3 Describing two categorical variables

In Section 2 we introduced several measures for describing one variable. In this and the following sections we will introduce some methods for describing two variables simultaneously. In this section, in particular, we consider the case of two categorical variables.

Contingency tables

The simplest, but usually quite telling, tool for describing two categorical variables is a contingency table. Contingency tables are also known as cross tables.

Definition 38. Let x_i and y_i be the values of two categorical variables associated to the ith individual in the population U ($i = 1, 2, \cdots, N$), where x takes K_x different categories and y takes K_y different categories. A *contingency table* is a matrix-like table that shows, in the cell (k_x, k_y), the frequency of elements taking the k_xth category of x and the k_yth category of y simultaneously (for $k_x = 1, 2, \cdots, K_x$ and $k_y = 1, 2, \cdots, K_y$). □

In simple words, a contingency table is a table that shows in each cell the frequency of one category of one variable x and one category of the second variable y. This is one of the many situations in which things are simpler than they sound: aqui

Example 39. Let U be the population of $N = 120$ students taking a course in statistics. Let x_i be the grade in the first assignment (Pass or Fail) and y_i be the grade in the exam (A, B, C, D, E, F) for the ith student ($i = 1, 2, \cdots, N$). Table 18 shows the values of x and y in the population of students.

x	y	x	y	x	y	x	y	x	y	x	y	x	y	x	y
Fail	F	Pass	F	Pass	F	Pass	F	Pass	E	Pass	D	Pass	C	Pass	B
Fail	F	Pass	F	Pass	F	Pass	F	Pass	E	Pass	D	Pass	C	Pass	B
Fail	F	Pass	F	Pass	F	Pass	F	Pass	E	Pass	D	Pass	C	Pass	B
Fail	F	Pass	F	Pass	F	Pass	F	Pass	D	Pass	D	Pass	C	Pass	B
Fail	F	Pass	F	Pass	F	Pass	F	Pass	D	Pass	D	Pass	C	Pass	B
Fail	F	Pass	F	Pass	F	Pass	F	Pass	D	Pass	D	Pass	C	Pass	B
Fail	F	Pass	F	Pass	F	Fail	E	Pass	D	Pass	D	Pass	C	Pass	B
Fail	F	Pass	F	Pass	F	Pass	E	Pass	D	Pass	D	Pass	C	Pass	B
Fail	F	Pass	F	Pass	F	Pass	E	Pass	D	Pass	D	Pass	C	Pass	B
Pass	F	Pass	F	Pass	F	Pass	E	Pass	D	Pass	D	Pass	C	Pass	B
Pass	F	Pass	F	Pass	F	Pass	E	Pass	D	Pass	D	Pass	C	Fail	A
Pass	F	Pass	F	Pass	F	Pass	E	Pass	D	Fail	C	Pass	C	Pass	A
Pass	F	Pass	F	Pass	F	Pass	E	Pass	D	Pass	C	Pass	C	Pass	A
Pass	F	Pass	F	Pass	F	Pass	E	Pass	D	Pass	C	Pass	C	Pass	A
Pass	F	Pass	F	Pass	F	Pass	E	Pass	D	Pass	C	Fail	B	Pass	A

Table 18: Results of $N = 120$ students in an assignment and an exam in statistics

Note that x takes $K_x = 2$ different categories and y takes $K_y = 6$ different categories. Therefore these values can be summarized in a contingency table of size 2×6 as shown below: □

		y						
		A	B	C	D	E	F	Total
x	Pass	4	10	17	23	11	42	107
	Fail	1	1	1	0	1	9	13
	Total	5	11	18	23	12	51	120

Note that we added the totals to the rows and to the columns. These totals are known as the *marginals*. They show the univariate distributions of each variable, i.e. the values of each variable disregarding the other.

Joint relative distribution

Dividing each entry in the contingency table by N we obtain the proportion of elements that fall in each cell. This is known as the *joint relative distribution*.

