BASIC DATA ANALYSIS

Patrick Boily
Data Action Lab | uOttawa | Idlewyld Analytics
pboily@uottawa.ca

[with files from Jen Schellinck | Sysabee]

DATA ANALYSIS PIPELINE

Data modeling and conceptual analysis

Data collection

Data transformation

Data storage

Data exploration

Data analysis

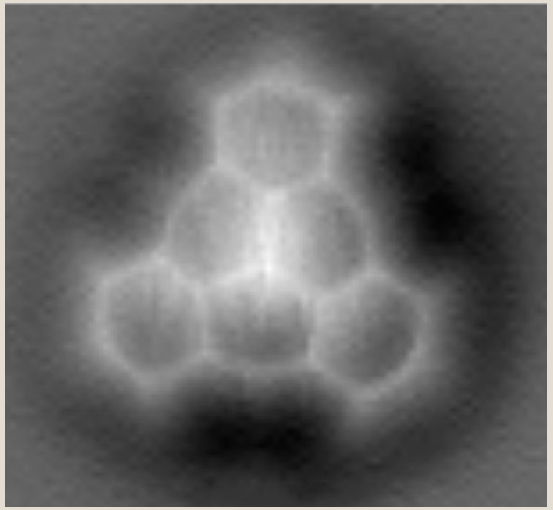
Data presentation

INSIGHTS & # CRUNCHING CORE CONCEPTS

BASIC DATA ANALYSIS

[https://upload.wikimedia.org/wikipedia/commons/d/do/TOAT_AFM.png]

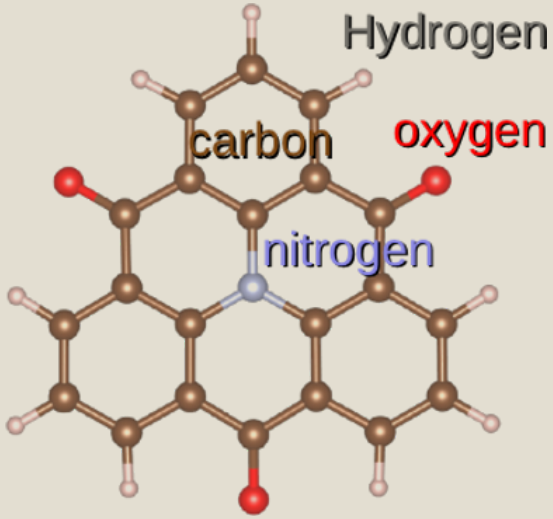
PATTERNS, GENERALIZATIONS, STRUCTURE



Pattern: a predictable, repeating regularity

Structure: an organization of elements in a system

Generalization: creation of more general or abstract concepts from more specific concepts or instances



Underlying goal during analysis: find patterns or structure in the data and **draw conclusions** *via* these patterns or structures.

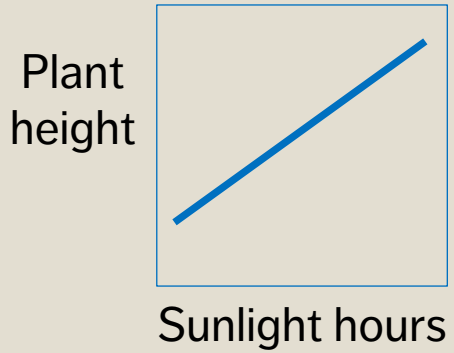
Finding patterns and structure is not pointless, per se, but it is how these discoveries are used to **draw insights** that is important.

DEPENDENT VS. INDEPENDENT VARIABLES

In an *experimental setting*:

- **control/extraneous variables:** we do our best to keep these controlled and unchanging while other variables are changed
- **independent variables:** we control their values as we suspect they influence the dependent variables
- **dependent variables:** we do not control their values; they are generated in some way during the experiment, and presumably are dependent on everything else

How do these translate over to other datasets?



DATA TYPES

Numerical data: integers or continuous numbers

- 1, 7, 34.654, 0.000004

Text data: strings of text – may be restricted to a certain number of characters

- “Welcome to the park”, “AAAAA”, “345”, “45.678”

Categorical data: a fixed number of values, may be numeric or represented by strings. **There is no specific or inherent ordering**

- (‘red’,’blue’,’green’), (‘1’,’2’,’3’)

Ordinal data: categorical data with an inherent ordering. Unlike integer data, the spacing between values is **not** defined

- (very cold, cold, tepid, warm, super hot)

CATEGORICAL → NUMERICAL

Categorical data can be turned into numerical data by generating **frequency counts** of the different values of the categorical variable.

This in turn allows us to apply numerical analysis techniques.

House colour	Frequency
red	40
blue	13
green	2

SPECIAL ROLE OF CATEGORICAL DATA

Categorical data plays a special role:

- in data science, **categorical variables** come with a **pre-defined** set of values
- in experimental science, a **factor** is an independent variable with its levels being defined (it may also be viewed as a category of treatment)
- in business analytics, these are **dimensions** (with members) vs. measures

However they are labeled, they can be used to **subset** or **roll up/summarize** the data.

HIERARCHICAL/NESTED/MULTILEVEL DATA/MODELS

When a categorical variable has multiple levels of abstraction, new categorical variables can be created from these levels.

The ‘new’ categorical variable has pre-defined relationships with the more detailed level.

We can often zoom in and out with time/space variables.

Granularity of the data: what is the ‘maximum zoom’?

Year	Quarter	Count_Q
2012	1	34
2012	2	12
2012	3	52
2012	4	0
2013	1	21
2013	2	9
2013	3	112
2103	4	8

Year	Count_Y
2012	98
2013	150

INSIGHTS & # CRUNCHING CORE TECHNIQUES

BASIC DATA ANALYSIS

DATA SUMMARIZING

Min: smallest value

Max: largest value

Median: “middle” value

Mode: most frequent value

Unique Values: list of unique values

etc.

Signal	Type
4.31	Blue
5.34	Orange
3.79	Blue
5.19	Blue
4.93	Green
5.76	Orange
3.25	Orange
7.12	Orange
2.85	Blue

ROLLING-UP DATA

We can perform operations over a set (or subset) of the data, typically over its **columns**.

Such an operation is akin to **compressing** or '**rolling-up**' the many data values into a single representative value.

Examples: 'mean', 'sum', 'count', 'variance', etc.

We can apply the same roll-up function to many different columns, providing a **mapping** (list) of columns to values.

Signal	Type
4.31	Blue
5.34	Orange
3.79	Blue
5.19	Blue
4.93	Green
5.76	Orange
3.25	Orange
7.12	Orange
2.85	Blue

Count	Signal avg	Signal stdev	Type mode
9	4.73	1.33	Blue/ Orange

CONTINGENCY/PIVOT TABLES

Contingency table: a table which examines the relationship between two categorical variables *via* their relative (**cross-tabulation**).

Pivot table: a table generated by applying operations (sum, count, mean, etc.) to variables, possibly based on another (categorical) variable. Contingency tables as special cases of pivot tables.

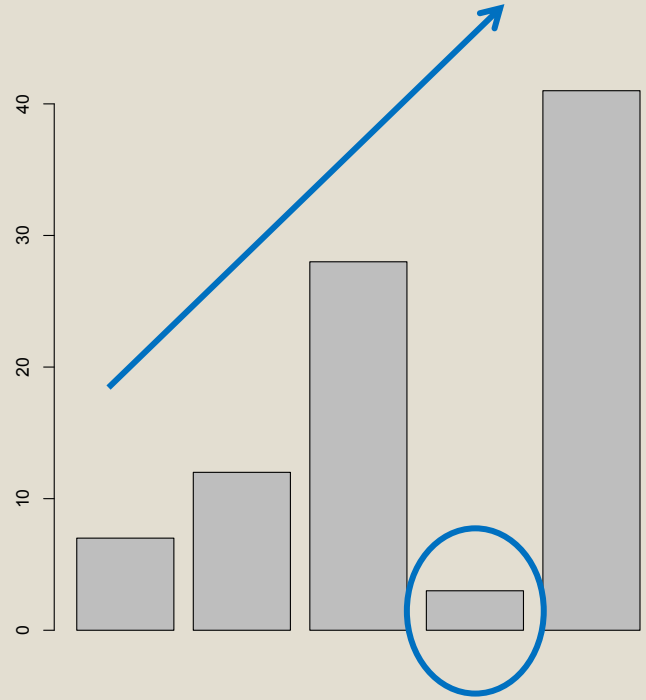
	Large	Medium	Small
Window	1	32	31
Door	14	11	0

Type	Count	Signal avg	Signal stdev
Blue	4	4.04	0.98
Green	1	4.93	N.A.
Orange	4	5.37	1.60

ANALYSIS THROUGH VISUALIZATION

Analysis (broad definition):

- identifying patterns or structure
- adding meaning to these patterns or structure by **interpreting** them in the context of the system.



Option 1: use analytical methods to achieve this.

Option 2: visualize the data and use the brain's analytic power (perceptual) to reach meaningful conclusions about these patterns.

We will discuss further.

DATA DESCRIPTIONS (IN-DEPTH)

BASIC DATA ANALYSIS

DATA DESCRIPTIONS (REPRISE)

In a sense, the underlying reason for any analysis is to reach **data understanding**.

Studies and experiments give rise to **units**, which are typically described with **variables** (and measurements).

Variables are either **qualitative** (categorical) or **quantitative** (numerical):

- categorical variables take values (levels) from a finite set of **classes**
- numerical variables take values from a (potentially infinite) set of **quantities**

EXAMPLES

- **Age** is a numerical variable, measured in years, although is is often reported to the nearest year integer, or in an age range of years (in which case it is ordinal).
- Typical numerical variables include distance in m, volume in cm^3, etc.
- **Disease diagnosis** is a categorical variable with 2 categories (positive/negative).
- **Compliance with a standard** is a categorical variable: there could be 2 levels (compliant/non-compliant) or more (compliance, minor non-compliance issues, major non-compliance issues).
- **Count variables** are numerical variables.

NUMERICAL SUMMARIES

In a first pass, a variable can be described along 2 dimensions: **centrality** & **spread** (**skew** and **kurtosis** are also used sometimes).

Centrality measures include:

- **median**, **mean**, **mode** (less frequently)

Spread (or **dispersion**) measures include:

- **standard deviation** (sd), **variance**, **quartiles**, **inter-quartile range** (IQR), **range** (less frequently).

The median, range and the quartiles are easily calculated from an **ordered list** of data.

MEDIAN

The **median** of a quantitative variable with n observations is a value which splits the ordered data into 2 equal subsets: half the observations are below (or equal to) the median, and half above (or equal to) it.

If n is **odd**, then the median is the $\frac{n+1}{2}$ —ordered observation.

