2.1.5 More on descriptive statistics

In this subsection we will discuss the ideas of shifting and scaling, z-values and the so-called "empirical rule".

Shifting a variable

Let $x_1, x_2, \cdots, x_N$ be the observations of a variable x in a population U, let a be a constant and let

$$y_i = x_i + a \qquad \text{for all } i = 1, 2, \cdots, N.$$

In other words, y is simply the same as x but adding some constant. Let us consider, for instance, the age (in years) of ten individuals as of December 31, 2018 ($= x$), and the age (in years) of the same individuals as of December 31, 2023 ($= y$) given in Table 9 and illustrated by dotplots in Figure 14. Thus $y_i = x_i + 5$.

x	26	29	31	32	34	37	38	39	40	46
y	31	34	36	37	39	42	43	44	45	51

Table 9: Age of ten individuals as of December 31, 2018 (x) and December 31, 2023 (y)

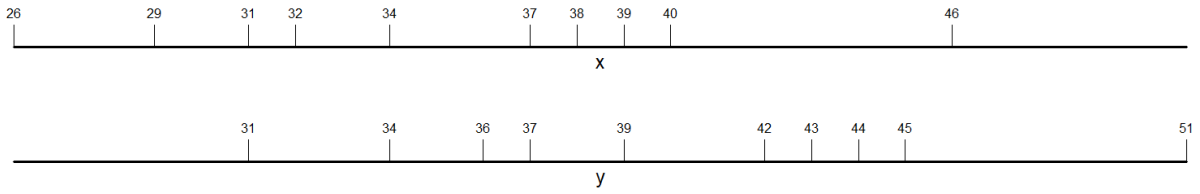

Figure 14: Dotplots of the age of ten individuals in Table 9.

By looking at Figure 14 we can see the effect that adding a constant to a variable has: it *shifts* the observations a units. It also gives an intuition about what will happen with the different parameters that we have studied so far. Intuitively, all location measures are equally shifted, whereas variability and shape measure will remain unaffected. It turns out that this intuition is correct. Table 10 shows the different parameters that we have studied for both x and y in Table 9.

Type	Parameter	x	y
Location	First quartile	31.25	36.25
	Mean	35.2	40.2
	Median	35.5	40.5
	Third quartile	38.75	43.75
Variability	Range	20	20
	IQR	7.5	7.5
	Variance	35.29	35.29
	Standard deviation	5.94	5.94
Shape	Skewness	0.16	0.16

Table 10: Descriptive statistics of ten individuals in Table 9

Scaling a variable

Let us now consider a second situation. Let $x_1, x_2, \cdots, x_N$ be the observations of a variable x in a population U, let b be a constant and let

$$y_i = b\, x_i \qquad \text{for all } i = 1, 2, \cdots, N.$$

In other words, y is simply the same as x but multiplied by a constant. Let us consider, for instance, the price of ten cell phones in a particular store in Swedish Krona SEK ($= x$) and in Czech Koruna CZK ($= y$) given in Table 11 and illustrated by dotplots in Figure 15. Taking into account that (today) one Czech Koruna is equivalent to 2.17 Swedish Kronor, we have $y_i = 2.17\, x_i$.

x	2000	7000	8500	9800	11500	14500	16000	16500	17500	20500
y	4340	15190	18445	21266	24955	31465	34720	35805	37975	44485

Table 11: Price of ten cell phones in SEK (x) and CZK (y)

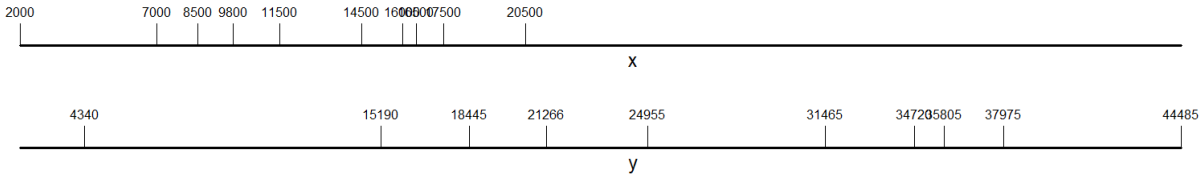

Figure 15: Dotplots of the price of ten cell phones Table 11.

