

INTRODUCTORY BIOSTATISTICS for BIOLOGISTS

Descriptive Statistics

Ana Luisa Papoila, Faculty of Medical Sciences, UNL and Center of Statistics and Applications, University of Lisbon

Fernanda Diamantino, Faculty of Sciences, Center of Statistics and Applications, University of Lisbon

Carina Silva-Fortes, Higher School of Technologies and Health of Lisbon and Center of Statistics and Applications, University of Lisbon

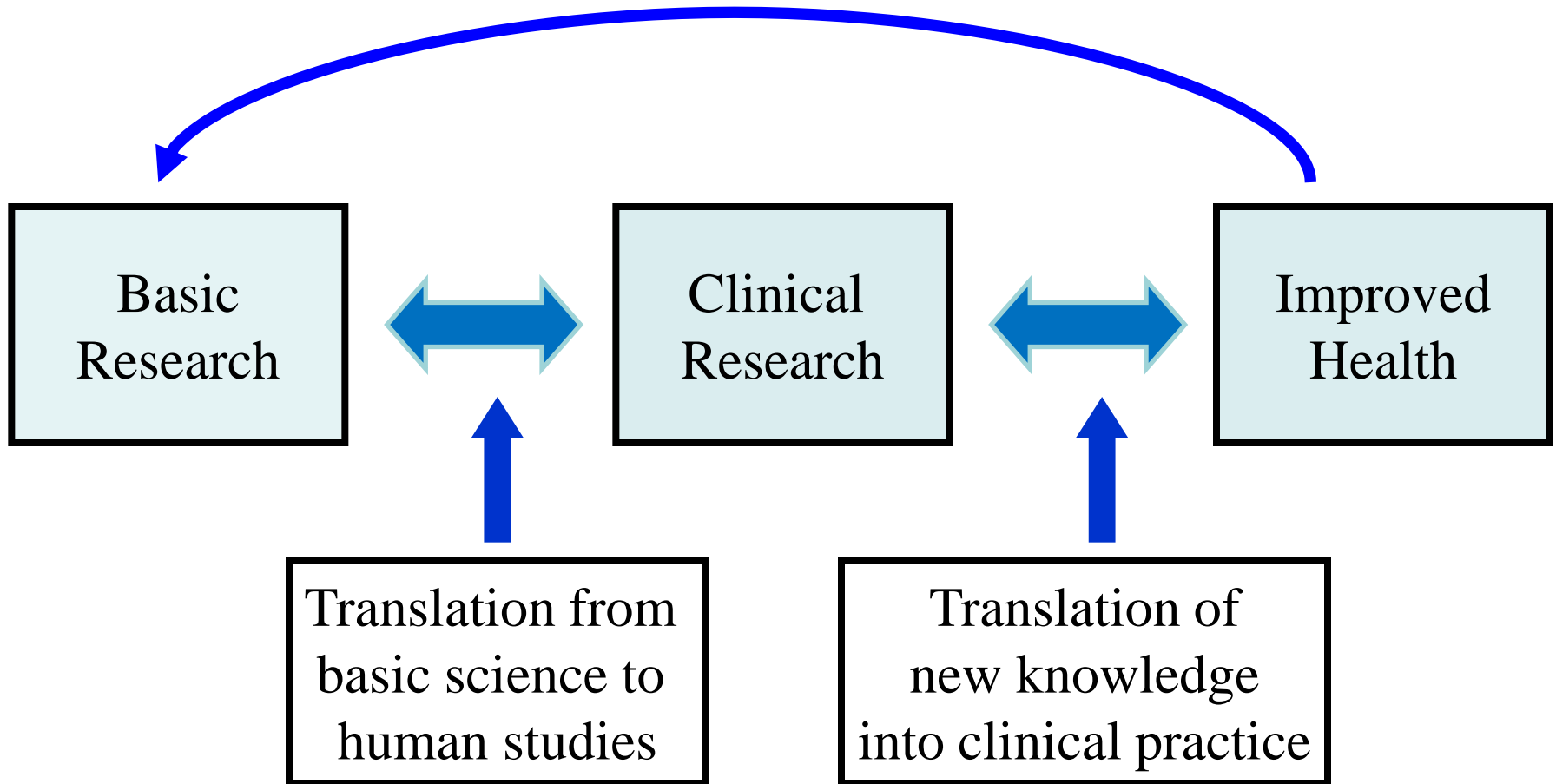
Planning a research study

- ✓ Identification of a research question
- ✓ Literature review
- ✓ Aim and objectives
- ✓ Methodology
 - Variables
 - Design
 - Identification of sources of bias and potential confounders
 - Sampling
 - ...
- ✓ Collection and data analysis
- ✓ Presenting results

Ethical issues: Confidentiality and Anonymity

Types of research studies

Basic and Clinical Research



Adapted from Sung et al. (2003) JAMA, 289, 1278-89.

Copyright © (2003) American Medical Association. All Rights reserved.

Control group

Research study example: experimental trial

Purpose of a control group

- To allow discrimination of subjects outcomes (for example, changes in symptoms, signs, or other morbidity) caused by the test “treatment” from outcomes caused by other factors, such as the natural progression of the disease, observer or subject expectations, or other “treatment”.
- The control group experiment tells us what would have happened to subjects if they had not received the test “treatment” or if they had received a different “treatment” known to be effective.

Control group

Research study example: experimental trial

A control group in a scientific experiment is a group where the factor being tested is not applied so that it may serve as a standard for comparison against another group where the factor is applied.



A substance that takes part in a biochemical reaction (e.g. blood-clotting factors) or a biological process (e.g. growth factors)

Research study example: experimental trial

A control group is one that is treated in exactly the same way as the experimental group except for the factor that is being investigated.

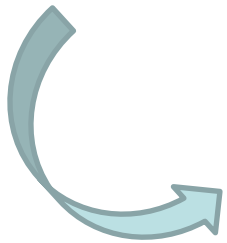
The test and control groups should be similar with regard to all baseline and on-treatment variables that could influence the outcome, except for the study “treatment” or factor.

Research study example: experimental trial

Testing plant fertilizer:

only half the plants in a garden will receive the fertilizer

Control group



the plants that receive no fertilizer: they establish the baseline level of growth that the fertilizer-treated plants will be compared against.

Without a control group, the experiment cannot determine whether the fertilizer-treated plants grow more than they would have if untreated.

Randomisation

Research study example: experimental trial

The test and control groups should be similar with regard to all baseline and on-treatment variables that could influence the outcome, except for the study “treatment” or factor.



Randomisation

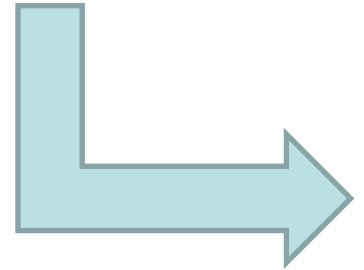
Research study example: experimental trial

Reasons for using **randomisation** to allocate treatments to subjects in a **controlled** trial:

- ✓ to prevent biases;
- ✓ to produce treatment groups in which the distribution of factors that may influence the outcome, known and unknown, are similar.

Bias

A systematic tendency of any factors associated with the design, conduct, analysis and evaluation of the results of an experimental trial to make the estimate of a “treatment” effect deviate from its true value.



Bias

Research study example: experimental trial

➤ Selection bias (very important!!!!)

Can result when the selection of subjects into a study leads to a result that is different from what you would have gotten if you had enrolled the entire target population

Selection bias refers to systematic differences between baseline characteristics of the groups that are compared.

In experimental trials, selection bias occurs when researchers recruit subjects in ways that abuse the data (interferences from researchers to assign the subjects into groups).

...

Bias

Research study example: experimental trial

➤ Performance bias

It refers to systematic differences between groups in the care that is provided, or in exposure to factors other than the interventions of interest (eliminated by using blinding).

➤ Assessment bias

It occurs in studies that rely on subject self-assessment or researcher assessment of subject status (eliminated by using blinding or objective outcome measures).

Accuracy and Precision

Accuracy and Precision

Refers to how close a measurement is to a “true” value.

Reproducibility

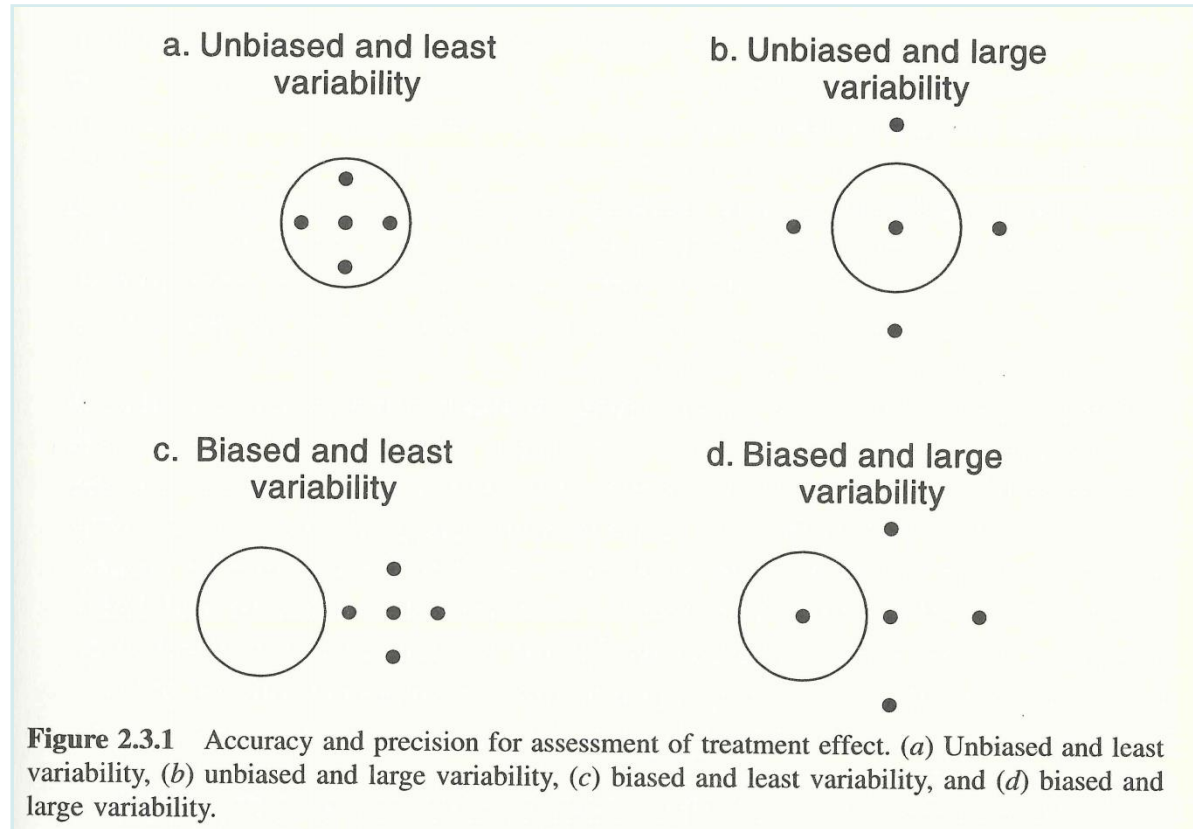
is the degree to which repeated measurements under unchanged conditions show the same results

is a measure of the spread of different readings;

refers to how closely individual measurements agree with each other;

“repeatable, reliable, getting the same measurement each time.”

