# Data exploration

Prof. Crescenzio Gallo

*Aggregate Professor of Computer Science and Processing Systems*

Department of Clinical and Experimental Medicine

**crescenzio.gallo@unifg.it**

# Introduction

The exploratory analysis aims to highlight:

- characteristics of attributes
- existing relationships between attributes

for a given dataset.

The process of exploratory analysis can be summarized in three main steps

- *Univariate*: the properties of each attribute are studied, evaluated individually (UNI-variate),
- *Bivariate*: pairs of attributes are considered (BI-variate) measuring the existing link between them,
- *Multivariate*: study the links existing between several attributes (MULTI-variate).

# Univariate analysis

It allows you to study the behavior of an attribute by treating it as independent from the rest of the attributes in the dataset. It studies the propensity of attribute values to position themselves near a *central value* and the propensity to take values in a wider or narrower *range around the central value*.

Useful in that learning models assume assumptions about the behavior of an attribute to ensure that they will lead to reliable results.

$$D = \{x_1,\ldots,x_m\}$$

dataset containing "$m$" observations, $x_k$ is the $k$-th observation.

$$X = (x^1{}_k,\ldots,x^n{}_k)$$

"$n$" attributes are considered, $x^j{}_k$ is the $j$-th attribute of the $k$-th observation.

# Univariate analysis: categorical attributes

Being a univariate analysis, to simplify the notation, we will write $\underline{X}$ instead of $\underline{X}^j$ and

$$\underline{X} = \{x^j_1,\ldots,x^j_m\}$$

to indicate the "$m$" realizations belonging to the dataset "$D$" of the attribute $\underline{X}^j$.

## Graphical analysis of categorical attributes

Different forms of representation of the empirical attribute distribution are used.

$$V = \{v_1,\ldots,v_H\}$$

is called *support* of the attribute, the set of values that the categorical attribute can take.

In this case we assume that the attribute $X$ can take "$H$" distinct values.

# Univariate analysis: categorical attributes

Consider a database that contains information about car sales for the following fields:

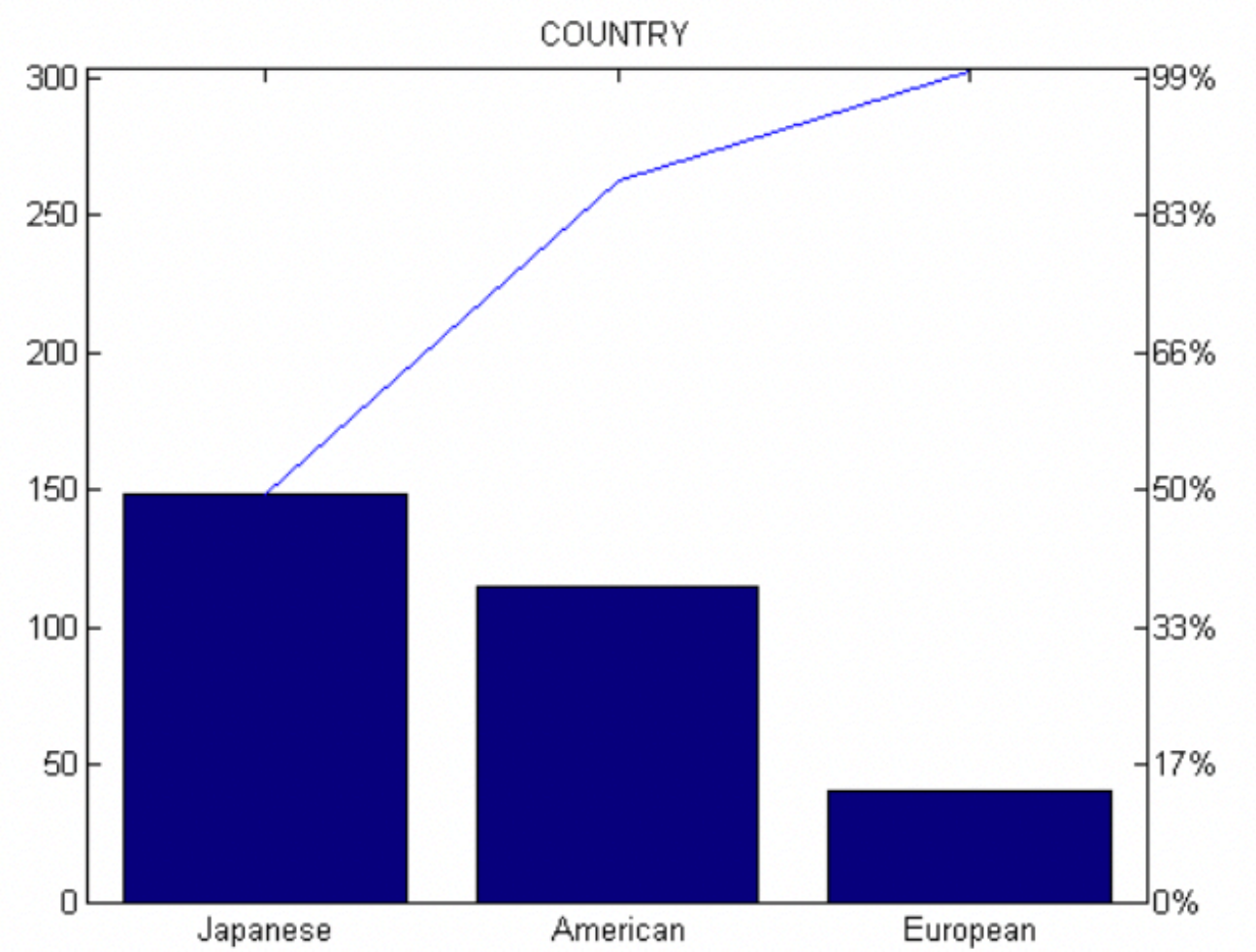
- GENDER: gender of the buyer
- MARITAL STATUS: marital status of buyer
- AGE: buyer's age
- COUNTRY: geographical area of origin of the car
- TYPE: type of the car
- SIZE: size of the car

and pay attention to the field (attribute) COUNTRY. This attribute can assume only the following values:

$V$ = {American, European, Japanese}

# Univariate analysis: categorical attributes

A graphic representation to analyze a categorical attribute is undoubtedly offered by a vertical bar chart which shows the empirical frequency for each value of the support of the variable considered.



# Univariate analysis: categorical attributes

We denote by $e_h$ the frequency of the $h$-th value for the considered attribute, for how many of the "$m$" observations in the dataset the considered attribute takes on the $h$-th value.

Given the frequency $e_h$, the relative empirical frequency or empirical density $f_h$ is defined as follows:

$$f_h = e_h / m$$

If the available sample size is sufficiently large, according to the central limit theorem, the empirical relative frequency function is a good approximation of the probability density of the attribute $X$. More precisely, it is possible to affirm that for a sample of sufficiently high number the following relationship is valid:

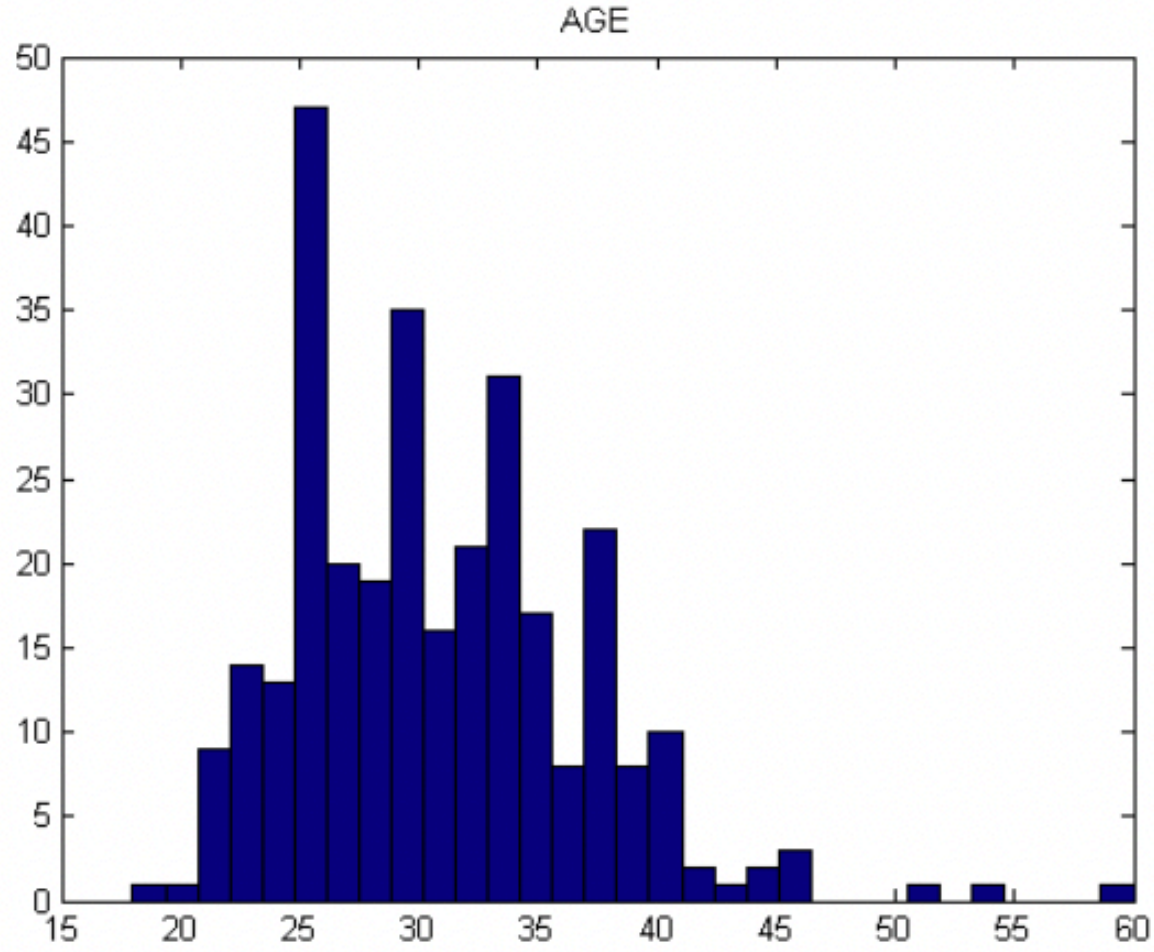
$$f_h \approx p_h = \Pr\{X=v_h\}$$

where $p_h$ represents the probability that the attribute $X$ assumes the value $v_h$.

# Univariate analysis: **numerical attributes**

## Graphical analysis of numerical attributes

For *discrete numeric attributes* that take on a *finite and limited number of values*, it is possible to use a *representation using bar charts* as shown for categorical attributes.



# Univariate analysis: **numerical attributes**

For *continuous or discrete attributes that have infinite values*, it is impossible to use the same representation; the abscissa axis, corresponding to the values of the attribute, is divided into intervals, usually of equal amplitude, that are assimilated to distinct classes.

In other words, we proceed to the *discretization* (***histogram***) of a continuous attribute through a finite and limited number of classes, introducing a certain degree of approximation.

It is appropriate to proceed to a subdivision in "$R$" intervals as narrow as possible, consistent with the need to keep limited the number of distinct classes generated.

Once determined the subdivision, we proceed to count the number of observations "$e_r$", $r$=1, ..., $R$, that fall in each interval.

# Univariate analysis: numerical attributes

**Procedure - Histograms for empirical density**

We choose the number of classes "$R$", depending on the number "$m$" of observations of the sample, we try to obtain a number of classes between 5 and 20 with the caveat that for each class there are at least 5 observations.

We define the total range and the amplitude "$I_r$" of each class, usually we divide the difference between maximum and minimum value of the attribute for the number of intervals that we want to obtain, in order to obtain intervals of equal amplitude.