If n is **even**, then the median is any value between the $\frac{n}{2}$ and $\frac{n}{2}+1$ ordered observations (we usually take their average).

The **procedure** is simple: order the data and follow the even/odd rules to the letter.

MEDIAN

1. Imagine a quantitative variable with $n = 5$ observations, taking the values: $4,6,1,3,7$.

Start by ordering the values: $1,3,4,6,7$.

$n = 5$ is odd, so use the first rule, i.e. look for the $\frac{n+1}{2} = \frac{5+1}{2} = 3^{\text{rd}}$ observation, which is **4**.

Note that there are 2 observations below 4 $(1,3)$ and 2 observations above 4 $(6,7)$.

---

2. Imagine a quantitative variable with $n = 6$ observations, taking the values: $4,6,1,3,7,23$.

Start by ordering the values: $1,3,4,6,7,23$.

$n = 6$ is even, so use the second rule, i.e. look for any value between the $\frac{n}{2} = \frac{6}{2} = 3^{\text{rd}}$ and the $\frac{n}{2} + 1 = \frac{6}{2} + 1 = 4^{\text{th}}$ observation, say **5.2**.

Note that there are 3 observations below 5.2 $(1,3,4)$ & 3 observations above 5.2 $(6,7,23)$.

MEAN

The **mean** of a sample is simply the **arithmetic average** of its observations:

$$\text{mean} = \frac{x_1 + \cdots + x_n}{n}$$

Other means exist, such as the **harmonic** mean and the **geometric** mean.

Examples:

- $\text{mean}(4,6,1,3,7) = \frac{4+6+1+3+7}{5} = \frac{21}{5} = \mathbf{4.2} \approx 4 = \text{median}(4,6,1,3,7)$
- $\text{mean}(4,6,1,3,7,23) = \frac{4+6+1+3+7+23}{6} = \frac{44}{6} = \mathbf{7.3} \approx 5.2 = \text{median}(4,6,1,3,7,23)$

MEAN OR MEDIAN?

Which measure of centrality should be used to report on the data?

The mean is **theoretically supported** (CLT, which won't be discussed here).

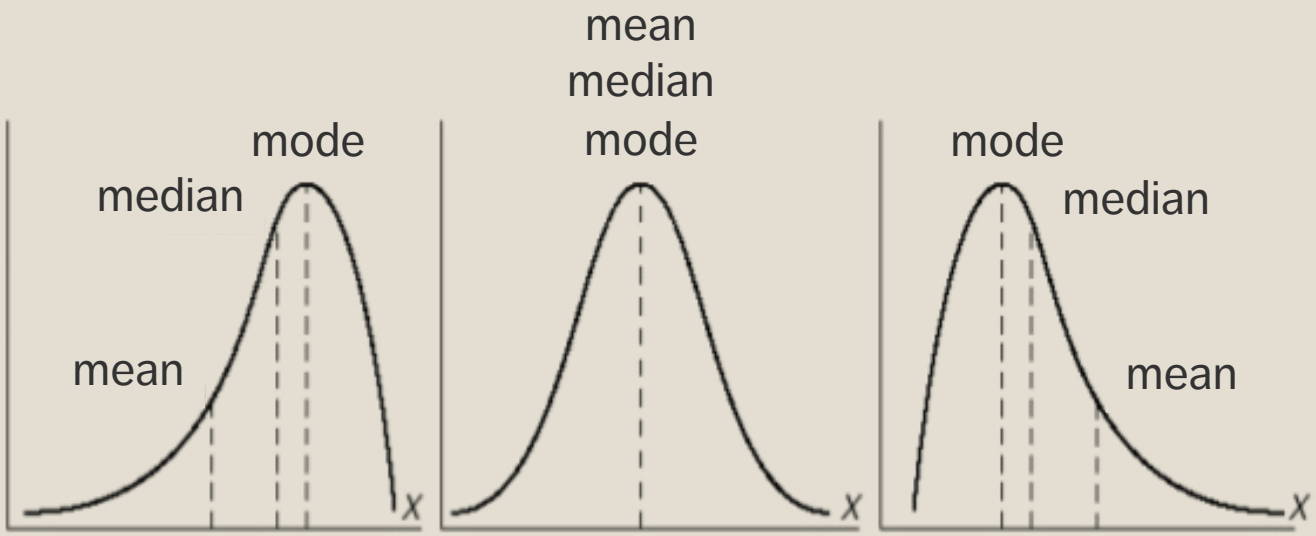
If the data distribution is roughly **symmetric** then both values will be near one another.

If the data distribution is **skewed,** then the mean is **pulled toward the long tail** and as a result gives a distorted view of the true centre.

Consequently, medians are generally used for house prices, incomes etc.

The median is **robust** against outliers and incorrect readings whereas the mean is not.

MEAN OR MEDIAN?

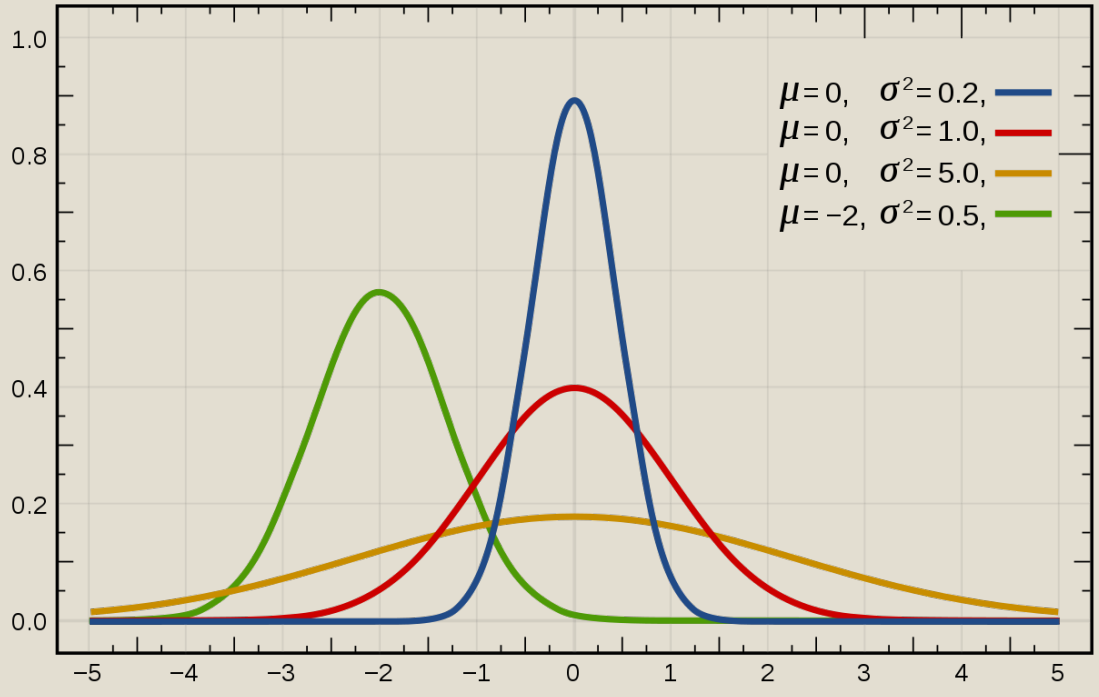


[Wikipedia]

STANDARD DEVIATION

The centrality measures provide an idea as to where the variable's values are "**massed**".

The standard deviation (sd) provides a notion of its **spread**; higher sd means higher spread.



STANDARD DEVIATION

The **sd** is built from a **fancy average** of the variable's observations:

$$\text{sd} = \sqrt{\frac{(x_1 - \text{mean})^2 + \cdots + (x_n - \text{mean})^2}{n}}.$$

Examples:

- $\text{sd}(4,6,1,3,7) = \sqrt{\frac{(4-4.2)^2+(6-4.2)^2+(1-4.2)^2+(3-4.2)^2+(7-4.2)^2}{5}} \approx \mathbf{2.14}$
- $\text{sd}(4,6,1,3,23) = \sqrt{\frac{(4-7.3)^2+(6-7.3)^2+(1-7.3)^2+(3-7.3)^2+(7-7.3)^2+(23-7.3)^2}{6}} \approx \mathbf{3.98}$

QUANTILES

Another way to provide information about the spread of the data is with the help of **centiles**, **deciles**, or **quartiles**.

The **lower quartile** Q_1 of a column with n entries is a numerical value which splits the ordered data into 2 unequal subsets: 25% of the observations are **below** (or at) Q_1 and 75% of the observations are **above** (or at) Q_1 .

Similarly, the **upper quartile** Q_3 splits the ordered data into 75% of the observations **below** (or at) Q_3 and 25% of the observations **above** (or at) Q_3.

The median can be interpreted as the **middle quartile** Q_2 of the data, the minimum as Q_0, and the maximum as Q_4; $(Q_0, Q_1, Q_2, Q_3, Q_4)$ represent the **5-pt summary** of the data.

QUANTILES

Centiles p_i, where $i = 0, \dots, 100$, and **deciles** d_j, where $j = 0, \dots, 10$, run through different **splitting percentages**

$$p_{25} = Q_1, p_{75} = Q_3, p_{50} = d_5 = Q_2, \text{etc.}$$

Procedure:

1. **sort** the n observations in **increasing order**

   $$y_1 \leq y_2 \leq \cdots \leq y_n.$$

   The smallest y_1 has **rank** 1 and the largest y_n has **rank** n.

2. a value between the observations of ranks $\left\lfloor \frac{n}{4} \right\rfloor$ and $\left\lfloor \frac{n}{4} \right\rfloor + 1$ is a **lower quartile** Q_1
3. a value between the observations of ranks $\left\lfloor \frac{3n}{4} \right\rfloor$ and $\left\lfloor \frac{3n}{4} \right\rfloor + 1$ is an **upper quartile** Q_3

QUANTILES

4. a value between the observations of ranks $\left\lfloor \frac{in}{100} \right\rfloor, \left\lfloor \frac{in}{100} \right\rfloor + 1$ is a **centile** $p_i, i = 1, \dots, 99$
5. a value between the observations of ranks $\left\lfloor \frac{jn}{10} \right\rfloor, \left\lfloor \frac{jn}{10} \right\rfloor + 1$ is a **decile** $d_j, j = 1, \dots, 9$

In practice, we take the average value between $\left\lfloor \frac{kn}{m} \right\rfloor$ and $\left\lfloor \frac{kn}{m} \right\rfloor + 1$ to obtain a unique m −quantile of order k for the data, where $k = 1, \dots, m - 1$.

