There are some percentiles that are of particular interest, these are the *quartiles*.

Definition 12. The *quartiles* are the percentiles that divide the population into four quarters, so the first quartile is the 25th percentile, $\breve{x}_{25,U}$; the second quartile is the 50th percentile, $\breve{x}_{50,U}$; and the third quartile is the 75th percentile, $\breve{x}_{75,U}$. Note that the second quartile coincides with the median, $\breve{x}_U$. □

Example 13. Consider the population of ten students in Examples 5 and 7 and the x-ordered population shown in Table 4. In order to calculate the first quartile, which is the 25th percentile, first we find that

$$c = (N-1)p + 1 = (10-1)0.25 + 1 = 3.25,$$

therefore $a = 3$ and $b = 0.25$, and we get

$$\breve{x}_{0.25,U} = (1-b)x_{(a)} + bx_{(a+1)} = (1-0.25)x_{(3)} + 0.25x_{(3+1)} = 0.75 \cdot 9 + 0.25 \cdot 15 = 10.5.$$

In the same way we find that the second quartile is $\breve{x}_{0.50,U} = 24$, which is the median found in Example 7; and the third quartile is $\breve{x}_{0.75,U} = 31.5$. The three quartiles are illustrated in Figure 8, which in loose words indicates that one quarter of the population takes values smaller than 10.5, one quarter takes values between 10.5 and 24, one quarter takes values between 24 and 31.5, and one quarter takes values larger than 31.5.

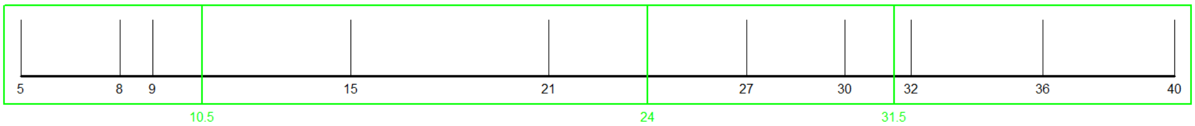

Figure 8: The quartiles of the exam points of the students.

Being a nominal variable, the percentiles and quartiles of y are not defined. □

2.1.2 Measures of variability

All the parameters that we have introduced so far (mean, median, mode and percentiles) are useful for indicating where the "mass" of the data lies. It is interesting that just one number summarizes quite well some of the information provided by an entire dataset. However, (of course) it does not show the whole picture. Consider the three variables shown in Table 5 and illustrated with dotplots in Figure 9. They all have the same mean (50), they also have the same median and the same mode (50), but it is evident that they are quite different: the first one is completely concentrated on the value 50; the last one may be centered around 50 but its values vary quite a lot; and the second one is somewhere in between. In this section we will introduce some descriptive parameters that allow for measuring the variability in the observations.

x_1	x_2	x_3
50	40	01
50	44	20
50	47	30
50	49	40
50	50	50
50	50	50
50	50	50
50	51	60
50	53	70
50	56	80
50	60	99

Table 5: Three variables with the same mean, median and mode

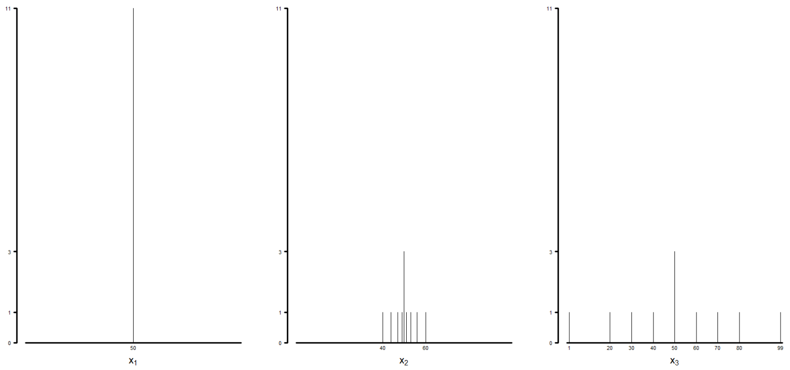

Figure 9: Dotplots of the three variables in Table 5.

