Statistics: descriptive statistics, describing the data

S. Maset

Dipartimento di Matematica e Geoscienze, Università di Trieste

PEM 2018-2019

Outline

Introduction

- Suppose that there are data at our disposal.

  It is very important that such data are presented clearly and concisely and in a manner such that their important features can be immediately seen.

  This is particularly useful when we have a lot of data.

  Data are presented by means of frequency tables or graphical representations of frequency tables.

- We always assume that the data are an $n-$tuple

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$$

with components $x_1, x_2, \ldots, x_n$. Although we could say that $x_1, x_2, \ldots, x_n$ are the data, from now on, we say that $\boldsymbol{x}$ is the data, n is the size of the data and $x_1, x_2, \ldots, x_n$ are the components of the data.

Example. Consider, a classroom of 25 students.

The data $\mathbf{x}$ *could be a* $n-$*tuple with* $n = 25$, *where the component* x_i *is the height, the weight or the favorite type of ice-cream for the* $i-$*th student,* $i \in \{1, \ldots, 25\}$.

It is assumed that the students are ordered in some manner, for example in alphabetic order.

Frequency tables

- Assume that the components $x_1, x_2, \ldots, x_n$ of $\boldsymbol{x}$ belong to a set $\{v_1, \ldots, v_l\}$, where $v_1, \ldots, v_l$ are distinct elements called the **data values.**

  Example. For
  $$\boldsymbol{x} = (-1, 0, 1, -1, -1, 0, -1)$$
  we have the data values $-1, 0, 1$*.*

  Not necessarily the data values have to be numbers as in the example of the favorite type of ice-cream for the students.

  The **frequency table** for the data $\boldsymbol{x}$ associates to each data value v_j, $j \in \{1, \ldots, l\}$, its **frequency** f_j in $\boldsymbol{x}$, i.e. the number of times that v_j appears as a component of $\boldsymbol{x}$: we have
  $$f_j = \left|\{i \in \{1, \ldots, n\} : x_i = v_j\}\right| .$$

Example. *For*

$$\boldsymbol{x} = (-1, 0, 1, -1, -1, 0, -1)$$

we have the frequency table

Data value v_j	*Frequency* f_j
-1	4
0	2
1	1

A frequency table is useful when $I \ll n$ and I is not a large number.

In fact, in this case, in few lines (I lines) we can describe the long data $\boldsymbol{x}$ (of size n): we describe the data $\boldsymbol{x}$ by giving $2I$ numbers (the data values and their frequencies) rather than n numbers (the components).

Note that

$$\sum_{j=1}^{I} f_j = n.$$

In fact, the components of the data $\boldsymbol{x}$ can be partitioned by their values: we count f_1 components with value v_1, f_2 components with value v_2 and so on, for a total of n components.

Example. Let

$$\boldsymbol{x} = (2, 2, 0, 0, 5, 8, 3, 4, 1, 0, 0, 7, 1, 7, 1, 5, 4, 0, 4, 0, 1, 8, 9, 7, 0, 1, 7, 2, 5, 5, 4, 3, 3, 0, 0, 2, 5, 1, 3, 0, 1, 0, 2, 4, 5, 0, 5, 7, 5, 1)$$

where x_i, $i \in \{1, \ldots, 50\}$, *is the number of sick days over the last six weeks of the* $i-$*th worker in a certain company.*

A set of data values is $\{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$ *and the frequency table is*

Data value v_j	*Frequency* f_j
0	12
1	8
2	5
3	4
4	5
5	8
6	0
7	5
8	2
9	1

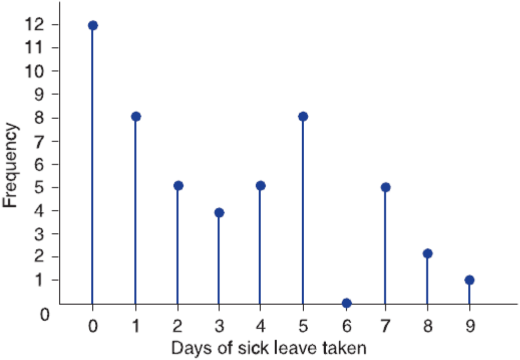
Observe that the data value 6 *does not appear in the components of* $\boldsymbol{x}$. *Its frequency is* 0.

- A frequency table can be graphically pictured by:
  - a **line graph**;
  - a **bar graph**;
  - a **frequency polygon**.

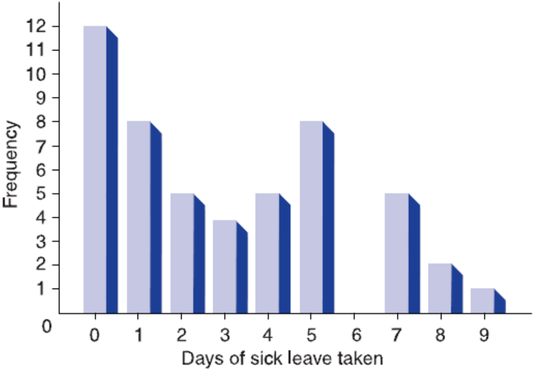
  These graphs have the data values in abscissas and the frequencies in ordinates.