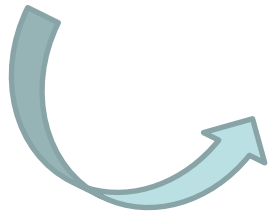
Accuracy and Precision



Some epidemiological concepts

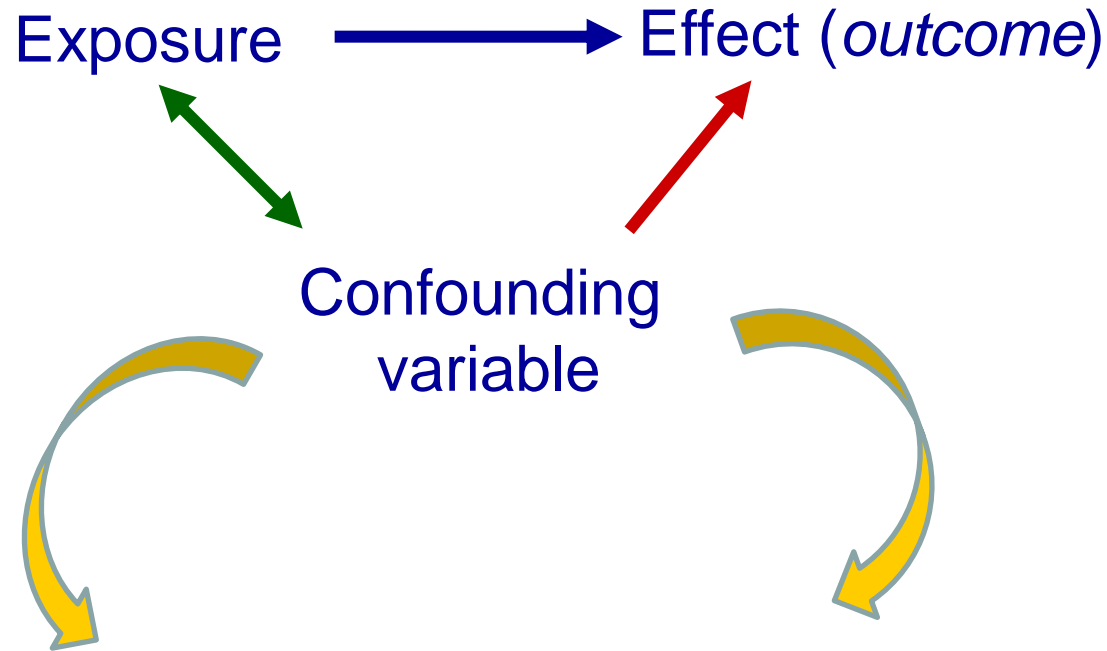
Confounding

A situation in which a measure of association or relationship between exposure and outcome is distorted by the presence of another variable.



Some epidemiological concepts

Confounding

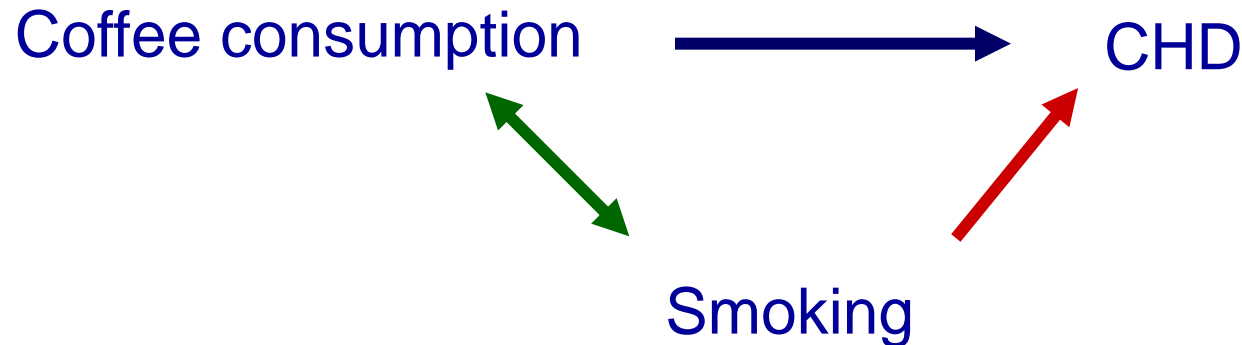


Must not be a consequence of the exposure

Must be associated to the outcome **independently** of the exposure

Some epidemiological concepts

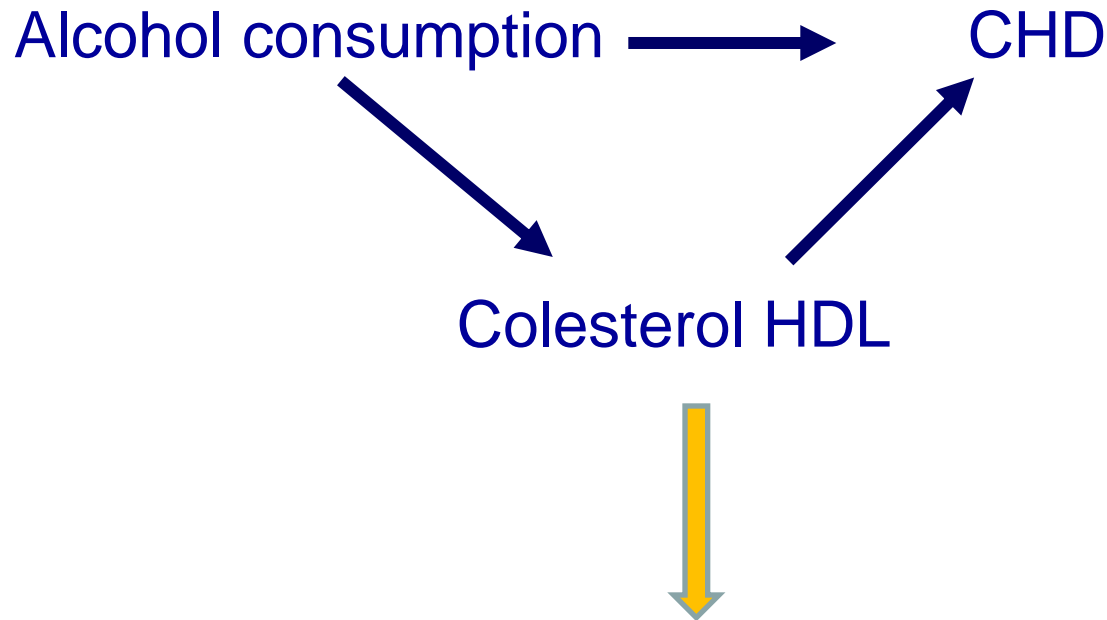
Confounding



- smoking is associated with coffee consumption and with CHD
- it is not a consequence of coffee consumption
- it is a risk factor even for those who do not drink coffee.

Some epidemiological concepts

Confounding



Cholesterol HDL is in the causal path between alcohol consumption and CHD.

Some epidemiological concepts

Confounding

How to control confounding?

Study design:

- Randomization
- Matching
- Restriction

Data analysis:

- Stratification
- Multivariable regression analysis

Some epidemiological concepts

Control variables (also known as "constant variables")

In the design of experiments and data analysis, control variables are those variables that are not changed throughout the trials in an experiment because the experimenter is not interested in the effect of that variable being changed for that particular experiment. Control variables are extraneous factors, possibly affecting the experiment, that are kept constant so as to minimize their effects on the outcome.

Example:

keeping the pressure constant in an experiment designed to test the effects of temperature on bacterial growth

Control variables, if not monitored, may lead to confounding

Some epidemiological concepts

Interaction

There is **interaction** when the difference between the levels of a factor depends from the level of other factors.

Effect modification occurs in biomedical research when a measure of statistical association between an exposure and an outcome differs according to the levels of a third variable—the **effect modifier**.

Biological interaction between two causes occurs if the effect of one depends on the presence of the other.

Some epidemiological concepts

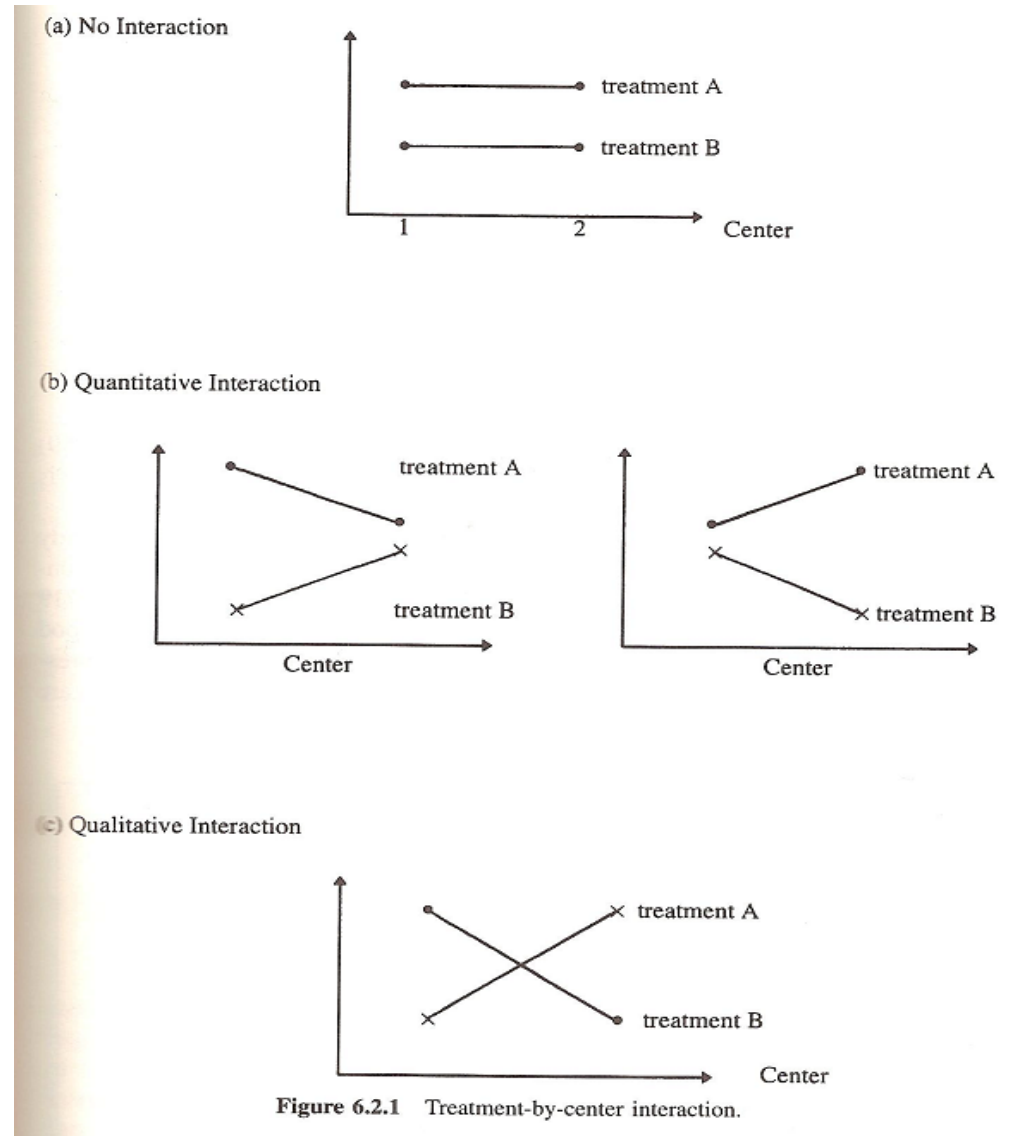
Interaction

Quantitative interaction:

The magnitude of the treatment effect is not the same across the levels of other factors but the direction of the treatment remains the same for all levels of other factors.