Example 40. Now, we divide each entry of the contingency table in Example 39 by $N = 120$ to obtain the joint relative distribution: □

		y						
		A	B	C	D	E	F	Total
x	Pass	0.0333	0.0833	0.1417	0.1917	0.0917	0.35	0.8917
	Fail	0.0083	0.0083	0.0083	0	0.0083	0.075	0.1083
	Total	0.0417	0.0917	0.15	0.1917	0.1	0.425	1.0000

One way for represent graphically the information provided by a contingency table or the joint relative distribution is through *mosaic plots*. In a mosaic plot each cell of the table is represented by a rectangle, whose area is proportional to the value of the cell. Figure 25 shows a mosaic plot of the grade in the assignment and the exam for the population of 120 students. Note, for instance, that as the number of students who passed the assignment and got A in the exam is four times bigger than the number of students who failed the assignment and got A in the exam, therefore, in the mosaic plot, the former is represented by a rectangle that is four times bigger than the latter.

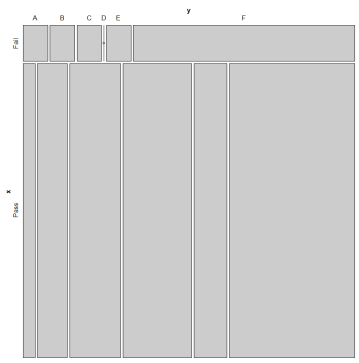

Figure 25: Mosaic plot of the grades in a home assignment and an exam of 120 students.

Conditional distributions

Dividing each cell by the row-totals gives the conditional distribution of y given x.

Example 41. In order to obtain the conditional distribution of the grade in the exam y conditioned on the grade on the assignment x, we divide each entry by the corresponding row-total:

		y						
		A	B	C	D	E	F	Total
x	Pass	0.0374	0.0935	0.1589	0.215	0.1028	0.3925	1.000
	Fail	0.0769	0.0769	0.0769	0	0.0769	0.6923	1.000
	Total	0.0417	0.0917	0.15	0.1917	0.1	0.425	1.000

This table allows us to see some facts that are not so evident from the previous tables, for instance, among students who fail the assignment, almost 70% fail the exam too; while among students who pass the assignment, only around 40% fail the exam. □

In this case, we are obtaining the distribution of y for each value of x. For instance, in the first row we are considering only the students who passed the assignment. Among them 3.74% got A in the exam, 9.17% got B and so on. In the second row we are considering only the students who failed the assignment. Among them 7.69% got B in the exam, 7.69 got B and so on.

Conditional distributions can be represented graphically through *stacked bar charts* as follows. As we are conditioning on x we create K_x bars of length one. Then, each bar is subdivided according to the conditional frequencies of the categories of y. Figure 26 shows a stacked bar chart for the distribution of the grades in the exam conditioned on the grade in the assignment.

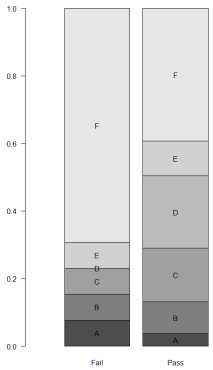

Figure 26: Stacked bar chart for the distribution of grades in the exam conditioned on the grade in the assignment

Dividing each cell by the column-totals gives the conditional distribution of x given y.

Example 42. In order to obtain the conditional distribution of the grade in the assignment x conditioned on the grade in the exam y, we divide each entry by the corresponding column-total:

		y						
		A	B	C	D	E	F	Total
x	Pass	0.8	0.9091	0.9444	1	0.9167	0.8235	0.8917
	Fail	0.2	0.0909	0.0556	0	0.0833	0.1765	0.1083
	Total	1.00	1.00	1.00	1.00	1.00	1.00	1.00

This table says, for instance, that considering only the students who got A in the exam, 80% of them passed the assignment too whereas 20% failed it; considering only the students who got B in the exam, 91% of them passed the assignment too whereas 9% failed it; etc.