Examples:

- $Q_1(1,3,4,6,7) = \frac{1}{2}\left(y_{\lfloor 5/4 \rfloor} + y_{\lfloor 5/4 \rfloor + 1}\right) = \frac{1}{2}\left(y_{\lfloor 1.25 \rfloor} + y_{\lfloor 1.25 \rfloor + 1}\right) = \frac{1}{2}(y_1 + y_2) = \frac{1}{2}(1 + 3) = \mathbf{2}$
- $d_7(1,3,4,6,7,23) = \frac{1}{2}\left(y_{\lfloor 7 \cdot 6/10 \rfloor} + y_{\lfloor 7 \cdot 6/10 \rfloor + 1}\right) = \frac{1}{2}\left(y_{\lfloor 4.2 \rfloor} + y_{\lfloor 4.2 \rfloor + 1}\right) = \frac{1}{2}(y_4 + y_5) = \frac{1}{2}(6 + 7) = \mathbf{13/2}$

OTHER MEASURES

Centrality:

- the **mid-range** of a variable is $\frac{\min+\max}{2} = \frac{Q_0+Q_4}{2}$.
- the **tri-mean** of a variable is $\frac{Q_1+2Q_2+Q_3}{4}$.

Dispersion:

- the **range** of a variable is $\max - \min = Q_4 - Q_0$.
- the **inter-quartile range** of a variable is $\mathrm{IQR} = Q_3 - Q_1$.

In general, we can gleam a better understanding of a variable through **multiple** measures.

VISUAL SUMMARIES – BOXPLOT

The **boxplot** is a quick way to present a graphical summary of a univariate distribution.

Draw a box along the observation axis, with endpoints at Q_1 and Q_3, and with a "belt" at the median.

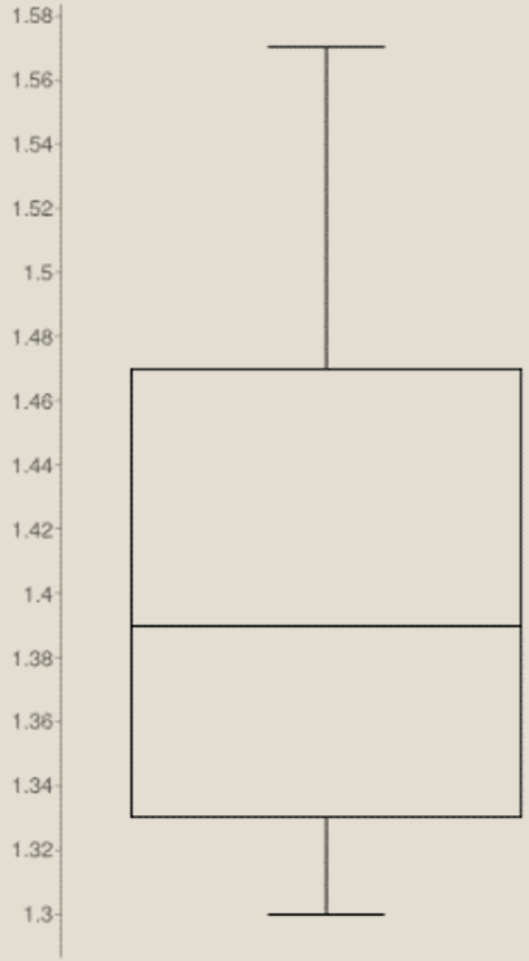
Plot a line extending from Q_1 to the smallest obs. less than $1.5 \times \mathrm{IQR}$ to the left of Q_1.

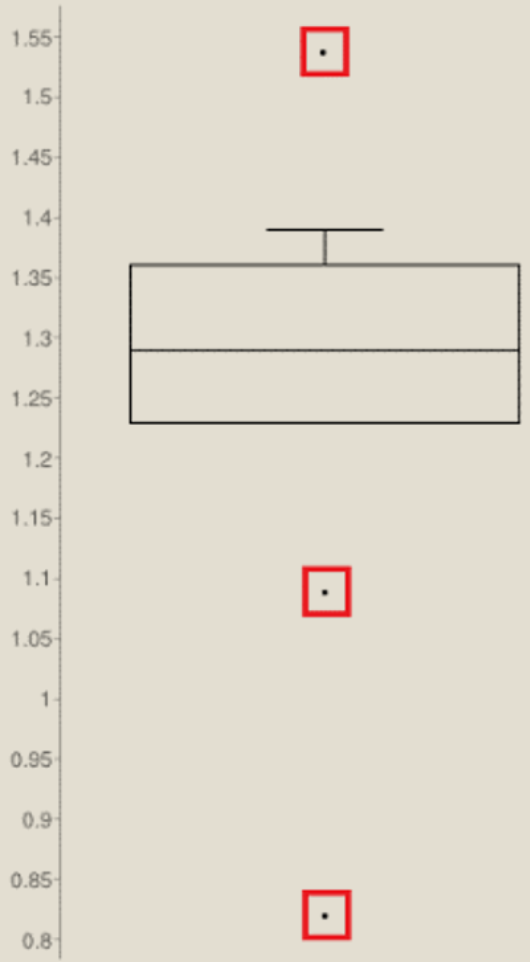
Plot a line extending from Q_3 to the smallest obs. more than $1.5 \times \mathrm{IQR}$ to the right of Q_3.

Any suspected outlier is plotted separately.

Queuing dataset: arrival rates (left), processing rates (right)

EXAMPLES





VISUAL SUMMARIES – HISTOGRAM

Histograms can also provide an indication of the distribution of a variable.

They should include/contain the following information:

- the range of the histogram is $r = Q_4 - Q_0$;
- the number of bins should approach $k = \sqrt{n}$, where n is the number of obs.;
- the bin width should approach r/k, and
- the frequency of observations in each bin should be added to the chart.

EXAMPLE

Consider the daily number of car accidents in Sydney over a 40-day period:

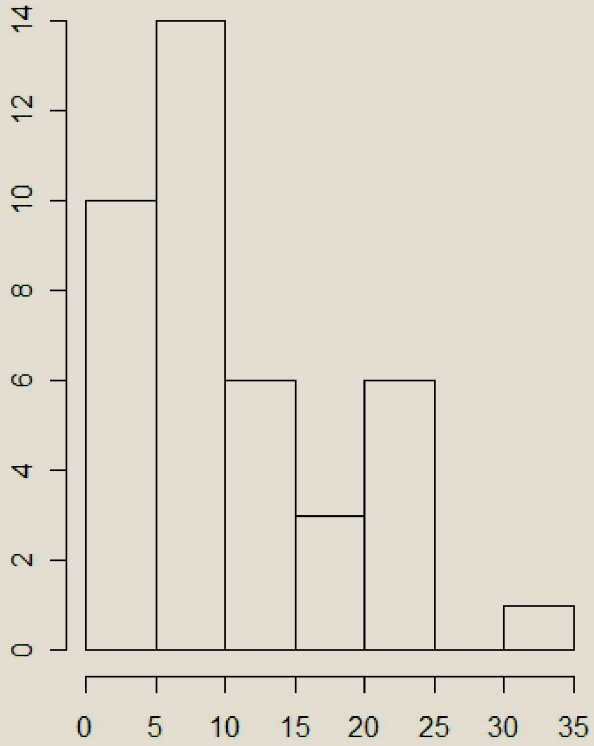
6, 3, 2, 24, 12, 3, 7, 14, 21, 9, 14, 22, 15,
2, 17, 10, 7, 7, 31, 7, 18, 6, 8, 2, 3, 2, 17,
7, 7, 21, 13, 23, 1, 11, 3, 9, 4, 9, 9, 25

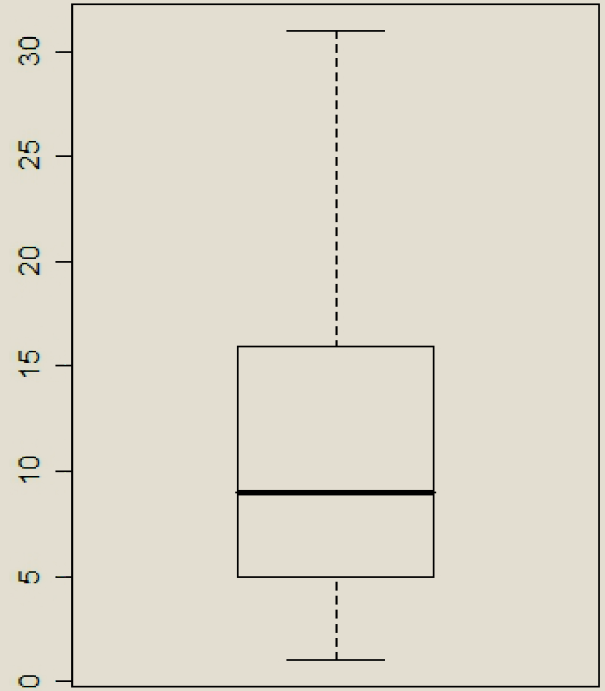
The sorted values are:

1 2 2 2 2 3 3 3 3 4 6 6 7 7 7 7 7
7 8 9 9 9 9 10 11 12 13 14 14 15 17
17 18 21 21 22 23 24 25 31

min	Q_1	med	Q_3	max
1	5.5	9	15.5	31

Is it more likely that one would see between 5-15 accidents on a given day, or between 25-35?





SKEWNESS

If the data distribution is **symmetric** then median = mean, and Q_1 and Q_3 are equidistant from the median: $Q_3 - Q_2 = Q_2 - Q_1$.

Otherwise:

- if $Q_3 - Q_2 > Q_2 - Q_1$, then the data distribution is **skewed to the right**
- If $Q_3 - Q_2 < Q_2 - Q_1$, then the data distribution is **skewed to the left**

In the previous example,

$$Q_3 - Q_2 = 15.5 - 9 = \mathbf{6.5} > \mathbf{3.5} = 9 - 5.5 = Q_2 - Q_1,$$

so the distribution is skewed to the right.

SKEWNESS

skewed to the left | symmetric | skewed to the right

The **shape of a dataset** can be used to suggest an analytical model for the situation of interest.