Qualitative interaction:

The direction of the treatment effect changes in some levels of other factors.



Descriptive Statistics

Basic concepts:

Statistical Analysis typically involves data from a sample to make inferences about the characteristics of a population.

Population: Can be defined as including all items with the characteristic one wishes to understand (also called universe).

Sample: Subset of the population, often chosen randomly and preferably representative of the population as a whole.

Descriptive Statistics

Basic concepts:

Variable: is a characteristic or condition that changes or has different values for different individuals.

Random Variable: is a function that assigns a real number $X(s)$, to each outcome s in a sample. Random Variables are represented by capital letters (*e.g.* X_1, X_2).

Observation (or case): is a concretization of a variable. For example, the weight of a randomly chosen rat is such an observation (observations are represented by lowercase, *e.g.* x_1, x_2).

Descriptive Statistics

Basic concepts:

Parameter: is a numeric quantity, usually unknown, that describes a certain population characteristic. For example, the population mean, μ , is a parameter that is often used to indicate the mean value of a quantity (usually are represented by Greek letters, *e.g.* μ , σ .)

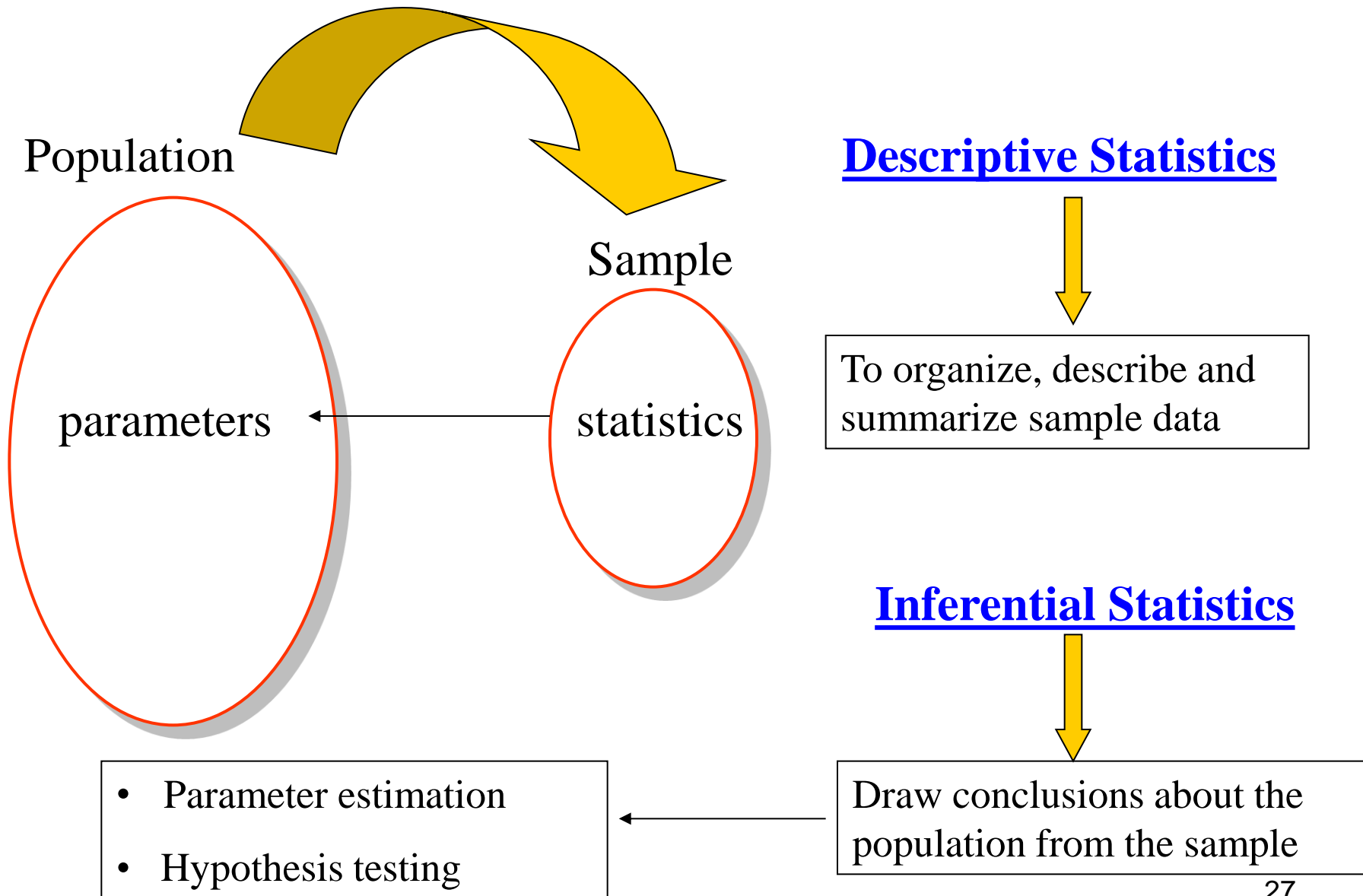
Descriptive Statistics

Basic concepts:

Estimator/Statistic: is a statistic (that is, a function of the data) that is used to infer the value of an unknown parameter in a statistical model.

Estimate/Statistic: An estimate is an indication of the value of an unknown quantity based on observed data. More formally, **an estimate is the particular value of an estimator that is obtained from a particular sample of data and used to indicate the value of a parameter.**

Descriptive Statistics



Descriptive Statistics

Types of data:

Qualitative: data describing the attributes or properties of the data (it can't be measured).

Examples:

- marital status
- stages of a disease
- eye color
- hair color
- ...

Descriptive Statistics

Types of data:

Quantitative: data that can be expressed numerically (it can be measured).

Examples:

- temperature
- height
- weight
- ...

Descriptive Statistics

Types of data:

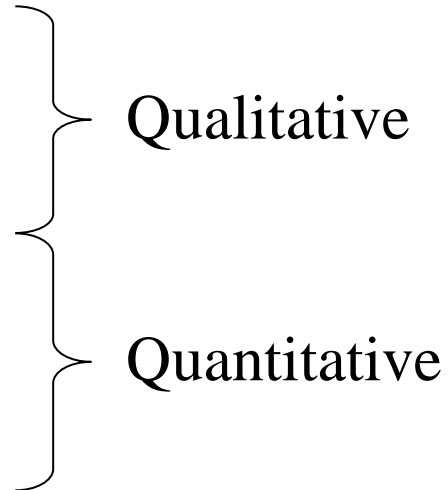
Measurement scales

✓ **Nominal**

✓ **Ordinal**

✓ **Intervalar**

✓ **Ratio**



Descriptive Statistics

Types of data:

Measurement scales

Nominal:

It is classified into categories that describe the characteristic of interest. Each value on the measurement scale has a unique meaning.

Examples:

- ❖ eye color
- ❖ gender
- ❖ marital status

Descriptive Statistics

Types of data:

Measurement scales

Ordinal:

Each value on the measurement scale has a unique meaning and have an ordered relationship to one another.

Examples:

- ❖ socioeconomic status
- ❖ stages of a disease
- ❖ military positions

Descriptive Statistics

Types of data:

Measurement scales

Interval:

It is an ordinal scale in which intervals have the same interpretation throughout. The zero point is arbitrary.

Example:

Fahrenheit scale of temperature:

The difference between 10 degrees and 30 degrees represents the same temperature difference as the difference between 50 degrees and 70 degrees: this is because each 20-degree interval has the same physical meaning

Descriptive Statistics

Types of data:

Measurement scales

Ratio:

It is an interval scale with the additional property that its zero position indicates the absence of the quantity being measured (**a minimum value of zero**).

OBS: The scale has a true zero point below which no values exist.

Example:

Height  A height of 180 cm doubles a height of 90 cm

Descriptive Statistics

Types of data:

Discrete variable:



A variable that can take a finite or infinite countable number of numerical values.

Example:

Number of ...

Obs: nominal and ordinal variables are discrete

Descriptive Statistics

Types of data:

Continuous variable:



A variable that can take an infinite number of values.

Examples:

Height, age, weight, urea levels ...