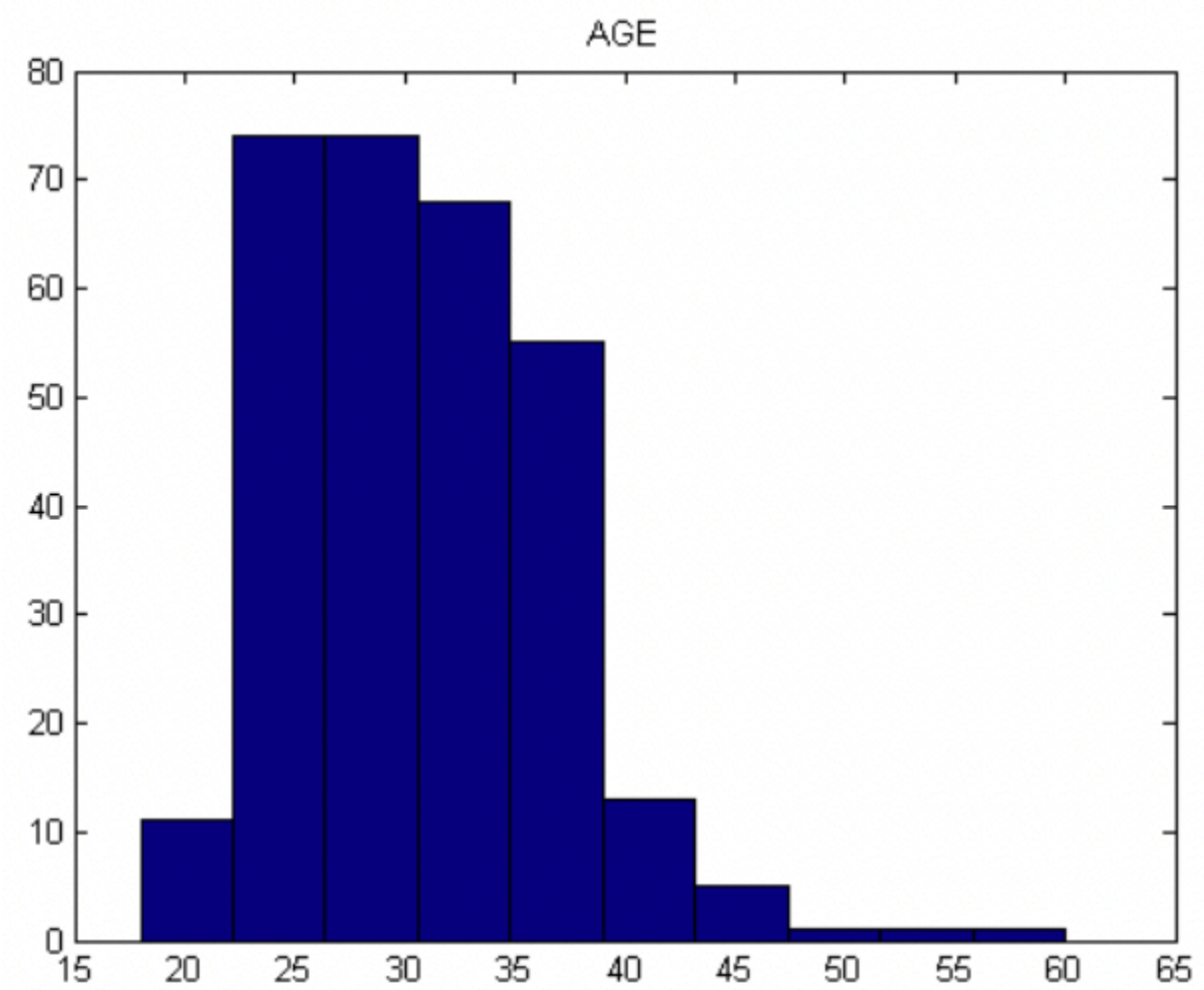
The boundaries of each class are assigned to make them disjoint, avoiding that there are values that belong simultaneously to contiguous classes. Each interval is closed on the left and open on the right.

We count the number of observations in each interval and assign to the corresponding rectangle a height equal to the empirical density "$f_r$" defined as

$$f_r = e_r / m$$

# Univariate analysis: **numerical attributes**

Considering the **AGE** attribute, and setting the number of intervals equal to ten ($R$=10) we obtain the histogram shown below.



# Univariate analysis: **numerical attributes**

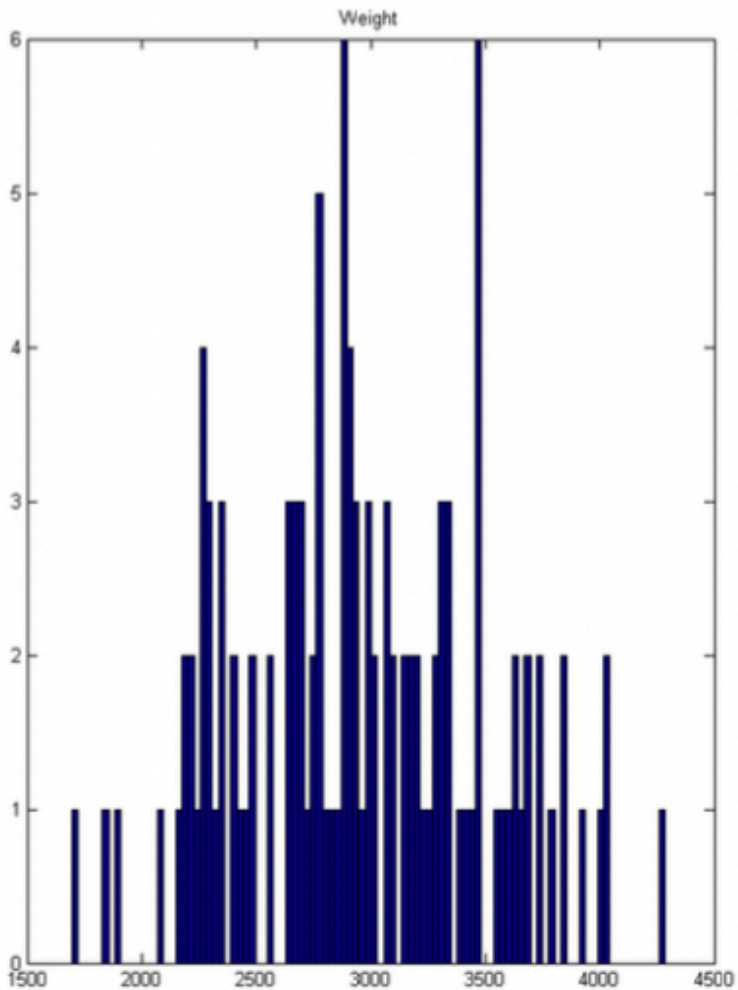
Consider a database that contains the following fields:

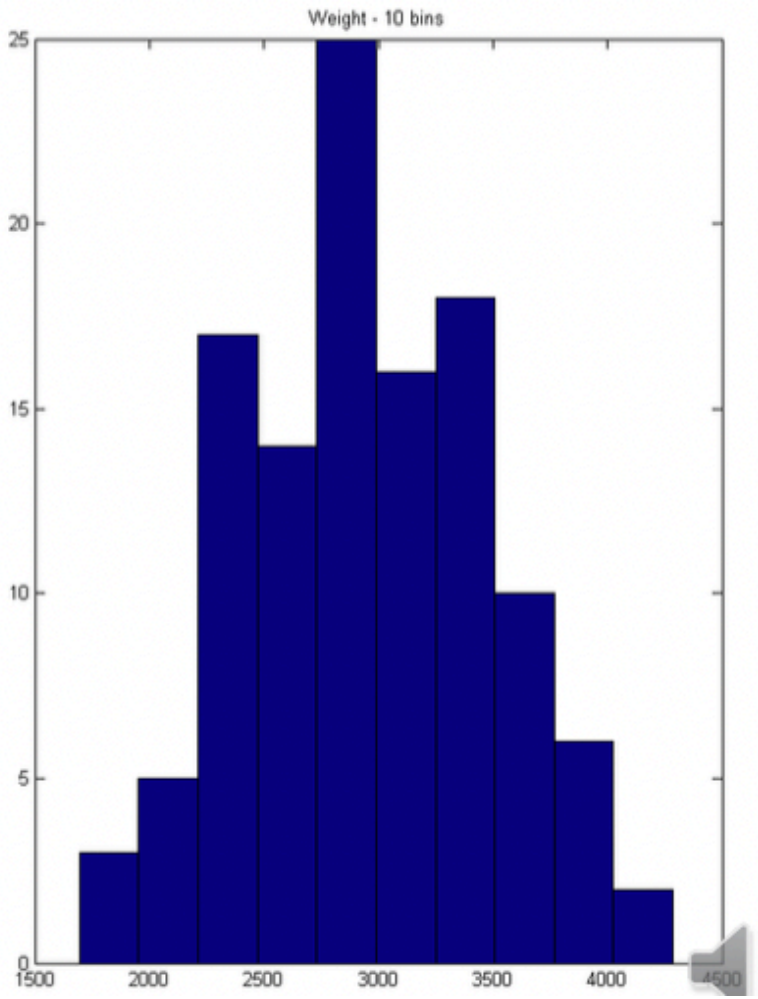
- Model: car model
- Country: geographical area of origin of the car
- Type: type of the car
- Weight: car weight
- Turning Circle: technical characteristic of the car
- Displacement: engine displacement
- Horsepower: engine power
- Gas Tank Size: tank capacity

related to different cars.

# Univariate analysis: **numerical attributes**

Let's consider the **Weight** attribute, for which we report below the bar chart in two versions.



# Univariate analysis: **numerical attributes**

## Central positioning indices for numerical attributes

Here are the main indices of central tendency:

- Mean $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$
- Median $x^{med} = \begin{cases} x_{(m+1)/2} & \text{if } m \text{ is } odd \\ \frac{x_{(m/2)} + x_{(m/2+1)}}{2} & \text{if } m \text{ is } even \end{cases}$
- Mode $x^{mod}$ most frequent value
- Mid-range $x^{midr} = \frac{x^{max} + x^{min}}{2}$
- Geometric mean $x^{geom} = \sqrt[m]{\prod_{i=1}^{m} x_i}$

# Univariate analysis: **numerical attributes**

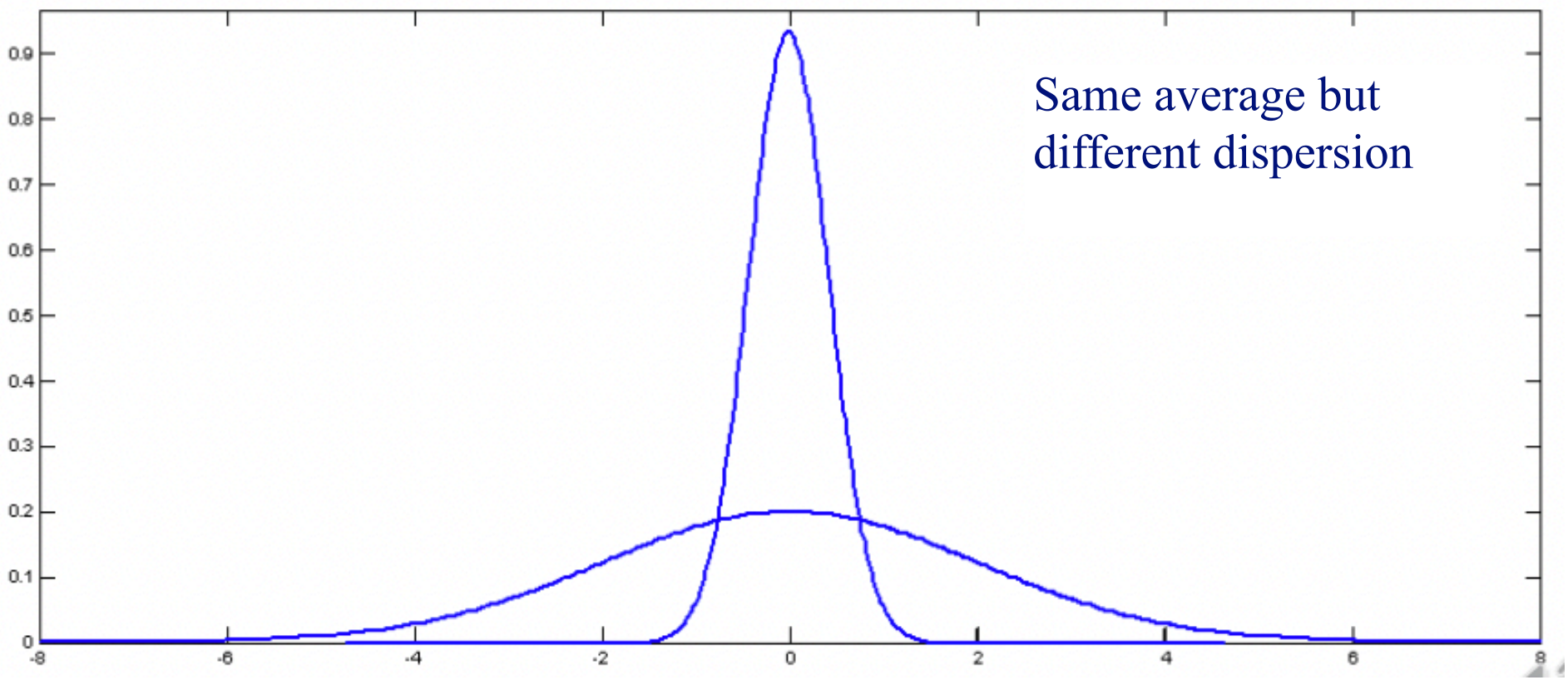
For the **Weight** attribute, the values of the central tendency indices are shown below:

- mean = 2957.6
- median = 2920
- mode = 2885
- midrange = 2990
- geomean = 2908.5