The range

The first parameter we introduce is the *range*.

Definition 14. Let x be a variable that is at least ordinal, the *range* is the difference between the maximum and the minimum of x, i.e.

$$\text{range}_{x,U} = x_{(N)} - x_{(1)}. \qquad \square \tag{4}$$

Example 15. Let us consider the population of $N = 10$ students in Examples 5 to 11 and let us compute the range of $x =$ "number of points obtained by the students in the exam". We get

$$\text{range}_{x,U} = x_{(N)} - x_{(1)} = 40 - 5 = 35.$$

The range is illustrated on the left panel of Figure 10 as the length of the red segment that spans from the minimum to the maximum. □

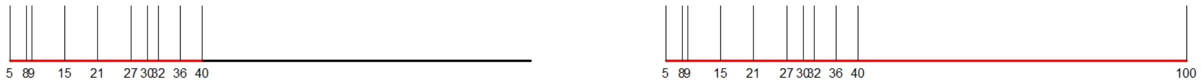

Figure 10: Dotplot of the exam points of the students. (The range in red)

Probably the biggest advantage of the range as a measure of variability is its simplicity. On the other hand, the range has some limitations. In particular, it completely ignores the distribution of the data and it is very sensitive to extreme observations, in other words, the range is a non-robust parameter. To see this, note that if a new student scores 100 points in the exam, the range becomes 95 (see the right panel of Figure 10, where the range is represented as the length of the red segment).

The interquartile range

One parameter that is similar to the range but is less sensitive to extreme observations is the *interquartile range.*

Definition 16. Let x be a variable that is at least ordinal, the *interquartile range* of x in the population U, $\mathrm{IQR}_{x,U}$, is the difference between the third and the first quartiles of x, i.e.

$$\mathrm{IQR}_{x,U} = \breve{x}_{75,U} - \breve{x}_{25,U}. \qquad \square \tag{5}$$

Recall that the first quartile is the value that leaves one quarter of the population under it and the third quartile is the value that leaves three quarters of the population under it. Therefore the central half of the population lies between the first and the third quartiles. The interquartile range is just the range spanned by this central part.

Example 17. In Example 13 we found that the first and the third quartiles of the number of points in the exam for our population of students are, respectively, $\breve{x}_{0.25,U} = 10.5$ and $\breve{x}_{0.75,U} = 31.5$. Using this, we obtain the interquartile range as

$$\mathrm{IQR}_{x,U} = \breve{x}_{0.75,U} - \breve{x}_{0.25,U} = 31.5 - 10.5 = 21.$$

The IQR is illustrated on the left panel of Figure 11 as the length of the red segment that goes from the first to the third quartile. As indicated before, the IQR is the range spanned by the central half of the population. In this way, the central half of the values of x span over a range of length 21 units. □

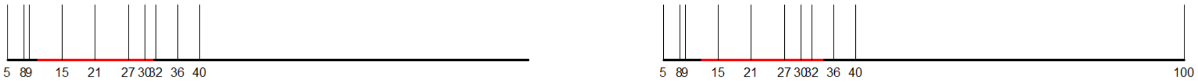

Figure 11: Dotplot of the exam points of the students. (The IQR in red)

Unlike the range, the IQR is robust with respect to extreme observations. As we saw before a student scoring 100 points in the exam would change the range from 35 to 95. The IQR, on the other hand, would change from 21 to 22. So the extreme observation is having only a minor impact on the IQR.

The variance

Both the range and interquartile range are measures of variability that are based only on two observations of the entire population. A parameter that makes use of all the observations is the variance.