Example of the sick days.

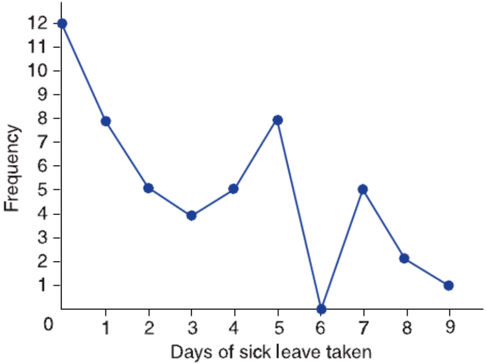
Line graph:



Bar graph:


Frequency
12
11
10
9
8
7
6
5
4
3
2
1
0
0 1 2 3 4 5 6 7 8 9
Days of sick leave taken

Frequency polygon:



- Suppose that the data values $v_1, \ldots, v_l$ are numbers.

  The data $\boldsymbol{x}$ is called **symmetric** about a number c if, for any data value v_{j_1}, $j_1 \in \{1, \ldots, l\}$, there exists a data value v_{j_2}, $j_2 \in \{1, \ldots, l\}$, such that

  $$v_{j_2} - c = -\left(v_{j_1} - c\right) \text{ and } f_{j_2} = f_{j_1},$$

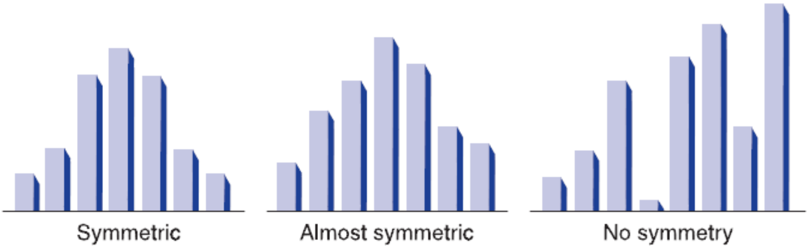
  i.e. v_{j_2} and v_{j_1} are at the same distance from c on the opposite side and have the same frequency.

  Example. A data with the frequency table

Data value v_j	*Frequency* f_j
0	1
1	0
2	2
3	3
4	2
5	0
6	1

  is symmetric about 3*.*

Examples of bar graphs related to the notion of a symmetric data:



- The **relative frequency** of the data value v_j, $j \in \{1, \ldots, l\}$, in the data $\boldsymbol{x}$ of size n is

$$\widehat{f}_j := \frac{f_j}{n}, \; j \in \{1, \ldots, l\}.$$

Note that

$$\widehat{f}_j \in [0, 1], \; j \in \{1, \ldots, l\},$$

(this follows by $0 \leq f_j \leq n$) and

$$\sum_{j=1}^{l} \widehat{f}_j = \sum_{j=1}^{l} \frac{f_j}{n} = \frac{1}{n} \sum_{j=1}^{l} f_j = \frac{1}{n} \cdot n = 1.$$

The **relative frequency table** for the data $\boldsymbol{x}$ associates to each data value v_j, $j \in \{1, \ldots, l\}$, its relative frequency $\widehat{f}_j$.

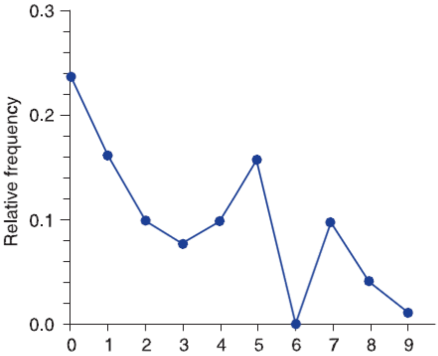
A relative frequency table can be graphically pictured by a line graph, a bar graph or a relative frequency poligon.

Example of the sick days.

Relative frequency table

Data value v_j	*Relative frequency* $\widehat{f}_j$
0	$\frac{12}{50} = 24\%$
1	$\frac{8}{50} = 16\%$
2	$\frac{5}{50} = 10\%$
3	$\frac{4}{50} = 8\%$
4	$\frac{5}{50} = 10\%$
5	$\frac{8}{50} = 16\%$
6	0
7	$\frac{5}{50} = 10\%$
8	$\frac{2}{50} = 4\%$
9	$\frac{1}{50} = 2\%$

Relative frequency polygon



- In MATLAB, a frequency table is obtained by the function $\mathrm{tabulate}$:

$$\mathrm{table} = \mathrm{tabulate}(\mathrm{x})$$

gives, for the data in the vector x, the matrix table whose first column contains the data values, the second column the frequencies and the third column the relative frequencies (in percentage). For a better display use the format $\mathrm{short\ g}$.

