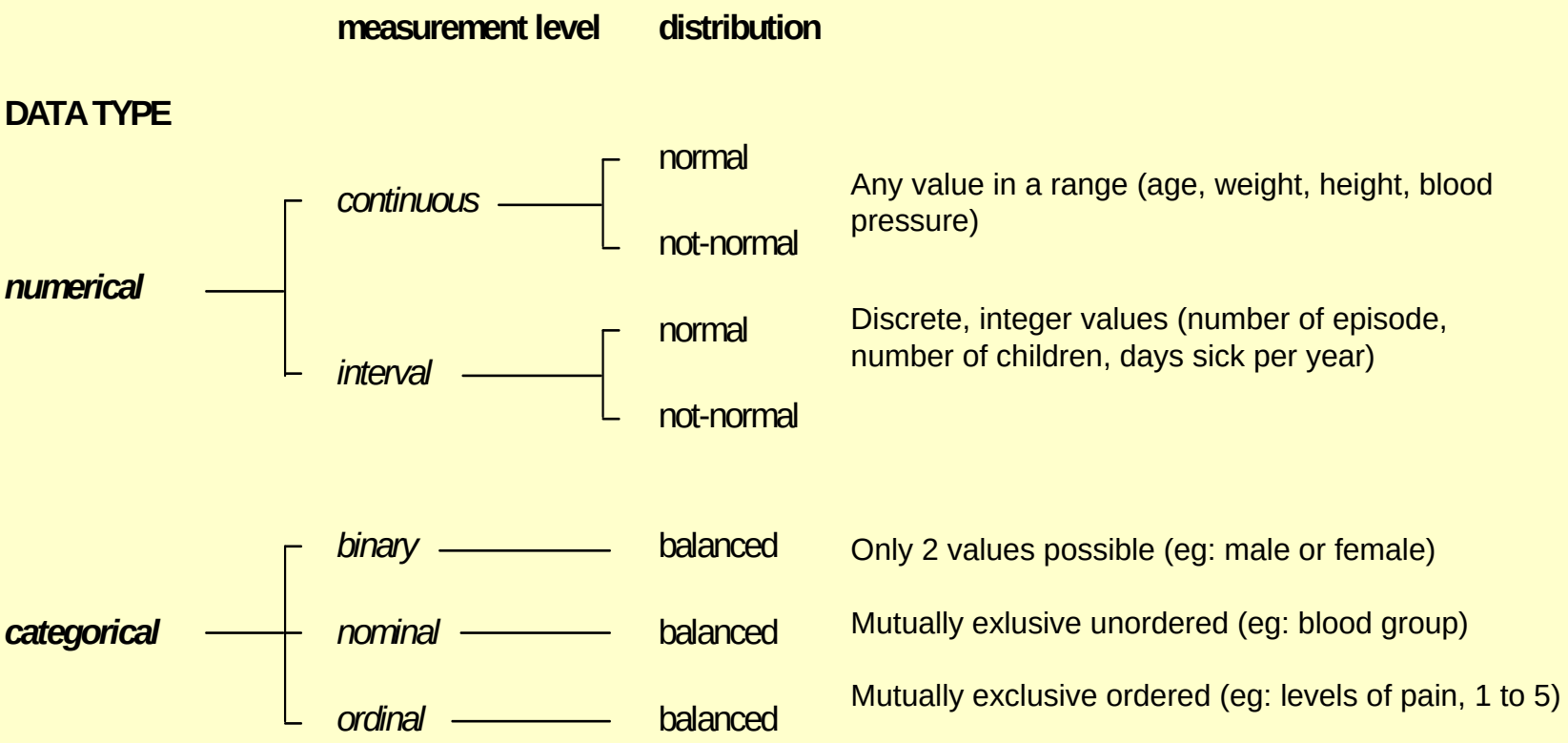
☒ **Data**: contains information about a particular area of research

☒ Data comprise **Observations**: one or more variables

☒ A **variable** is any quantity that varies

Petrie chapter 1: Types of data

Types of data (2)



Types of data (3)

☒ descriptive statistic, Statistical method and presentation depending on type of data and data distribution

Pitfall: variable with numerous ordered categories: e.g. a 7-point pain score,

Is this ordinal or discrete ?