Definition 18. There are two slightly different definitions of the *variance.* The first one (that is more intuitive) is

$$S'^2_{x,U} \equiv \frac{1}{N} \sum_U (x_i - \bar{x}_U)^2 , \tag{6}$$

which is simply the mean of the square distances from each observation to the mean. The second definition uses $N-1$ instead of N in the denominator, i.e.

$$S^2_{x,U} \equiv \frac{1}{N-1} \sum_U (x_i - \bar{x}_U)^2 . \qquad \square \tag{7}$$

Although the reason for using $N-1$ instead of N in the denominator may not be clear at this point, it will become more clear later in the course. Many authors call $S'^2_{x,U}$ the "population variance" and $S^2_{x,U}$ the "sample variance". We will try to avoid those labels. Instead we will simply say that if the purpose is to *describe* the variability of the observations of a variable x in the population U, $S'^2_{x,U}$ *must* be preferred over $S^2_{x,U}$. For large populations, however, the difference between $S'^2_{x,U}$ and $S^2_{x,U}$ becomes negligible.

Furthermore, if x is a dummy variable with mean $\bar{x}_U = P_x$, we have

$$S'^2_{x,U} = P_x(1-P_x) \qquad \text{and} \qquad S^2_{x,U} = \frac{N}{N-1}P_x(1-P_x). \tag{8}$$

Example 19. Let us consider the population of $N = 10$ students in Examples 5 to 11 and let us compute the variance $S^2_{x,U}$ of $x =$ "number of points obtained by the students in the exam" using (6). In Example 5 we found that the mean is $\bar{x}_U = 22.3$, therefore $S^2_{x,U}$ is

$$S^2_{x,U} = \frac{1}{N-1}\sum_U (x_i - \bar{x}_U)^2 = \frac{1}{10-1}\left((8-22.3)^2 + (15-22.3)^2 + \cdots + (27-22.3)^2\right) =$$
$$\frac{1}{9}\left(-14.3^2 + -7.3^2 + \cdots + 4.7^2\right) = \frac{1}{9}(204.49 + 53.29 + \cdots + 22.09) = \frac{1}{9}1412.10 = 156.9.$$

Now, let us calculate the variance of the dummy variable $y =$ "sex of the students". In Example 5 we found that the mean is $P_y = \bar{y}_U = 0.6$, therefore $S^2_{y,U}$ is

$$S^2_{y,U} = \frac{1}{N-1}\sum_U (y_k - \bar{y}_U)^2 = \frac{1}{10-1}\left((1-0.6)^2 + (1-0.6)^2 + \cdots + (0-0.6)^2\right) =$$
$$\frac{1}{9}\left(0.4^2 + 0.4^2 + \cdots + -0.6^2\right) = \frac{1}{9}(0.16 + 0.16 + \cdots + 0.36) = \frac{1}{9}2.4 = 0.2667.$$

Let us find the variance of y using the alternative expression given by (8), i.e.

$$S^2_{y,U} = \frac{N}{N-1}P(1-P) = \frac{10}{10-1}0.6 \cdot 0.4 = 0.2667.$$

which is the same value obtained before. □

Some properties of the variance:

- by construction, variances cannot be negative;
- if a constant is added to each observation, the variance of the resulting variable is simply the variance of the original variable. More formally, let $y_i = x_i + c$ for $i = 1, 2, \cdots, N$, then $S^2_{y,U} = S^2_{x,U}$ and $S'^2_{y,U} = S'^2_{x,U}$;
- if each observation is multiplied by a constant, the variance of the resulting variable is simply the variance of the original variable multiplied by the constant squared. More formally, if $y_i = cx_i$ for $i = 1, 2, \cdots, N$, then $S^2_{y,U} = c^2 S^2_{x,U}$ and $S'^2_{y,U} = c^2 S'^2_{x,U}$.

In R, the function `var()` can be used to compute $S^2_{x,U}$.

Standard deviation

Due to the squared terms, the units of the variance are the squared units of the variable of interest (whatever this is). For instance, in Example 19, the variance of the scores of the students in the exam is 141.21 points squared. This fact makes it difficult to interpret the variance. The *standard deviation* tries to solve this problem.