Exercise. Use the function $\mathrm{tabulate}$ on the example of the sick days. The frequency table does not contain the data value 6 with frequency 0. By using the command help for the function $\mathrm{tabulate}$, find how to include such value.

Let

$$\mathrm{v} = \mathrm{table}(:, 1),\ \ \mathrm{f} = \mathrm{table}(:, 2),\ \ \mathrm{fhat} = \mathrm{table}(:, 3).$$

The bar graph and the relative bar graph are obtained by

$$\mathrm{bar}(\mathrm{v}, \mathrm{f}) \text{ and } \mathrm{bar}(\mathrm{v}, \mathrm{fhat}).$$

Exercise. By using the command help for the function bar, find how to modify the width of the bars.

The line graph and the relative line graph are obtained by

$$\mathrm{bar}(\mathrm{v},\mathrm{f},0),\ \ \mathrm{hold\ on},\ \ \mathrm{plot}(\mathrm{v},\mathrm{f},'\,\mathrm{o}\,')$$

and

$$\mathrm{bar}(\mathrm{v},\mathrm{fhat},0),\ \ \mathrm{hold\ on},\ \ \mathrm{plot}(\mathrm{v},\mathrm{fhat},'\,\mathrm{o}\,')$$

The frequency polygon and the relative frequency polygon are obtained by

$$\mathrm{plot}(\mathrm{v},\mathrm{f},'-\mathrm{o}')$$

and

$$\mathrm{plot}(\mathrm{v},\mathrm{fhat},'-\mathrm{o}')$$

Exercise. Construct by MATLAB the line graph, the bar graph and the frequency polygon, as well as their relative counterparts, for the sick days example.

- Exercise. When there were two points for the win in football matches, it was important, beside the number of points obtained in a championship, also the so called "english average". The english average is constructed in the following way: in each match a team obtains
  - 1 point for an away win;
  - 0 point for a home win or an away tie;
  - -1 point for a home tie or an away defeat;
  - -2 point for a home defeat.

  Maintaining a zero english average during the tournament was considered a good manner for win the championship.
  - Explain why nowadays, with three points for the win, the english average is less important.
  - Prove that, in a going and return tournament with n teams, the final english average is given by $s - 3(n - 1)$, where s is the final number of points obtained in the tournament (with two points for the win).
  - Collect in Wikipedia the english average of the Serie A winners for the seasons from 1946-47 to 1993-1994 (seasons with two points for the win);
  - Construct the relative frequency table for the data and the relative frequency bar graph.

Pie charts

- A **pie chart** is another graphical picture of a relative frequency table: we partition a circle in sectors by associating at each data value v_j, $j \in \{1, \ldots, l\}$, a sector of area

$$\widehat{f_j} \cdot \text{Area of the circle}$$

or, equivalently, of angle

$$\widehat{f_j} \cdot 360^\circ.$$

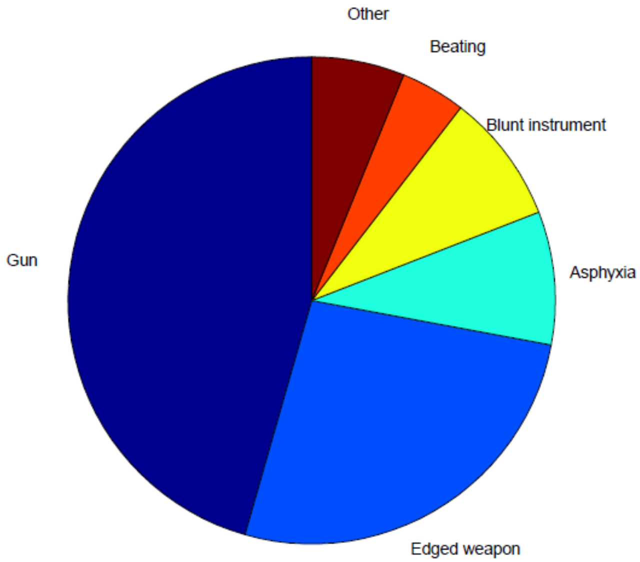
Pie charts are often used to plot relative frequencies when the data values are nonnumeric.

Example. Consider the frequency table of types of murders in Italy during 2010 ($n = 487$)

Type of murder	*Frequency*	*Relative frequency*
Gun	*222*	$\frac{222}{487} = 45.59\%$
Edged weapon	*129*	$\frac{129}{487} = 26.49\%$
Asphyxia	*43*	$\frac{43}{487} = 8.83\%$
Blunt instrument	*42*	$\frac{42}{487} = 8.62\%$
Beating	*21*	$\frac{21}{487} = 4.31\%$
Other	*30*	$\frac{11}{487} = 6.16\%$

Exercise. Describe the data $\boldsymbol{x}$ *relevant to this frequency table.*

Pie chart:



- In MATLAB, a pie chart is obtained by the function pie:

$$\mathrm{pie(f)}$$

produces, for a frequencies vector f, the corresponding pie chart.