Descriptive Statistics

Types of data:

Exercise:

<input type="checkbox"/> Gender	Qualitative, nominal
<input type="checkbox"/> Heart pressure	Quantitative, ratio
<input type="checkbox"/> Temperature (Centigrade)	Quantitative, interval
<input type="checkbox"/> N° of visits to emergency	Quantitative, ratio
<input type="checkbox"/> Illness severity	Qualitative, ordinal

Descriptive Statistics

Types of data:

Exercise:

☐ **Urea levels**

Quantitative, ratio

☐ **N° of white blood cells**

Quantitative, ratio

☐ **Glasgow score**

Qualitative, ordinal

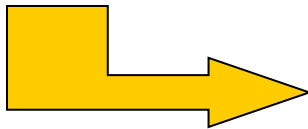
☐ **Potassium levels**

Quantitative, ratio

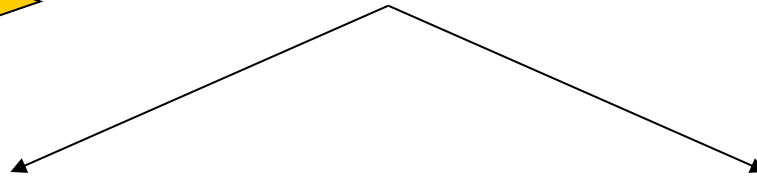
Descriptive Statistics

Organizing data:

Discrete data:



Frequency table



Absolute frequency of each
different value

Relative frequency =

$$\frac{\text{absolute frequency}}{\text{sample size}}$$

Descriptive Statistics

Organizing data:

Example: response to a treatment

response	good	moderate	poor
Nº patients	50	36	15

Frequency table

classes	Abs. freq.	Rel. freq.	Cum. freq.
god	50	0.495	0.495
moderate	36	0.356	0.851
poor	15	0.149	1.000
Total	101	1.000	

Descriptive Statistics

Organizing data:

Continuous data:



Frequency distribution



grouping data into classes

1. Find the range
2. Select the number of classes (e.g. Sturges' Rule)
3. Determine the class width
4. Define the class limits
5. Count the number of observations in each class
6. Calculate relative and cumulative frequencies of each class

Descriptive Statistics

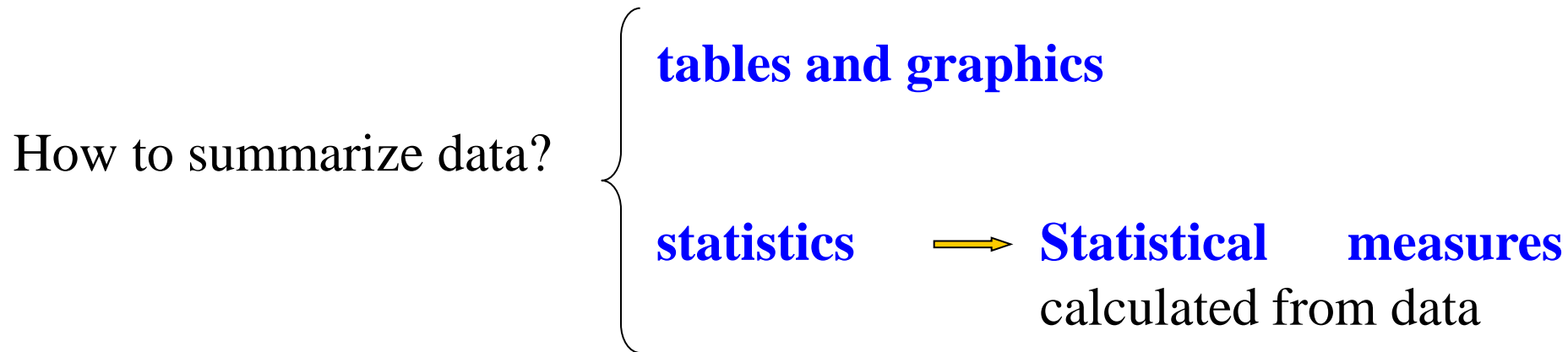
Organizing data:

Example:

Changes in serum cholesterol levels after administration of a drug which aims to reduce cholesterol levels.

Classes	Abs. freq.	Rel. freq.
[-100.5, -80.5[1	0.01
[-80.5, -60.5[6	0.04
[-60.5, -40.5[16	0.10
[-40.5, -20.5[31	0.20
[-20.5, -0.5[40	0.26
[-0.5, 19.5[43	0.28
[19.5, 39.5[16	0.10
[39.5, 59.5[3	0.02
Total	156	1

Descriptive Statistics



Types of statistical measures:

- measures of location
- measures of dispersion (variability)
- measures of shape

Descriptive Statistics

Measures of location

❖ **Sample Mean** $\longrightarrow \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

❖ **Mode**

is the value that occurs most frequently in the data. The mode is a useful way to summarize qualitative data.

❖ **Median**

is the value in the middle when observations are arranged in ascending order. With an odd number of observations, the median is the middle value. With an even number of observations the median is the average of the two middle observations.

Descriptive Statistics

Measures of location

❖ Quantiles

Each of any set of values of a variable which divides a frequency distribution into equal groups, each containing the same fraction of the total population.

Descriptive Statistics

Measures of location

❖ Quantiles - **Examples**

Quartiles (Q):

one of the values of a variable that divides the distribution of the variable into four groups having equal frequencies

Deciles (D):

one of the values of a variable that divides the distribution of the variable into ten groups having equal frequencies

Percentiles (P):

one of the values of a variable that divides the distribution of the variable into 100 groups having equal frequencies

Descriptive Statistics

Measures of location

- First (lower) decile = 10th percentile (P_{10})
- First (lower) quartile, Q_1 = 25th percentile (P_{25})
- Second (middle) quartile, Q_2 = 50th percentile (P_{50})
- Third quartile, Q_3 = 75th percentile (P_{75})
- Ninth (upper) decile = 90th percentile (P_{90})

Descriptive Statistics

Measures of location

the mean is not a resistant measure of central tendency



It is influenced by extremely high or low data values (*outliers*)



alternative

median

Descriptive Statistics

Measures of location

mean or median?

- when the distribution is symmetric, mean and median are equal
- the median is not so sensitive to *outliers*
- the mean reflects the value of all the observations

Descriptive Statistics

Measures of dispersion

Are statistical measures that summarise the amount of spread or variation in the distribution of values in a variable.

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

❖ sample standard deviation $\longrightarrow s$

❖ ranges (*e.g.* range, interquartile range)



Difference between the maximum
and the minimum



Difference between Q_3 and Q_1

Descriptive Statistics

Measures of dispersion

the standard deviation is not a resistant measure



It is influenced by extremely high or low data values (*outliers*)



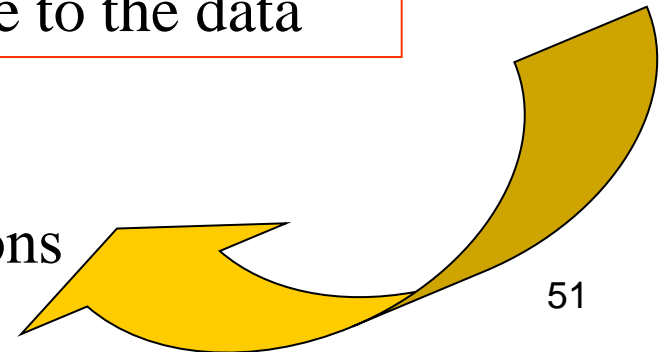
alternative

Interquartile
range

The interquartile range is a **more resistant** measure than the standard deviation because it is not so sensitive to the data

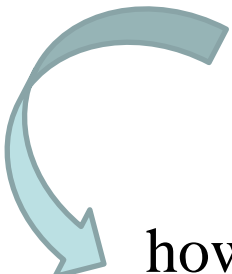
however

It doesn't reflect the value of all the observations



Descriptive Statistics

Measure of relative dispersion


$$\text{coefficient of variation} = \frac{\text{standard deviation}}{\text{mean}} \times 100\%$$

how big is the SD relative to the mean?

- the value of the CV is independent of the unit in which the measurement has been taken, so it is a dimensionless (unitless) number
- useful for comparing the variability of data with different units or widely different means

Descriptive Statistics

Concluding:

When the mean is used as a location measure  use the standard deviation as the dispersion measure

When the median is used as a location measure  use the interquartile range as the dispersion measure

Descriptive Statistics

Measures of Shape

❖ **Skewness** measures departure from symmetry

Positively skewed (skewed to the right)
Symmetric
Negatively skewed (skewed to the left)

❖ **Kurtosis** measures the “peakedness” of a distribution

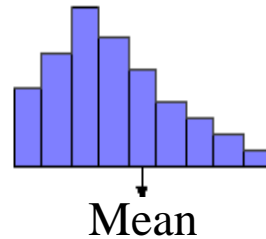
Leptokurtic
Mesokurtic
Platykurtic

Descriptive Statistics

Measures of Shape - Skewness

Positively skewed

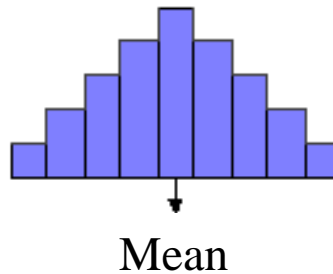
(mean > median > mode)



Skewness coefficient > 0

Symmetric

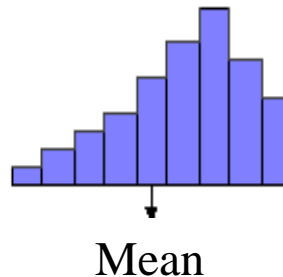
(mean = median = mode)



Skewness coefficient = 0

Negatively skewed

(mean < median < mode)

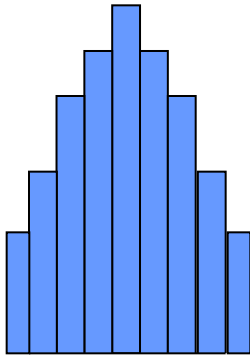


Skewness coefficient < 0

Descriptive Statistics

Measures of Shape - Kurtosis

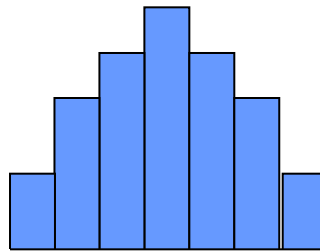
Leptokurtic



Kurtosis
coefficient > 0

sharper than a Gaussian
distribution, with values
concentrated around the mean

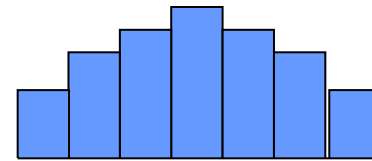
Mesokurtic



Kurtosis
coefficient $= 0$

like Gaussian
distribution

Platykurtic



Kurtosis
coefficient < 0

flatter than a Gaussian
distribution with a wider peak

Descriptive Statistics

Graphing data:

Discrete data  *e.g.* Barplots

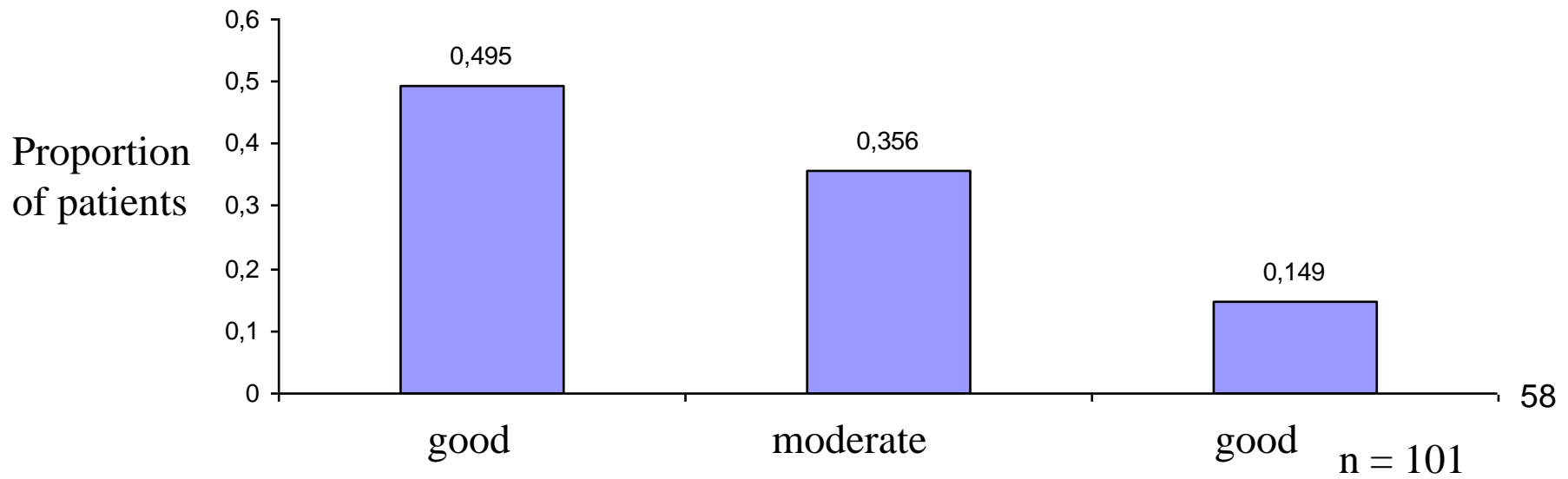
- A barplot draws a bar with a height proportional to the count in the table. The height could be given by the frequency, or by the proportion. The graph will look the same, but the scales may be different.
- The bars are not contiguous (touching one another) nor do the areas of the strips have a meaning; rather, the heights of the rectangles are proportional to the frequency.

Descriptive Statistics

Graphing data:

Example: response to a treatment

response	good	moderate	poor
N° patients	50	36	15



Descriptive Statistics

Graphing data:

Continuous data



Histogram



It reflects the shape of the population distribution

Example:

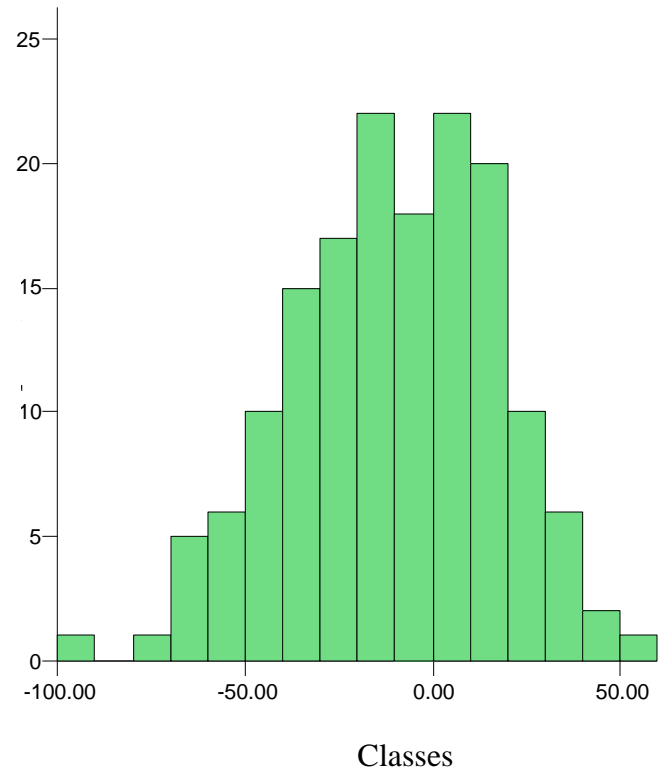
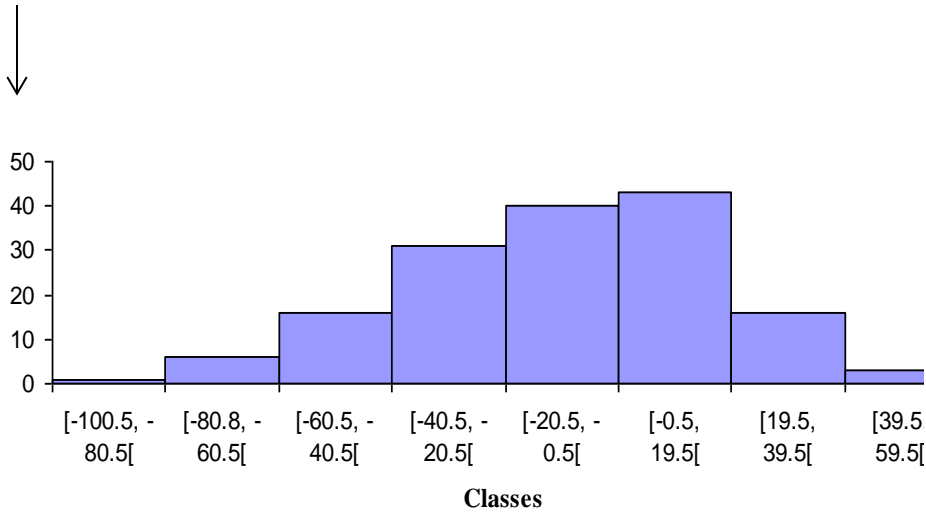
Changes in serum cholesterol levels after administration of a drug which aims to reduce cholesterol levels.

Descriptive Statistics

Graphing data:

Histogram

Absolute frequency



n = 156

Descriptive Statistics

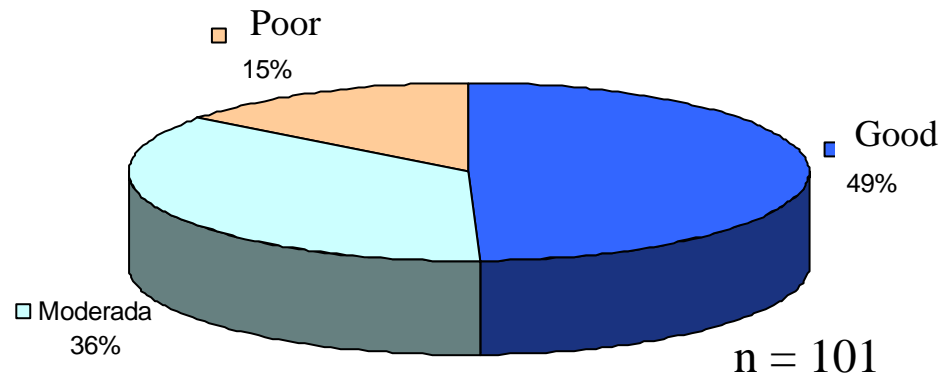
Graphing data:

Pie chart:

Each class corresponds to each slice, and the slice areas are proportional to the frequencies.

Example:

response to a treatment



Descriptive Statistics

Graphing data:

Error bars:

- Figures with error bars can, if used properly, give information describing the data (descriptive statistics), or information about what conclusions, or inferences, are justified (inferential statistics).
- These two basic categories of error bars are depicted in exactly the same way, but are actually fundamentally different.

Descriptive Statistics

Graphing data:

Error bars:

Table I. Common error bars

Error bar	Type	Description	Formula
Range	Descriptive	Amount of spread between the extremes of the data	Highest data point minus the lowest
Standard deviation (SD)	Descriptive	Typical or (roughly speaking) average difference between the data points and their mean	$SD = \sqrt{\frac{\sum (X - M)^2}{n - 1}}$
Standard error (SE)	Inferential	A measure of how variable the mean will be, if you repeat the whole study many times	$SE = SD/\sqrt{n}$
Confidence interval (CI), usually 95% CI	Inferential	A range of values you can be 95% confident contains the true mean	$M \pm t_{[n-1]} \times SE$, where $t_{[n-1]}$ is a critical value of t . If n is 10 or more, the 95% CI is approximately $M \pm 2 \times SE$.

Descriptive Statistics

Graphing data:

Descriptive error bars:

Range and standard deviation (SD) are used for descriptive error bars because they show how the data are spread.

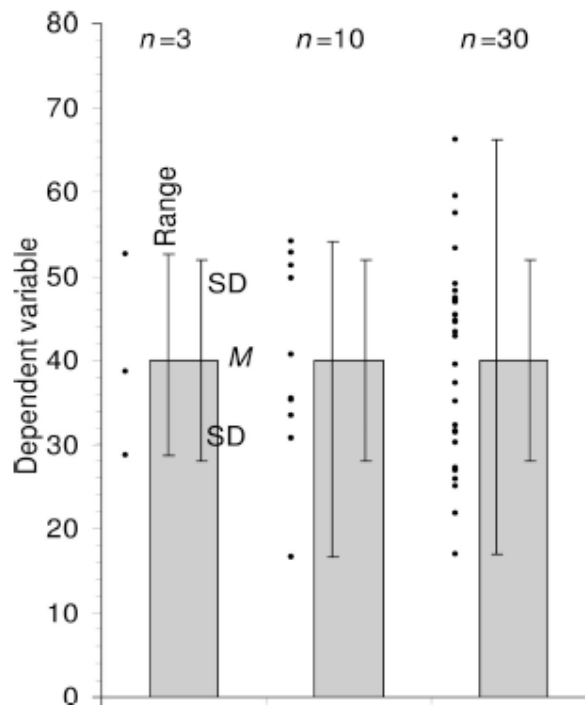


Figure 1. **Descriptive error bars.** Means with error bars for three cases: $n = 3$, $n = 10$, and $n = 30$. The small black dots are data points, and the column denotes the data mean M . The bars on the left of each column show range, and the bars on the right show standard deviation (SD). M and SD are the same for every case, but notice how much the range increases with n . Note also that although the range error bars encompass all of the experimental results, they do not necessarily cover all the results that could possibly occur. SD error bars include about two thirds of the sample, and $2 \times SD$ error bars would encompass roughly 95% of the sample.

Descriptive Statistics

Graphing data:

Inferential error bars:

In experimental biology it is more common to be interested in comparing samples from two groups, to see if they are different.

For example, you might be comparing wild-type mice with mutant mice, or drug with placebo, or experimental results with controls. To make inferences from the data (*i.e.*, to make a judgment whether the groups are significantly different, or whether the differences might just be due to random fluctuation or chance), a different type of error bar can be used. These are **standard error (SE)** bars and **confidence intervals (CIs)** bars.