By looking at Figure 15 we can see the effect that multiplying a variable by a constant has: it *scales* the observations by a factor of b. It also gives an intuition about what will happen with the different parameters that we have studied so far. Intuitively, all location measures are equally scaled, variability measures are also scaled by the same factor b, with one exception: the variance. The variance is scaled by a factor of b^2. The skewness remain unaffected.Table 12 shows the different parameters that we have studied for both x and y in Table 11.

Type	Parameter	x	y
Location	First quartile	8825	19150
	Mean	12380	26860
	Median	13000	28210
	Third quartile	16375	35530
Variability	Range	18500	40150
	IQR	7550	16380
	Variance	31 770 000	149 600 000
	Standard deviation	5636	12230
Shape	Skewness	-0.3094	-0.3094

Table 12: Descriptive statistics of the price of ten cell phones in Table 11

Shifting and scaling a variable

Let us now consider a third situation in which we combine the two situations above. Let $x_1, x_2, \cdots, x_N$ be the observations of a variable x in a population U, let a and b be two constants and let

$$y_i = b(x_i + a) \qquad \text{for all } i = 1, 2, \cdots, N.$$

In other words, y the same as x but adding a constant and then multiplied by another constant. Let us consider, for instance, the temperatures in a weather station in Sweden measured at twelve different time points over a year in Fahrenheit ($= x$) and Celsius (y) given in Table 13 and illustrated by dotplots in Figure 16. Remember that $y_i = \frac{5}{9}(x_i - 32)$.

x	-0.4	19.4	26.6	32.0	44.6	64.4	73.4	71.6	66.2	51.8	30.2	23.0
y	-18	-7	-3	0	7	18	23	22	19	11	-1	-5

Table 13: Temperature in a weather station at twelve time points in Fahrenheit ($= x$) and Celsius (y)

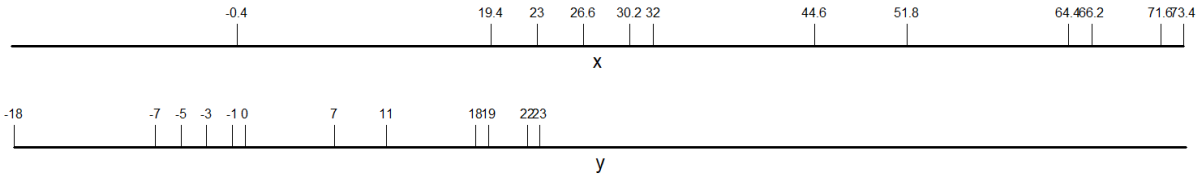

Figure 16: Dotplots of the temperature at twelve time points in Table 13.

By looking at Figure 16 we can see the effect that adding a constant a and multiplying by another one b has on a variable x: first, the observations are *shifted* a units and then they are *scaled* by a factor of b. It also gives an intuition about what will happen with the different parameters that we have studied so far. Intuitively, all location measures are equally shifted and scaled, variability measures will be scaled by the factor b, with one exception: the variance. The variance is scaled by a factor of b^2. The skewness remains unaffected. Table 14 shows the different parameters that we have studied for both x and y in Table 13.