Exercise. Construct by MATLAB the previous pie chart for the types of murders. For the labels, consult the command help for the function pie. Moreover, find how to insert the percentages of each data values in the labels.

Class intervals and histograms

- Using a frequency table is effective when the set of the data values is not large.

  But the set of data values can be large, or even infinite. Indeed, there are situation where we can consider as set of the data values an interval of real numbers.

  An example of this last situation is given by a data containing heights of individuals. We can consider as set of data values the interval of real numbers from 100 cm to 250 cm.

  When the set of data values is large or infinite, it is better to partition the set of data values into subsets, called **classes** or **bins**, and then consider the number of components of the data $\boldsymbol{x}$ falling in each class.

- Now, we assume that the set of data values is an interval I of real numbers and use subintervals of I as classes. These subintervals are called **class intervals**.

  Since the class intervals have to be a partition of I, they have to be adjacent without no gaps between them.

  It is common to choose class intervals of the same length.

  The endpoints of the class intervals are called the **class boundaries**.

  We adopt the **left-end inclusion convention**: the class intervals are intervals of type $[a, b)$. We write $a - b$ for $[a, b)$.

How many class intervals have we to choose?

The number of chosen class intervals should be a trade-off between:

- choosing too few intervals at a cost of losing too much information about the data: the extreme case is when all the components are in a unique interval;
- choosing too many intervals, which will result in small frequencies of the components of the data in each class and the impossibility to see a pattern in the data: the extreme case is when each interval contains zero or one component.

From 5 to 10 class intervals are typical.

- The **frequency table of the class intervals** for the data $\boldsymbol{x}$ associates to each class interval its frequency, i.e. the number of components of $\boldsymbol{x}$ falling in that class interval.

  Example. Level of blood cholesterol for $n = 40$ *first year universitary students.*

Table 2.5 Blood Cholesterol Levels

213	174	193	196	220	183	194	200
192	200	200	199	178	183	188	193
187	181	193	205	196	211	202	213
216	206	195	191	171	194	184	191
221	212	221	204	204	191	183	227

Table 2.6 Blood Cholesterol Levels in Increasing Order

171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195,
196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, 227

We use the six adjacent class intervals of length 10

$170-180,\ 180-190,\ 190-200,\ 200-210,\ 210-220,\ 220-230.$

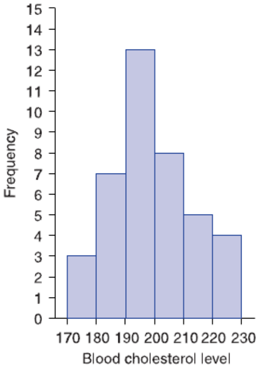
The frequency table of the class intervals

Table 2.7 Frequency Table of Blood Cholesterol Levels

Class intervals	Frequency	Relative frequency
170–180	3	$\frac{3}{40} = 0.075$
180–190	7	$\frac{7}{40} = 0.175$
190–200	13	$\frac{13}{40} = 0.325$
200–210	8	$\frac{8}{40} = 0.20$
210–220	5	$\frac{5}{40} = 0.125$
220–230	4	$\frac{4}{40} = 0.10$

- A bar graph for the (absolute or relative) frequencies of the class intervals is called an **histogram**.

  The histogram for blood cholesterol levels:



Observe that in an histogram the bars are touching one another, this is because the class intervals are adjacent.

Example. Birth rate (for $1\mathrm{K}$ population in 1 year) in each of 50 states in US

Table 2.8 Birth Rates per 1000 Population

State	Rate	State	Rate	State	Rate
Alabama	14.2	Louisiana	15.7	Ohio	14.9
Alaska	21.9	Maine	13.8	Oklahoma	14.4
Arizona	19.0	Maryland	14.4	Oregon	15.5
Arkansas	14.5	Massachusetts	16.3	Pennsylvania	14.1
California	19.2	Michigan	15.4	Rhode Island	15.3
Colorado	15.9	Minnesota	15.3	South Carolina	15.7
Connecticut	14.7	Mississippi	16.1	South Dakota	15.4
Delaware	17.1	Missouri	15.5	Tennessee	15.5
Florida	15.2	Montana	14.1	Texas	17.7
Georgia	17.1	Nebraska	15.1	Utah	21.2
Hawaii	17.6	Nevada	16.5	Vermont	14.0
Idaho	15.2	New Hampshire	16.2	Virginia	15.3
Illinois	16.0	New Jersey	15.1	Washington	15.4
Indiana	14.8	New Mexico	17.9	West Virginia	12.4
Iowa	13.1	New York	16.2	Wisconsin	14.8
Kansas	14.2	North Carolina	15.6	Wyoming	13.7
Kentucky	14.1	North Dakota	16.5		

Source: Department of Health and Human Services.