Definition 20. The *standard deviation* is the positive square root of the variance. As we have two different definitions of the variance (Equations (6) and (7)), we also have two different definitions of the standard deviation:

$$S'_{x,U} \equiv \sqrt{S'^2_{x,U}} = \sqrt{\frac{1}{N}\sum_U (x_i - \bar{x}_U)^2} \qquad \text{and} \qquad S_{x,U} \equiv \sqrt{S^2_{x,U}} = \sqrt{\frac{1}{N-1}\sum_U (x_i - \bar{x}_U)^2}. \qquad \square \tag{9}$$

A loose interpretation of the standard deviation is as the average distance from the observations to the mean.

Example 21. In Example 19 we found that the variance of the variable x = "scores of the students in the final exam" is $S^2_{x,U} = 156.9$ and the variance of the dummy variable y = sex of the students is $S^2_{y,U} = 0.2667$. It is straightforward to verify that the standard deviations are $S_{x,U} = 12.53$ and $S_{y,U} = 0.5164$, respectively. In other words, on average, the exam points deviate 12.53 units from the average $\bar{x}_U = 22.3$. This is illustrated on the top panel of Figure 12 where the red segment extends one standard deviation on each side of the average, in other words, on the average the observations deviate this much from the mean. □

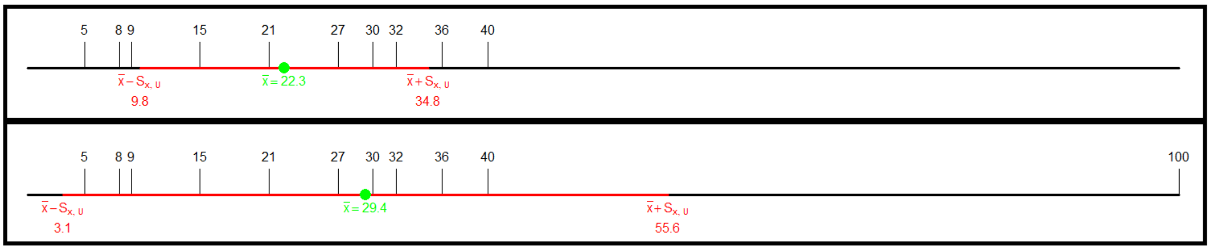

Figure 12: The standard deviation as the average distance to the mean.

As it was the case with the mean, the standard deviation is a non-robust parameter. Consider the new student who scores 100 points in the exam. The standard deviation becomes 26.27, so this extreme observation has a strong impact on the standard deviation. This is illustrated on the lower panel of Figure 12.

Some properties of the standard deviation:

- by construction, standard deviations cannot be negative;
- if a constant is added to each observation, the standard deviation of the resulting variable is simply the standard deviation of the original variable. More formally, let $y_i = x_i + c$ for $i = 1, 2, \cdots, N$, then $S_{y,U} = S_{x,U}$ and $S'_{y,U} = S'_{x,U}$;
- if each observation is multiplied by a constant, the standard deviation of the resulting variable is simply the standard deviation of the original variable multiplied by the constant. More formally, if $y_i = cx_i$ for $i = 1, 2, \cdots, N$, then $S_{y,U} = c\,S_{x,U}$ and $S'_{y,U} = c\,S_{x,U}$.

In R, the function `sd()` can be used to compute $S_{x,U}$.

2.1.3 Measures of shape

In Sections 2.1.1 and 2.1.2 we introduced several parameters that allow for describing where the mass of observations is located and how spread they are. In this subsection we discuss two characteristics that are related to the shape of the observations of a variable: skewness and outliers.

Skewness

The *skewness* is a measure that indicates whether the observations of a variable are symmetric or not. There are several slightly different definitions of the skewness. But for the purposes of the course, the following definition will suffice.