Figure 27 shows a stacked bar chart for the distribution of the grades in the assignment conditioned on the grade in the exam. □

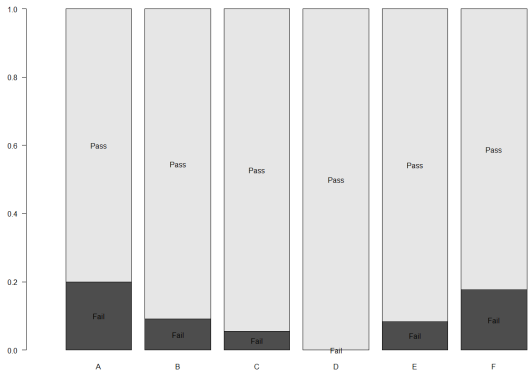

Figure 27: Stacked bar chart for the distribution of grades in the assignment conditioned on the grade in the exam

4 Describing one numerical and one categorical variable

In Section 2 we introduced several measures (numerical and graphical) for describing a numerical variable. If we want to describe simultaneously one numerical variable x and one categorical variable y, basically what we do is to calculate the descriptive measures of x for each category of y. In other words, we calculate descriptive measures of x *conditioned* on y.

We will illustrate the idea with an example. Let the population U of interest be the days in 2012. We are interested in describing the relationship between the average wind speed ($= x$, in miles per hour, mph) and the month of the year (y). Note that the wind speed x is a numerical variable, whereas the month y is a categorical variable.

Although 2012 was a leap year, we have no measurements for December 31, thus the size of our population is $N = 365$.

Just for completeness and as a kind of reminder, we will start by summarizing the numerical variable *wind speed* alone. The third column of Table 19 shows the descriptive parameters of the wind speed. We can see that the daily wind speed spanned from 0 to 6.8 although the central part of the measurements spanned between 0.5 and 2.4. The daily wind speed has a mean of 1.55, and half of the days it was below 1.10. On average, the observations deviate from the mean by 1.32 mph. Also, the wind speed exhibits some positive skewness.

Parameter	Notation	With outliers	Without outliers
Min	$x_{(1)}$	0	0
First quartile	$\breve{x}_{0.25,U}$	0.5	0.5
Mean	$\bar{x}_U$	1.55	1.48
Median	$\breve{x}_U$	1.10	1.1
Mode	$\dot{x}_U$	0.5	0.5
Third quartile	$\breve{x}_{0.75,U}$	2.4	2.35
Max	$x_{(N)}$	6.8	5.1
Range	$\text{range}_{x,U}$	6.8	5.1
Interquartile range	$IQR_{x,U}$	1.9	1.85
Standard deviation	$S_{x,U}$	1.32	1.20
Skewness	$Sk_{x,U}$	1.14	0.88

Table 19: Descriptive parameters of the daily average wind speed in 2012 (in mph)

Figure 28 shows some graphical summaries of the daily wind speed. On the top panel we find a dotplot, on the central–left panel we find the histogram, a boxplot is shown on the central–right panel and a time–series plot is shown in the bottom panel. A time–series plot is a convenient way for illustrating data that is measured in time (as the variable *average wind speed*) it is constructed by placing the time (days, in our case) along the horizontal axis and the measurements along the vertical axis. The data is skewed towards the right: most observations take "small values" and a few observations take "large values". This is, of course, in agreement with the skewness which equal 1.14. This is evidenced by the "tail" that extends towards the right in the dotplot and the histogram or by the top half of the boxplot being much longer that the bottom half. The fact that most observations take "small values" and a few observations take "large values" can also be evidenced in the time–series plot.