Describing the Data

Descriptive statistics

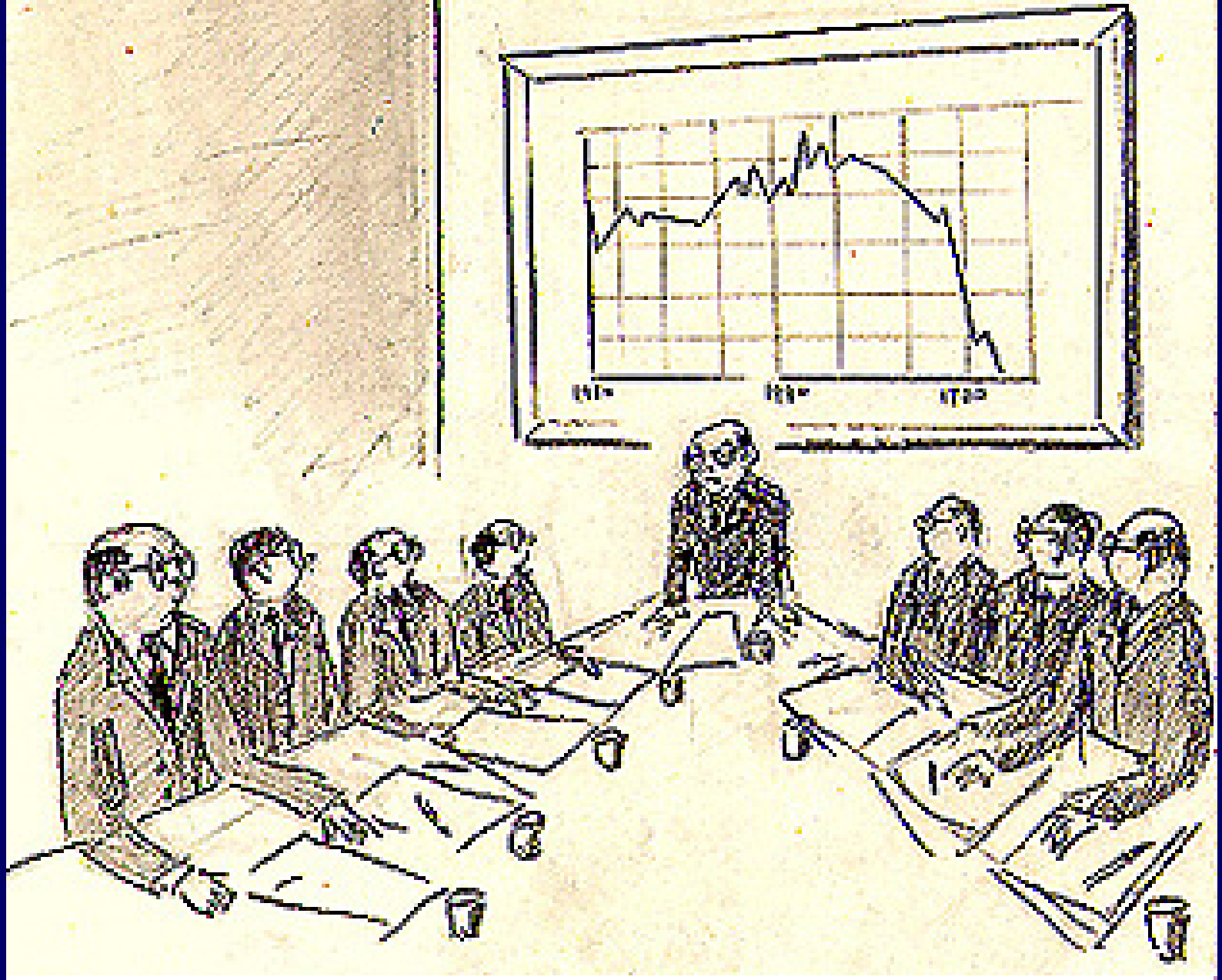
To summarize a large amount of observations with a minimum of loss of information into numbers, figures or tables

Describing the Data (1)

☒ **Categorical data (binary, nominal, ordinal)** can be summarized as:

- Number (N)
- Percentage (%)
- Proportion (0 to 1)

Binary: Only 2 values possible (eg: male or female)
Nominal: Mutually exlusive unordered (eg: blood group)
Ordinal: Mutually exclusive ordered (eg: level of pain, 1 to 5)

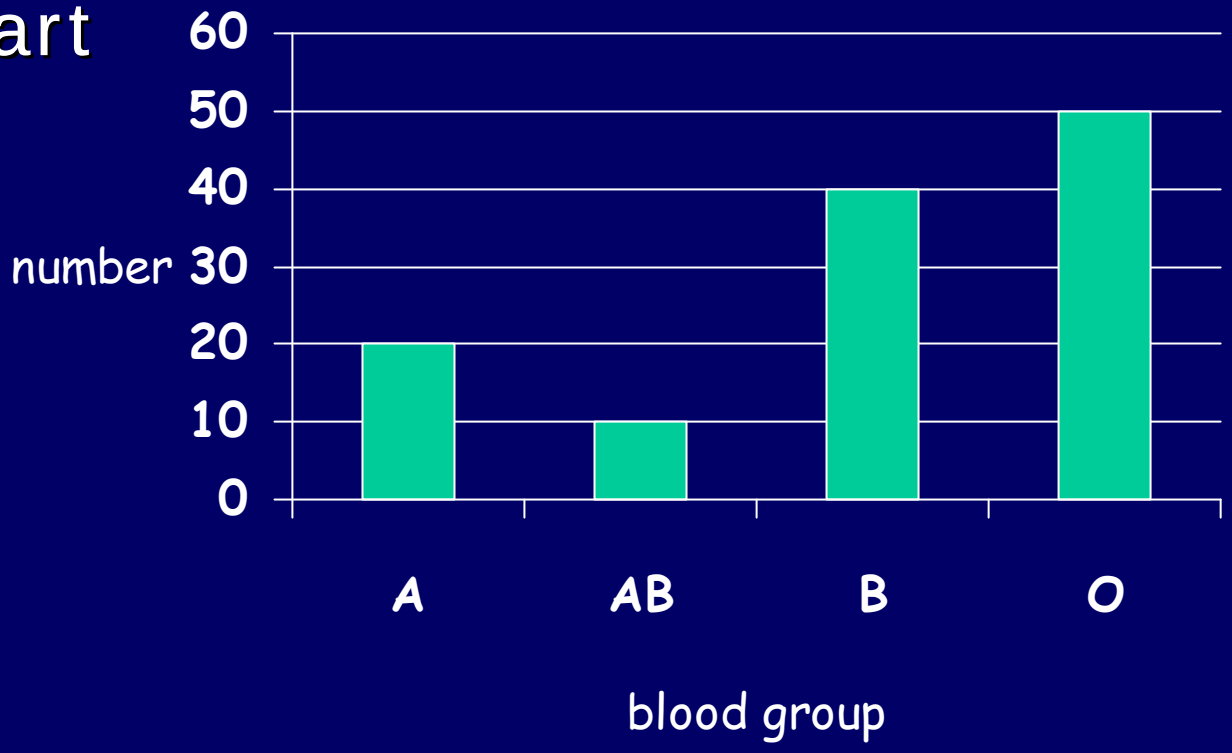


Gentlemen, I am sorry to say that a short graphic tells more than a long talk

Describing the Data (2)