# Univariate analysis: **numerical attributes**

## Dispersion indices for numerical attributes

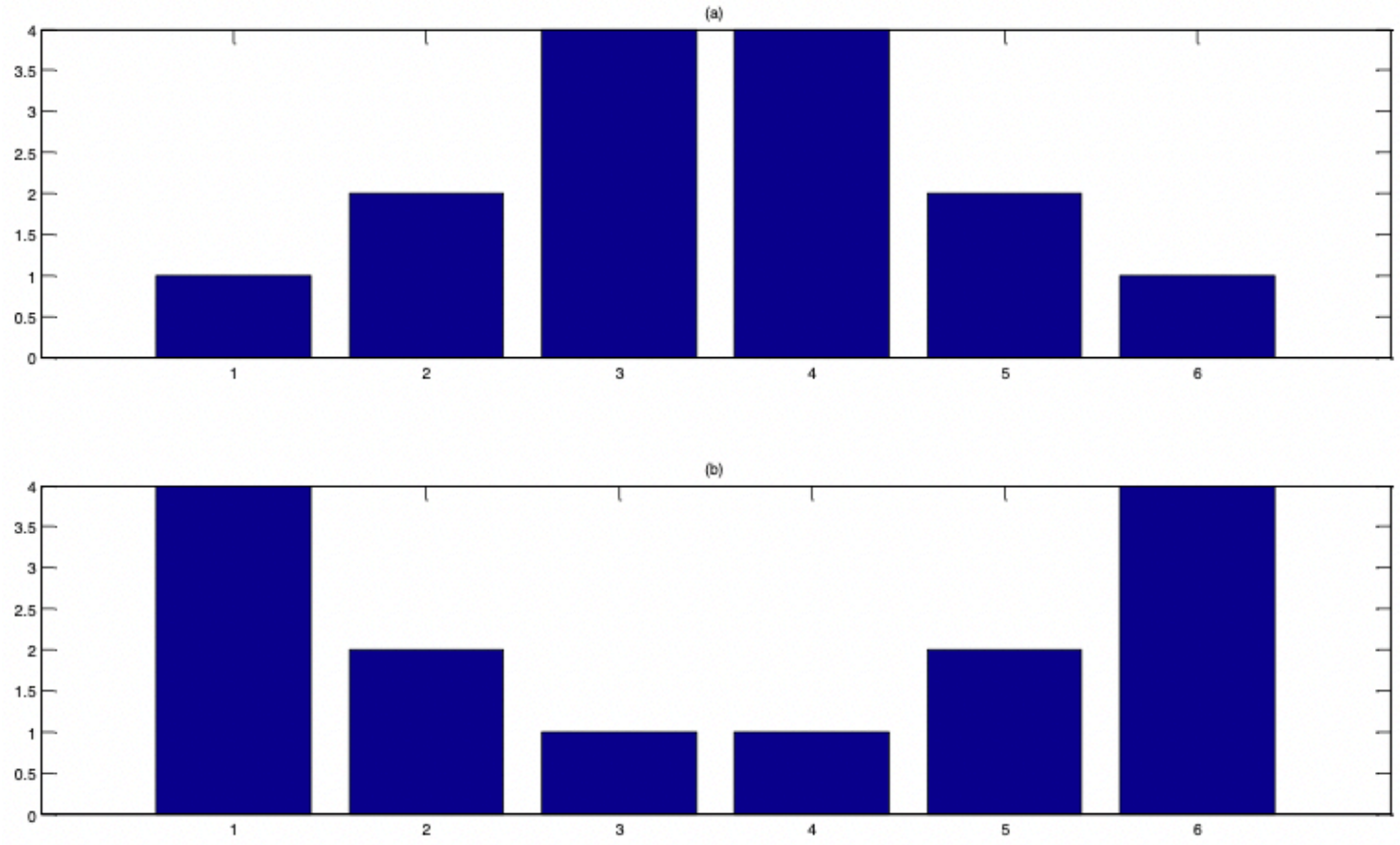
In addition to the indices of central tendency, it is relevant to define other indices that inform us about the level of *dispersion of the data*, i.e., the degree of *variability* that the observations manifest with *respect to* the *central values*.



# Univariate analysis: **numerical attributes**

A first measure of dispersion is represented by the *Range* defined as follows:

$$x^{\text{range}} = x^{\max} - x^{\min}$$




Ignore the actual dispersion of the data, equal range but very different dispersions.

# Univariate analysis: **numerical attributes**

The Deviation or Gap of a value is defined as the difference with sign from the arithmetic sample mean

$$s_i = x_i - \bar{x}$$

The relationship is valid

$$\sum_{i=1}^{m} s_i = 0$$

The dispersion of observations around the sample mean is expressed through the *Mean Absolute Deviation* (**MAD**), defined as follows

$$\text{MAD} = \frac{1}{m}\sum_{i=1}^{m}|s_i| = \frac{1}{m}\sum_{i=1}^{m}|x_i - \bar{x}|$$

Another important measure is the *Sample Variance*

$$S^2 = \frac{1}{m-1}\sum_{i=1}^{m} s_i^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})^2$$

and its root, the *Sampling Standard Deviation*: $S = \sqrt{S^2}$

# Univariate analysis: **numerical attributes**

In the particular case in which the distribution of the values of the attribute considered is *Normal* it is possible to remember that the

68.27% of the observations will belong to the interval $[\underline{x}-S, \underline{x}+S]$

95.45% of the observations will belong to the interval $[\underline{x}-2S, \underline{x}+2S]$

99.74% of the observations will belong to the interval $[\underline{x}-3S, \underline{x}+3S]$

A further index of dispersion is represented by the *Coefficient of Variation*

$$CV = 100\ S/\underline{x}$$

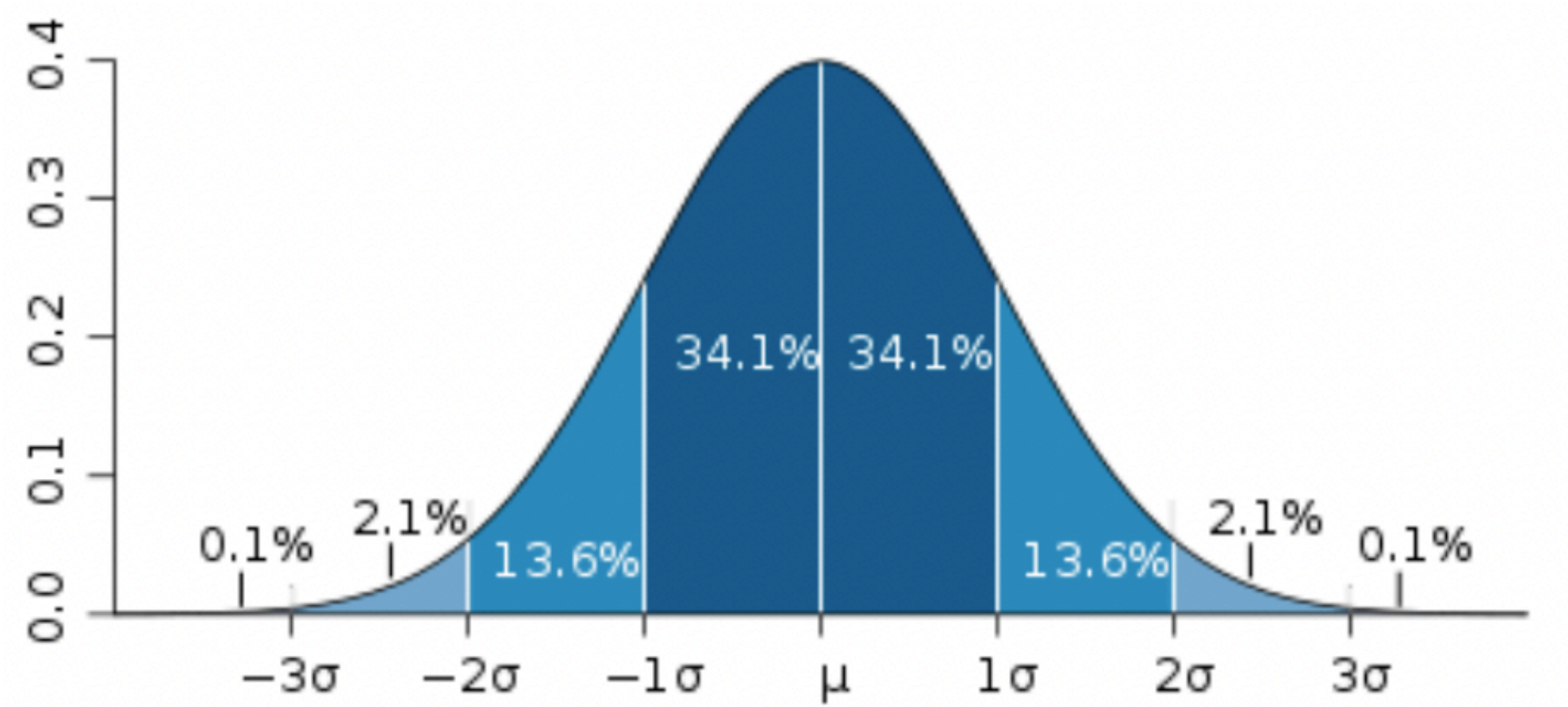
The values of the indices are reported below for the variable **Weight**

range: 2590; variance: 286940

mad: 433.7589 — stdev: 535.6635

cv: 18.1112

# Univariate analysis: **numerical attributes**



# Univariate analysis: **numerical attributes**

## Relative positioning indices for numeric attributes

These are indices used to locate a value relative to other values in the sample.

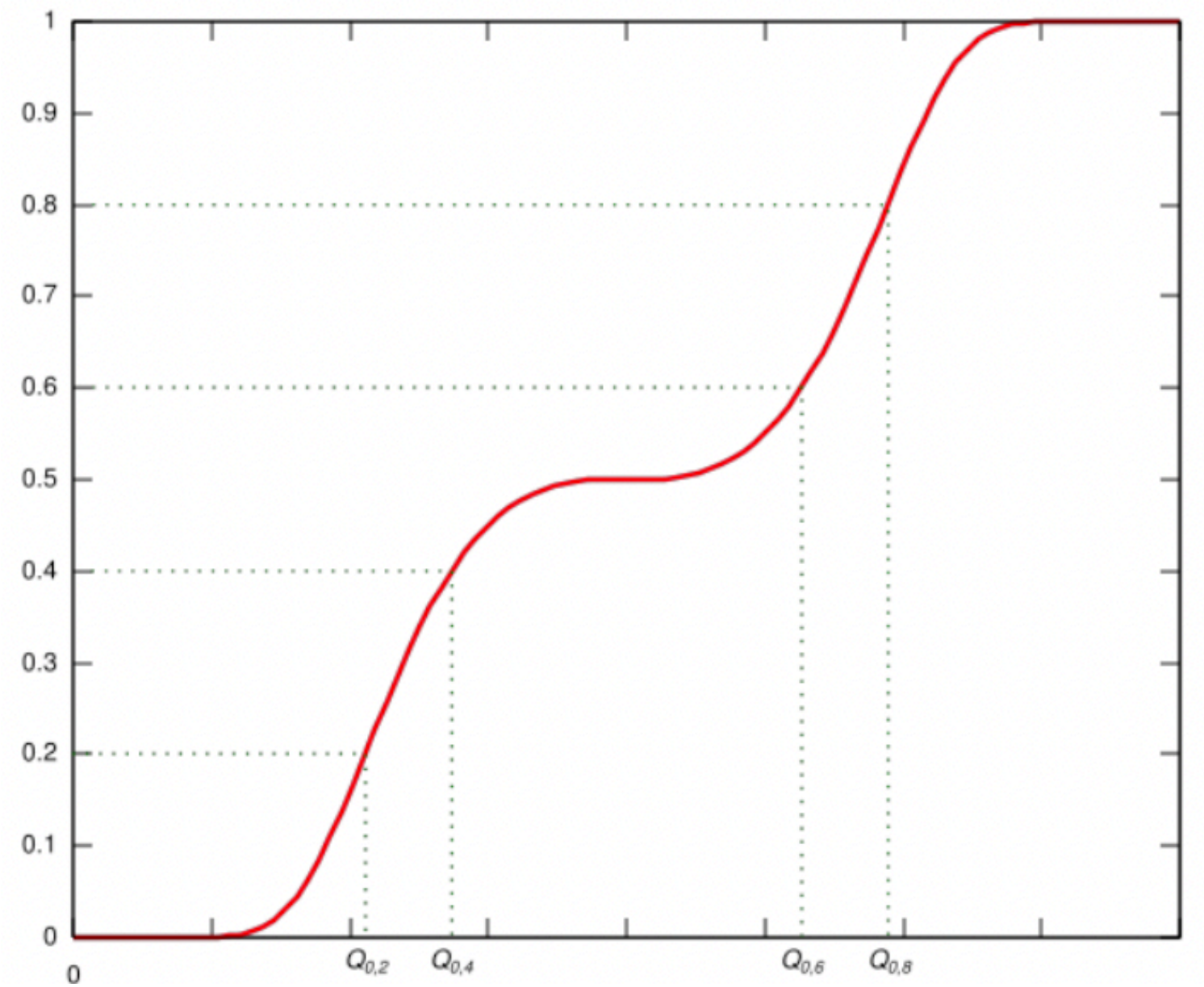
### *Quantiles*

Suppose we have arranged the "$m$" values of the attribute $\{x_1,\ldots,x_m\}$

in non-decreasing order.