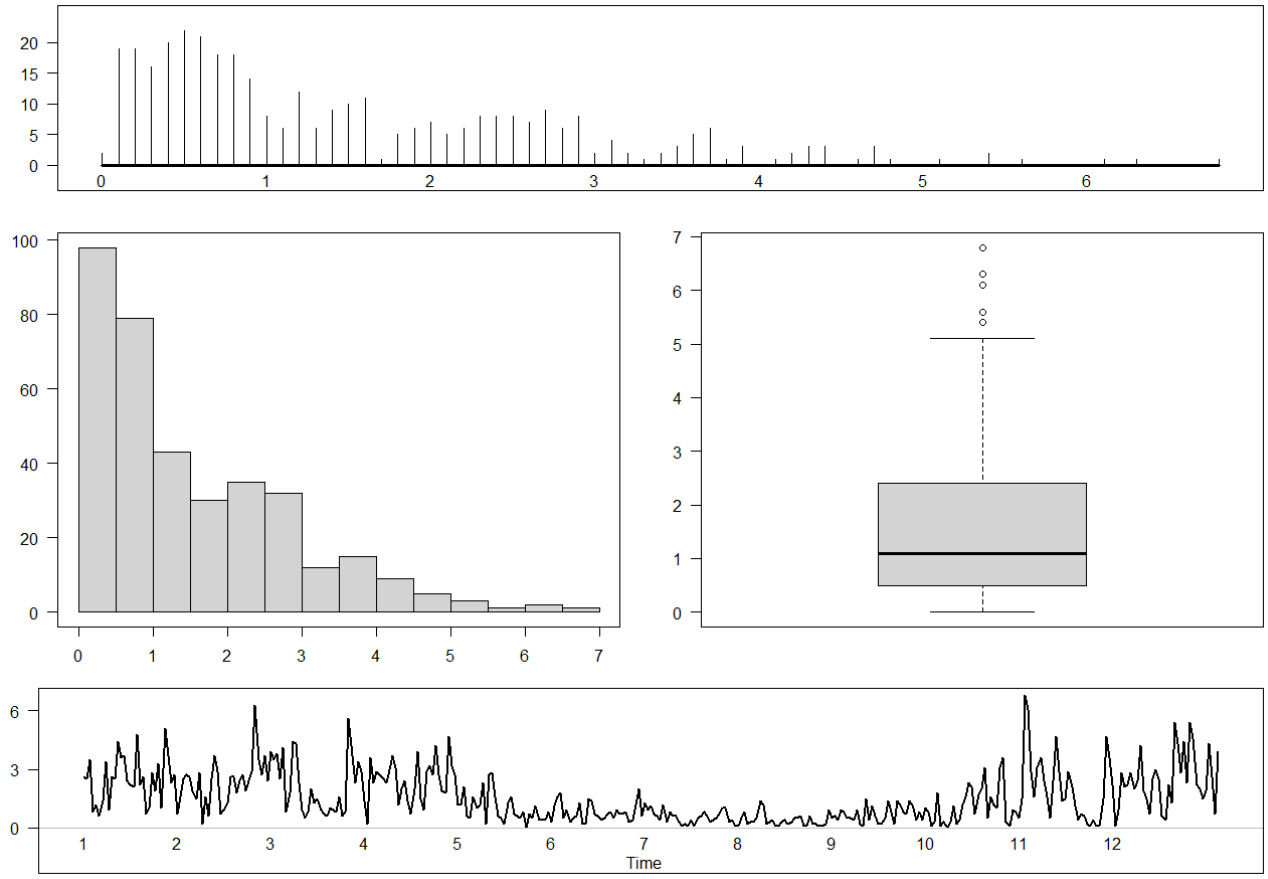

Figure 28: Dotplot (top panel), histogram (center–left), boxplot (center–right) and time–series plot (bottom) of the daily average wind speed in Hopkins Memorial Forest during 2012.

The boxplot in Figure 28 highlights some points that can be regarded as outliers. These observations correspond to the 25th of February, the 26th of March, the 29th and 30th of October and the 16th and 21th of December. All summary measures were recalculating after removing these observations from the dataset. The results are shown in the last column of Table 19. We can see that, although the outliers have some impact in some parameters it is not very large, thus we will keep working with the whole set of observations.

The time–series plot in Figure 28 shows another interesting fact: during summer (june to september) the wind speed is slower than during the rest of the year. It also varies less. Let us dig more into this fact by analyzing simultaneously the average wind speed and the month of the year.