Definition 22. The *skewness* of a numerical variable x in the population U is defined as

$$Sk_{x,U} \equiv \frac{\frac{1}{N}\sum_U (x_i - \bar{x}_U)^3}{S_{x,U}^3} \tag{10}$$

where $\bar{x}_U$ is the mean (1) and $S_{x,U}$ is the standard deviation (9). □

A skewness equal to zero indicates that the values of the variable are symmetric around the mean. A skewness larger than zero indicates that the variable has a tail that extends to the right. In this case we say that the variable is skewed to the right or positively skewed. Analogously, a skewness smaller than zero indicates that the variable has a tail that extends to the left. In this case we say that the variable is skewed to the left or negatively skewed.

x_1	x_2	x_3
37	18	53
41	23	74
49	26	78
50	28	79
58	29	82
65	37	84
84	38	90
85	44	91
86	68	94

Table 6: Three variables with different skewness

Example 23. Table 6 shows three variables whose dotplots can be found in 13. Note that the values of x_1 are more or less well spread over its range, therefore the skewness will be small. The values of x_2 are highly concentrated at the left side with one large value at the right, so we expect the skewness to be positive. On the other hand, x_3 takes values concentrated at the right side with one small value at the left, so we expect the skewness to be negative.

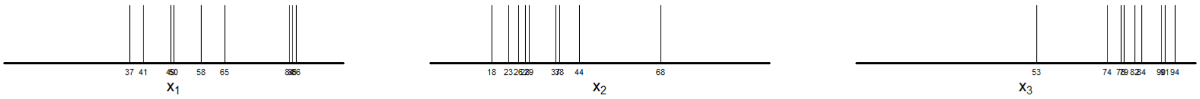

Figure 13: Dotplots of the three variables in Table 6.

Let us find the skewness of the first variable, x_1, in Table 6. We have that $N = 9$ and $\bar{x}_{1U} = 61.7$, so the numerator is

$$\frac{1}{N}\sum_U (x_{1i} - \bar{x}_{1,U})^3 = \frac{1}{9}\left((37-61.7)^3 + (41-61.7)^3 + \cdots + (86-61.7)^3\right) =$$

$$\frac{1}{9}\left(-24.7^3 + -20.7^3 + \cdots + 24.3^3\right) = \frac{1}{9}\left(-15008 + -8827 + \cdots + 14408\right) =$$

$$\frac{1}{9}10783 = 1198.$$

In order to calculate the denominator, first we calculate the variance $S^2_{x1,U}$,

$$\frac{1}{N-1}\sum_U (x_{1i} - \bar{x}_{1,U})^2 = \frac{1}{9-1}\left((37-61.7)^2 + (41-61.7)^2 + \cdots + (86-61.7)^2\right) =$$

$$\frac{1}{8}\left(-24.7^2 + -20.7^2 + \cdots + 24.3^2\right) = \frac{1}{8}(608.4 + 427.1 + \cdots + 592.1) = \frac{1}{8}2992 = 374.$$

So $S^3_{x_1,U} = (S^2_{x_1,U})^{3/2} = 382.75^{3/2} = 7233$ and the skewness is

$$Sk_{x_1,U} = \frac{\frac{1}{N}\sum_U (x_{1i} - \bar{x}_{1,U})^3}{S^3_{x_1,U}} = \frac{1198}{7233} = 0.166,$$

which is close to zero, indicating that x_1 is almost symmetric but slightly skewed to the right.

It is left as an exercise to verify that the skewness of x_2 is $Sk_{x_2,U} = 1.043$, which means that it is skewed to the right; and the skewness of x_3 is $Sk_{x_3,U} = -1.009$, which means that it is skewed to the left. □

In R, the function `skewness()` from the package `e1071` can be used to compute the skewness.

Outliers

We have mentioned *extreme* observations several times over the previous pages. In the statistical jargon, extreme observations are called *outliers.* An outlier is an element whose value stands far from where most of the other values are, either being too large or being too small.

There is no formal definition of an outlier that is generally accepted by statisticians. It could be said that no definition is satisfactory enough. Just for the sake of completeness we provide the following definition which is due to John Tukey, but we insist in that no definition of outlier is accepted as satisfactory.

Definition 24. The ith element of the population U is an outlier with respect to the variable x if its value x_i satisfies

$$x_i < \breve{x}_{0.25,U} - 1.5IQR_{x,U} \qquad \text{or} \qquad x_i > \breve{x}_{0.75,U} + 1.5IQR_{x,U}$$

where $\breve{x}_{0.25,U}$, $\breve{x}_{0.75,U}$ and $IQR_{x,U}$ are, respectively, the first quartile, the third quartile and the interquartile range. □

In words, we have defined an outlier as an element whose value is larger than the third quartile plus 1.5 times the IQR or smaller than the first quartle minus 1.5 times the IQR.