Type	Parameter	x	y
Location	First quartile	25.70	-3.50
	Mean	41.90	5.50
	Median	38.30	3.50
	Third quartile	64.85	18.25
Variability	Range	73.80	41.00
	IQR	39.15	21.75
	Variance	563.5	173.9
	Standard deviation	23.74	13.19
Shape	Skewness	-0.0984	-0.0984

Table 14: Descriptive statistics of the temperature at twelve time points in Table Table 13

Let us summarize the results of this section in the following result:

Result 28. Let $x_1, x_2, \cdots, x_N$ be the observations of a variable x in a population U, let a and b be two constants and let

$$y_i = b(x_i + a) \qquad \text{for all } i = 1, 2, \cdots, N.$$

We have

$$\bar{y}_U = b(\bar{x}_U + a) \qquad \dot{y}_U = b(\dot{x}_U + a) \quad \breve{y}_{p,U} = b(\breve{x}_{p,U} + a) \quad Sk_{y,U} = Sk_{x,U}$$
$$\text{range}_{y,U} = b\,\text{range}_{x,U} \quad \text{IQR}_{y,U} = b\,\text{IQR}_{x,U} \quad S_{y,U} = b\,S_{x,U} \quad S^2_{y,U} = b^2\,S^2_{x,U} \qquad \square$$

Standardization and the z-scores

The special case of Result 28 when $a = -\bar{x}_U$ and $b = 1/S_{x,U}$ is so important that we present it as another result:

Result 29. Let $x_1, x_2, \cdots, x_N$ be the observations of a variable x in a population U and let

$$z_i = \frac{x_i - \bar{x}_U}{S_{x,U}} \qquad \text{for all } i = 1, 2, \cdots, N.$$

then

$$\bar{z}_U = 0 \qquad \text{and} \qquad S_{z,U} = 1.$$

The variable z is called the *standard form* of x and the process of substracting the mean to a variable and then dividing by its standard deviation is called *standardization*. The resulting z-values are called *standardized values* or simply the z-scores. $\square$

z-scores indicate the distance from the different values to the mean in standard deviation units. For instance, a z-score of 1 means that the observation is one standard deviation above than the mean and a z-score of -2 indicates that the observation is two standard deviation below the mean. In this way, z-scores allow to measure how big or small (in terms of distance to the mean) an observation is with respect to other.

Example 30. Let us consider again our population of ten students and their points in an exam (x). The first row of Table 15 reproduces the number of points.

We found before that the mean is $\bar{x}_U = 22.3$ and the standard deviation is $S_{x,U} = 12.53$. Let us find the z-score for the first student:

$$z_1 = \frac{8 - 22.3}{12.53} = -1.14.$$

Which means that this student's result is 1.14 standard deviations below the mean. The remaining z-scores are obtained in an analogous way. They are shown in the second row of Table 15. The highest score (40) is 1.41 standard deviations above the mean, whereas the smallest score (5) is 1.38 below the mean. So it could be said that the result of 40 points is more "remarkable" than the result of 5 in the sense that it is farther away from the mean. $\square$

i	1	2	3	4	5	6	7	8	9	10
x_i	8	15	5	36	40	30	9	21	32	27
z_i	-1.14	-0.58	-1.38	1.09	1.41	0.61	-1.06	-0.10	0.77	0.38

Table 15: Points of ten students in an exam in Statistics and their z-scores

Another use of z-scores is for comparing observations from different variables, possibly from different populations. For instance, let us say that the next year, the course of statistics was taken by eight students. One of the students got 41 points in the exam. In absolute terms, evidently, this value is higher than 40, which was the maximum score during the previous year, but it may be that the exam was easier, right? By standardizing both populations we can establish "how good was each student *with respect to their own populations*".

Example 31 (Continuation of Example 30)**.** Let U_2 be the population of eight students who took the master course in statistics the next year. The first row of Table 16 shows their points in the exam. The average number of points during this year was $\bar{x}_{U_2} = 35$ and the standard deviation was $S_{x,U_2} = 7.01$. Thus, we obtain the z-scores shown in the second row of Table 16.

i	1	2	3	4	5	6	7	8
x_i	48	29	26	32	41	37	35	32
z_i	1.85	-0.86	-1.28	-0.43	0.86	0.29	0.00	-0.43

Table 16: Points of eight students in an exam in Statistics and their z-scores

The mean during the second year was much larger than during the first year. A result of 40 points in the exam during the first year is 1.41 standard deviations over the mean, but a result of 41 during the second year is only 0.86 standard deviations over the mean. Thus, compared to their own populations, the student with 40 points performed better than the one with 41 points.