Month	x_1	$\breve{x}_{0.25}$	$\bar{x}$	$\breve{x}$	$\dot{x}$	$\breve{x}_{0.75}$	x_N	range$_x$	IQR_x	S_x	Sk_x
Jan	0.6	1.30	2.43	2.50	2.6	3.35	5.1	4.5	2.05	1.24	0.27
Feb	0.2	1.60	2.34	2.50	2.4, 2.6, 2.7	2.70	6.3	6.1	1.10	1.17	0.95
Mar	0.5	0.85	2.12	1.60	0.8	3.45	5.6	5.1	2.60	1.45	0.65
Apr	0.2	1.83	2.45	2.60	2.7	3.08	4.7	4.5	1.25	1.06	-0.06
May	0.0	0.50	0.98	0.70	0.5	1.25	2.8	2.8	0.75	0.73	1.03
Jun	0.2	0.50	0.83	0.70	0.5, 0.7	1.12	2.0	1.8	0.62	0.48	0.80
Jul	0.1	0.30	0.52	0.50	0.1, 0.6	0.70	1.2	1.1	0.40	0.31	0.41
Aug	0.1	0.20	0.41	0.30	0.1, 0.2	0.50	1.4	1.3	0.30	0.32	1.42
Sep	0.1	0.40	0.69	0.65	0.8	0.98	1.5	1.4	0.58	0.43	0.37
Oct	0.0	0.45	1.61	1.20	0.3	2.05	6.8	6.8	1.60	1.63	1.66
Nov	0.1	0.52	1.74	1.55	0.1	2.70	4.7	4.6	2.17	1.36	0.54
Dec	0.4	1.90	2.62	2.35	2.2	3.67	5.4	5.0	1.77	1.39	0.41

Table 20: Descriptive parameters of the daily average wind speed in 2012 (in mph) by month

Table 20 shows some descriptive measures of the daily average wind speed by month. Let us take a look at the mean and the median: they take the smallest values during August and they increase month after month until the end of the year. The year starts with similar values as those in December and then the average wind speed decreases until July, with one exception: March. It is uncertain if March 2012 was particularly calm or if that is pattern that repeats itself every year. Something similar happens with the variability, the smallest variability occurs during July/August and then it increases until the end of the year. Figure 29 shows side–by–side boxplots of the wind speed by month. They allow to see this behavior in a graphical way.

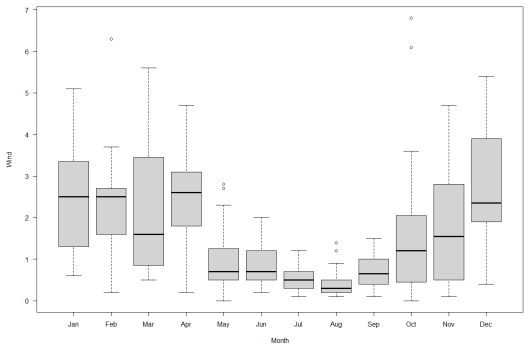

Figure 29: Boxplots of the daily average wind speed in Hopkins Memorial Forest during 2012 by month.

One last thing to highlight before we close this section. Note that when we analyze all observations together, we identified six outliers. When we analyze the data by month some values that were considered as outliers before are not considered as outliers any longer. Also, some values that were not outliers before, are now considered to be outliers. This is due to the fact that a value is considered as an outlier with respect to a population. In a different population, an outlier may not be an outlier any longer.

5 Measures of relationships between two variables

All the parameters that we have introduced up to this point are measures for describing one variable at a time. In this section we introduce three measures that allow for describing the association between two variables.

5.1 The correlation coefficient

In fact, the measure that we introduce in this subsection is more formally known as Pearson's correlation coefficient. A measure due to Karl Pearson.

Definition 43. Let x_i and y_i be the values of two variables associated to the ith element in U $(i = 1, 2, \cdots, N)$. The *correlation coefficient* (or simply, the correlation) between x and y is defined as

$$r_{xy,U} \equiv \frac{\sum_U (x_i - \bar{x}_U)(y_i - \bar{y}_U)}{\left(\sum_U (x_i - \bar{x}_U)^2 \sum_U (y_i - \bar{y}_U)^2\right)^{1/2}}. \qquad \square \tag{11}$$

By construction, the correlation coefficient ranges from -1 to +1. The correlation indicates both the strength and the sign of the linear association between x and y:

- it takes the value of zero when there is no linear relationship between x and y;
- the absolute value of the correlation indicates the strength of the linear association between x and y: values further away from zero indicate a stronger association. A correlation of +1 or -1 indicates a perfect linear association;
- the sign of the correlation indicates the type of association. A positive correlation indicates that increments in one variable are associated to increments in the other one. A negative correlation indicates that increments in one variable are associated to decrements in the other one.

It should be noted that the correlation of a variable with itself is always equal to 1, i.e. $r_{xx,U} = 1$.