Displaying **categorical data** graphically:

- Bar chart

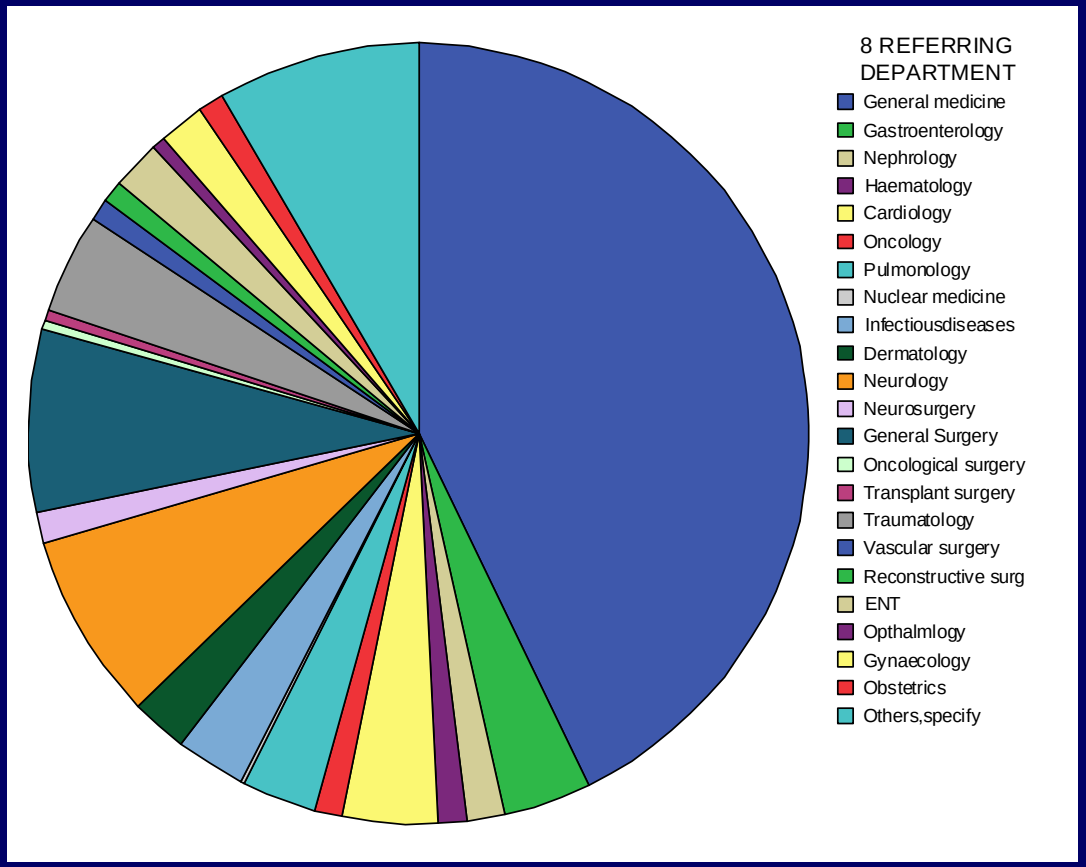


Petrie chapter 4: Displaying data graphically

Describing the Data (3)

Displaying **categorical data** graphically:

- Pie chart



Describing the Data (4)

☒ **Continuous Data** can be summarized as:

- Mean and Standard Deviation

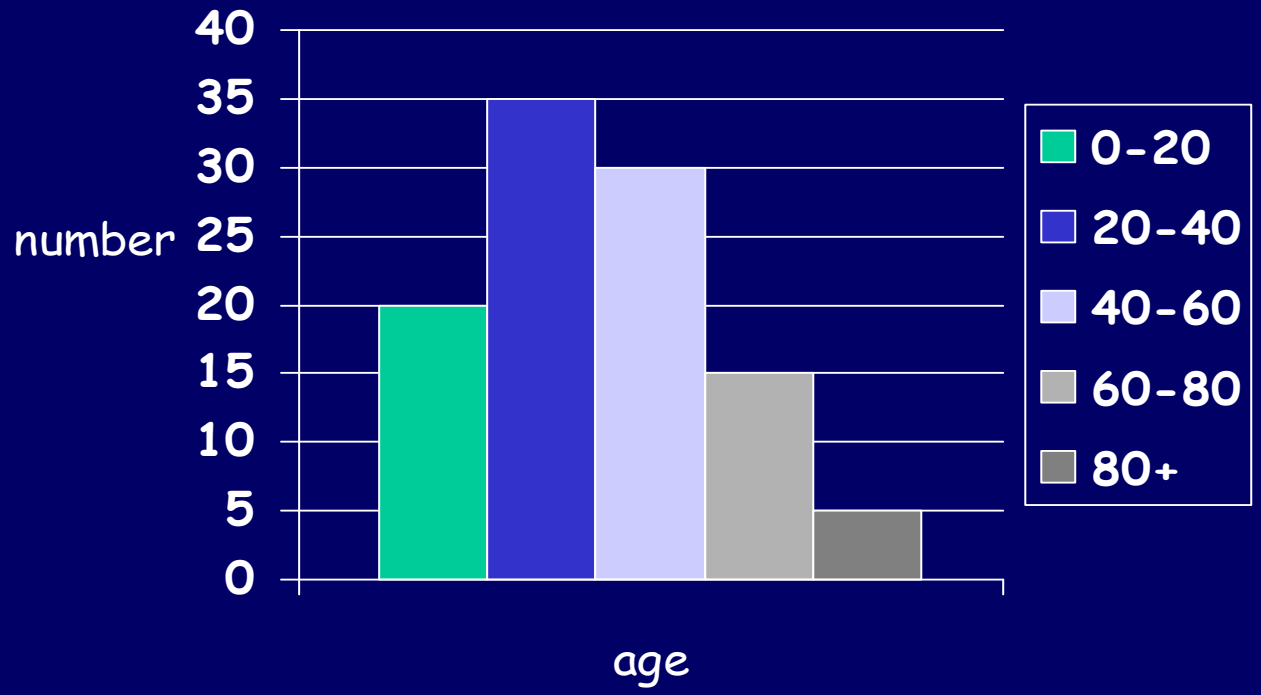
- Median and P25 to P75 / Range

- Mode

Describing the Data (5)

Displaying **Continuous data** graphically:

- Histogram

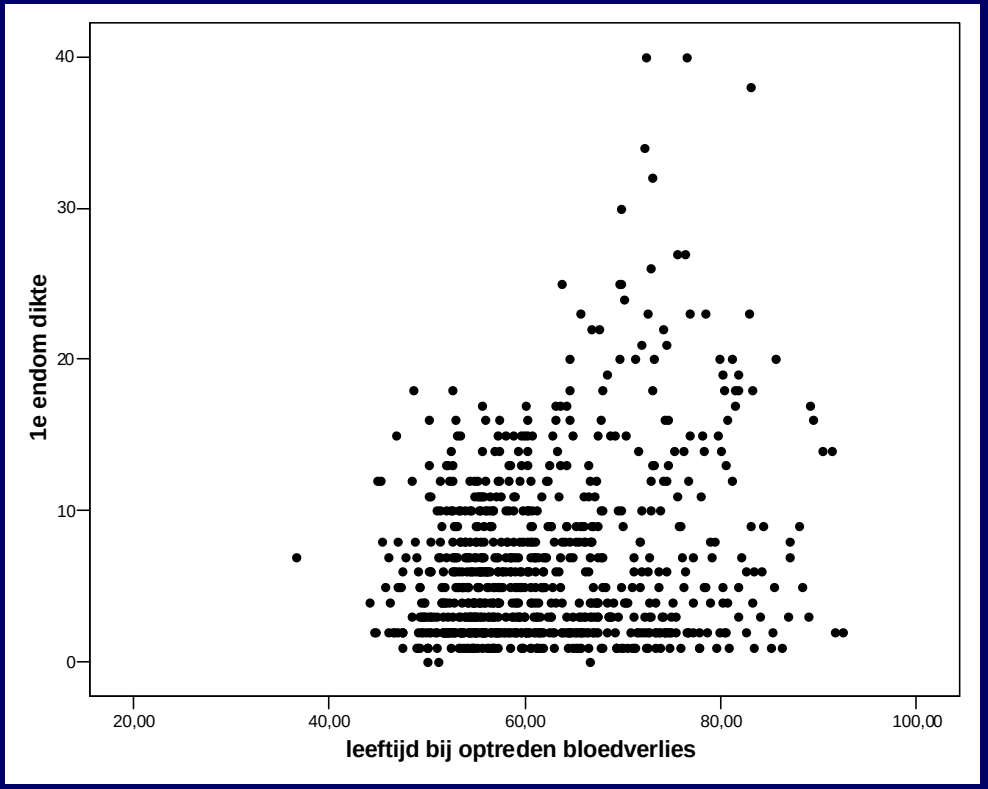


Petrie chapter 4: Displaying data graphically

Describing the Data (6)

Displaying **Continuous data** graphically:

- Scatter plot



Describing the Data (6)

☒ In continuous data: distribution of the data is key in determining which descriptive statistics one should use:

- Normal distribution
- Not normal distribution or skewed data

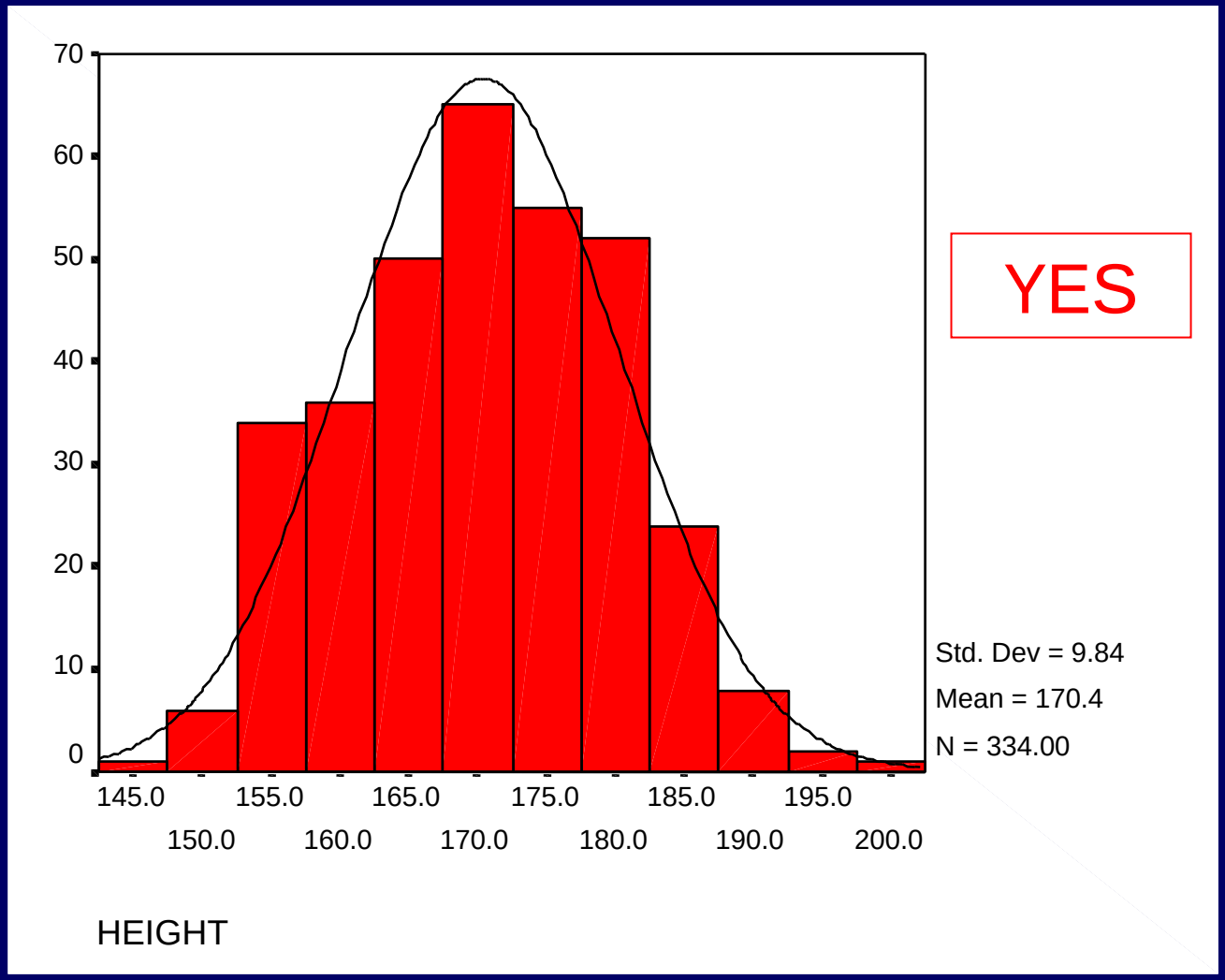
Petrie chapter 7: Theoretical distributions: the normal distribution

Normal distribution (1)

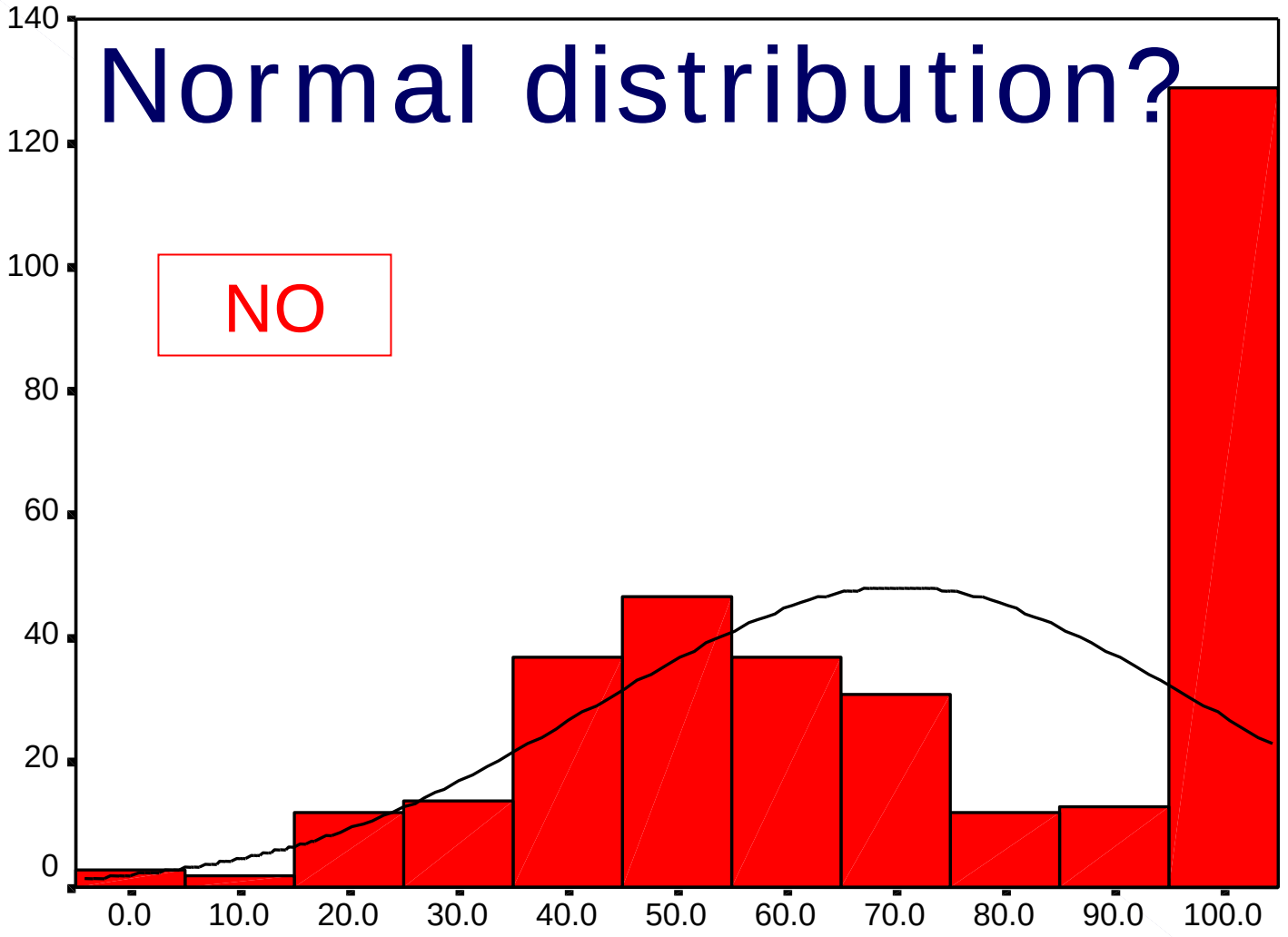
- ☒ Bell-shaped
- ☒ One top
- ☒ Symmetrical around its mean
- ☒ Mean and median are equal