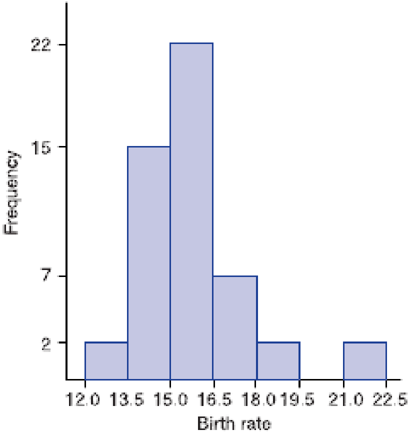
Data values range from 12.4 *in West Virginia to* 21.9 *in Alaska.*

We use class intervals of length 1.5*, starting with* $12 - 13.5$ *and finishing with* $21 - 22.5$*, with a total of* 7 *classes.*

The frequency table:

Class intervals	Frequency	Class intervals	Frequency
12.0–13.5	2	18.0–19.5	2
13.5–15.0	15	19.5–21.0	0
15.0–16.5	22	21.0–22.5	2
16.5–18.0	7		

The histogram:



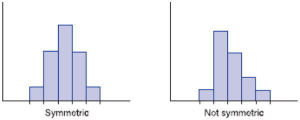
- The frequency table of the class intervals, or the histogram based on this table, does not contain all the information about the data that the frequency table has.

  In fact, the single values of the components in each class are lost: only their total number is reported in the histogram.

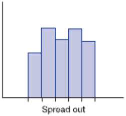
  However, an histogram can uncover important features of the data.

For example:

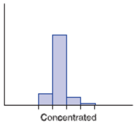
- How much symmetric the data is



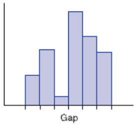
- How much spread out the data is



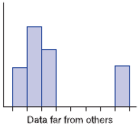
- Whether there are intervals with high levels of data concentration



- Whether there are gaps in the data values



- Whether some data values are far apart from others



- In MATLAB a frequency table of class intervals is obtained by the function $\mathrm{histcounts}$:

$$\mathrm{N} = \mathrm{histcounts}(\mathrm{x}, \mathrm{boundaries})$$

gives, for the data in the vector x and the class boundaries

$$a_0 < a_1 < \cdots < a_{m-1} < a_m$$

in the vector $\mathrm{boundaries}$, the frequency table vector N.

The vector N has length m, where m is the number of class intervals, and it is such that, for any $k \in \{1, \ldots, m-1\}$, $\mathrm{N}(k)$ is the number of components of x falling in the k-th class interval $a_{k-1} - a_k = [a_{k-1}, a_k)$ and $\mathrm{N}(m)$ is the number of components of x falling in the last class interval $a_{m-1} - a_m = [a_{m-1}, a_m)$ or equal to a_m.

Exercise. By using the MATLAB instruction $\mathrm{x} = \mathrm{rand}(100, 1)$, construct a random vector with components in $[0, 1]$. Then, construct the frequency table of the class intervals $0 - 0.1, 0.1 - 0.2, \ldots, 0.9 - 1$ for the data x.

The relative frequency table is then given by

$$\mathrm{Nhat} = \mathrm{N}/\mathrm{sum}(\mathrm{N}).$$

The hystogram is then obtained by

$$\mathrm{bar}(\mathrm{middlepoints}, \mathrm{N},'\mathrm{hist}') \text{ or } \mathrm{bar}(\mathrm{middlepoints}, \mathrm{Nhat},'\mathrm{hist}')$$

where

$$\mathrm{middlepoints} = 1/2 * (\mathrm{boundaries}(1 : \mathrm{end} - 1) + \mathrm{boundaries}(2 : \mathrm{end}))$$

is the vector of the middle points of the class intervals.

Exercise. Construct by MATLAB the histogram of the frequency table of the previous exercise as well as the histograms of the two previous examples.

- An histogram is a bar graph for representing the frequency table of the class intervals.

  Alternatively, we could use a line graph or a frequency polygon. In this case, on the horizontal axis, each class interval is represented by its middle point.

  Representation of a frequency table of class intervals by a frequency polygon is important when we want to compare two data.

Example. Systolic blood pressure (the maximum blood pressure) of two groups of male workers: one group is $30 - 40$ *years old and the other is* $50 - 60$ *years old.*

Absolute and relative frequency tables for class intervals

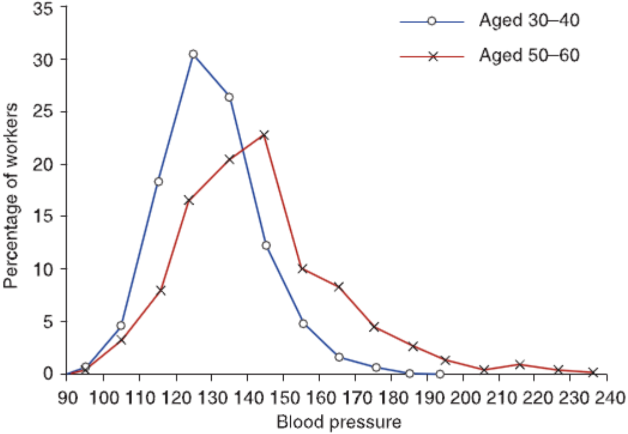
Table 2.9 Class Frequencies of Systolic Blood Pressure of Two Groups of Male Workers