Descriptive Statistics

Graphing data:

Inferential error bars:

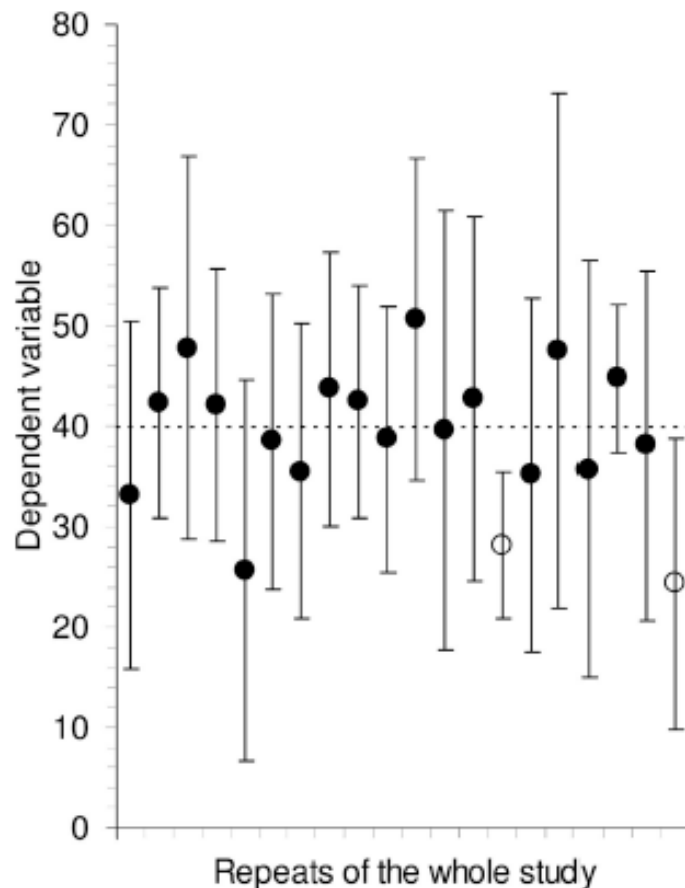


Figure 2. **Confidence intervals.** Means and 95% CIs for 20 independent sets of results, each of size $n = 10$, from a population with mean $\mu = 40$ (marked by the dotted line). In the long run we expect 95% of such CIs to capture μ ; here 18 do so (large black dots) and 2 do not (open circles). Successive CIs vary considerably, not only in position relative to μ , but also in length. The variation from CI to CI would be less for larger sets of results, for example $n = 30$ or more, but variation in position and in CI length would be even greater for smaller samples, for example $n = 3$.

Descriptive Statistics

Graphing data:

Error bars:

- **Rule 1:** when showing error bars, always describe in the figure legends what they are.
- **Rule 2:** the value of n (i.e., the sample size, or the number of independently performed experiments) must be stated in the figure legend. It is essential that n (the number of independent results) is carefully distinguished from the number of replicates, which refers to repetition of measurement on one individual in a single condition, or multiple measurements of the same or identical samples.

Descriptive Statistics

Graphing data:

Error bars - replicates:

Consider trying to determine whether deletion of a gene in mice affects tail length. We could choose one mutant mouse and one wild type, and perform 20 replicate measurements of each of their tails. We could calculate the means, SDs, and SEs of the replicate measurements, but these would not permit us to answer the central question of whether gene deletion affects tail length, because n would equal 1 for each genotype, no matter how often each tail was measured.

For replicates, $n = 1$, and it is therefore inappropriate to show error bars or statistics.

Descriptive Statistics

Graphing data:

Error bars - replicates:

Rule 3: error bars and statistics should only be shown for independently repeated experiments, and never for replicates. If a "representative" experiment is shown, it should not have error bars or p-values, because in such an experiment, $n = 1$.

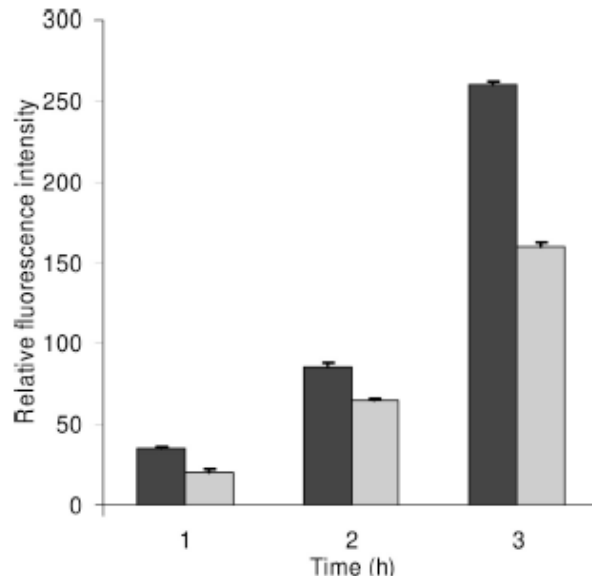


Figure 3. **Inappropriate use of error bars.** Enzyme activity for MEFs showing mean + SD from duplicate samples from one of three representative experiments. Values for wild-type vs. $-/-$ MEFs were significant for enzyme activity at the 3-h timepoint ($P < 0.0005$). This figure and its legend are typical, but illustrate inappropriate and misleading use of statistics because $n = 1$. The very low variation of the duplicate samples implies consistency of pipetting, but says nothing about whether the differences between the wild-type and $-/-$ MEFs are reproducible. In this case, the means and errors of the three experiments should have been shown.

Descriptive Statistics

Graphing data:

Error bars:

Rule 4: because experimental biologists are usually trying to compare experimental results with controls, it is usually appropriate to show inferential error bars, such as SE or CI, rather than SD. However, if n is very small (for example $n = 3$), rather than showing error bars and statistics, it is better to simply plot the individual data points.

Descriptive Statistics

Graphing data:

Error bars - some considerations:

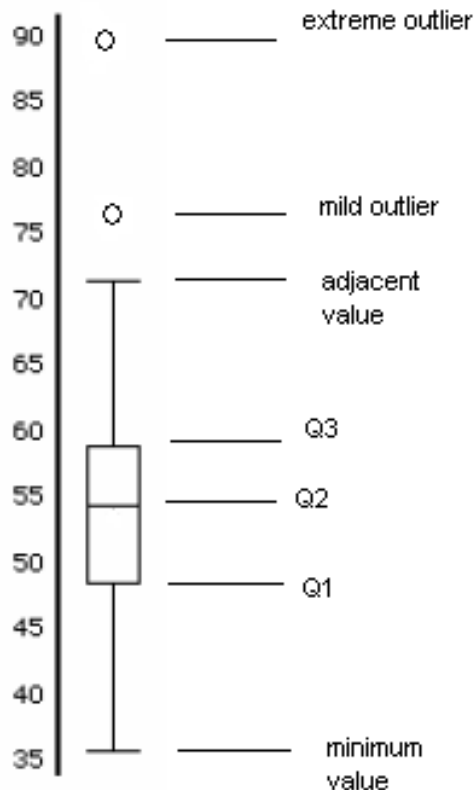
When first seeing a figure with error bars, ask yourself: "What is n ? "Are they independent experiments, or just replicates?" and, "What kind of error bars are they?"

If the figure legend gives you satisfactory answers to these questions, you can interpret the data, but remember that error bars and other statistics can only be a guide: you also need to use your biological understanding to appreciate the meaning of the numbers shown in any figure.

Descriptive Statistics

Graphing data:

Boxplot:



Interquartile Range: $IQR = Q_3 - Q_1$

Inner fences: $\{Q_1 - 1.5(IQR), Q_3 + 1.5(IQR)\}$

Outer fences: $\{Q_1 - 3(IQR), Q_3 + 3(IQR)\}$

This is a pictorial display that provides the main descriptive measures of the measurement set:

L - the largest measurement inside the inner fences

Q_3 - The upper quartile

Q_2 - The median

Q_1 - The lower quartile

S - The smallest measurement inside the inner fences

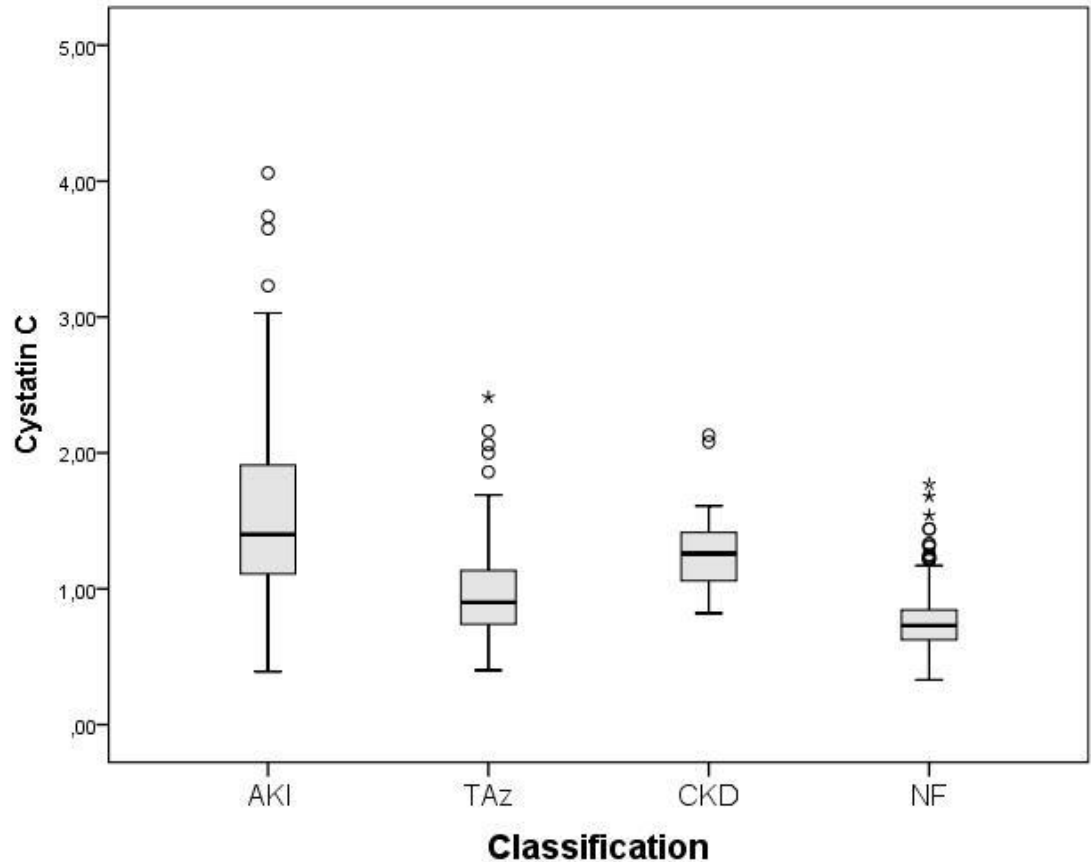
A *potential (mild) outlier* is a value located at a distance of more than $1.5IQR$ from the box. An *(extreme) outlier* is a value located at a distance of more than $3IQR$ from the box.

Descriptive Statistics

Graphing data:

Parallel Boxplot:

It is often useful to compare data from two or more groups by viewing box plots from the groups side by side.