Given any value $p$, $0 \leq p \leq 1$, a *quantile of order* "$p$" is a value "$q_p$" such that "$p \cdot m$" observations fall to the left of "$q_p$" and the remaining "$(1-p)m$" observations fall to its right.

# Univariate analysis: **numerical attributes**



# Univariate analysis: **numerical attributes**

Quantiles are also known as *percentiles*.

The *quantile for p=0.5* is equivalent to the *median*.

The *quantile for p=0.25* is referred to as the *lower quartile* ($q_L$) while the *quantile for p=0.75* is known as the *upper quartile* ($q_U$).

In some cases it may be useful to know the value of the *interquantile* distance, defined as follows:

$$D_q = q_U - q_L = q_{0.75} - q_{0.25}$$

In order to reduce the dependence of the mean on extreme observations, it is possible to resort to indicators of central positioning based on quantiles, which are usually more robust than those introduced above.

The *semi-mean*, obtained by averaging the attribute values between the lower and upper quartile.

# Univariate analysis: **numerical attributes**

The *truncated mean* is a generalization of the semi-mean, since it uses for the calculation of the average only the values between the quantiles of order "$p$" and "1-$p$". Usually it uses $p$=0.05.

The *winsorized mean* does not exclude from the calculation the values that belong to the tails, but encodes them according to an intuitive rule: the values less than the quantile of order "$p$" are increased to "$q_p$", while the values higher than the quantile of order "1-$p$" are decreased to "$q_{1-p}$".

The *z-score* is defined for the "$i$-th" observation $x_i$, as follows

$$z_i = \frac{x_i - \bar{x}}{s}$$

# Univariate analysis: **numerical attributes**

## Identification of outliers for numerical attributes

An *outlier* is intuitively defined as an observation that has an anomalous characteristic with respect to the set of remaining observations for the attribute under consideration.

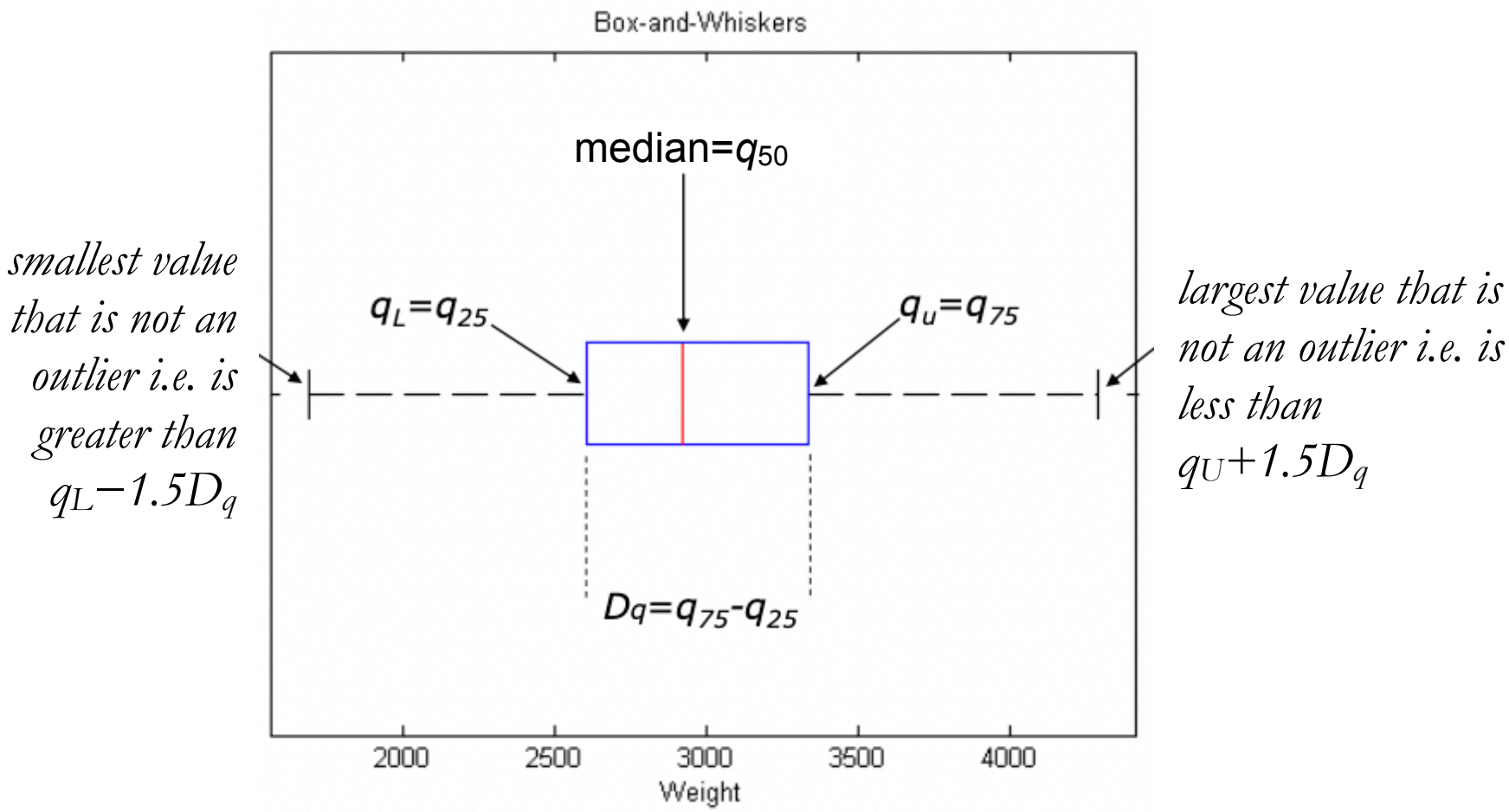
One way to identify an outlier is offered by the z-score, specifically, observations that show a *z-score value greater than 3 in absolute value* are considered as *suspected outliers*, and as s*trongly suspected outliers* those observations for which this value is *much greater than 3*, always in absolute value.

A second way to identify the outliers is offered by the graphical representation that goes under the name of *box-and-whisker diagram* that is based on the representation in graphical form of the median and the lower and upper quartiles. An observation is identified as an outlier if it falls outside the following boundaries:

- upper outer edge $q_L - 3D_q$
- lower outer edge $q_L - 1.5D_q$
- upper inner edge $q_U + 1.5D_q$
- upper outer edge $q_U + 3D_q$

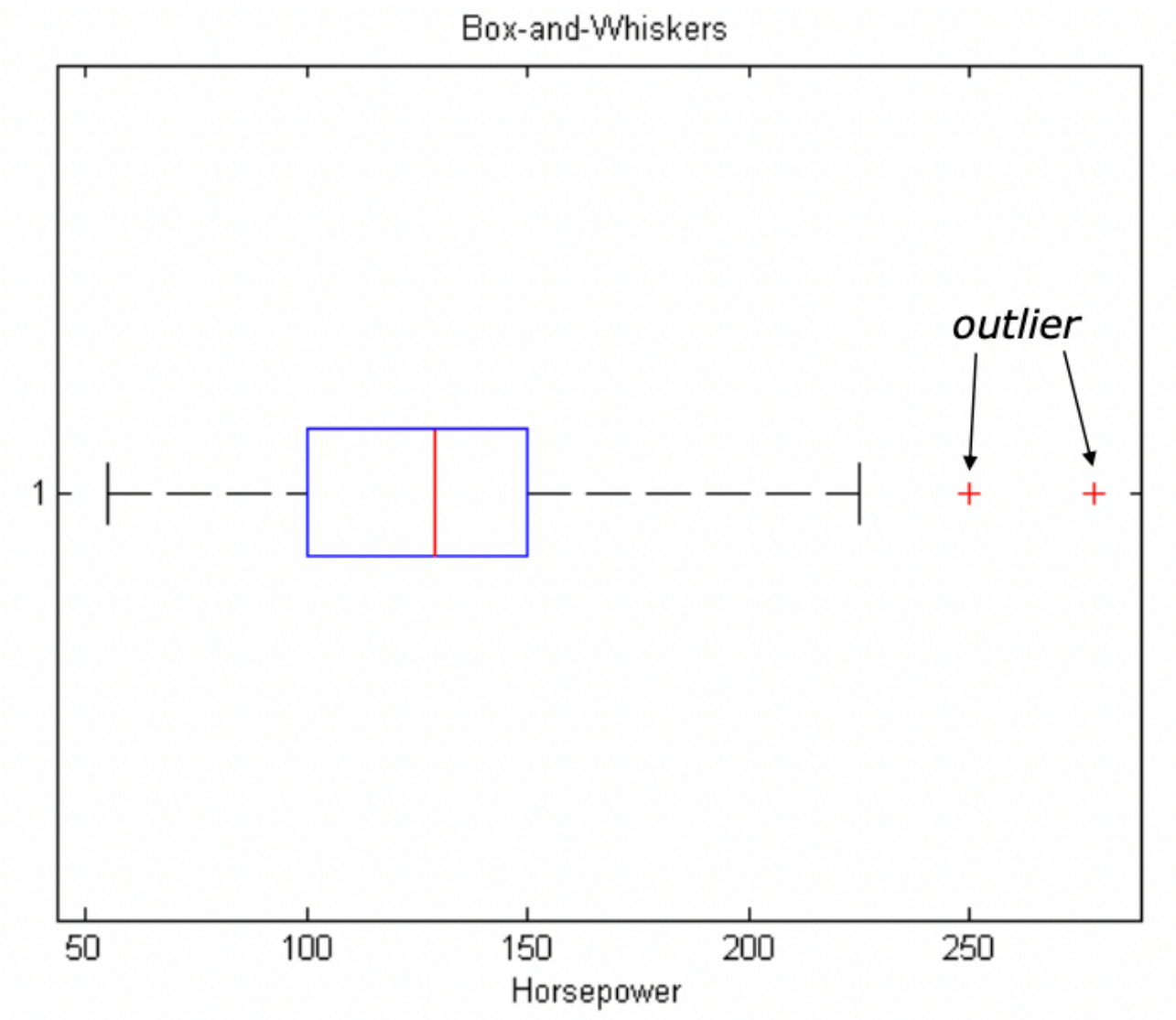
# Univariate analysis: **numerical attributes**

The advantage over the z-score is that the procedure identifies outliers independently of the extreme values of the observations, in fact the median and quartiles enjoy this independence. In contrast, the *z-score is strongly influenced by extreme values.*



# Univariate analysis: **numerical attributes**

Two outliers, identified by the red "+" symbol.



# Univariate analysis: categorical attributes

## Heterogeneity indices for categorical attributes

In the case of categorical attributes the indicators of central and relative positioning, together with the measures of dispersion that we introduced earlier are not applicable.

For a categorical attribute it is generally preferred to use indicators that express the degree of regularity with which the data

$$\{x_1,\ldots,x_m\}$$

are arranged within the set of "$H$" distinct values.

The *maximum heterogeneity* is obtained in the case in which the *relative empirical frequencies* "$f_h$" are *equal for all classes* "$h$".

*Minimum heterogeneity* is obtained if "$f_g$=1" for one "$g$" class and "$f_h$=0" for the remaining "$h$" classes.

We describe below two indices of heterogeneity for categorical attributes, the Gini index and Entropy.

# Univariate analysis: categorical attributes

## Gini Index

Defined as

$$G = 1 - \sum_{h=1}^{H} f_h^2$$

It takes on a *null value* in the case of *minimum heterogeneity*, that is when only one class takes on a value with a frequency equal to 1 and all the other classes have a frequency equal to zero.

If all classes assume equal value of the relative empirical frequency ($1/H$), the Gini index assumes its *maximum value* equal to

$$G = 1 - \sum_{h=1}^{H} f_h^2 = 1 - \sum_{h=1}^{H} \left(\frac{1}{H}\right)^2 = 1 - \frac{H}{H^2} = \frac{H-1}{H}$$

You can normalize the index so that it takes values in the range [0,1]:

$$G_{rel} = \frac{G}{\left(\frac{H-1}{H}\right)}$$

# Univariate analysis: categorical attributes

## Entropy Index

Defined as

$$E = -\sum_{h=1}^{H} f_h \log_2 f_h$$

It takes on a *null value* in the case of *minimum heterogeneity*, that is when only one class takes on a value with a frequency equal to 1 and all other classes have a frequency equal to zero.