Blood pressure	Number of workers	
	Aged 30–40	Aged 50–60
Less than 90	3	1
90–100	17	2
100–110	118	23
110–120	460	57
120–130	768	122
130–140	675	149
140–150	312	167
150–160	120	73
160–170	45	62
170–180	18	35
180–190	3	20
190–200	1	9
200–210		3
210–220		5
220–230		2
230–240		1
Total	**2540**	**731**

Table 2.10 Relative Class Frequencies of Blood Pressures

Blood pressure	Percentage of workers	
	Aged 30–40	Aged 50–60
Less than 90	0.12	0.14
90–100	0.67	0.27
100–110	4.65	3.15
110–120	18.11	7.80
120–130	30.24	16.69
130–140	26.57	20.38
140–150	12.28	22.84
150–160	4.72	9.99
160–170	1.77	8.48
170–180	0.71	4.79
180–190	0.12	2.74
190–200	0.04	1.23
200–210		0.41
210–220		0.68
220–230		0.27
230–240		0.14
Total	**100.00**	**100.00**

Relative frequency polygons for the two groups



- In MATLAB, the line graph and the relative line graph for the frequency table of class intervals are obtained by

$$\mathrm{bar(middlepoints, N, 0),\ \ hold\ on,\ \ plot(middlepoints, N,' o')}$$

and

$$\mathrm{bar(middlepoints, Nhat, 0),\ \ hold\ on,\ \ plot(middlepoints, Nhat,' o').}$$

The frequency polygon and the relative frequency polygon are obtained by

$$\mathrm{plot(middlepoints, N,' -o')}$$

and

$$\mathrm{plot(middlepoints, Nhat,' -o').}$$

Exercise. Construct by MATLAB the two frequency polygons for the previous example of the workers blood pressure.

Stem-and-leaf plots

- We can describe a data $\boldsymbol{x}$, of size n non-large, by a **stem-and-leaf plot**, where we divide the value of each component of $\boldsymbol{x}$ into two parts: its **stem,** more significant, and its **leaf**, less significant.

Example. Consider

$$\boldsymbol{x} = (11, 14, 18, 23, 27, 32, 36, 37, 44, 52, 58, 60, 65)\,.$$

By using the tens digit as stem and the ones digit as leaf, we obtain the stem-and-leaf plot

$$\begin{array}{r|l} 1 & 1,4,8 \\ 2 & 3,7 \\ 3 & 2,6,7 \\ 4 & 4 \\ 5 & 2,8 \\ 6 & 0,5 \end{array}$$

Example. Pro capita personal income in each state in US.

Table 2.11 Per Capita Personal Income (Dollars per Person), 2002

State name		State name		State name	
United States	30,941	Kentucky	25,579	Ohio	29,405
Alabama	25,128	Louisiana	25,446	Oklahoma	25,575
Alaska	32,151	Maine	27,744	Oregon	28,731
Arizona	26,183	Maryland	36,298	Pennsylvania	31,727
Arkansas	23,512	Massachusetts	39,244	Rhode Island	31,319
California	32,996	Michigan	30,296	South Carolina	25,400
Colorado	33,276	Minnesota	34,071	South Dakota	26,894
Connecticut	42,706	Mississippi	22,372	Tennessee	27,671
Delaware	32,779	Missouri	28,936	Texas	28,551
District of Columbia	42,120	Montana	25,020	Utah	24,306
Florida	29,596	Nebraska	29,771	Vermont	29,567
Georgia	28,821	Nevada	30,180	Virginia	32,922
Hawaii	30,001	New Hampshire	34,334	Washington	32,677
Idaho	25,057	New Jersey	39,453	West Virginia	23,688
Illinois	33,404	New Mexico	23,941	Wisconsin	29,923
Indiana	28,240	New York	36,043	Wyoming	30,578
Iowa	28,280	North Carolina	27,711		
Kansas	29,141	North Dakota	26,982		

The stem-and-leaf plot: the thousands digits as stem and the other as leaf

22	372
23	512, 688, 941
24	706
25	020, 057, 128, 400, 446, 575, 579
26	183, 894, 982
27	671, 711, 744
28	240, 280, 551, 731, 821, 936
29	141, 405, 567, 596, 771, 923
30	001, 180, 296, 578
31	319, 727
32	151, 677, 779, 922, 996
33	276, 404
34	071, 334
36	043, 298
39	244, 453
42	120, 706

- The choice of the stem and the leaf has to be done in order to have not too many stems and not too many leaves.