Example 25. Let us consider the population of eleven students and their grades on the exam again. The grades are shown below:

5	8	9	15	21	27	30	32	36	40	100

It is evident that the value 100 stands out from the other values in the population. We have discussed several times the effect that this observation has on the different parameters that we have introduced so far. Let us verify that this student satisfies our definition of an outlier.

The first and third quartiles of the number of points are, respectively, $\breve{x}_{0.25,U} = 12$ and $\breve{x}_{0.75,U} = 34$, therefore the interquartile range is $IQR_{x,U} = \breve{x}_{0.75,U} - \breve{x}_{0.75,U} = 34 - 12 = 22$. Thus any student with less points than $\breve{x}_{25,U} - 1.5IQR_{x,U} = 12 - 1.5 \cdot 22 = -21$ or more than $\breve{x}_{75,U} - 1.5IQR_{x,U} = 34 + 1.5 \cdot 22 = 67$ is considered an outlier. Indeed we see that the student with 100 points in the exam can be regarded as an outlier in this population. □

Once we have identified outliers, there are several recommendations about how to proceed:

- Probably the most important recommendation is to not ignore them by simply keeping on with our analysis. Sometimes the outliers are, in fact, errors in the dataset: are those 100 points only 10 in reality but someone typed an extra 0 by mistake? Try to verify that this is not the case!

- If possible, make use of robust parameters to reduce the impact that outliers will have in the final results (e.g. use the median as a measure of location and the IQR as a measure of variability).

- When may it not be possible to make use of robust parameters? In some situations it may have been established beforehand that some parameters (like the mean or the standard deviation) should be reported. If this is the case make sure to let the reader know that the population has outliers.

- Is it possible to remove the outliers from the population and carry out the analysis again without them? What is the intention with your analysis? Let us say that we are analyzing the income of the inhabitants of a city in which there is a billionaire. Including this billionaire in the analysis will affect our results, can we analyze the data without this value?

- Sometimes the data are analyzed with and without the outliers, just to make evident the impact they have on the results. Personally, I consider this to be the "best" choice. On the one hand, the outliers are still part of the population and all parameters still have some meaning (e.g. the mean is still the center of gravity). On the other hand, the remaining part of the population may be easier to describe without them.

2.1.4 Frequency distribution tables

Up to this point we have considered the case where the observations of the variable of interest are shown in an exhaustive manner, i.e. the data are listed for every element of the population. Sometimes this is not the case and data is shown in a grouped manner. One reason for grouping the data is because it saves space, another reason is that presenting the observations in this way allows to identify some characteristics of the population at once. Consider, for instance, a population of $N = 97$ start-ups, below we present the number of employees of each company:

5	4	7	6	5	4	5	5	5	5
8	6	2	1	4	7	6	6	3	6
8	8	8	5	2	4	4	1	4	6
2	1	6	3	4	7	5	3	9	11
6	4	7	6	4	7	10	7	2	3
7	9	9	4	7	3	7	2	1	3
2	2	4	5	7	4	2	2	3	3
6	2	7	4	10	7	3	4	5	7
4	2	6	7	4	8	6	6	4	9
4	4	2	3	5	6	4			

It would take some time to find out that the mean number of employees is $\bar{x}_U = 4.97$. Maybe using the x-ordered population would make things easier:

1	1	1	1	2	2	2	2	2	2
2	2	2	2	2	2	3	3	3	3
3	3	3	3	3	3	4	4	4	4
4	4	4	4	4	4	4	4	4	4
4	4	4	4	4	4	5	5	5	5
5	5	5	5	5	5	5	6	6	6
6	6	6	6	6	6	6	6	6	6
6	7	7	7	7	7	7	7	7	7
7	7	7	7	7	8	8	8	8	8
9	9	9	9	10	10	11			