Petrie chapter 7: Theoretical distributions: the normal distribution

Normal distribution?



Normal distribution?



Summarizing continuous data (1)

☒ Summarizing **normally distributed** data:

Mean: Adding up all the values and dividing this sum by the number of values

e.g.: 5 patients ages: 20, 22, 25, 28, 30

mean age = 125/5 = 25 yrs.

Petrie chapter 5: describing the data

Summarizing continuous data (2)

☒ Summarizing **not normally distributed** data:

Median: Arrange data in order of magnitude, starting with the smallest ending with the largest value, the Median is the middle value of this ordered set

e.g.: 5 patients age: 20, 22, 25, 63, 75

median age = 25 yrs

Summarizing continuous data (3)

☒ Summarizing **continuous** data:

mode: Value that occurs most frequently in the dataset

If each value occurs only once = there is no mode

Summarizing continuous data (3)

☒ Distinction between the mean & median:

Normally distributed data: e.g.: 5 patients age:

20, 22, 25, 28, 30 yrs

Mean age = 25 yrs; median age = 25 yrs

Not Normally distributed data:

20, 22, 25, 63, 78 yrs

Mean age = 42 yrs; median age = 25 yrs

Describing the data (6)

Mean

- advantage
  Uses all the data values
- Disadvantage
  Distorted by outliers and skewed data

Median

- advantage
  Not distorted by outliers or skewed data
- Disadvantage
  Less informative when distribution is normal

Spread (1)

☒ Mode: None

☒ Median:

- Range (min to max value)
  - disadvantage: distorted by outliers
- Percentiles:
  - 1% of the scores is below the first percentile (P1)
  - 25% of the scores are below the 25th percentile (P25) (quartiles)
  - 50% of the scores are below the 50th percentile (P50) (median)

Spread (2)

- Mean:
  - The extent to which each observation deviates from the mean, represents variation within the sample
  - Square each deviation and calculate the mean of these squared deviations: **Variance**
  - The square root of the variance is the **standard deviation**

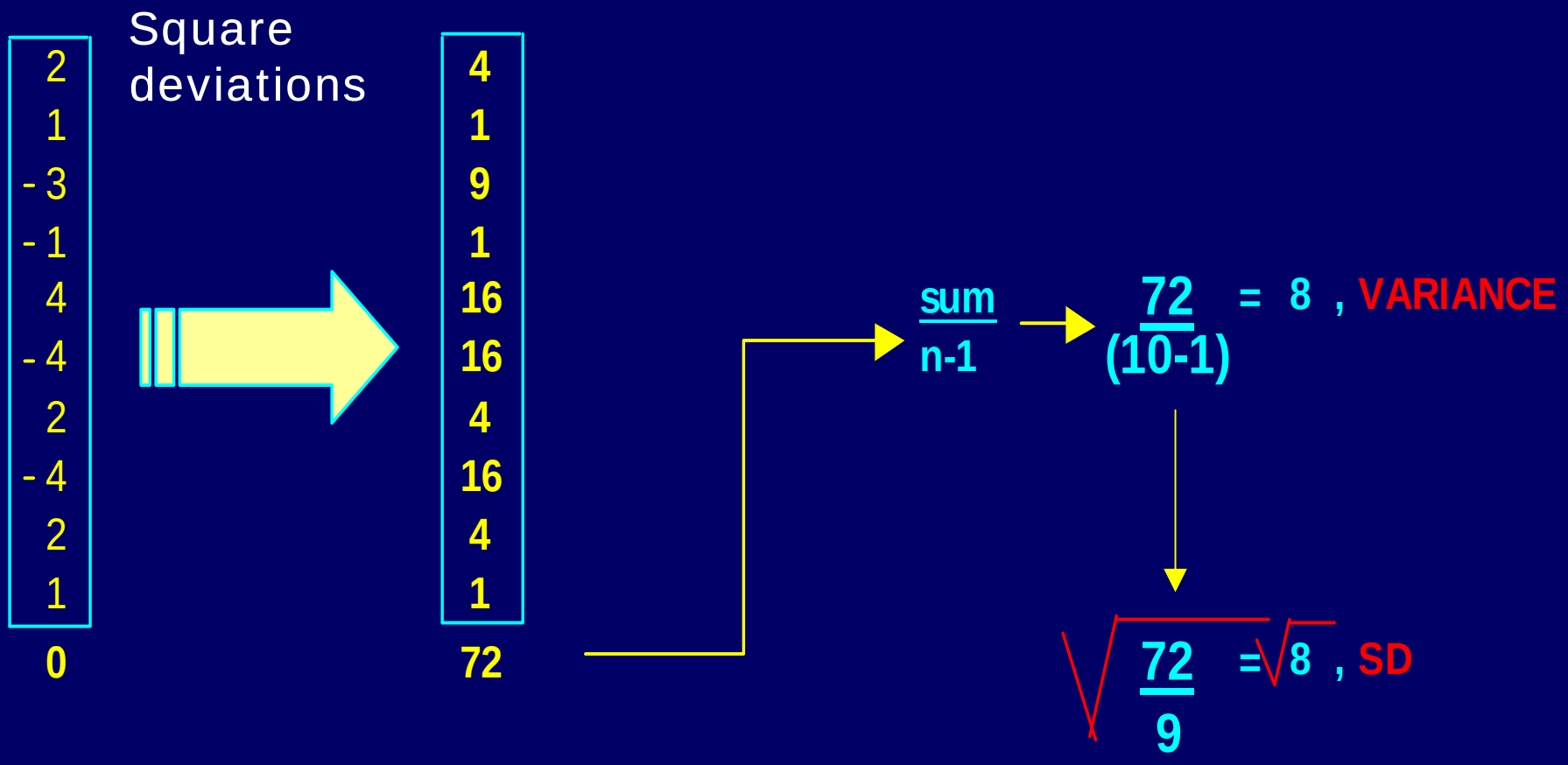
Spread (3)