CORRELATION

BASIC DATA ANALYSIS

MOTIVATING EXAMPLE

Consider the following data, consisting of $n = 20$ paired measurements (x_i, y_i) of hydrocarbon levels (x) and pure oxygen levels (y) in fuels:

```
x:  0.99  1.02  1.15  1.29  1.46  1.36  0.87  1.23  1.55  1.40
y: 90.01 89.05 91.43 93.74 96.73 94.45 87.59 91.77 99.42 93.65

x:  1.19  1.15  0.98  1.01  1.11  1.20  1.26  1.32  1.43  0.95
y: 93.54 92.52 90.56 89.54 89.85 90.39 93.25 93.41 94.98 87.33
```

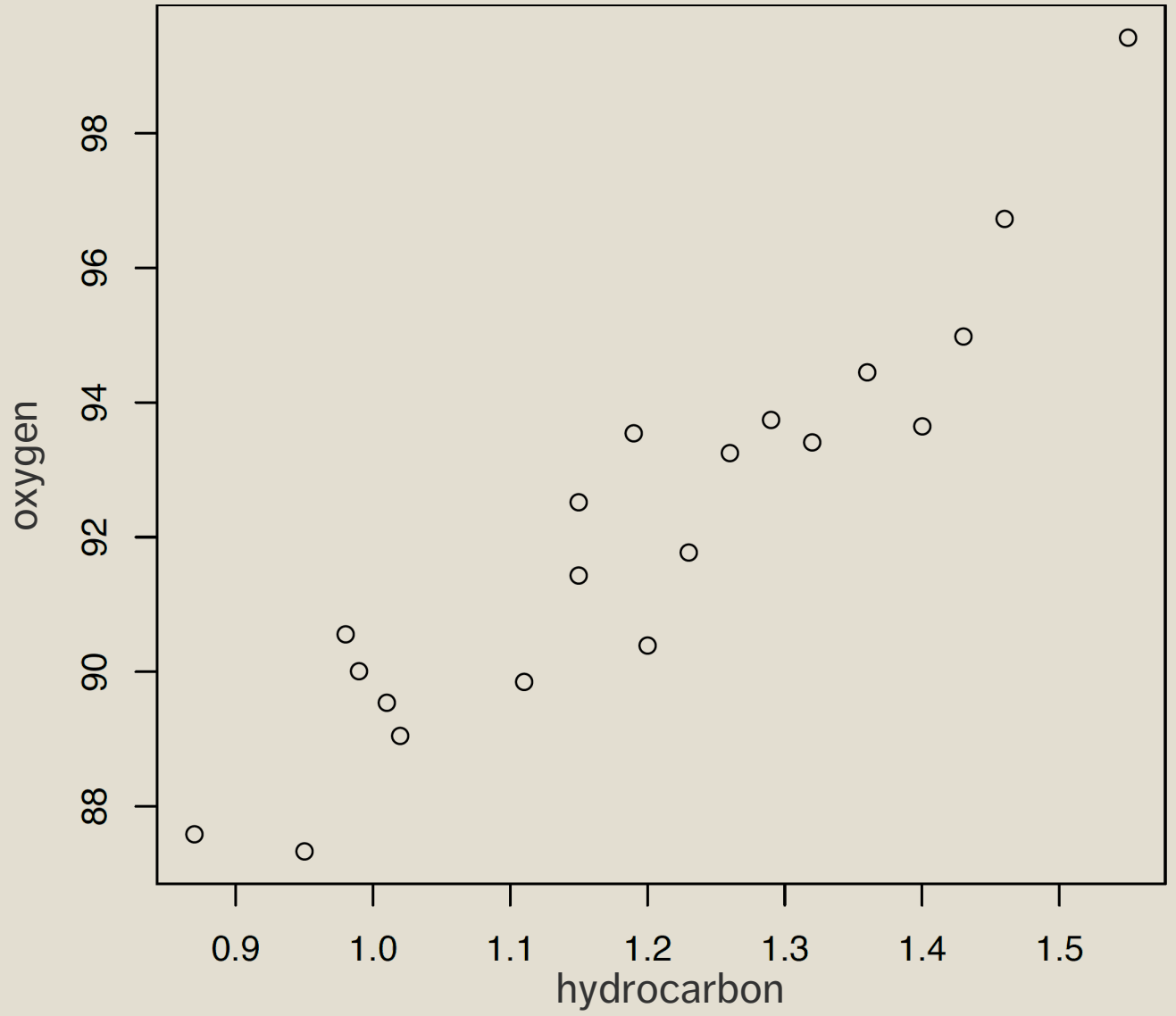
Goals:

- measure the **strength of association** between x and y
- **describe** the relationship between x and y

MOTIVATING EXAMPLE

A graphical display provides an initial description of the relationship.

It seems that points lie around a hidden line!



COEFFICIENT OF CORRELATION

For paired data (x_i, y_i), $i = 1, \dots, n$, the **correlation coefficient** of x and y is

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

This correlation is defined only if $S_{xx}, S_{yy} \neq 0$, neither the x_i nor the y_i are constant.

The variables x and y are **uncorrelated** if $\rho_{XY} = 0$ (or is very small, in practice), and they are **correlated** if $\rho_{XY} \neq 0$ (or $|\rho_{XY}|$ is "large", in practice).

For the hydrocarbon data, $S_{xy} \approx 10.18$, $S_{xx} \approx 0.68$, $S_{yy} \approx 173.38$ et

$$\rho_{XY} = \frac{10.18}{\sqrt{0.68 \cdot 173.38}} \approx \mathbf{0.94} \text{ (high correlation).}$$

PROPERTIES AND INTERPRETATION

- ρ_{XY} is **unaffected by changes of scale or origin**. Adding constants to x does not change the values of $x_i - \bar{x}$, say, and multiplying x and y by constants changes both the numerator and denominator equally.

- ρ_{XY} is **symmetric** in x and y (i.e., $\rho_{XY} = \rho_{YX}$).

- $-1 \leq \rho_{XY} \leq 1$;

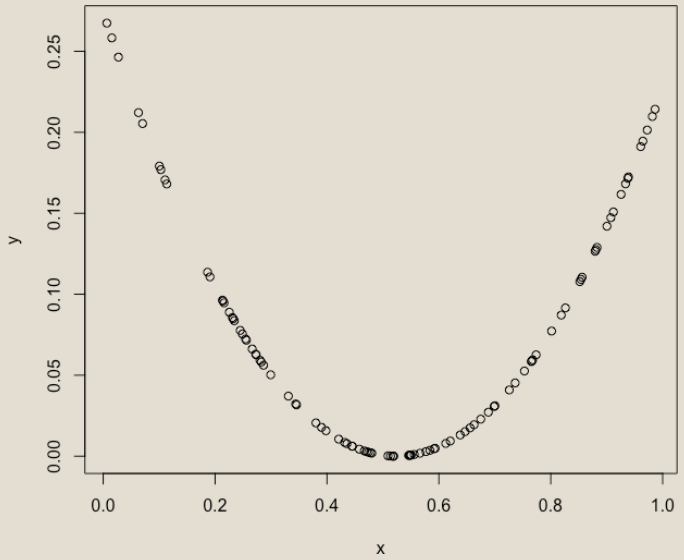
- if $\rho_{XY} = \pm 1$, then the observations (x_i, y_i) all lie on a straight line with a positive (negative) slope;

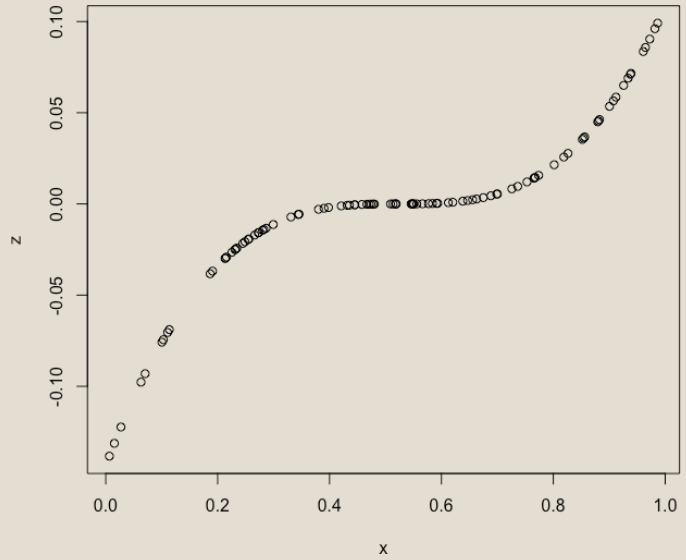
- the sign of ρ_{XY} reflects the trend of the points;

PROPERTIES AND INTERPRETATION

- **IMPORTANT:** a high correlation coefficient value $|\rho_{XY}|$ does not necessarily imply a **causal relationship** between the two variables;

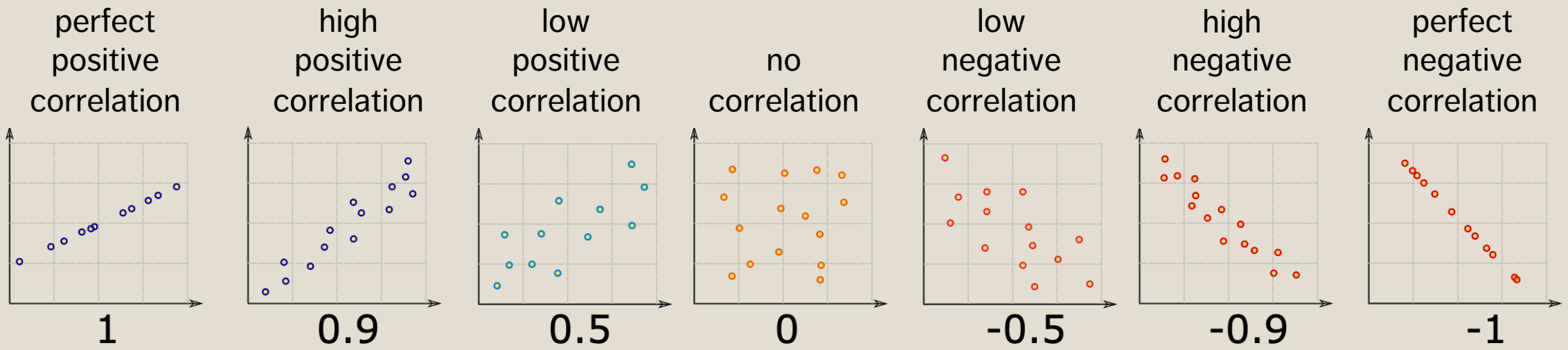
- note that x and y can have a very strong **non-linear** relationship without ρ_{XY} reflecting it (-0.12 on the left, 0.93 on the right).