If all classes assume equal value of the relative empirical frequency ($1/H$), the Entropy index assumes its *maximum value* equal to

$$E = -\sum_{h=1}^{H} f_h \log_2 f_h = -\sum_{h=1}^{H} \frac{1}{H} \log_2 \frac{1}{H} = -\frac{1}{H} \sum_{h=1}^{H} \log_2 \frac{1}{H} = \log_2 H$$

You can normalize the index so that it takes values in the range [0,1]:

$$E_{rel} = \frac{E}{\log_2 H}$$

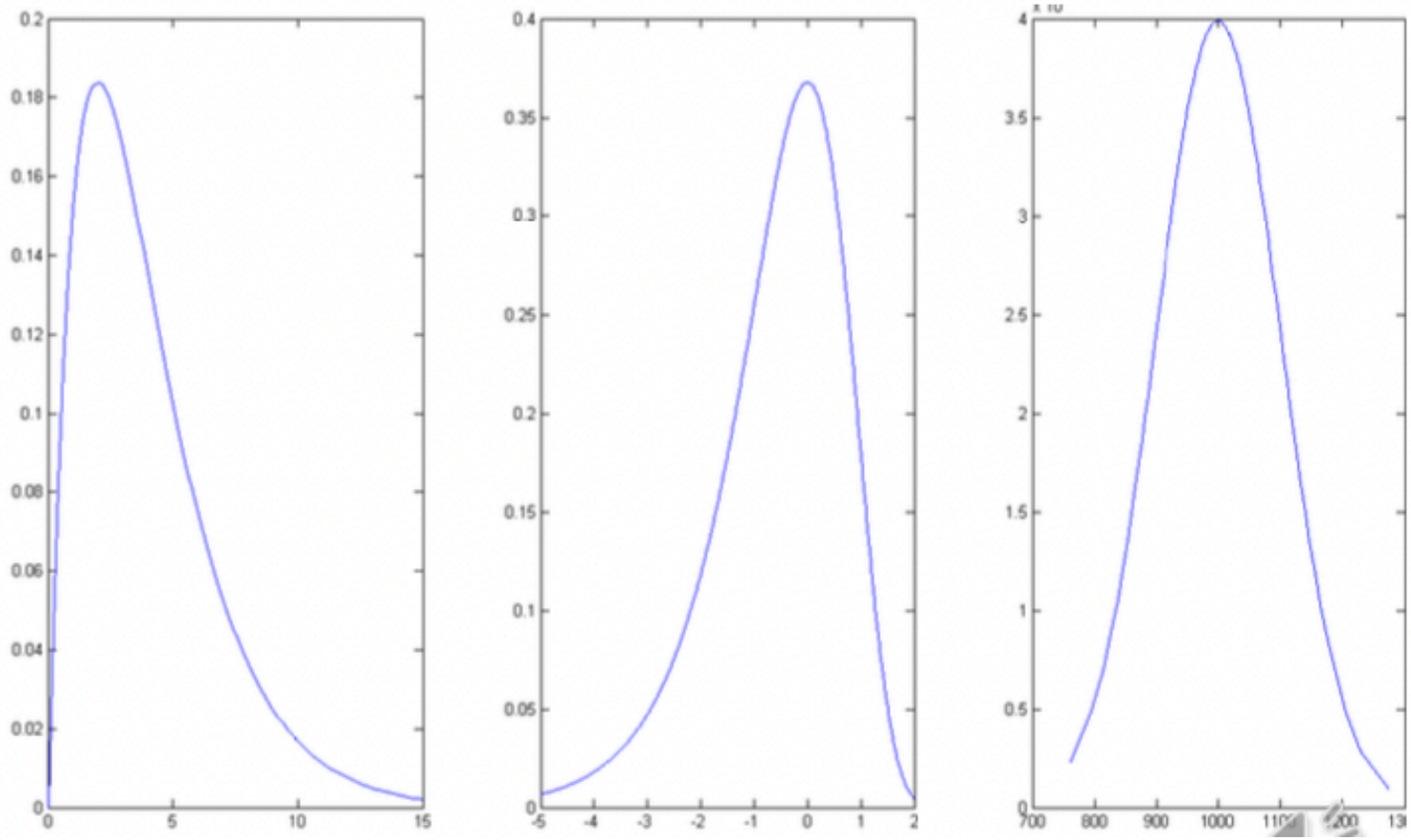
# Univariate analysis: **numerical attributes**

## Empirical Density Analysis

L'istogramma delle frequenze empiriche relative costituisce un importante strumento per l'analisi grafica di attributi sia categorici che numerici. È pertanto fondamentale disporre di indicatori di sintesi che consentano di studiare le proprietà della curva di densità empirica.

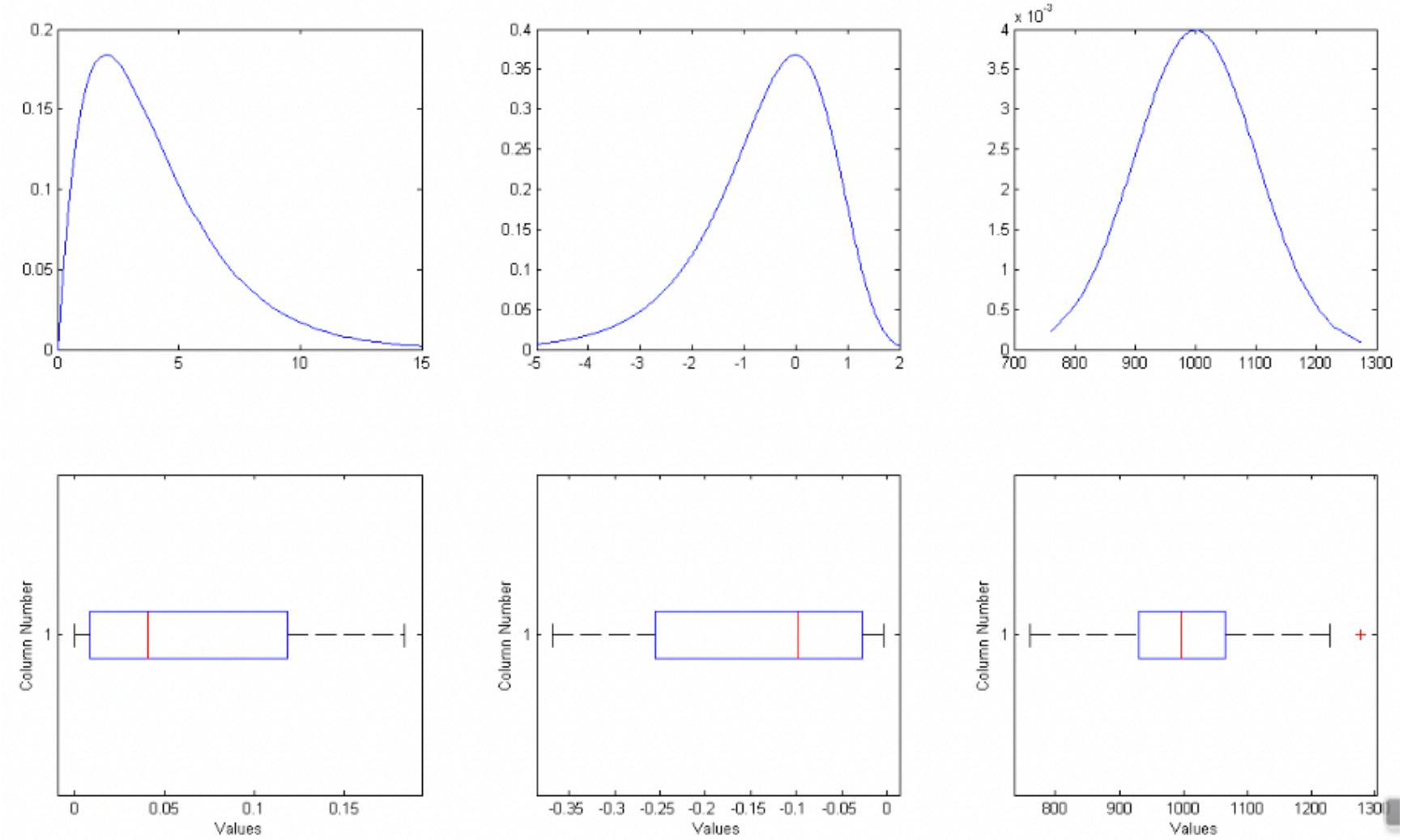
## Density curve asymmetry

A *density curve* is said to be *symmetrical* if the *mean coincides with the median*, otherwise it is said to be asymmetrical.



# Univariate analysis: **numerical attributes**

The Box-and-Whisker diagram allows a particularly intuitive identification of the nature and degree of asymmetry of an empirical density, as is illustrated in the figure below.



# Univariate analysis: **numerical attributes**

In the case of a numerical attribute, an asymmetry index based on the *third sample moment* can be introduced alongside the graphical analysis:

$$\bar{x}^3 = \frac{1}{m}\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^3$$

The *asymmetry index* is defined as follows:

$$I_{as} = \frac{\bar{x}^3}{S^3}$$

and is such that if the curve is

- *symmetrical* : $I_{as} = 0$
- *asymmetrical to the right* : $I_{as} > 0$
- *asymmetrical to the left* : $I_{as} < 0$

# Univariate analysis: numerical attributes

## Density curve kurtosis

Another relevant problem related to the density histogram concerns the nature of the theoretical probability distribution, usually unknown a priori, from which the observations are extracted.

This is a problem of estimating the unknown distribution from data, which is complex in its general form and which we will therefore not address.

There is, however, a distribution that recurs with some frequency for which elementary graphical and summary criteria are available to assess the degree of approximation to an assigned empirical density.

The first criterion, of a graphic nature, is based on the visual comparison of the histogram of the empirical density with a normal curve having mean and standard deviation coincident with those of the assigned density.

# Univariate analysis: **numerical attributes**

The *kurtosis index* concisely expresses the degree of approximation of an empirical density to the normal curve, it uses the *sample quarter moment*:

$$\bar{x}^4 = \frac{1}{m}\sum_{i=1}^{m}\left(x_i - \bar{x}\right)^4$$

and is defined as follows:

$$I_{curt} = \frac{\bar{x}^4}{S^2} - 3$$

If the *empirical frequency corresponds* perfectly to that of a *normal curve* we have