By definition: mean of the deviation is zero

	value	mean	deviation
	32	30	2
	31	30	1
	27	30	-3
	29	30	-1
	34	30	4
	26	30	-4
	32	30	2
	26	30	-4
	32	30	2
	31	30	1
mean	30		0

By definition: mean of this deviation is zero, the positive differences exactly cancel out the negative differences

Spread(4)



Variance = s^2

Standard Deviation (SD) = s

Spread (5)

☒ Standard deviation: is the square root of the variance

$$\sqrt{\frac{\sum_{i=n}^{n}\left(X_i-\overline{X}\right)^2}{n-1}}$$

Variance = s^2

Standard Deviation (SD) = s

Spread (6)

Normal

- Mean
- Standard Deviation:
  - Uses every observation
  - Easy to interpret
  - Sensitive to outliers

Not Normals

- Median
- Ranges:
  - Easily determined,
  - But easily distorted by outliers
- Percentiles (IQR / P25 to P75):
  - Not distorted by outliers
  - Could be applied for skewed data (not normally distributed)

Keep in Mind:

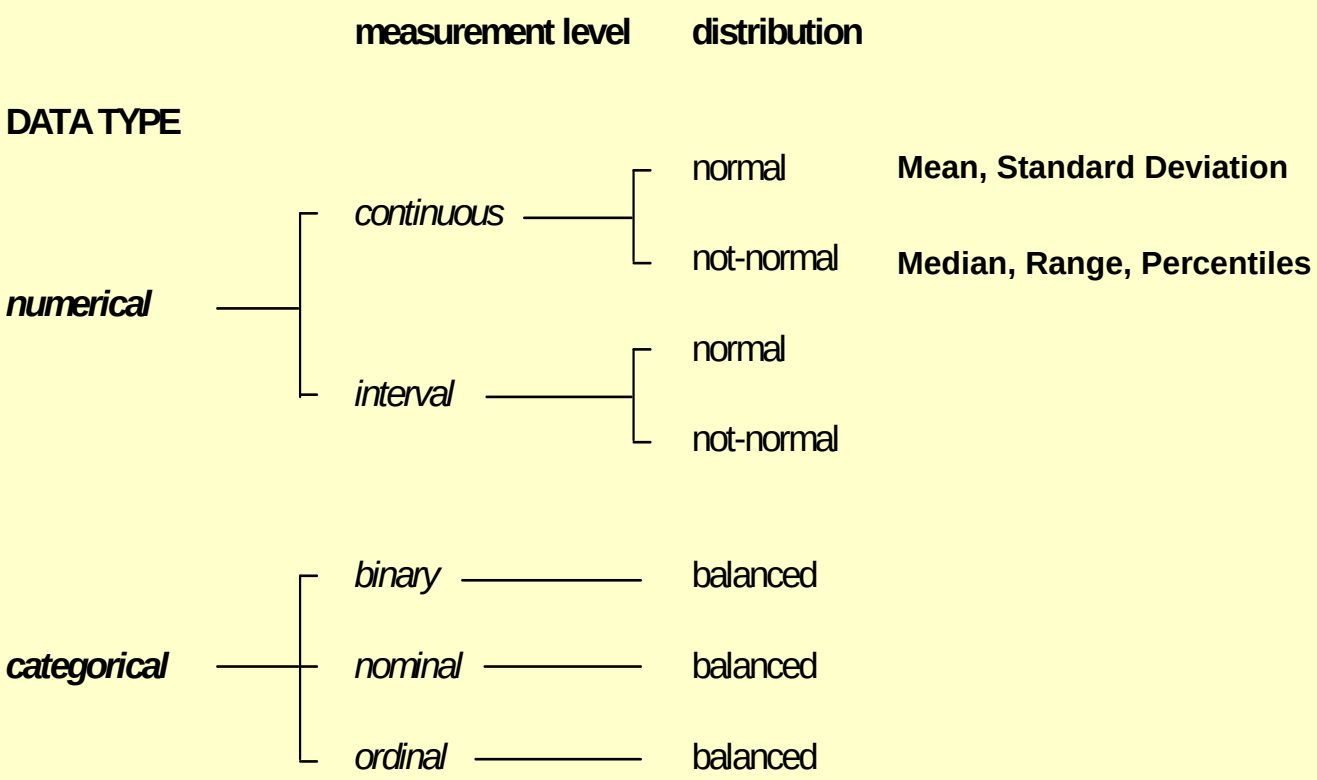
- ☒ For skewed data or not normally distributed: several transformations exist to transform the data into a Normal distribution: e.g. log-transformation, square root