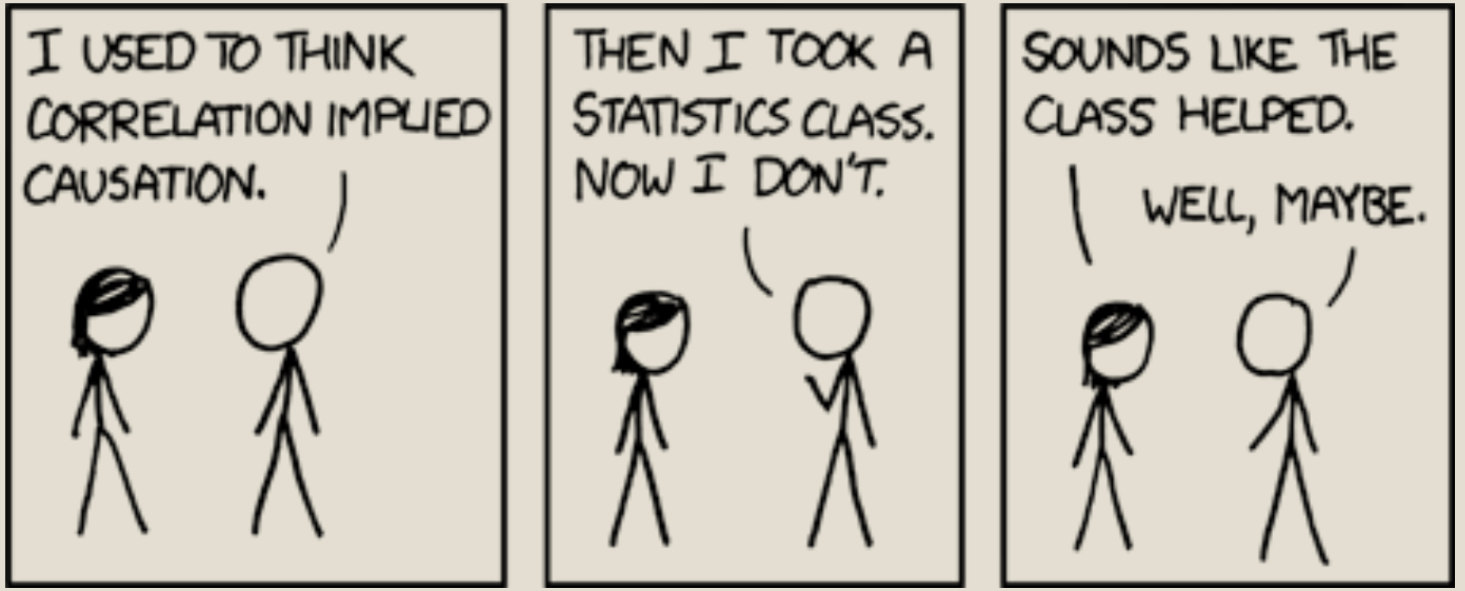


[Unknown author]

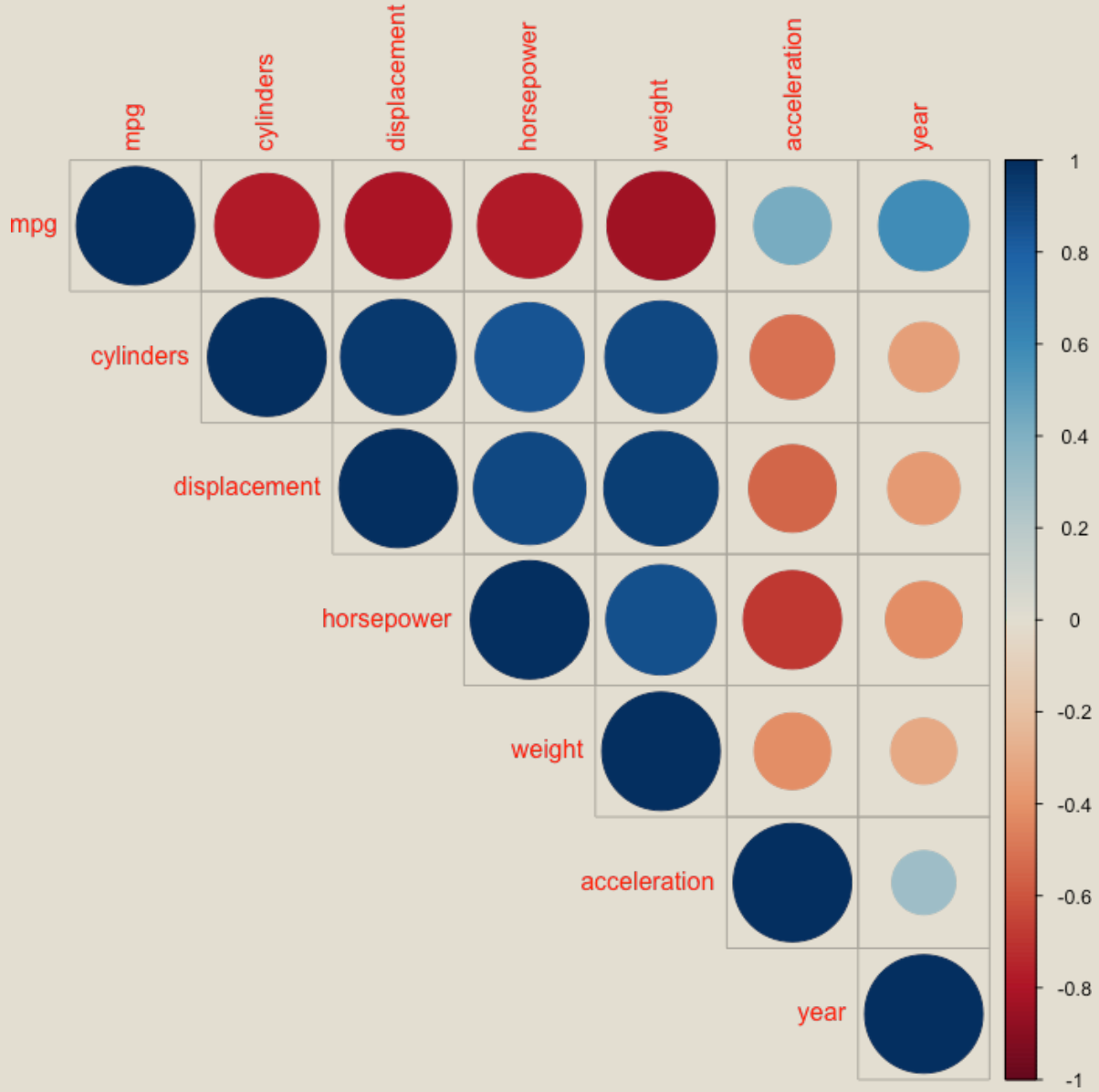
PROPERTIES AND INTERPRETATION



[https://xkcd.com/552/]



Correlation doesn't imply causation, but it does waggle its eyebrows suggestively and gesture furtively while mouthing 'look over there'.


mpg
cylinders
displacement
horsepower
weight
acceleration
year
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

REGRESSION ANALYSIS

BASIC DATA ANALYSIS

REGRESSION MODELING

The most common data modeling methods are regressions, both **linear** and **logistic.**

About 80% of real data applications end up using a simple regression as their final model, typically after very careful **data preparation**, **encoding**, and creation of variables.

There are several reasons for their frequent use:

- generally **straightforward** to understand and to train
- mean square error (MSE) objective function has a closed-form linear solution
- system of equations can usually be solved through matrix inversion or linear manipulation

REGRESSION MODELING

The data structure of a general modeling task is represented by

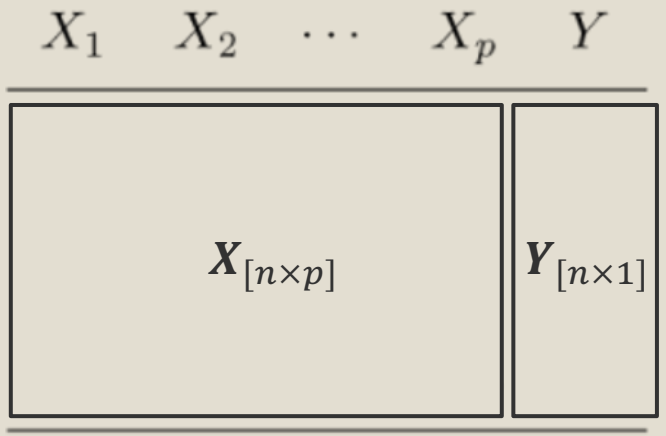
X_1	X_2	$\cdots$	X_p	Y
x_{11}	x_{12}	$\cdots$	x_{1p}	y_1
x_{21}	x_{22}	$\cdots$	x_{2p}	y_2
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_{n1}	x_{n2}	$\cdots$	x_{np}	y_n

We consider p **independent variables** X_i (the predictors), and we attempt to predict the **dependent variable** Y (the response).

In order to simplify the discussion in what follows, we introduce the matrix notation

$$\boldsymbol{X}_{[n\times p]}, \boldsymbol{Y}_{[n\times 1]}, \boldsymbol{\beta}_{[p\times 1]},$$

where n is the # of observations and p is the # of independent variables.



LINEAR REGRESSION

The basic assumption of linear regression is that the dependent variable y can be **approximated** by a linear combination of the independent variables as follows:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $\beta \in \mathbb{R}^p$ is to be determined based on the training set, and for which

$$\mathrm{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = 0, \qquad \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T|\boldsymbol{X}) = \sigma^2 I.$$

Typically, the errors are also assumed to be normally distributed, that is:

$$\boldsymbol{\varepsilon}|\boldsymbol{X} \sim N(0, \sigma^2 I).$$

LINEAR REGRESSION

If $\hat{\beta}_i$ is the estimate of the true coefficient β_i, the **linear regression** model associated with the data is