Although z-scores allow for comparing observations between different variables and different populations, we should be cautious with the interpretation. Let us take a look back at Examples 30 and 31. The mean during the first year was 22.3 points, whereas the exam during the second year was 35 points. We do not know if this is due to the exam being easier or to the students being better prepared for the exam. All we can say with the z-scores is that "with respect to their own population" a result of 40 during the first year was more remarkable than a result of 41 during the second one.

Empirical rule

In large populations, a variable x with mean $\bar{x}_U$ and standard deviation $S_{x,U}$ which is symmetric, unimodal and bell-shaped will satisfy:

- approximately 68% of the observations lie in the interval $[\bar{x}_U \pm S_{x,U}]$;
- approximately 95% of the observations lie in the interval $[\bar{x}_U \pm 2 \cdot S_{x,U}]$;
- almost all of the observations lie in the interval $[\bar{x}_U \pm 3 \cdot S_{x,U}]$.

Later in the course we will learn where does this empirical rule come from.

Example 32. Let us consider the population of $N = 97$ startups in Section 2.1.4. The mean and standard deviation of the number of employees, x, are $\bar{x}_U = 4.97$ and $S_{x,U} = 2.28$, respectively. Figure 17 shows the dotplot of x. We see that the variable is unimodal but it is not exactly symmetric. Nevertheless, let us see how well does the empirical rule work in this case.

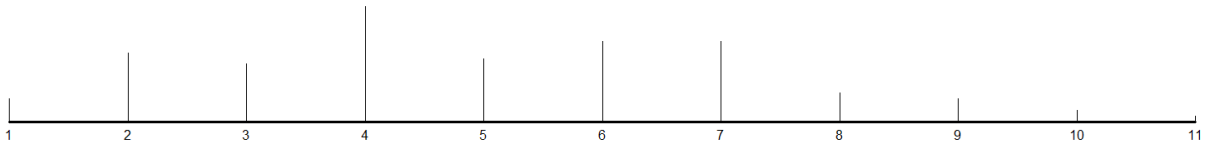

Figure 17: Dotplot of the number of employees of 97 startups

According to the empirical rule,

- approximately 68% of the observations lie in the interval $[4.97 - 2.28\,, 4.97 + 2.28] = [2.69\,, 7.24]$. One can verify that, in fact, 71.1% of the observations lie in this interval;
- approximately 95% of the observations lie in the interval $[4.97 - 2 \cdot 2.28\,, 4.97 + \cdot 2.28] = [0.42\,, 9.52]$. One can verify that 96.9% of the observations lie in this interval;
- almost all of the observations lie in the interval $[4.97 - 3 \cdot 2.28\,, 4.97 + 3 \cdot 2.28] = [-1.86\,, 11.79]$. One can verify that all of the observations lie in this interval.

Note that in this case even when the shape of the variable does not exactly satisfy the conditions for the empirical rule, it still works pretty well. □

2.2 Graphical description

In Section 2.1 we introduced several parameters that allow for describing different characteristics of variables measured in the elements of a population. It is often said that "a picture is worth a thousand words", accordingly, it is common practice to complement the numerical analysis of a variable by graphs. In this subsection we introduce several different graphs that allow for describing the information provided by one or two variables. However, one should be careful. Although a graph may be a nice way of presenting information, a poorly built graph may distort the reality. Throughout this section we try to point out some mistakes that should be avoided when constructing graphs.

There is a general rule that should be taken into account whenever you are graphing data: the so-called "area principle". This principle says that the *area* occupied by a part of the graph should be proportional to the value it represents. We will mention this rule repeatedly throughout this subsection.

2.2.1 Graphs to describe categorical variables

In this section we introduce two types of graphs that can be used for describing the information provided by a categorical variable.