  Petrie chapter 9: transformations

- ☒ Round off to the nearest two informative values

  0.4567 = 0.46

Summary


measurement level
distribution
DATA TYPE
numerical
continuous
normal
Mean, Standard Deviation
not-normal
Median, Range, Percentiles
interval
normal
not-normal
categorical
binary
balanced
nominal
balanced
ordinal
balanced

Introduction SPSS

3 types of files

1. Data file (*.SAV)
   - Spreadsheet (looks like excel)
2. Syntax file (*.SPS)
   - Contains your commands
3. Output file (*.SPO)
   - Result of the analysis

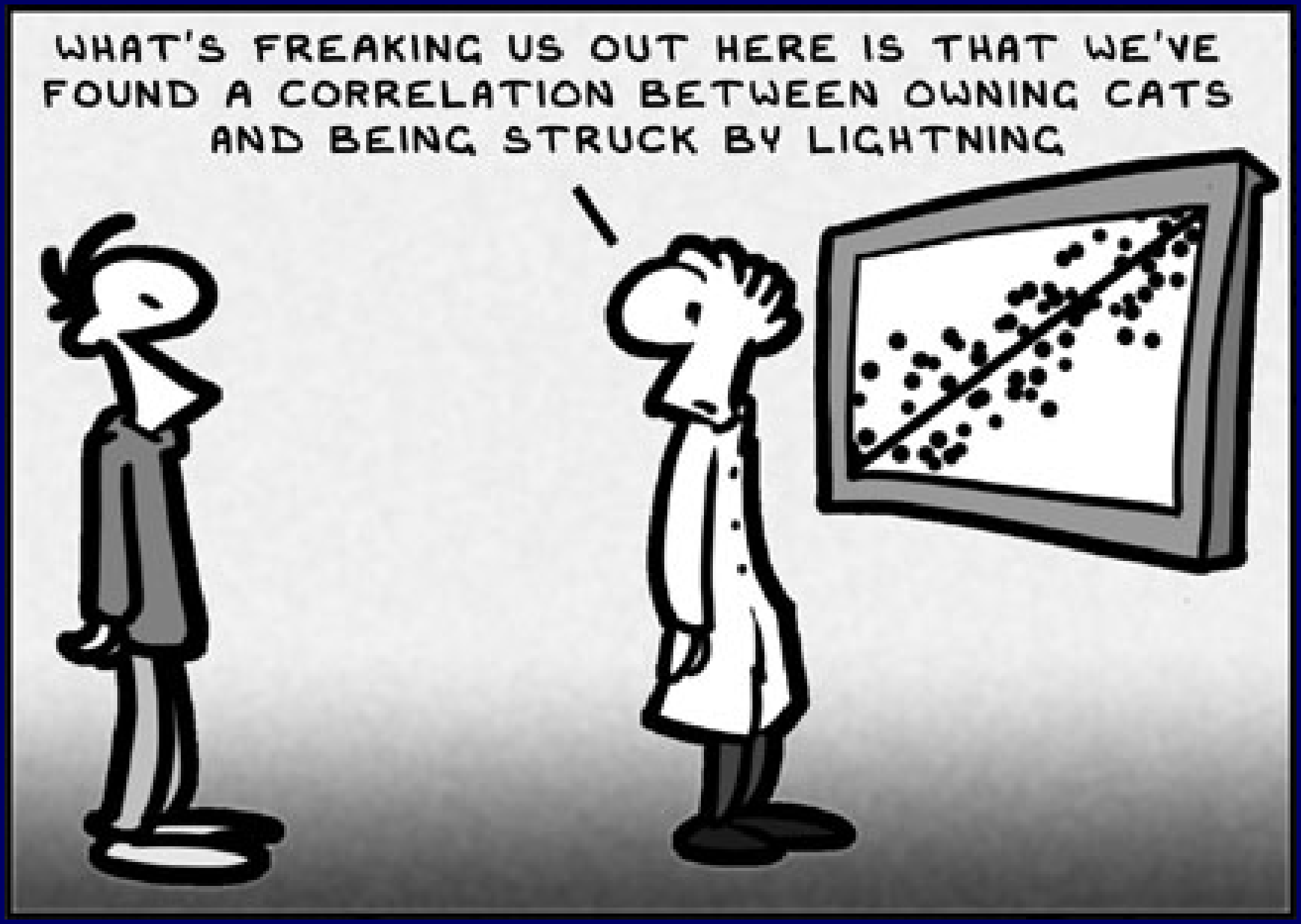
SPSS Demonstration

NB !!

☒ **SPSS will calculate anything,** including:

- A median for normal distributed data
- A mean for not normally distributed data

You are responsible for making sure that your results make sense


WHAT'S FREAKING US OUT HERE IS THAT WE'VE FOUND A CORRELATION BETWEEN OWNING CATS AND BEING STRUCK BY LIGHTNING

Excellent health statistics - smokers are less likely to die of age related illnesses.'