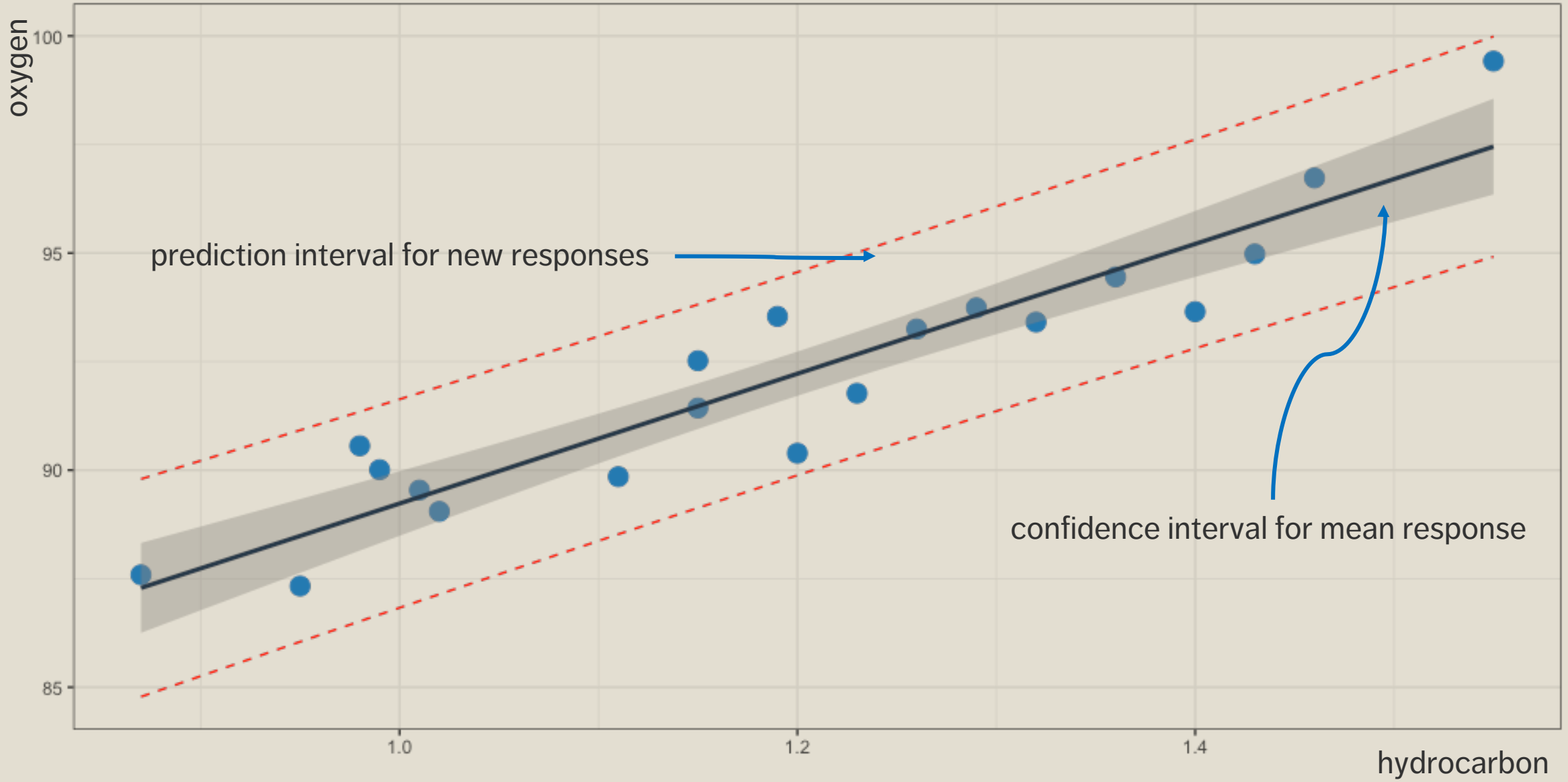
$$\widehat{\boldsymbol{Y}}(\boldsymbol{x}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$$

In matrix form, the regression problem requires a solution $\widehat{\boldsymbol{\beta}}$ to the **normal equation** $\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}^T\boldsymbol{Y}$.

When the symmetric positive definite matrix $\boldsymbol{X}^T\boldsymbol{X}$ is invertible, the fitted coefficient is simply $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{Y})$.

Note that $\boldsymbol{X}^T\boldsymbol{X}$ is a $p \times p$ matrix, which makes the inversion “easier” to compute, relatively speaking, when n is large.

$$\text{oxygen} = 14.95 \times \text{hydrocarbon} + 74.28$$



TIME SERIES AND CONTROL CHARTS

BASIC DATA ANALYSIS

“NASA engineers did not identify the association between unexpectedly low launch-pad temperatures and O-ring failures in the space shuttle booster rockets.

They interpreted this critical signal as simply chance variation in the failure of the joints.

Lack of this insight was critical in the decision to launch the *Challenger* on its final and disastrous flight.”

Vaughan, D. [1997], *The Challenger Launch Decision: Risky Technology, Culture and Deviance at NASA*, p.383

STATISTICAL PROCESS MONITORING

Processes are often subject to **variability:**

- variability due the **cumulative effect** of many small, essentially unavoidable causes (a process that only operates with such **common causes** is said to be **in (statistical) control**);
- variability due to **special causes**, such as improperly adjusted machines, poorly trained operators, defective materials, etc. (the variability is typically much larger for special causes, and such processes are said to be **out of (statistical) control**).

The aim of **statistical process monitoring** (SPM) is to identify occurrence of special causes.

TIME SERIES

Consider some observations $\{x_1, \dots, x_n\}$, arising from some process.

In practice, the index i is often a **time index** or a **location index**, i.e. the x_i are observed in **sequence** or in **regions**.

In the first case, the observations form a **time series**.

The processes that generate observations couls change over time/location due to:

- **external factors** (war, pandemic, election, etc.), or
- **internal factors** (policy change, modification of manufacturing process, etc.).

TIME SERIES

The mean and standard deviation might not provide a useful summary of the situation.

To get a sense of what is going on, it could be preferable to **plot the data** in the **order that it has been collected** (or according to geographical regions).

The horizontal coordinate represents:

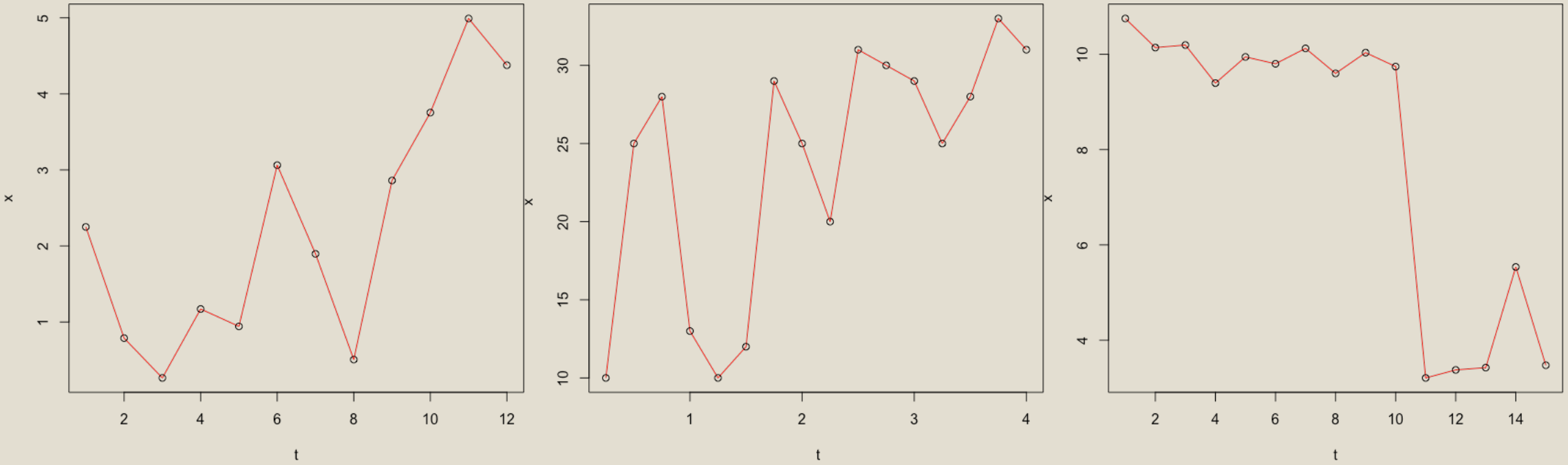
- the **time of collection** t (order, day, week, quarter, year, etc.), or
- the **location** i (country, province, city, branch, etc.).

The vertical coordinate represents the observations of interest x_t or x_i.

We then look for **trends**, **cycles**, **shifts**, etc.

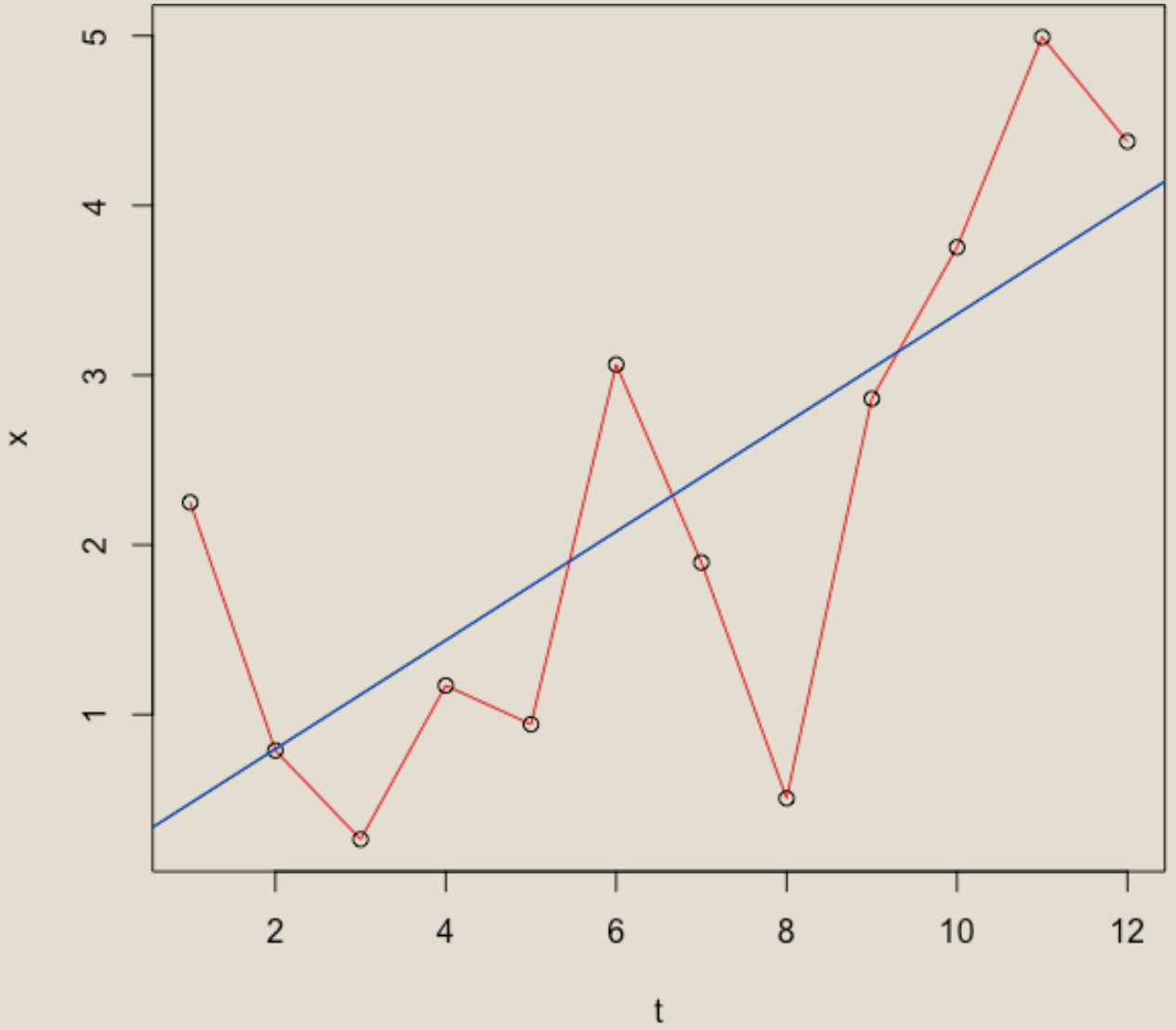
EXAMPLES

The following time series record sales x (in 10,000$) for 3 different products, against the passage of time t in years (left), quarters (middle), weeks (right). Is any action necessary?