Bar charts

In a bar chart the categories of the variable of interest are placed along one of the axes (typically the horizontal axis), with the another axis representing the frequency of each category. Bar charts are useful for draw attention to the frequency of the categories.

- You can plot either the absolute or the relative frequency. The resulting plot is identical except for the scale.
- The bars should not touch each other and it is important that all bars have the same width;

Value	Absolute frequency	Relative frequency
F	51	0.425
E	12	0.100
D	23	0.192
C	18	0.150
B	11	0.092
A	5	0.042
Total	120	1

Table 17: Frequency distribution table of the grades of 120 students in an exam in Statistics

- If the variable is ordinal, the categories must be sorted either in ascending or descending order.
- If the variable is not ordinal, it is common practice to sort the categories from the most frequent to the least frequent or vice versa.
- It is important to *always* label the axes, otherwise a reader may not know what is being shown in the chart.

Example 33. Let U be the population of $N = 120$ students taking a course in statistics. Let x_i be the grade in the exam (A, B, C, D, E, F) for the ith student ($i = 1, 2, \cdots, N$). Table 17 shows the frequency distribution table of x.

Figure 18 shows a bar plot of the grades in the exam of the $N = 120$ students. □

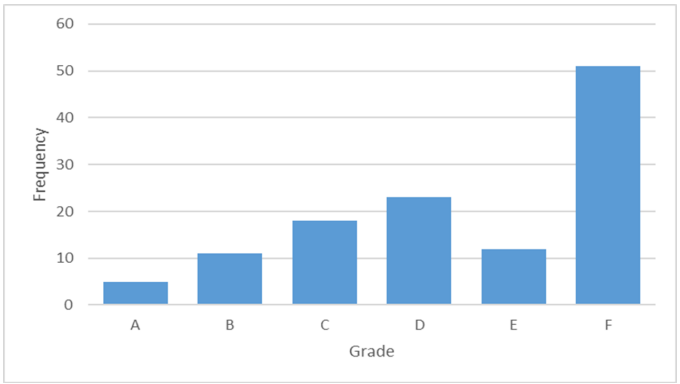

Figure 18: Bar chart of the grades of 120 students in an exam.

The bars should start at zero, otherwise the chart will be misleading as the differences between categories will look bigger than they actually are, thus we would be violating the area principle. If the intention is, precisely, to draw attention to these differences, the scale may be changed, but it is important to make this clear to the reader. For example, Figure 19 is a bar plot of the number of employees of a company by sex. By looking at the plot we get the impression that there are around three times more men than women in the company, but after a closer look we see that a misleading scale has been chosen. In fact there are 102 men and 98 women, so the relative difference is not as large as the plot may incorrectly suggest.

Figure 20 shows a bar plot of the same data with the vertical axis starting at zero.

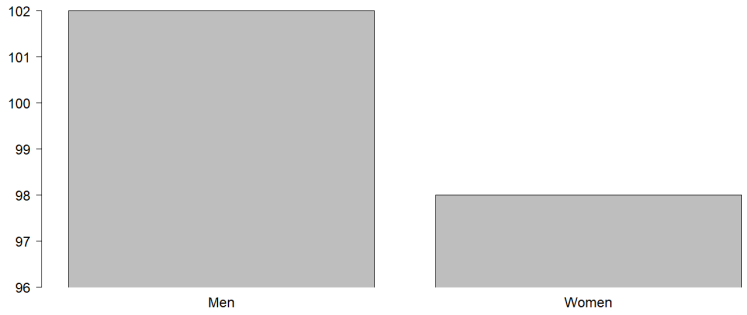

Figure 19: Bar chart of the the number of men and women in a company.