Example. Percentage of foreigner new births with respect to all births in the various municipalities in the province of Pordenone during the year 2010.

Andreis	0,0
Arba	25,0
Arzene	16,7
Aviano	23,4
Azzano Decimo	25,1
Barcis	50,0
Brugnera	22,8
Budoia	4,3
Caneva	10,7
Casarsa della Delizia	20,7
Castelnovo del Friuli	16,7
Cavasso Nuovo	21,1
Chions	24,6
Cimolais	0,0
Claut	0,0
Clauzetto	0,0
Cordenons	14,5
Cordovado	4,0
Erto e Casso	0,0
Fanna	45,5
Fiume Veneto	14,4
Fontanafredda	19,8
Frisanco	0,0
Maniago	23,9
Meduno	0,0
Montereale Valcellina	23,3
Morsano al Tagliamento	14,8
Pasiano di Pordenone	39,8
Pinzano al Tagliamento	14,3
Polcenigo	20,0

Continuation:

Porcia	11,2
Pordenone	38,1
Prata di Pordenone	46,8
Pravisdomini	58,1
Roveredo in Piano	14,0
Sacile	20,6
San Giorgio della Richinvelda	6,5
San Martino al Tagliamento	14,3
San Quirino	18,4
San Vito al Tagliamento	19,3
Sequals	4,5
Sesto al Reghena	4,4
Spilimbergo	25,0
Tramonti di Sopra	0,0
Tramonti di Sotto	0,0
Travesio	15,4
Valvasone	0,0
Vito d'Asio	0,0
Vivaro	0,0
Zoppola	18,5
Vajont	40,7
Provincia di Pordenone	**23,1**
TOTALE FVG	**17,0**

The stem-and-leaf plot: the stem is the tens digit, the leaf is the remainder

0	*0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,4.0,4.3,4.4,4.5,6.5*	*(17)*
1	*0.7,1.2,4.0,4.3,4.3,4.4,4.5,4.8,5.4,6.7,6.7,8.4,8.5,9.3,9.8*	*(15)*
2	*0.0,0.6,0.7,1.1,2.8,3.3,3.4,3.9,4.6,5.0,5.0,5.1*	*(12)*
3	*8.1,9.8*	*(2)*
4	*0.7,5.5,6.8*	*(3)*
5	*0.0,8.1*	*(2)*

The numbers in parentheses on the right represent the number of components of the data (i. e. the number of leaves) in each stem class.

If the stem was the integer part, not the tens digit, then there would be too many stems.

- In MATLAB, stem-and-leaf plots for a data in the vector *x* are obtained by

$$\mathrm{stemleafplot}(\mathrm{x}, \mathrm{p}).$$

  The stems are the 10$^{\mathrm{p}+1}$ digits of the components of x .

  Exercise. Construct by MATLAB the stem-and-leaf plots of the two previous examples.

- Exercise. Collect in http://www.alltime-athletics.com/ the best performance of sprinters who have broken the 9.90 seconds barrier in the 100 m men run. Construct a stem-and-leaf plot.

- A stem-and-leaf plot looks like a histogram turned on its side, with the advantage that it presents the components of the data falling in the class intervals. So, it contains all the information that the frequency table for the data can give.

  It is most helpful for data of moderate size: if the size of the data is very large, then the the leaves might be too many and the stem-and-leaf plot might be not more informative than a histogram.

  Exercise. Assume that in a stem-and-leaf plot, we are using as stems the 10^{p+1} digits. Describe the class intervals of the histogram relevant to the stem-and-leaf plot.

Paired data

- Now, we consider **paired data** given by two $n-$tuple $\boldsymbol{x}$ and $\boldsymbol{y}$: in other terms, the paired data is given by n pairs

$$(x_i, y_i)\,,\ i \in \{1, \dots, n\}.$$

Example. Consider $n = 30$ workers of a given company and their IQ score and salary: for $i \in \{1, \dots, 30\}$

- *x_i is the IQ score of the worker i;*
- *y_i is the salary of the worker i.*

Table 2.12 Salaries versus IQ

Worker i	IQ score x_i	Annual salary y_i (in units of $1000)	Worker i	IQ score x_i	Annual salary y_i (in units of $1000)
1	110	68	16	84	19
2	107	30	17	83	16
3	83	13	18	112	52
4	87	24	19	80	11
5	117	40	20	91	13
6	104	22	21	113	29
7	110	25	22	124	71
8	118	62	23	79	19
9	116	45	24	116	43
10	94	70	25	113	44
11	93	15	26	94	17
12	101	22	27	95	15
13	93	18	28	104	30
14	76	20	29	115	63
15	91	14	30	90	16

*Stem-and-leaf plot for **x***

12	4	(1)
11	0,0,2,3,3,5,6,6,7,8	(10)
10	1,4,4,7	(4)
9	0,1,1,3,3,4,4,5	(8)
8	0,3,3,4,7	(5)
7	6,9	(2)