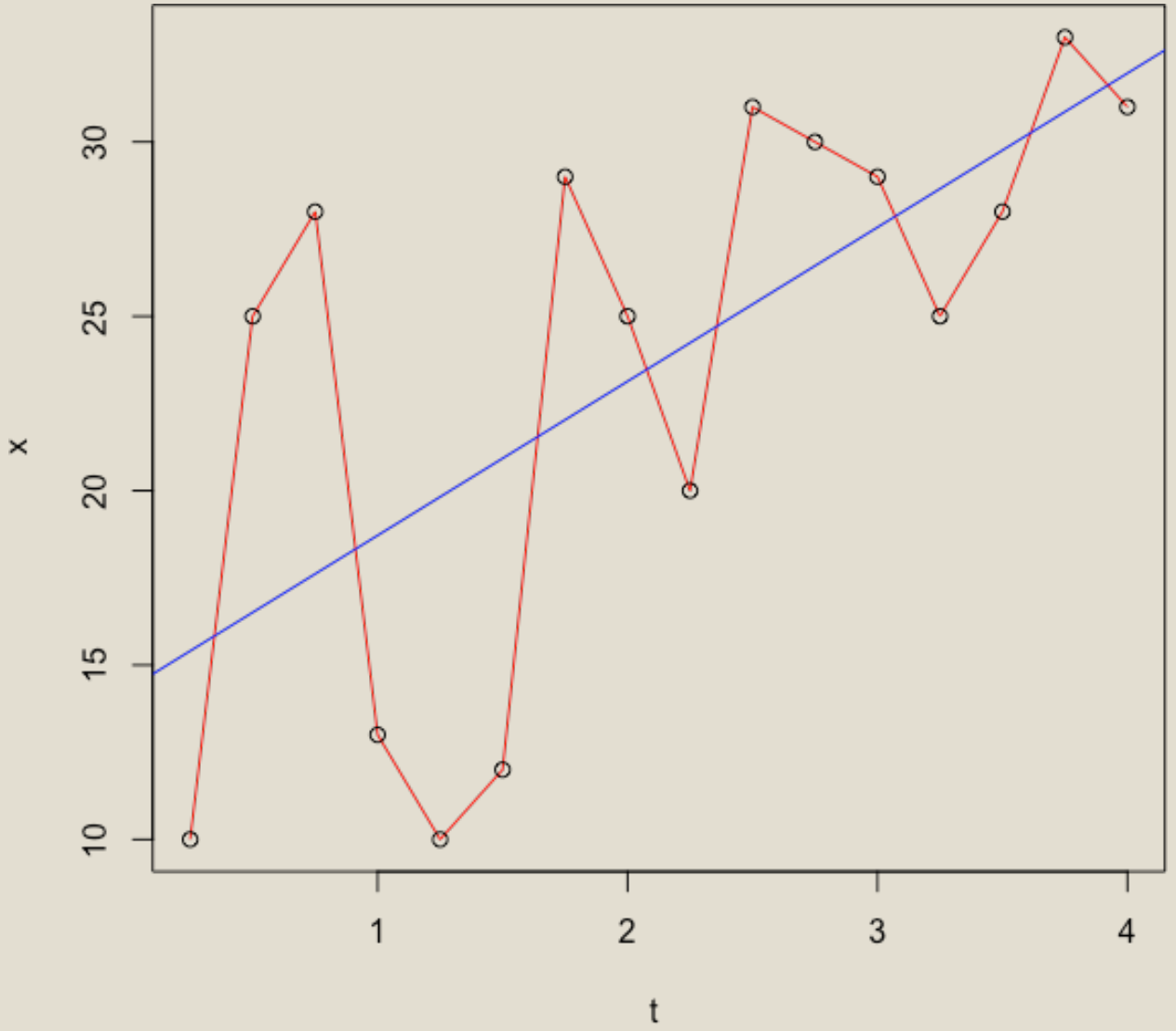
There are occasional drops in sales from one year to the next, but a clear upward trend.

If only the last two points are presented to stockholders, they might think that there are issues and that changes have to be made.



There is a cyclic effect with increases from Q1 to Q2, and from Q2 to Q3, but decreases from Q3 to Q4, and from Q4 to Q1.

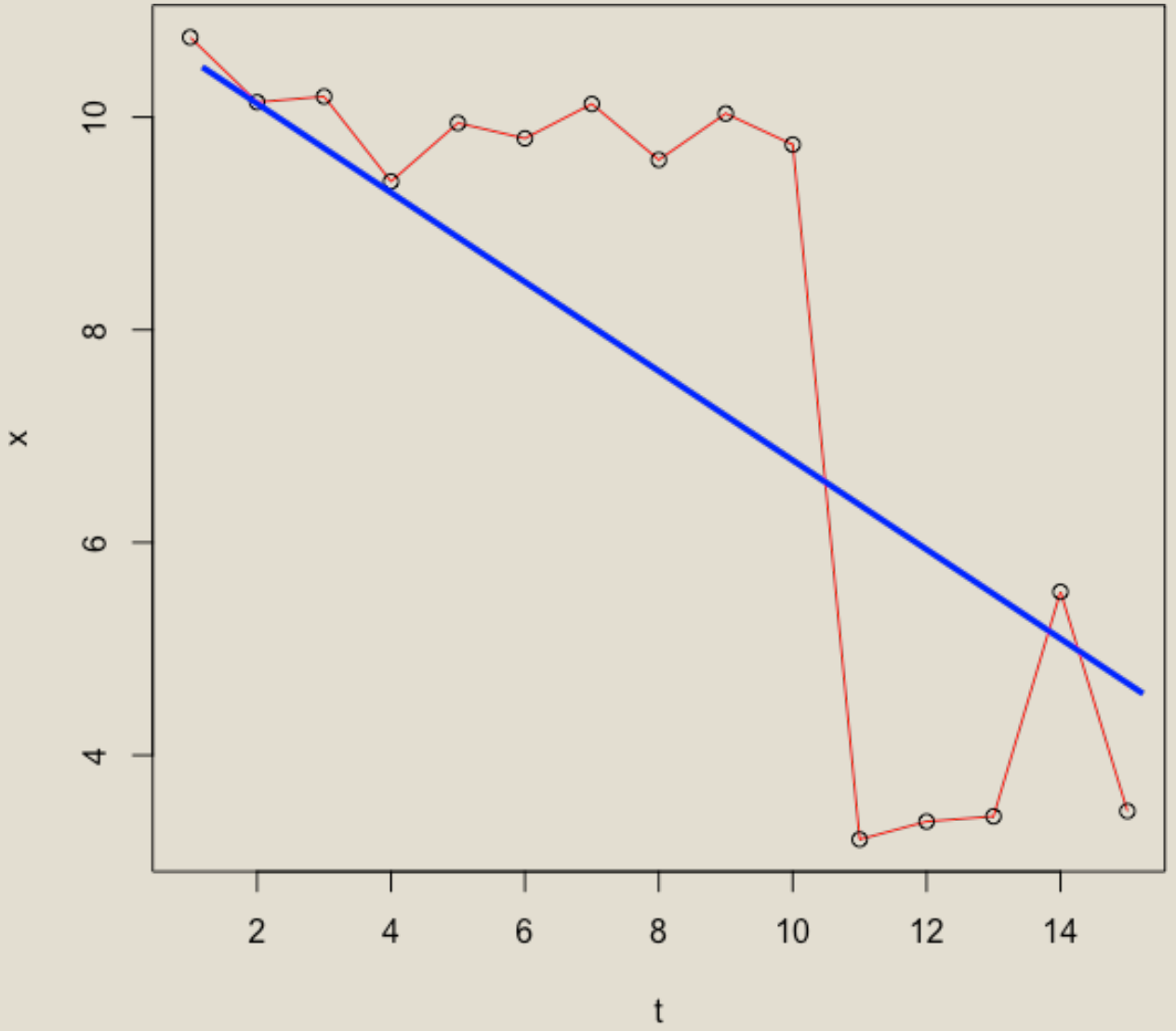
Overall, there seems to be an upward trend, as indicated by the line of best fit.



Clearly, something happened after the tenth week.

Whether the special causes are internal or external depend on the context (which we do not have at our disposal).

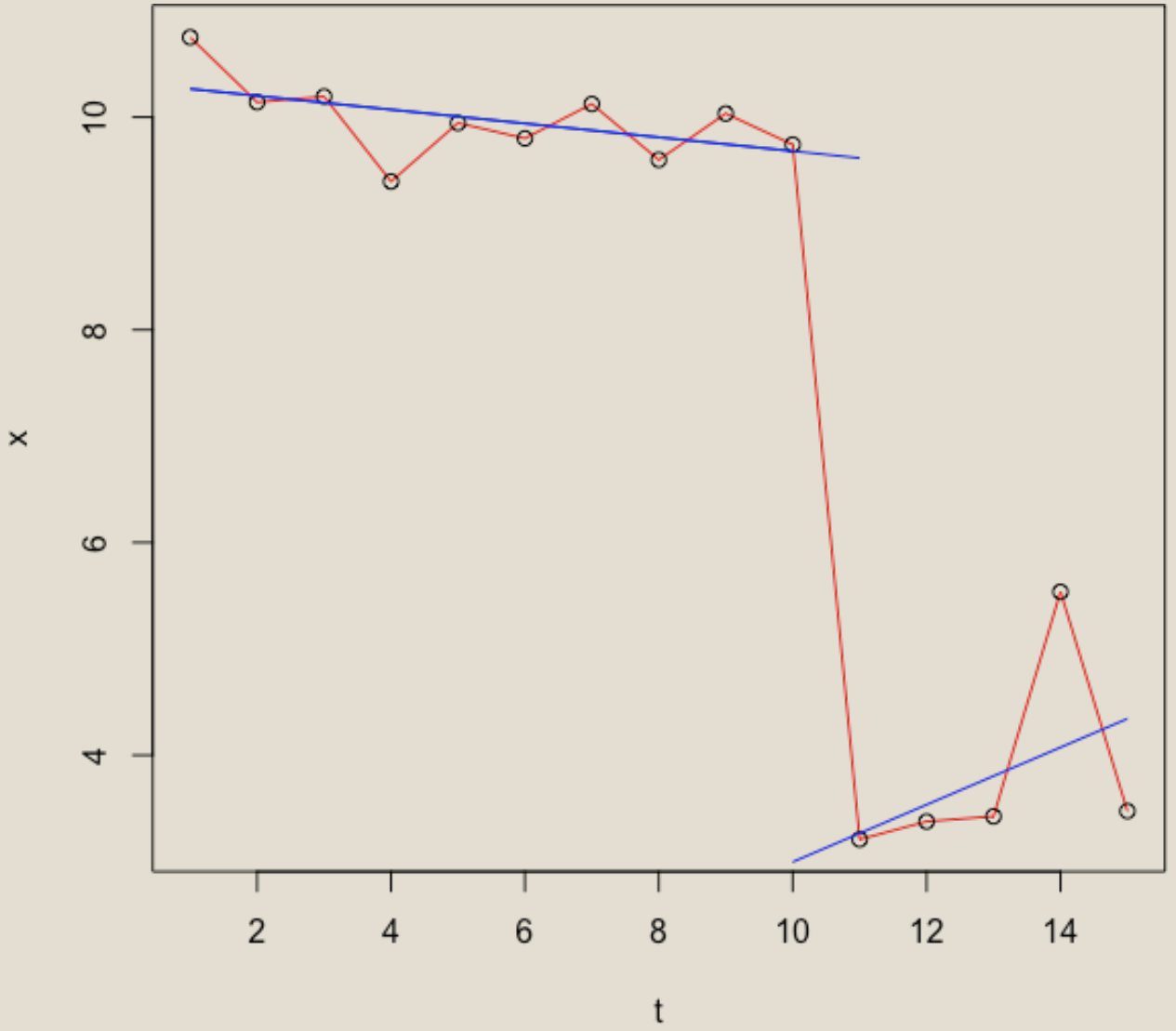
Action seems to be needed.