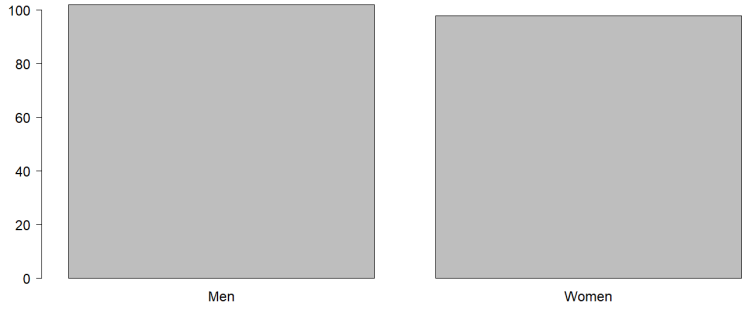

Figure 20: Bar chart of the the number of men and women in a company.

Pie charts

If we want to draw attention to the proportion of elements in each category, then we will probably use a pie chart to depict the division of a whole into its constituent parts. The circle (or "pie") represents the total, and the segments (or "pieces of the pie") cut from its center depict shares of that total. The pie chart is constructed so that the area of each segment is proportional to the corresponding frequency.

Example 34. Figure 21 shows a pie chart of the grades of the 120 students in Example 39 in a final exam. □

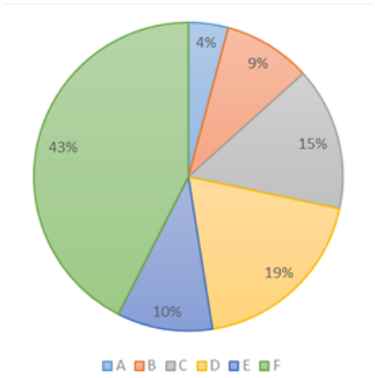

Figure 21: Pie chart of the grades of 120 students in an exam and an assignment.

It should be noted that the order of the categories is lost in a pie chart, therefore it may not be the best choice for ordinal variables. In this sense, instead of using a pie chart for the grades of the students as in the example above, a bar chart may be more adequate.

2.2.2 Graphs to describe numerical variables

In this section we introduce three types of charts that can be used to illustrate the information provided by one numerical variables.

Dot plots

By now, we should be familiar with dotplots. We have it extensively to illustrate the parameters that were introduced in Subsection 2.1.

In a dotplot the observations are represented as dots (or other symbols, like line segments) over a number line. If one value occurs multiple times, we just stack them over each other.

Dotplots are simple to create and (hopefully) easy to interpret. However, it is often said that they are useful for illustrating small to moderate populations. For instance, R documentation for the function `stripchart` (which allows for creating dotplots) says "These plots are a good alternative to boxplots when sample sizes are small" and Wikipedia's page says that dotplots "are suitable for small to moderate sized data sets[...] When dealing with larger datasets[...]dotplots may become too cluttered".

Example 35. The following are the number of points obtained by the 120 students in Example 33:

10	87	40	20	47	40	40	94	48	15
15	66	66	15	5	18	37	29	92	64
93	70	78	45	59	41	68	42	68	93
28	85	18	63	15	15	86	71	40	32
75	64	37	53	25	76	11	35	63	50
52	63	73	79	13	16	83	74	15	60
81	78	20	80	80	66	82	5	20	79
75	10	68	61	63	63	61	15	50	88
76	33	50	57	70	61	9	0	84	77
15	60	27	94	34	20	75	50	76	34
5	58	42	73	20	36	40	83	58	55
28	55	30	60	73	42	65	69	61	61

Figure 22 represents a dotplot of the number of points obtained by the 120 students in the final exam. □

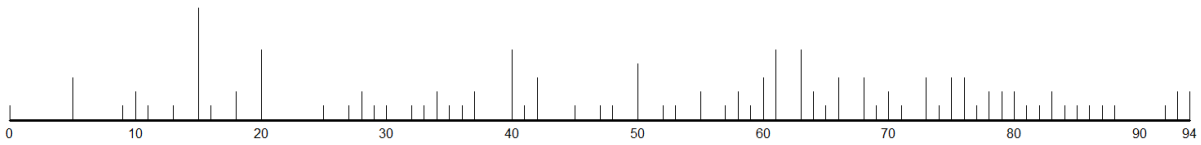

Figure 22: Dotplot of the number of points of 120 students in an exam.