*Stem-and-leaf plot for **y***

7	0,1	(2)
6	2,3,8	(3)
5	2	(1)
4	0,3,4,5	(4)
3	0,0	(2)
2	0,2,2,4,5,9	(6)
1	1,3,3,4,5,5,6,6,7,8,9,9	(12)

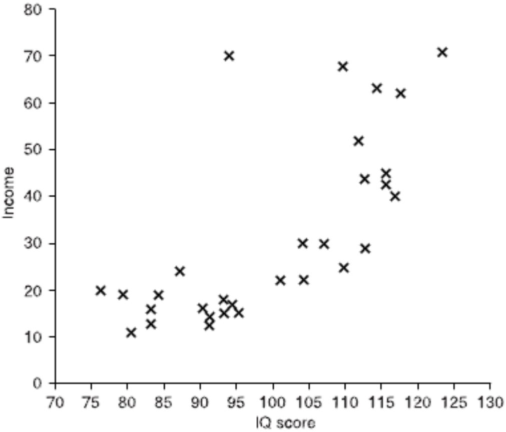
We want to learn whether higher IQ scores tend to go along with higher income, at this company.

Stem-and-leaf plots for $\boldsymbol{x}$ *and* $\boldsymbol{y}$*, which separately consider* $\boldsymbol{x}$ *and* $\boldsymbol{y}$*, are not useful for this aim: it is necessary to consider the pairs* (x_i, y_i)*,* $i \in \{1, \ldots, 30\}$*.*

- A useful way for representing paired data $\boldsymbol{x}$ and $\boldsymbol{y}$ is to plot each pair (x_i, y_i), $i \in \{1, \ldots, n\}$, as a point in the cartesian plane xy.

  Such a plot is called a **scatter diagram**.

  The scatter diagram for the paired data IQ and salary:



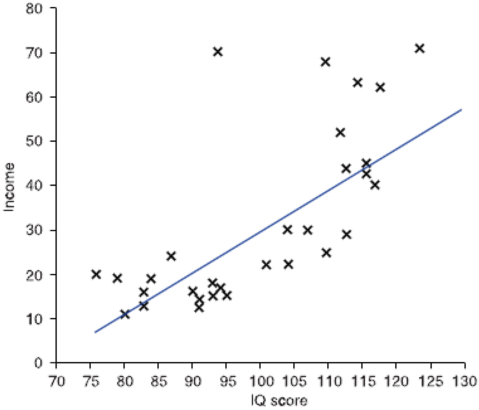
- In MATLAB the scatter diagram for data in the vectors x and y is obtained by

$$\mathrm{plot}(\mathrm{x}, \mathrm{y}, '\mathrm{x}')$$

Exercise. By using MATLAB, construct the scatter diagram for the paired data IQ and salary.

- It is clear from the scatter diagram of the example that higher IQ scores appear to go along with higher incomes.

  The general trend can be captured by the line shown below



called the regression line.

The **regression line** of a paired data $\boldsymbol{x}$ and $\boldsymbol{y}$ is the nonvertical line

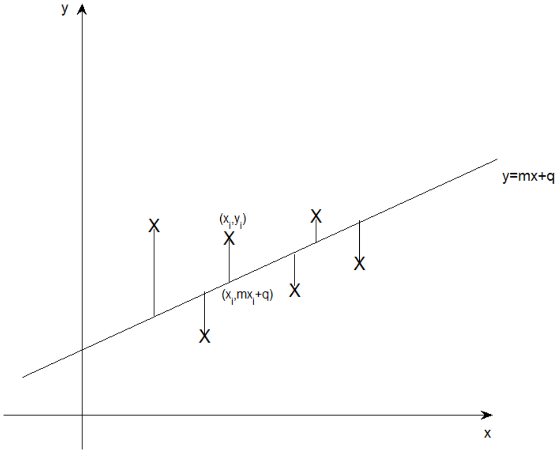
$$y = mx + q$$

that minimizes the **residual sum of the squares**

$$rss = \sum_{i=1}^{n} (y_i - mx_i - q)^2$$

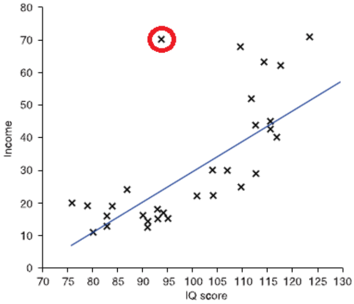
among all the possible nonvertical lines in the plane (among all $(m, q) \in \mathbb{R}^2$).

rss is the sum of the squares of the distances between the point (x_i, y_i) and the point $(x_i, mx_i + q)$ on the line $y = mx + q$, for $i \in \{1, \ldots, n\}$.



We will show later how to determine the regression line.

- A scatter diagram is also useful in detecting **outliers**, which are components (x_i, y_i) of the paired data that do not appear to follow the trend of the other components.



Having noted the outliers, we have to decide whether they are caused by an error in the collection of the data or they are genuine.