$$I_{curt} = 0$$

in case

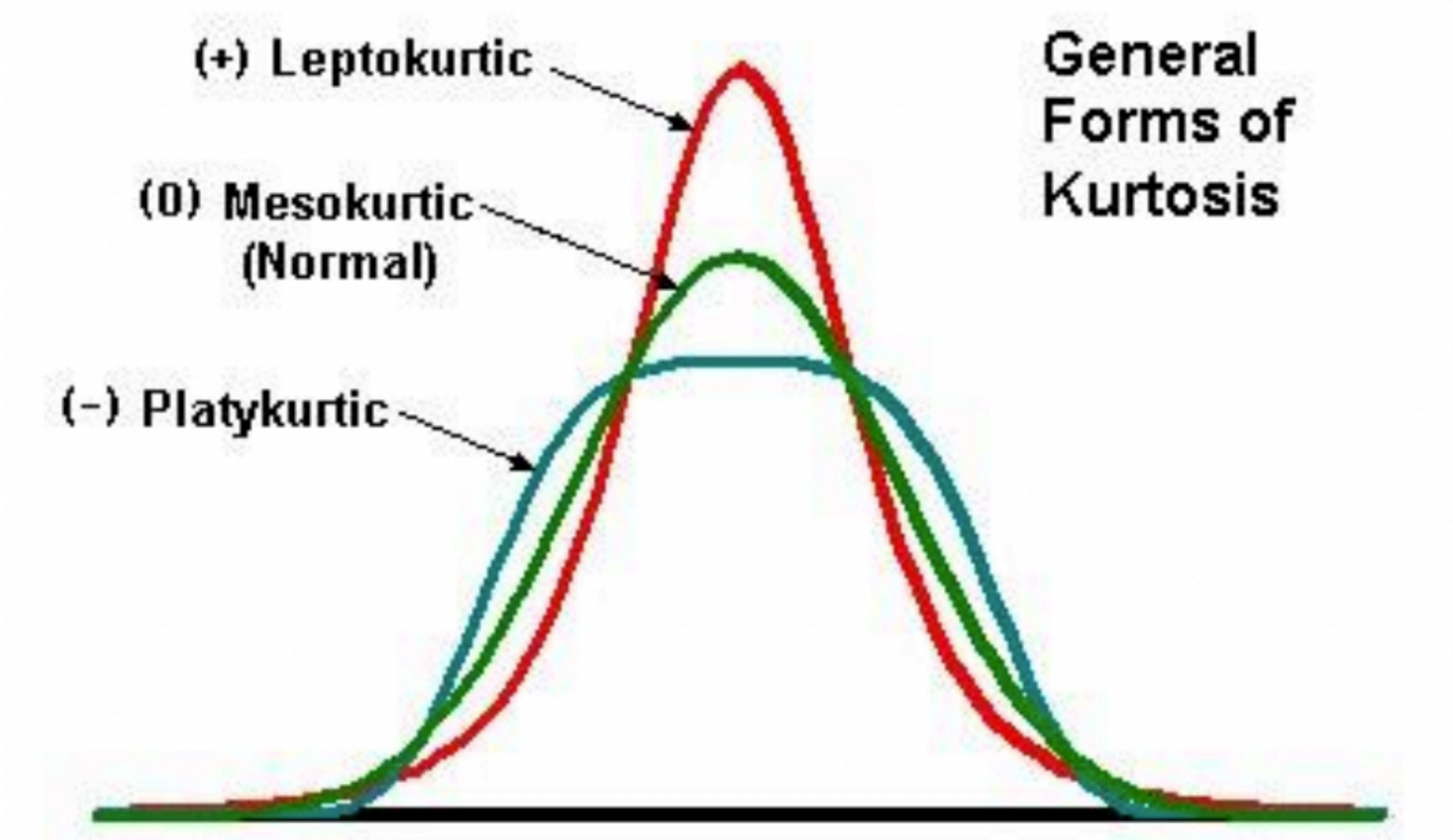
$$I_{curt} < 0$$

we speak of *hyponormal distribution*, while if

$$I_{curt} > 0$$

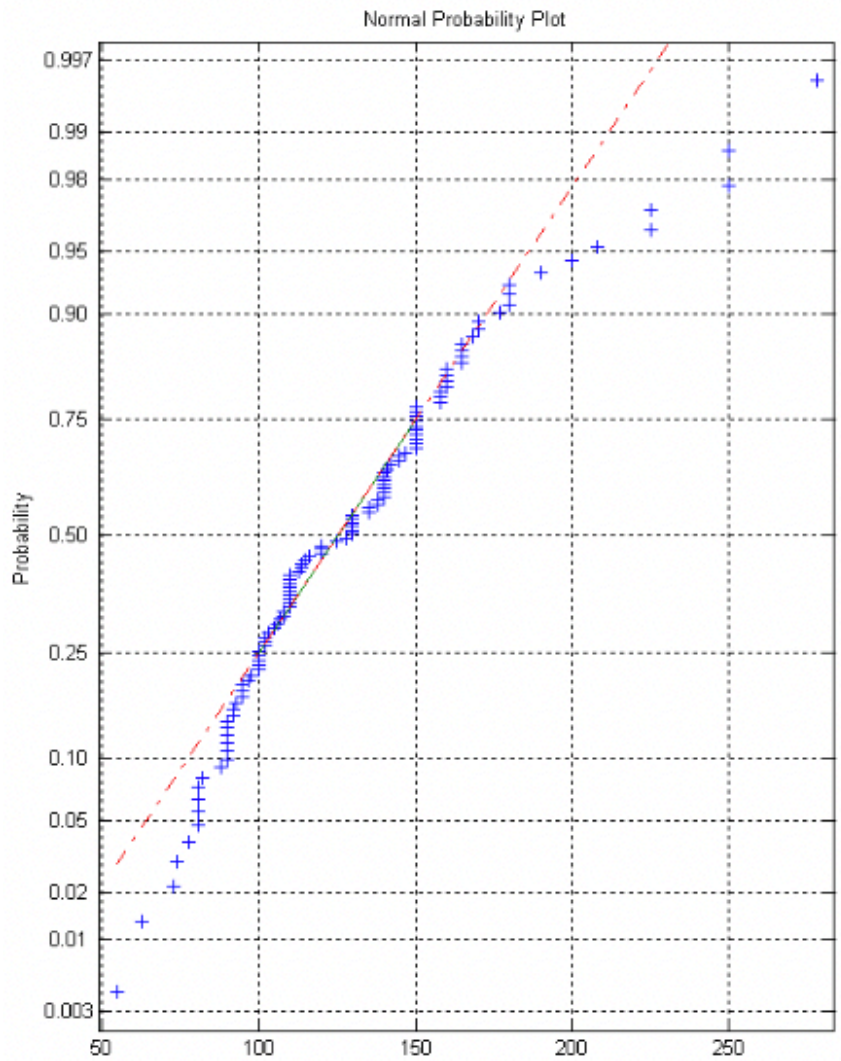
we talk about *hypernormal distribution*.

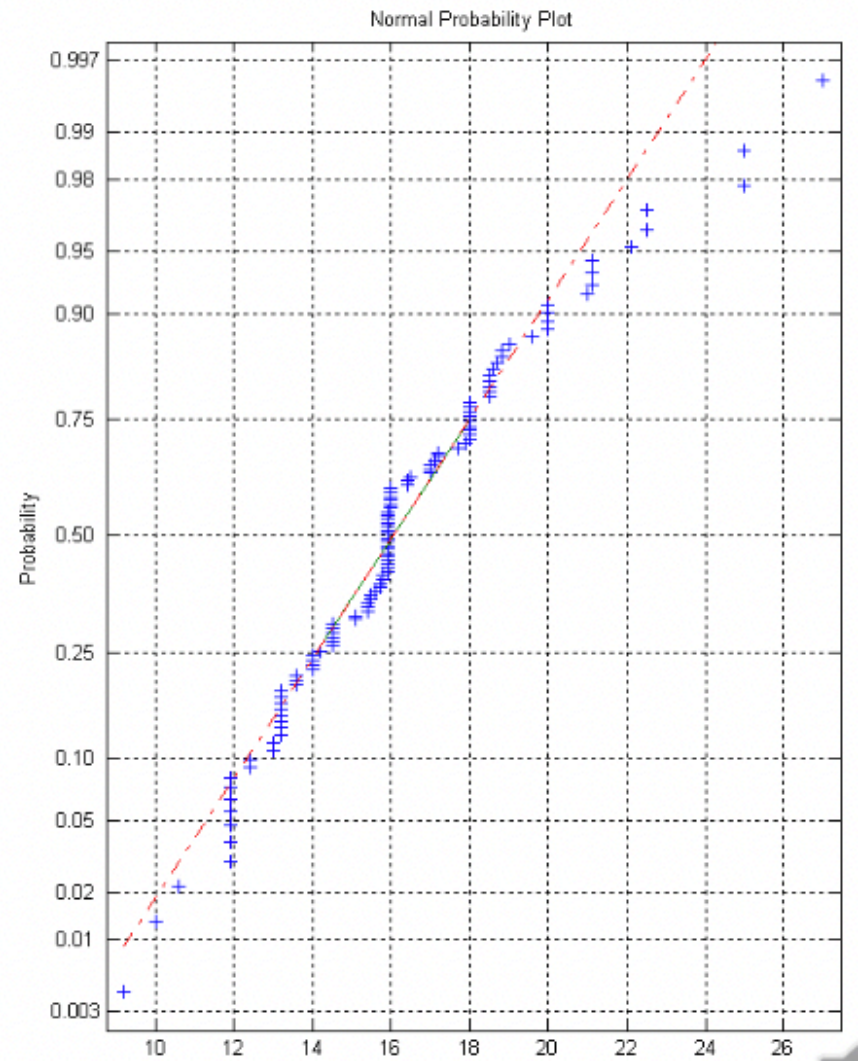
# Univariate analysis: **numerical attributes**



# Univariate analysis: **numerical attributes**

You can also use the *normal probability plot*, which is a special case of *quantile-quantile plot*.



# Bivariate analysis

Let us now analyze the relationships between pairs of attributes that we will generically indicate with $X^j$ and $X^k$. It is possible to distinguish three cases that can occur:

- *both attributes are numeric*
- *one attribute is numeric and the other is categorical*
- *both attributes are categorical*

In the course of the next slides we will use the following notation:

$$\underline{X}^j = \{x_1^j, \ldots, x_m^j\} \qquad \underline{X}^k = \{x_1^k, \ldots, x_m^k\}$$

indicating the "*m*" observations for the pair of attributes considered i.e. attribute $X^j$ and attribute $X^k$.

# Bivariate analysis: graphical analysis

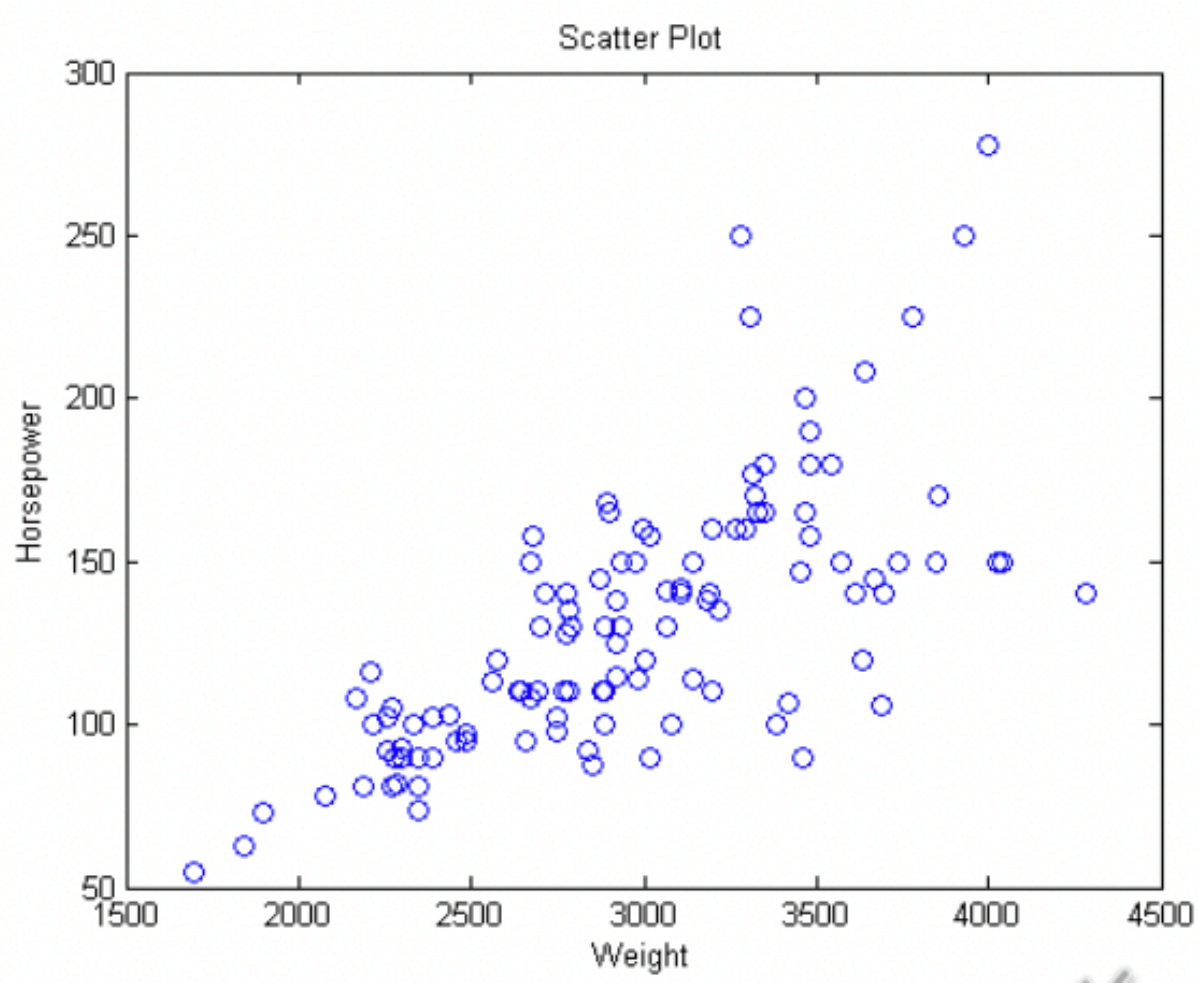
## Graphical Analysis

There are several types of graphical displays that allow you to study the relationship between two attributes.

## Scatter diagrams

A *scatter diagram* (scatterplot) is the most intuitive graphical representation of the link between two numerical attributes.