Example 44. Let U be the population of $N = 10$ students taking a Master course in statistics. Let x_i and y_i be, respectively, the scores in a home assignment and the final exam of the ith student $(i = 1, 2, \cdots, N)$. Table 21 shows the observed values.

i	1	2	3	4	5	6	7	8	9	10
x_i	0	9	2	24	25	23	4	20	28	5
y_i	8	15	5	36	40	30	9	21	32	27

Table 21: Score of ten students in a home assignment and an exam in statistics

It is convenient to introduce at this point a type of chart that allows for displaying the information of two numerical variables simultaneously, namely, the *scatter plot*. A scatter plot is constructed by simply placing the N points (x_1, y_1), (x_2, y_2), $\cdots$, (x_N, y_N) on a cartesian plane. If one of the variables can be considered to be dependent on the other one, this variable is typically taken to be y and, consequently, represented on the vertical axis. Unless there are good reasons for doing otherwise, the scale of the axes should be chosen in such a way that the points occupy the whole plot region.

Figure 30 shows a scatter plot of the scores in the assignment and the exam for the population of ten students. We see that, in general terms, students who score a higher number

of points in the assignment also score a higher number of points in the exam, so we see a positive association between both variables, therefore, we expect a positive correlation. (The function `plot()` allows to create scatter plots in R.)

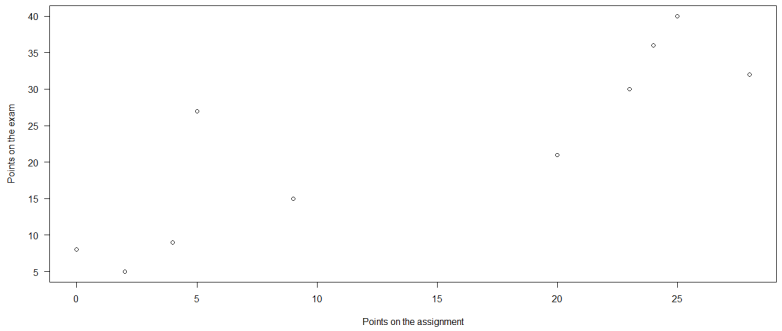

Figure 30: Scatter plot of the points of 10 students in an assignment and an exam.

Let us find the correlation coefficient (11) between x and y. The mean of the two variables is $\bar{x}_U = 14$ and $\bar{y}_U = 22.3$. Let us calculate the numerator first:

$$\sum_U (x_i - \bar{x}_U)(y_i - \bar{y}_U) = (0-14)(8-22.3) + (9-14)(15-22.3) + \cdots + (5-14)(27-22.3) = 200.2 + 36.5 + \cdots + -42.3 = 1064.$$

Now we calculate the first term in the denominator

$$\sum_U (x_i - \bar{x}_U)^2 = (0-14)^2 + (9-14)^2 + \cdots + (5-14)^2 = 196 + 25 + \cdots + 81 = 1080.$$

And the second term in the denominator is

$$\sum_U (y_i - \bar{y}_U)^2 = (8-22.3)^2 + (15-22.3)^2 + \cdots + (27-22.3)^2 = 204.49 + 53.29 + \cdots + 22.09 = 1412.1.$$

This gives

$$r_{xy,U} = \frac{\sum_U (x_i - \bar{x}_U)(y_i - \bar{y}_U)}{\left(\sum_U (x_i - \bar{x}_U)^2 \sum_U (y_i - \bar{y}_U)^2\right)^{1/2}} = \frac{1064}{(1080 \cdot 1412.1)^{1/2}} = 0.8616,$$

which means that there is a relatively high linear association between the number of points that students get in the assignment and the exam. □

It is worth repeating that the correlation measures the *linear association* between two variables. And both variables are very important for interpreting the correlation correctly:

- Correlation indicates association, nothing more than that. In particular, it does not measure causality. It is a common error to conclude causality from correlation. For instance, if we find that there is a high and positive correlation between the amount invested in advertisement and the amount of sales in a set of companies, it may be tempting to conclude that more investment in advertisement leads to higher sales. Although that may still be true, we cannot make that conclusion based on the correlation coefficient.

- Correlation measures linear association. It may be that both variables are associated in a different way and yet the correlation between them is small simply because the association is not linear.