Clearly, something happened after the tenth week.

Whether the special causes are internal or external depend on the context (which we do not have at our disposal).

Action seems to be needed.



CONTROL CHARTS

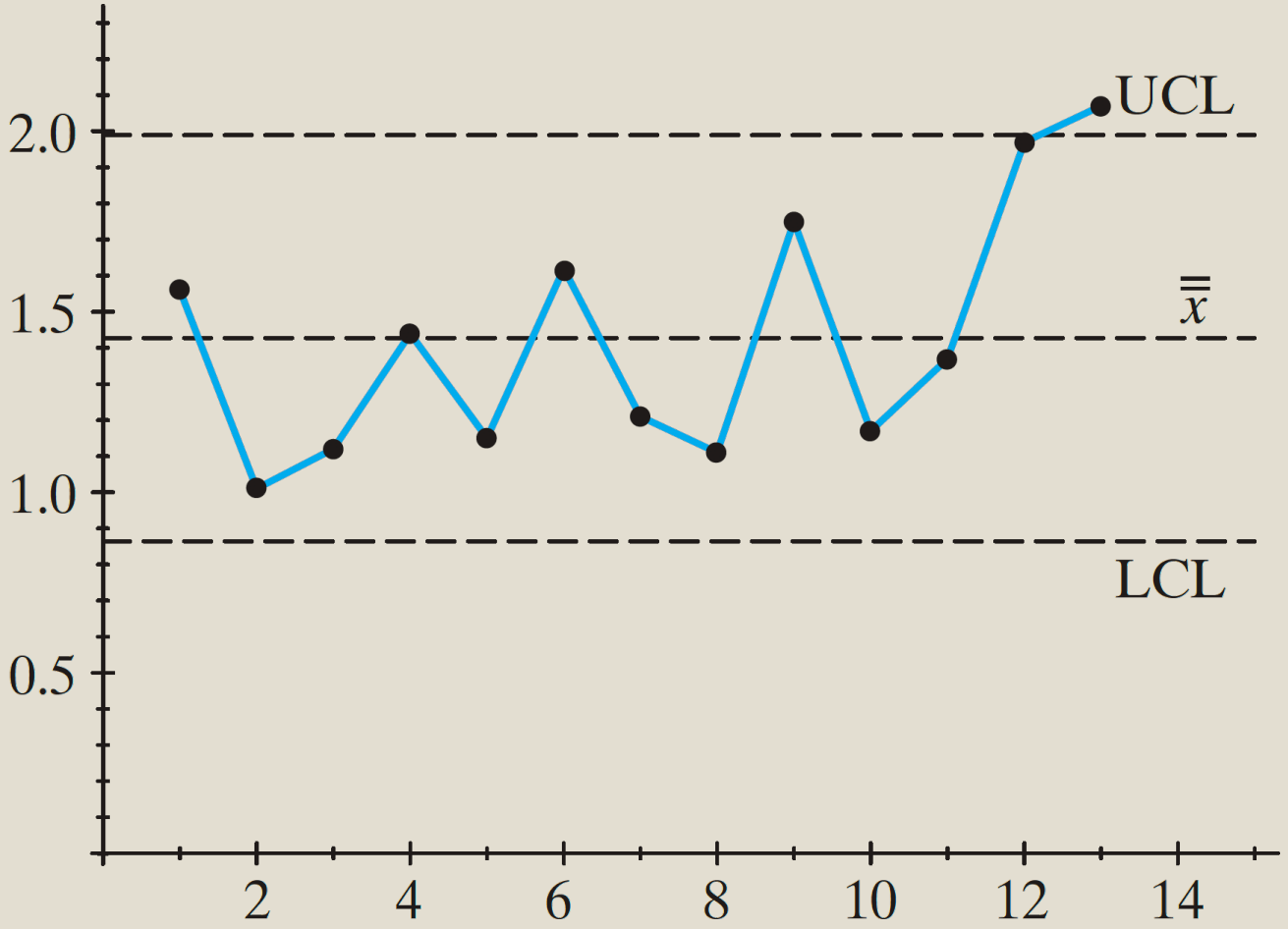
A **control chart** consists of observed values of a statistic, such as $\bar{x}$ or s, plotted as a time series.

The **upper control limit** (UCL) is the upper end of a confidence interval, the **lower control limit** (LCL) is the lower end of a confidence interval, anchored around a **central line** (CL), often the **grand mean** $\bar{\bar{x}}$.

How to compute UCL, LCL depends on the situation at hand (somewhat complicated).

For such charts, if we observe $\bar{x}_i > \mathrm{UCL}$ or $\bar{x}_i < \mathrm{LCL}$, this is an indication that the process is **unstable** and potentially **out of (statistical) control**.

[Tannis and Hogg]



In this control chart, $\bar{x}_{13}$ is outside the control limits in the 13th sampling period.

We suspect that the process has changed and some investigation/action is needed to correct this change (upward shift?).

CONTROL CHARTS (REPRISE)

A **control chart** consists of:

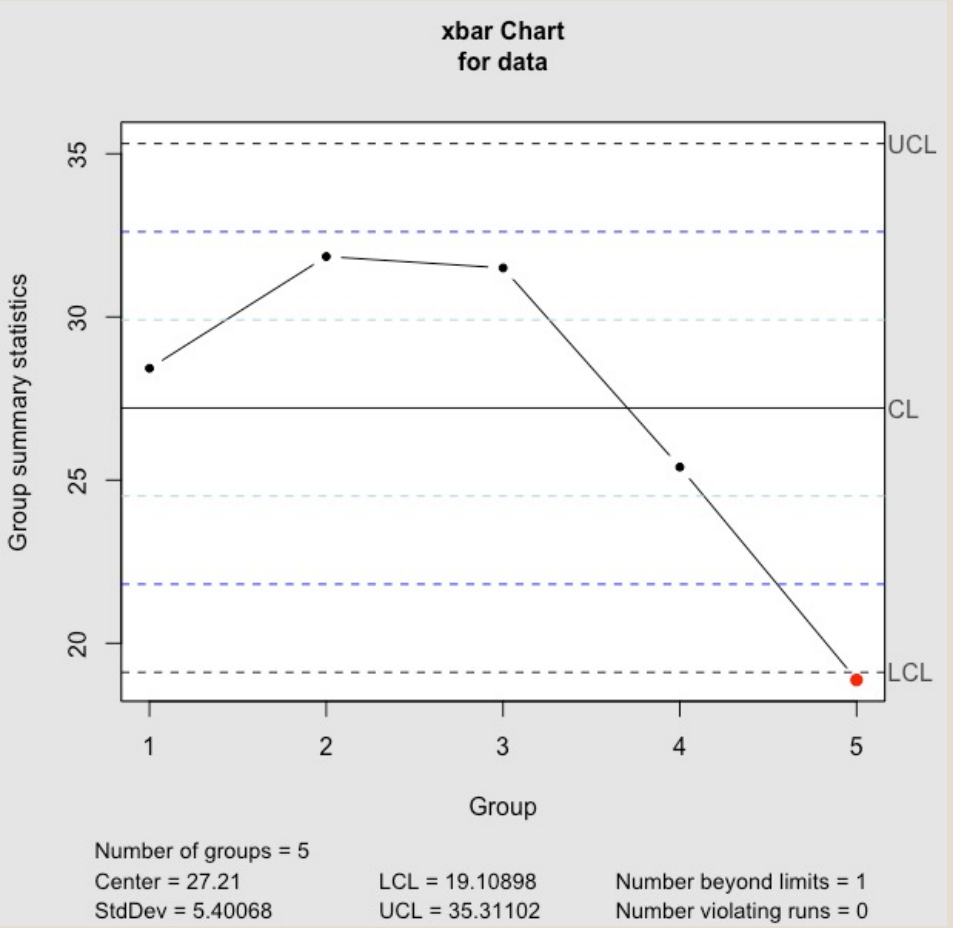
- points representing a sample statistic taken from the process at different times
- the grand mean and the mean standard deviation of the sample statistic, which is computed using all observations, and is used to determine:
  - the **center line**, which is drawn at the value of the grand mean, and
  - the **upper and lower control limits** which indicate the threshold at which the process output is considered statistically unlikely (typically 3 standard deviations away from the central line).
- optional features include:
  - **upper and lower warning limits**, drawn as separate lines, typically two standard deviations from the central line;
  - division into **zones**, with the addition of rules governing freq. of observations in each zone;
  - annotation for **events of interest**, from the point of view of process quality, and
  - action on special causes.

ILLUSTRATION

We are interested in the efficiency of $n = 4$ counters in $N = 5$ stations. The measurements are provided below.

Is there any reason to suspect that there are problems in any of the stations?

i	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$
1	27.1	29.4	27.2	30.0
2	30.6	32.5	32.4	31.9
3	25.7	35.5	30.0	34.8
4	31.1	23.2	25.0	22.3
5	24.1	34.2	15.2	2.0



CONTROL CHART RULESETS

Several rule sets exist for flagging and detecting process instability and process control loss; they should be clearly stated **prior** to the start of SPM, and strictly adhered to.

Example:

- a sample mean found outside the warning limits ⟹ flag as a **warning**
- a sample mean found outside the control limits ⟹ process **out of control**
- a run of 7 successive sample means all above or all below the central line:
  1. stop the production, quarantine and check
  2. adjust process and check new successive samples
  3. if all good, continue process; if not ⟹ **return to step 1**
- a run of 7 successive sample means all showing improvement or all showing a decrease ⟹ **follow previous sets of instructions**