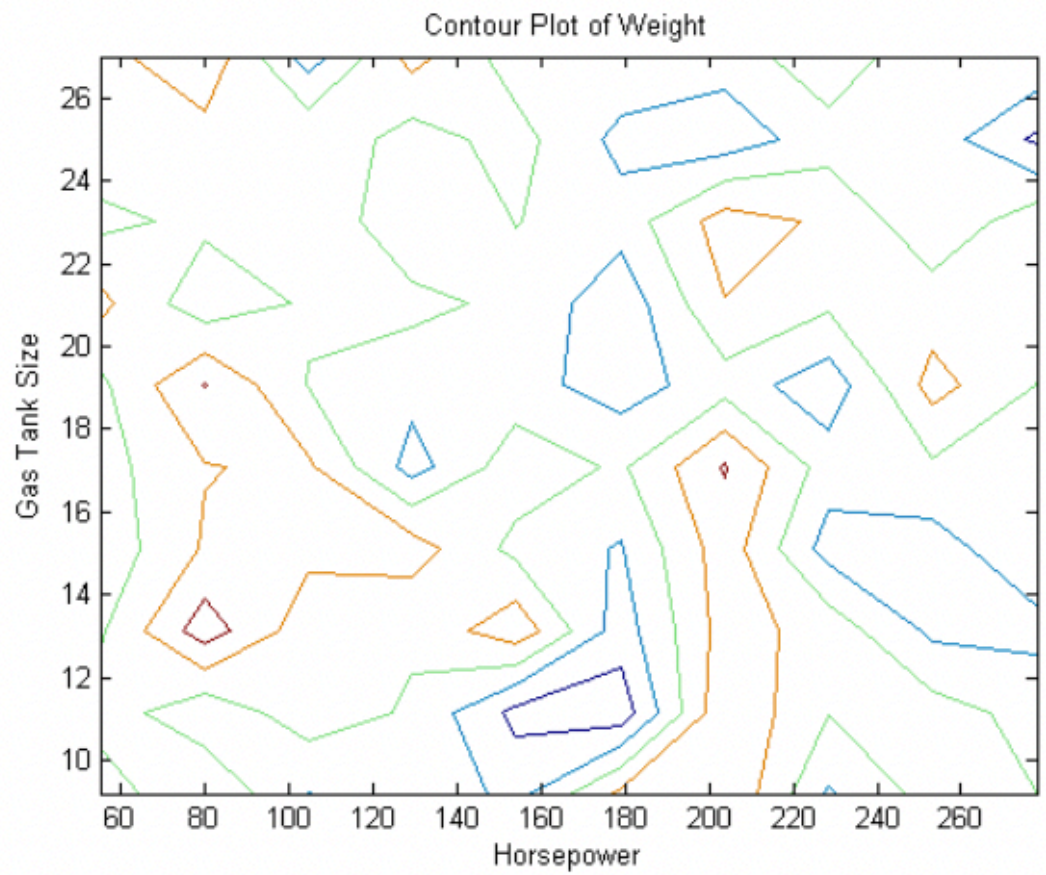
In the figure on the right the **Horsepower** attribute and the **Weight** attribute are considered.



# Bivariate analysis: graphical analysis

## Contour lines

They are an evolution of the dispersion diagrams and are therefore applicable only to numerical attributes. They are normally used to highlight the value of a third numerical attribute $X^z$ as the two attributes $X^j$ and $X^k$ on the axes of the diagram change.



# Bivariate analysis: graphical analysis

## Quantile-quantile diagrams

They allow to compare between them the distributions of the same attribute in correspondence of two different characteristics of the population or samples extracted from different populations (samples must have equal cardinality).

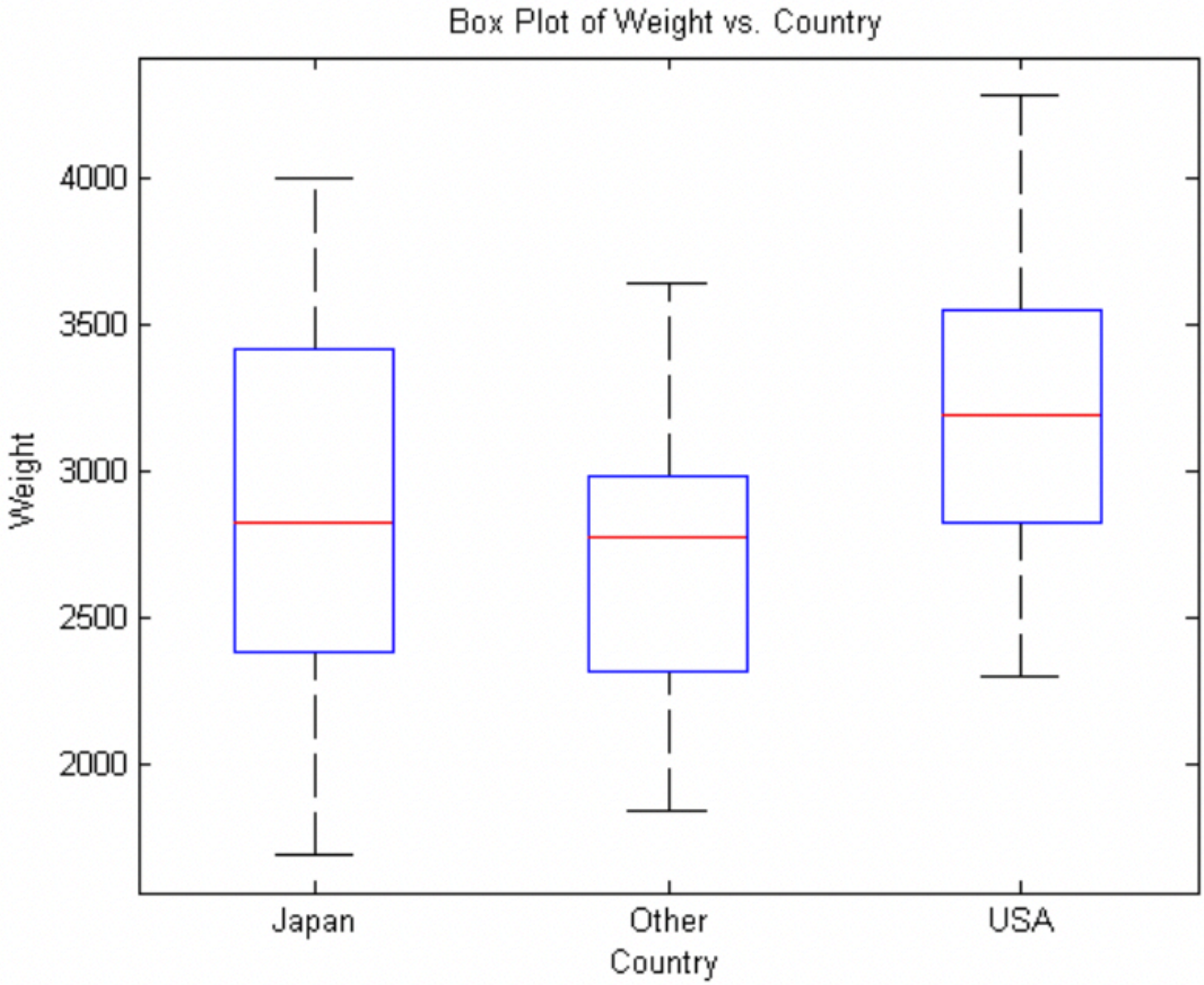
If the points are arranged along the line at 45° then there is reason to think that the two samples analyzed come from the same distribution.

## Box-and-Whisker diagrams

A further use of such diagrams is to compare the distributions of the same variable for observations belonging to distinct groups.

Technique applicable to a pair of attributes consisting of a continuous attribute and a categorical attribute.

# Bivariate analysis: graphical analysis



# Bivariate analysis: association indices

## Association indices for numerical attributes

As in the case of univariate analyses, it is useful to introduce alongside the graphical methods also synthetic indicators that express the nature and intensity of the link between numerical attributes.

### *Covariance*

Given the pair of attributes $X^j$ and $X^k$, the covariance is defined as follows:

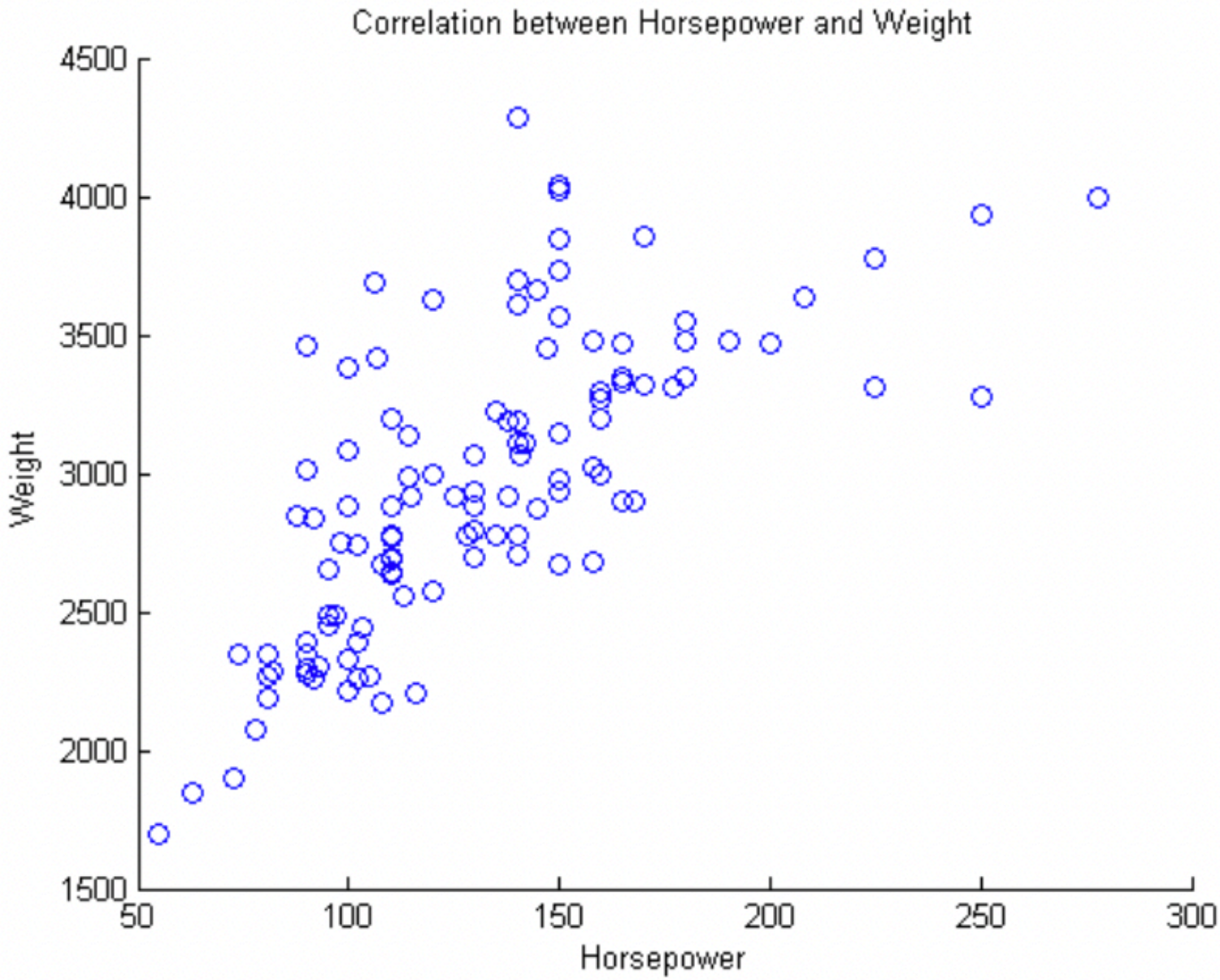
$$v^{jk} = \mathrm{cov}(X^j, X^k) = \frac{1}{m-2}\sum_{i=1}^{m}\left(x_i^j - \bar{x}^j\right)\left(x_i^k - \bar{x}^k\right)$$