OBS: Cystatin C is mainly used as a biomarker of kidney function

Descriptive Statistics

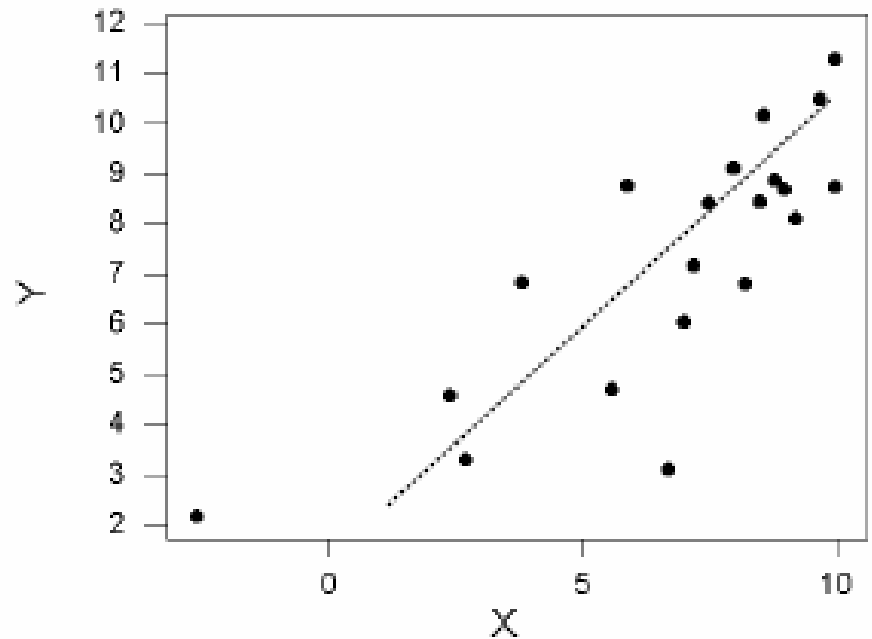
Graphing data:

Scatter Diagrams:

Often we are interested in the relationships between two quantitative variables.

Typical Patterns:

➤ **Positive linear** relationship:
if X increases then Y increases
and vice versa.



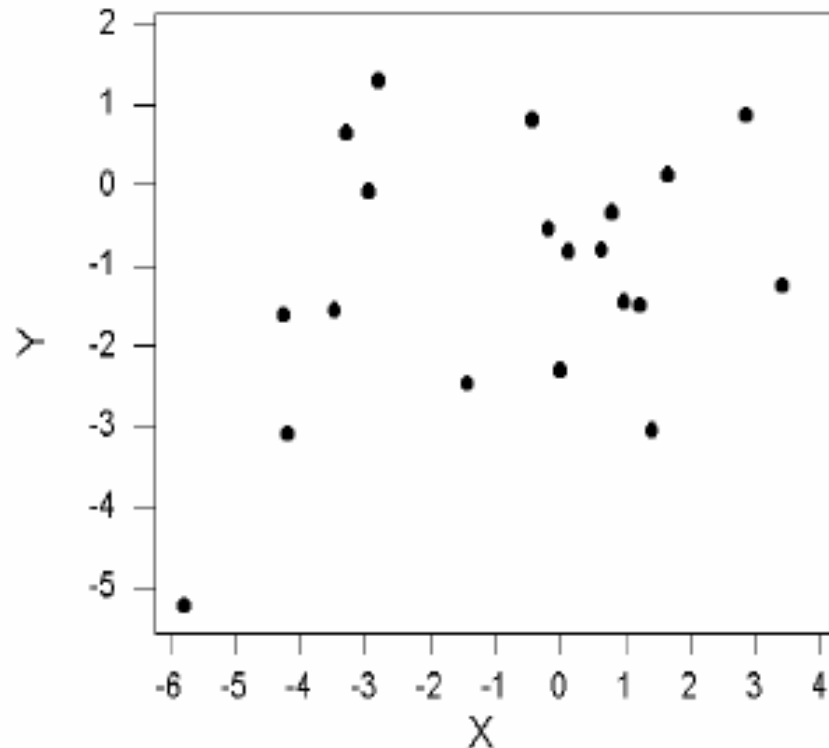
Descriptive Statistics

Graphing data:

Scatter Diagrams:

Typical Patterns:

➤ No relationship



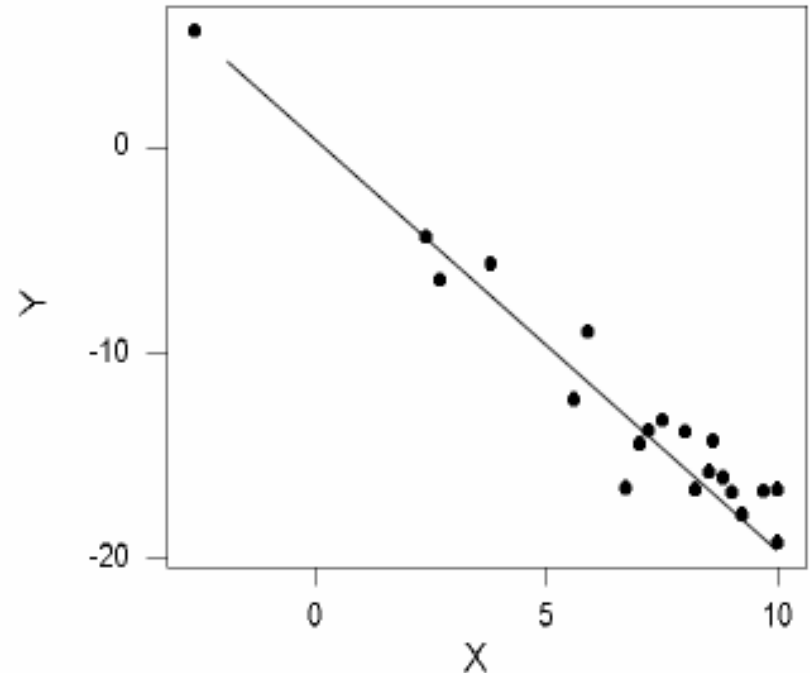
Descriptive Statistics

Graphing data:

Scatter Diagrams:

Typical Patterns:

➤ **Negative linear** relationship:
if X increases then Y decreases
and vice versa.



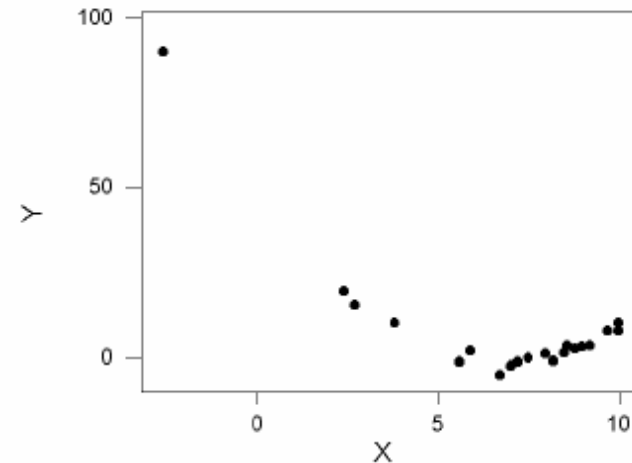
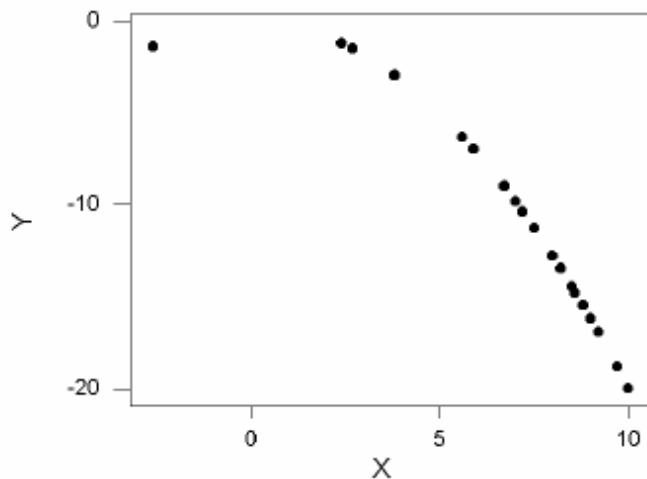
Descriptive Statistics

Graphing data:

Scatter Diagrams:

Typical Patterns:

- **Nonlinear** (*e.g.* concave) relationship (sometimes not so easy to identify)

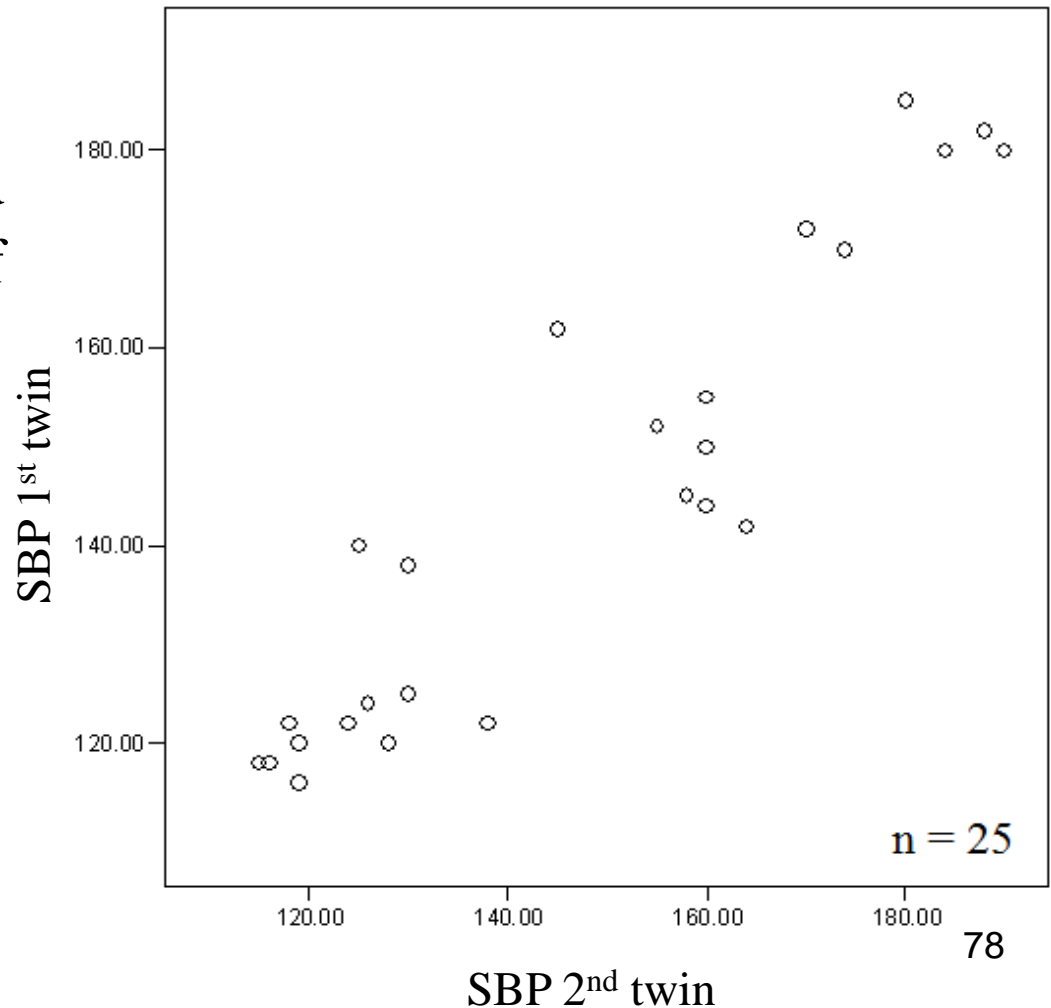


Descriptive Statistics

Graphing data:

Scatter Diagrams:

Example: systolic blood pressure of 25 pairs of identical twins



Descriptive Statistics

Measures of association:


Correlation  is there a linear association between two quantitative variables? How strong is this relationship?