The x-ordered population made things a bit easier (at least for me) as now, in order to compute the mean, we can see that four companies have one employee, twelve companies have two employees, and so on. So the mean becomes:

$$\bar{x}_U = \frac{1}{N}\sum_U x_i = \frac{1}{97}\left(\overbrace{1+1+1+1}^{\text{4 times}}+\overbrace{2+2+\cdots+2}^{\text{12 times}}+\cdots+11\right) =$$

$$\frac{1}{97}\left((4\times 1)+(12\times 2)+\cdots+(1\times 11)\right) = \frac{1}{97}482 = 4.97.$$

However, we still need to count the number of times each value appears in the population, i.e. we need to find the *frequency* of each value. Things would be much easier if the data was presented in a *frequency distribution table.*

Definition 26. Let x be a variable that has been measured on the N elements of the population U. Suppose that x takes K different values, $x_1, x_2, \cdots, x_K$ and each value occurs with a frequency $f_1, f_2, \cdots, f_K$. Furthermore, let $w_k = f_k/N$ be the proportion of elements taking the kth value ($k = 1, 2, \cdots, K$). The *frequency distribution table* of x in U is:

	Absolute	Relative
Value	frequency	frequency
x_k	f_k	w_k
x_1	f_1	w_1
x_2	f_2	w_2
$\vdots$	$\vdots$	$\vdots$
x_K	f_K	w_K
Total	N	1

Furthermore, if the variable is at least ordinal, it is common to add two columns to the table:

- the cumulative absolute frequency, F_k: which is the number of observations that are smaller or equal than x_k, i.e. $F_k = \sum_{l=1}^{k} f_l$ for all $k = 1, 2, \cdots, K$;
- the cumulative relative frequency, W_k: which is the proportion of observations that are smaller or equal than x_k, i.e. $W_k = \sum_{l=1}^{k} w_l$ for all $k = 1, 2, \cdots, K$.

Clearly, if the variable is nominal it does not make any sense to define the cumulative absolute frequency or the cumulative relative frequency. □

Example 27. Table 7 shows the frequency distribution table of the number of employees from the $N = 97$ start-ups. □

Value x_k	Absolute frequency f_k	Relative frequency w_k	Cumulative absolute F_k	Cumulative relative W_k
1	4	0.0412	4	0.0412
2	12	0.1237	16	0.1649
3	10	0.1031	26	0.2680
4	20	0.2062	46	0.4742
5	11	0.1134	57	0.5876
6	14	0.1443	71	0.7320
7	14	0.1443	85	0.8763
8	5	0.0515	90	0.9278
9	4	0.0412	94	0.9691
10	2	0.0206	96	0.9897
11	1	0.0103	97	1.0000
Total	97	1		

Table 7: Frequency distribution table of the number of employees in a population of $N = 97$ start-ups

This approach is quite convenient when we have to present data for large populations and the variable of interest takes only a few values. The number of employees in the example above takes only eleven different values. However, if the variable takes too many different values, the table would need too many rows. In these situations, the data is often grouped into intervals. For instance, Table 8 shows the number of employees grouped into $K = 3$ intervals: indicating that 46 companies have between one and four employees, 44 companies have between 5 and 8 employees and 7 companies have between 9 and 12. (If the variable is ordinal, we typically collapse the categories with the smallest frequencies.)

Value x_k	Absolute frequency f_k	Relative frequency w_k	Cumulative absolute F_k	Cumulative relative W_k
$(0,4]$	46	0.4742	46	0.4742
$(4,8]$	44	0.4536	90	0.9278
$(8,12]$	7	0.0722	97	1.0000
Total	97	1		

Table 8: Frequency distribution table of the number of employees in a population of $N = 97$ start-ups

Constructing tables with grouped data like Table 8 needs answering some questions, namely, the number of categories, K, to be defined and the width of each category. Regarding the first question, a rule of thumb (which I often use) is to set $K \approx \sqrt{N}$ classes. Regarding the second question, the class width 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.