Histograms

A histogram is a graph that consists of vertical bars constructed on a horizontal line that is marked off with intervals for the variable being displayed. The intervals correspond to the classes in a frequency distribution table. It is important that all intervals have the same width, otherwise the result may be misleading as we would be violating the area principle. If it is

not possible to create intervals with the same width (for instance, if the classes are given), it should be taken into account that the area of each bar must be proportional to its frequency.

Bars representing two categories that are adjacent should touch each other. As always, it is important to use labels for the axis.

In order to determine the number and the width of the categories we simply repeat the recommendations given in Section 2.1.4 when we introduced frequency distribution tables. Regarding the number of categories, a rule of thumb (which I often use) is to set $K \approx \sqrt{N}$ classes. Regarding the width of the classes, it can be defined as $\text{range}_{x,U}/K$, where $\text{range}_{x,U}$ is given by (14). However, good sense and some flexibility is needed for obtaining a "nice" presentation. Finally, it is very important to make sure that the categories are inclusive and nonoverlapping, so that every observation belongs to one and only one category.

Example 36. Figure 23 represents a histogram of the number of points obtained by the 120 students in the final exam. □

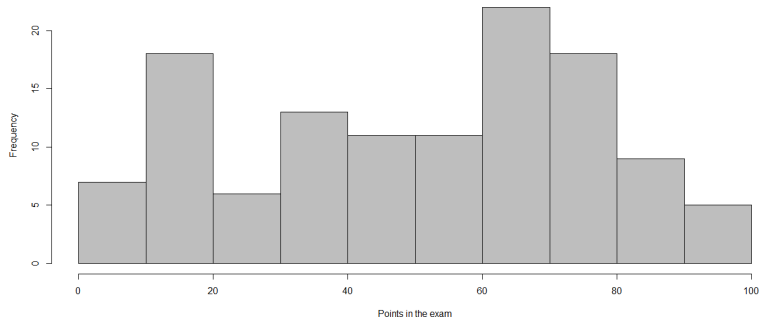

Figure 23: Histogram of the number of points of 120 students in an exam.

Box-and-Whisker plot

A box-and-whisker plot is a graph that describes the shape of a variable in terms of five parameters: the minimum value $x_{(1)}$, the first quartile (25th percentile) $\breve{x}_{25,U}$, the median $\breve{x}_U$, the third quartile (75th percentile) $\breve{x}_{75,U}$, and the maximum value $x_{(N)}$:

- A box of arbitrary width is drawn from the first to the third quartile. A line is drawn through the box at the median $\breve{x}_U$.
- There are two "whiskers":
  - one whisker is a line from the first quartile $\breve{x}_{25,U}$ to either the minimum $x_{(1)}$ or $\breve{x}_{25,U} - 1.5\, IQR_{x,U}$ (whichever is larger);
  - the other whisker is a line from $\breve{x}_{75,U}$ to either the maximum $x_{(N)}$ or $\breve{x}_{75,U} + 1.5\, IQR_{x,U}$ (whichever is smaller).
- If there are outliers (according to the definition in Subsubsection 2.1.3), they are presented as individual points.

Example 37. Figure 24 represents a box-and-whisker plot of the number of points obtained by the 120 students in the final exam. □

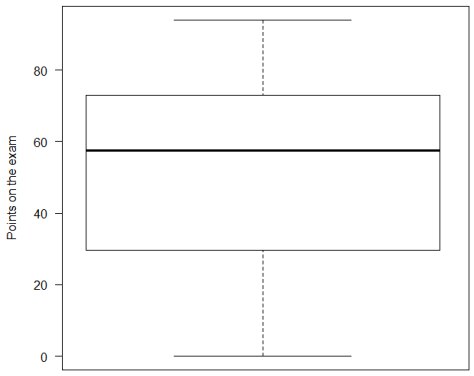

Figure 24: Box-and-whisker plot of the number of points of 120 students in an exam.