Correlation coefficient - ρ ($-1 \leq \rho \leq 1$)

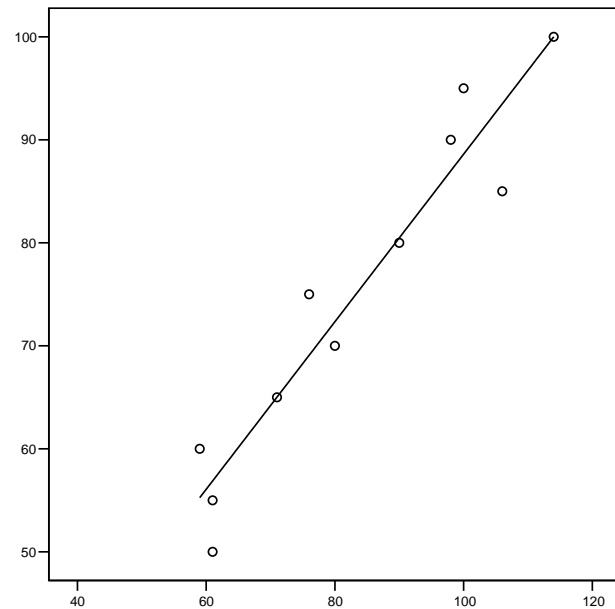
If $\rho > 0$  positive linear relationship

If $\rho < 0$  negative linear relationship

If $\rho = 0$  no linear relationship

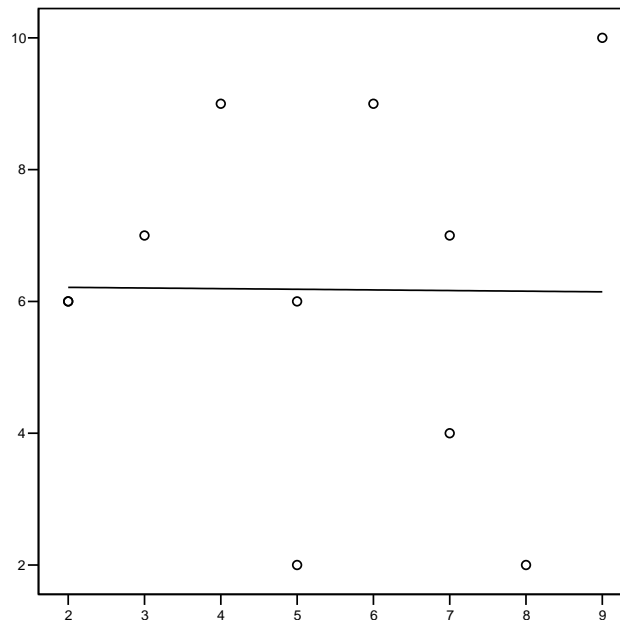
Descriptive Statistics

Measures of association:



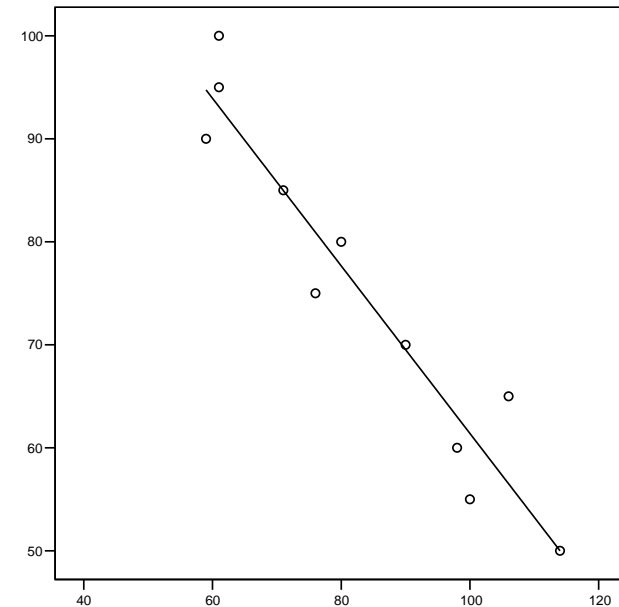
$$\rho > 0$$

Direct
relationship



$$\rho = 0$$

Linear
independence



$$\rho < 0$$

Inverse
relationship

Descriptive Statistics

Measures of association:

Correlation coefficients:

- ❖ Pearson's correlation coefficient (Normality assumed)
- ❖ Spearman's rank correlation coefficient*

* for ordinal categorical variables at least; when Normality assumption is not verified.

Descriptive Statistics

Measures of association:

Misuses of the correlation coefficient:

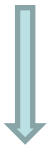
- It is incorrect to calculate a simple correlation coefficient for data which include more than one observation on some or all of the subjects, because such observations are not independent.
- Correlation is inappropriate for comparing alternative methods of measurements of the same variable, because it assesses association not agreement.
- Regression and correlation are separate techniques serving different purposes and need not automatically accompany each other.

Descriptive Statistics

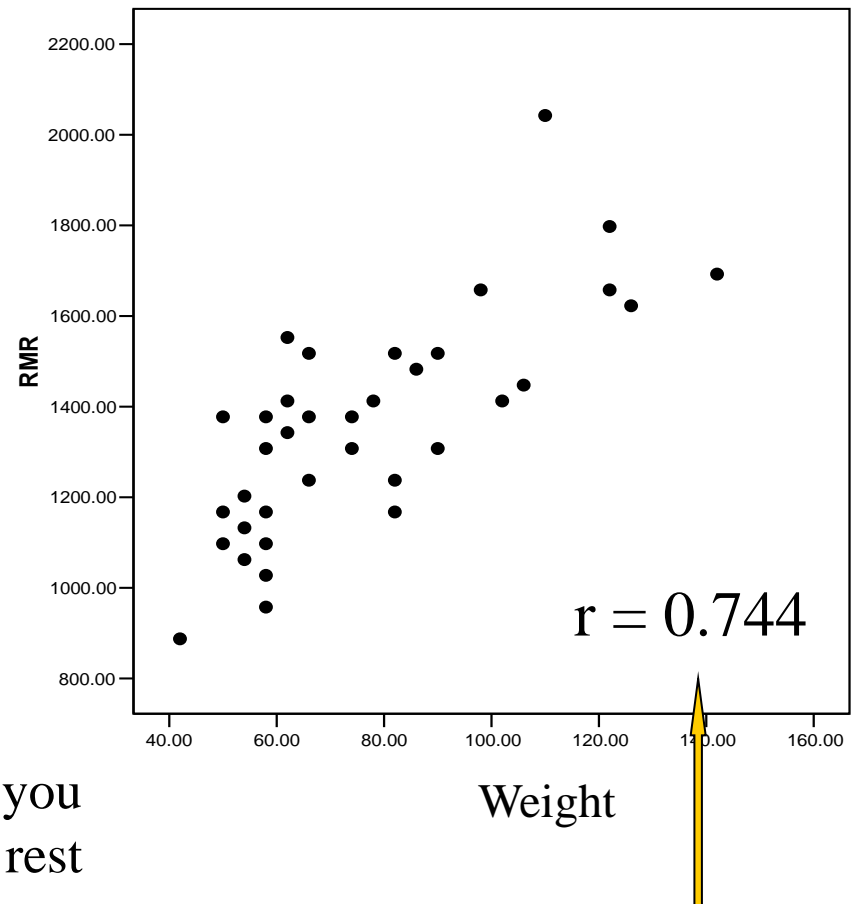
Measures of association:

Example:

Relationship between weight (kg)
and basal metabolic rate (RMR)
(kcal/24 hr)



RMR is an estimate of how many calories you would burn if you were to do nothing but rest for 24 hours. It represents the minimum amount of energy required to keep your body functioning, including your heart beating, lungs breathing, and body temperature normal.



Point estimate of ρ

Descriptive Statistics

Exercise:

The following data concerns weights (ounces) of malignant tumors from 57 patients. Perform a descriptive study with this data.

12	23	28	36	44	57
12	24	28	36	45	63
12	24	28	38	46	65
16	25	28	38	47	68
19	25	30	42	49	69
21	25	30	42	49	74
22	27	31	42	49	79
22	27	31	43	50	
23	27	32	43	51	
23	27	32	43	51	

Multiple choice questions

Multiple choice questions

Each statement is either true or false.

The mean of a large sample of size n

1. Is always greater than the median
2. It is calculated from the formula $\frac{\sum_{i=1}^n x_i}{n}$
3. Estimates the population mean with greater precision than the mean of a small sample.
4. Increases as the sample size increases
5. Is always greater than the standard deviation

Multiple choice questions

I. The following are measures of the spread of a distribution:

1. Interquantile range
2. Standard deviation
3. Range
4. Median
5. Mode

II. As the size of a random sample increases:

1. The standard deviation decreases
2. The standard error of the mean decreases
3. The mean decreases
4. The accuracy of the parameter estimates increases

Multiple choice questions

III. A correlation coefficient:

1. Always lies in the range 0-1
2. Can be use to predict one variable from another
3. Could be used to summarize the relationship between haemoglobin concentration and blood group in a sample of hospital patients
4. Is a measure of the extent to which two continuous variables are linearly related

Bibliography

- ❑ Altman, D. (1991). *Practical statistics for medical research*. First edition. Chapman & Hall, London.
- ❑ Bland, M. (2000). *An introduction to medical statistics*. Oxford University Press.
- ❑ Daniel, Wayne and Cross, Chad (2013). *Biostatistics: A Foundation for Analysis in the Health Sciences*. Third edition. Oxford University Press.
- ❑ Cumming, G., Fidler, F. and Vaux, D. L. (2007). Error bars in experimental biology. *The Journal of Cell Biology*, 177(1):7-11.