### *Correlation*

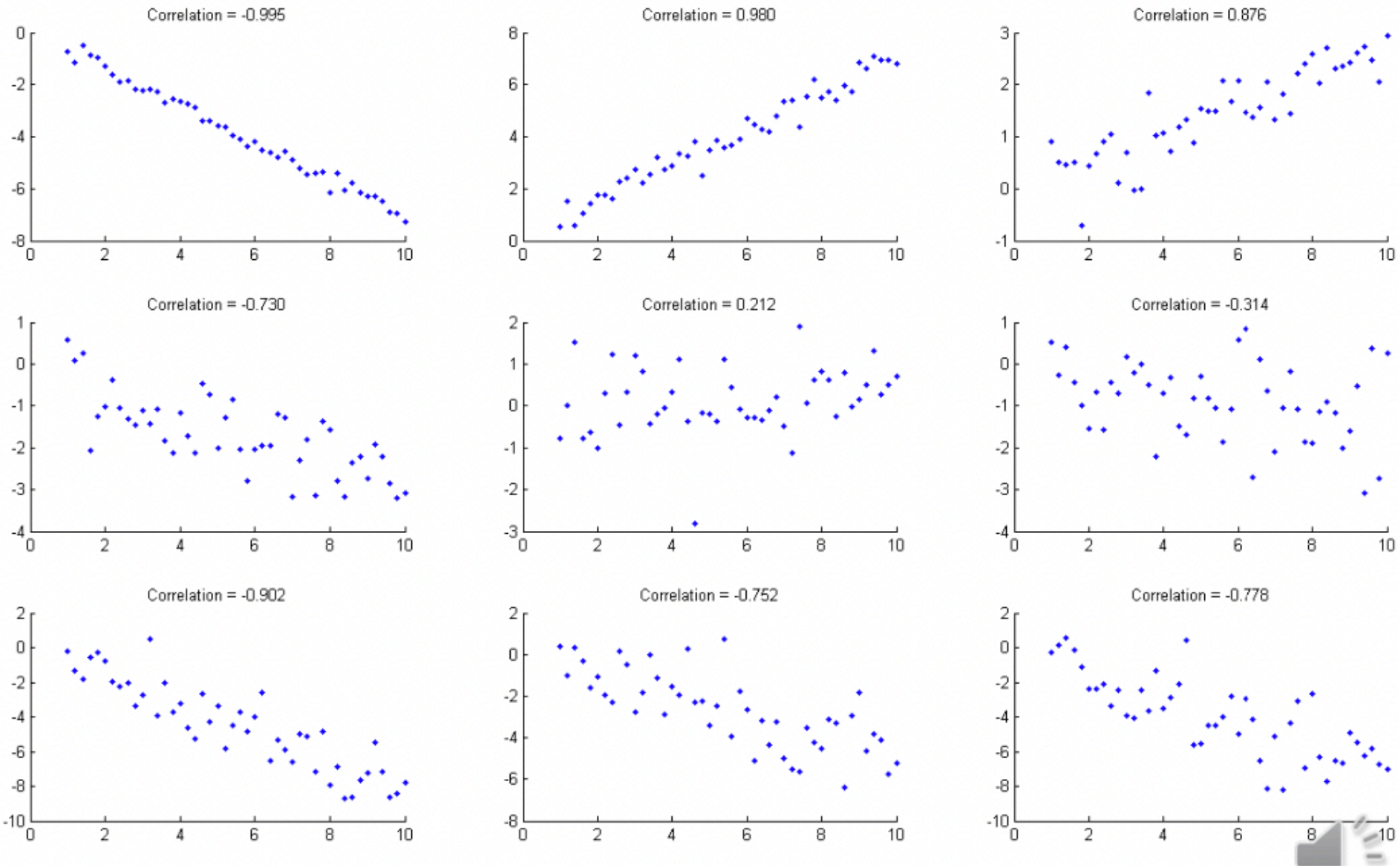
Given the pair of attributes $X^j$ and $X^k$, the correlation is defined as follows:

$$r^{jk} = \mathrm{corr}(X^j, X^k) = \frac{v^{jk}}{s^j s^k}$$

# Bivariate analysis: association indices



# Bivariate analysis: association indices



# Bivariate analysis: association indices

## Contingency tables for categorical attributes

We indicate with

$$V = \{v_1,\ldots,v_j\} \qquad U = \{u_1,\ldots,u_k\}$$

the sets of distinct values assumed by the categorical attributes $X^j$ and $X^k$.

It is possible to construct a *contingency table*, consisting of a matrix "$T$" whose generic element "$t_{rs}$" indicates the frequency with which the pair of values is present among the records of the dataset:

$$\{x_i^j = v_r\} \qquad \{x_i^k = u_s\}$$

| | | Type | | | | |
|---|---|---|---|---|---|---|
| | | *small* | *medium* | *compact* | *large* | *sporty* |
| Country | *Japan* | 7 | 6 | 3 | 4 | 10 |
| | *Other* | 12 | 8 | 12 | 1 | 4 |
| | *USA* | 3 | 16 | 7 | 12 | 11 |

# Bivariate analysis: association indices

You can calculate the sum of the values for each row and each column by obtaining the *marginal frequencies*:

$$f_r = \sum_{s=1}^{K} t_{rs} \qquad g_s = \sum_{r=1}^{J} t_{rs}$$

| | | Type | | | | | |
|---|---|---|---|---|---|---|---|
| | | *small* | *medium* | *compact* | *large* | *sporty* | |
| Country | *Japan* | 7 | 6 | 3 | 4 | 10 | **30** |
| | *Other* | 12 | 8 | 12 | 1 | 4 | **37** |
| | *USA* | 3 | 16 | 7 | 12 | 11 | **49** |
| | | **22** | **30** | **22** | **17** | **25** | **116** |

The attributes $X^j$ and $X^k$ are said to be *stochastically independent* if

$$\frac{t_{r1}}{g_1} = \frac{t_{r2}}{g_2} = \ldots = \frac{t_{rK}}{g_K}, \qquad r = 1,2,\ldots, J$$

**NOTE**: this is not exactly the case !!!

# Bivariate analysis: association indices

You can show that the previous condition is equivalent to the following one

$$\frac{t_{1s}}{f_1} = \frac{t_{2s}}{f_2} = \ldots = \frac{t_{Js}}{f_J}, \qquad s = 1,2,\ldots, K$$

In intuitive and informal terms *two attributes are stochastically independent if the analysis conducted on one of the two attributes, having the knowledge of the values assumed by the other attribute, leads to obtain the same results obtained without having the knowledge of the values assumed by the other attribute*.

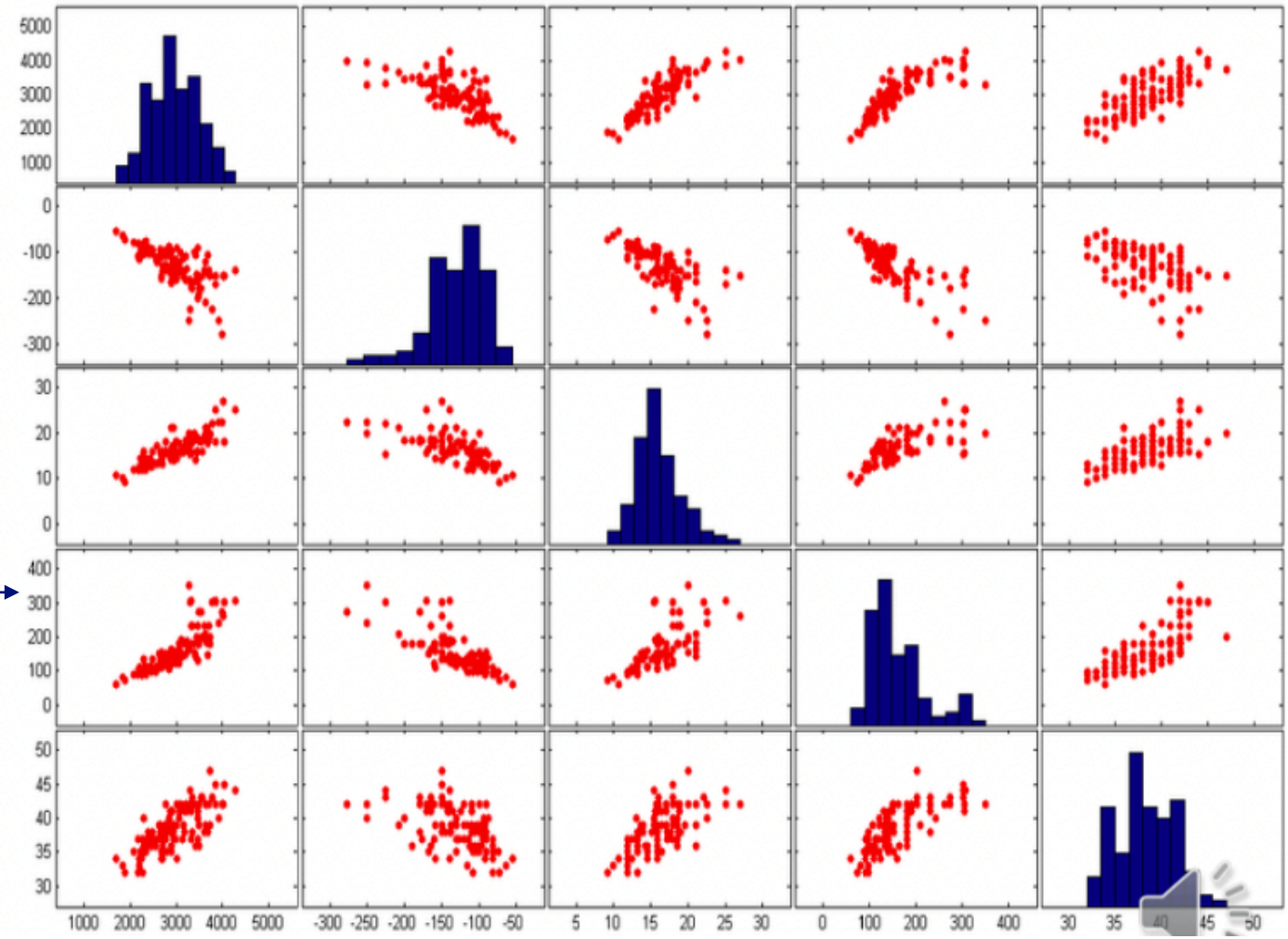
# Multivariate analysis

It is proposed to extend the notions introduced in the bivariate case to evaluate the relationships that exist between multiple attributes of a dataset.

## Graphical Analysis

All the graphical methods we present below apply only to numerical attributes.

*Matrix of scatter plots*



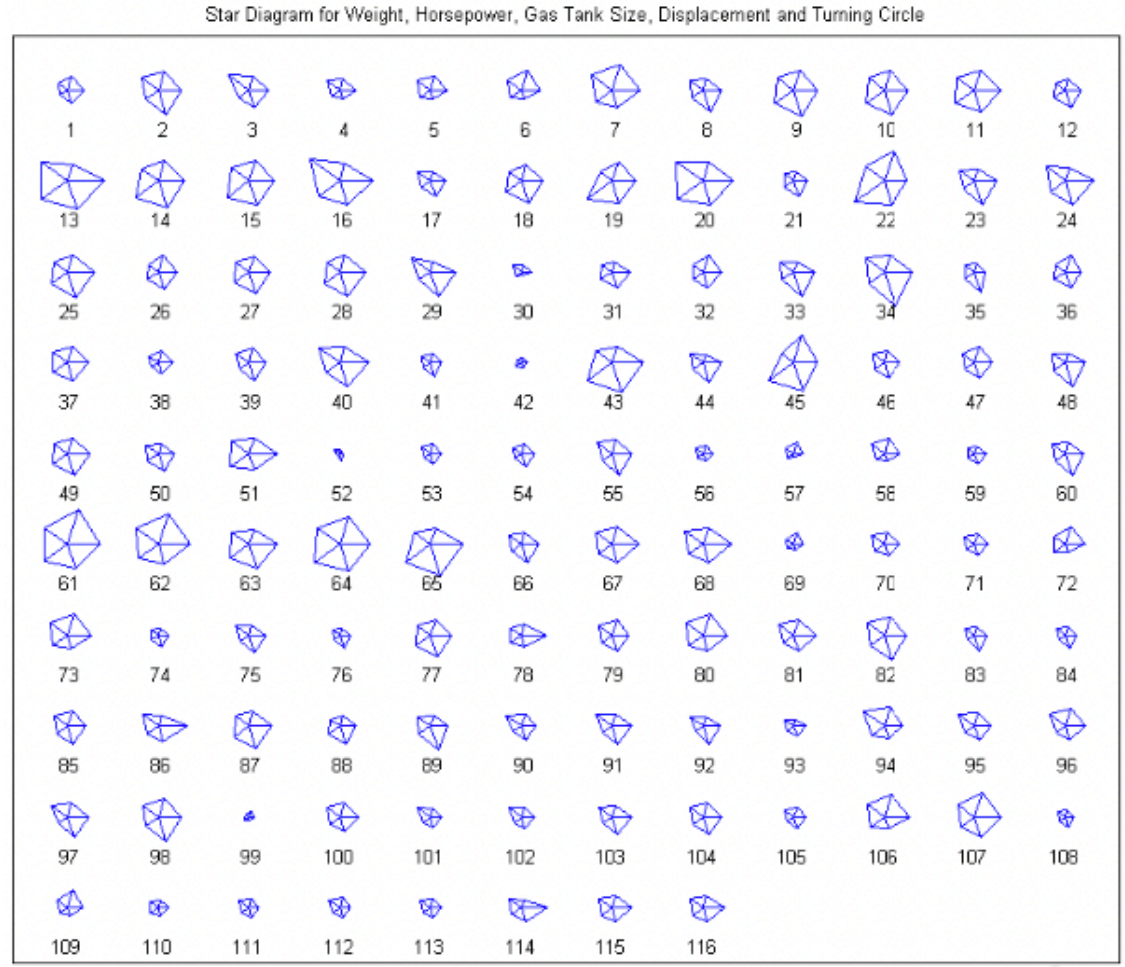
# Multivariate analysis: graphical analysis

## Star diagrams

They belong to the broader class of icon-based diagrams. They make it possible to intuitively highlight the different characteristics of each specific observation in the dataset.

Each observation corresponds to a star-shaped icon from the center of which depart *as many rays as* there are *attributes*.

The *length* of each *ray* is *equal to the value of the corresponding attribute, normalized in the interval* [0,1] in order to have a homogeneous representation of the different attributes.



# Multivariate analysis: graphical analysis