Chapter 2

Descriptive Statistics

The last two examples from the previous chapter were concerned with the problem of inferring information about something (whether the coin is fair or not, and the amount of its bias when it is unfair) based on statistical evidence (the recorded data from a coin flip experiment).

In practice, one often starts a statistical analysis with exploring the available data in order to obtain an intuitive understanding of the patterns, trends and anomalies. Plots (such as histograms, pie charts, scatter plots, ...) and summary statistics (such as mean, variance, correlation, ...) are used to visually and quantitatively describe the data.

In this chapter, we briefly review some of the common plots and summary statistics. These concepts are not only useful in describing data in practice, but they also provide intuition about some more abstract concepts which we will encounter later on, in the context of probability theory.

2.1 Distribution of a variable

2.1.1 Bar plots, dot plots and histograms

Let us start with an example of a data set. The file "`email50.txt`" contains data about 50 emails received at a mailbox.[1] Table 2.1 shows a few rows and columns from the data. Each row of the data set corresponds to one email received at the mailbox, and the columns represent some properties recorded for each email: whether the email was a spam, whether the word "winner" appeared in the email, the number of exclamation marks appeared in the email, and so on. The description of the columns is written below the table.

	spam	to_multiple	cc	attach	dollar	winner	viagra	num_char	line_breaks	exclaim_mess
1	0	0	0	0	0	no	0	21.70	551	8
2	0	0	0	0	0	no	0	7.01	183	1
3	1	0	4	2	0	no	0	0.63	28	2
4	0	0	0	0	0	no	0	2.45	61	1
5	0	0	0	0	9	no	0	41.62	1088	43
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
50	0	0	0	0	0	no	0	15.83	242	4

`spam`	Indicator for whether the email was spam.
`to_multiple`	Indicator for whether the email was addressed to more than one recipient.
`cc`	Indicator for whether anyone was CCed.
`attach`	Number of attached files.
`dollar`	Number of times a dollar sign or the word "dollar" appeared in the email.
`winner`	Indicates whether "winner" appeared in the email.
`viagra`	Number of times "viagra" appeared in the email.
`num_char`	Number of characters in the email, in thousands.
`line_breaks`	Number of line breaks in the email (does not count text wrapping).
`exclaim_mess`	Number of exclamation marks in the email message.

Table 2.1: A few rows and columns from the *emails* data set.

Each email is referred to as a *case* or a *data point*, and each column as a *variable*. Note that the data set contains different types of variables. Some are *numerical* (e.g., the number of characters in thousands, the number of exclamation marks) and others are *categorical* (e.g., whether the email is a spam or not,[2] whether the word "winner" appeared in the email). Among the numerical variables, one can distinguish between two

[1] Source: Data sets for *OpenIntro Statistics* (4th ed., 2019) by D. Diez, M. Çetinkaya-Rundel, C. D. Barr.

[2] Although the variable `spam` has values `0` and `1`, these values do not represent numbers but categories.

types: some can only take discrete values (e.g., the number of exclamation marks), while others can take any value within a certain range (e.g., the number of characters in thousands).[3] The first type is referred to as *discrete* variables and the second type as *continuous* variables.

In statistics, we primarily care about the values of the variables *in the aggregate*, and focus on the values of the variables for individual cases only occasionally, in order to pinpoint anomalies or *outliers*. For instance, we do not care about

- whether the word "winner" has appeared in a specific email in the data set,

but we may care about

- the fraction of the emails in which the word "winner" appeared, or
- the (statistical) co-dependence between the number of characters and the number of exclamation marks appearing in emails.

Ⓠ How can we describe the aggregate values of a single variable from a data set?

For categorical variables, this is quite straightforward: we only need to state the proportion of cases for which the variable takes each possible value. For instance, in the above data set, the variable `spam` takes value `1` in 10% of the cases (i.e., 5 cases out of 50) and value `0` in the remaining 90% of the cases, whereas the variable `winner` takes value "`yes`" in only 2% of the cases (i.e., in 1 case out of 50) and value "`no`" in the remaining 98% of the cases. We can also visualize this information with a *bar plot* (see Figure 2.1), or with a *pie chart*.

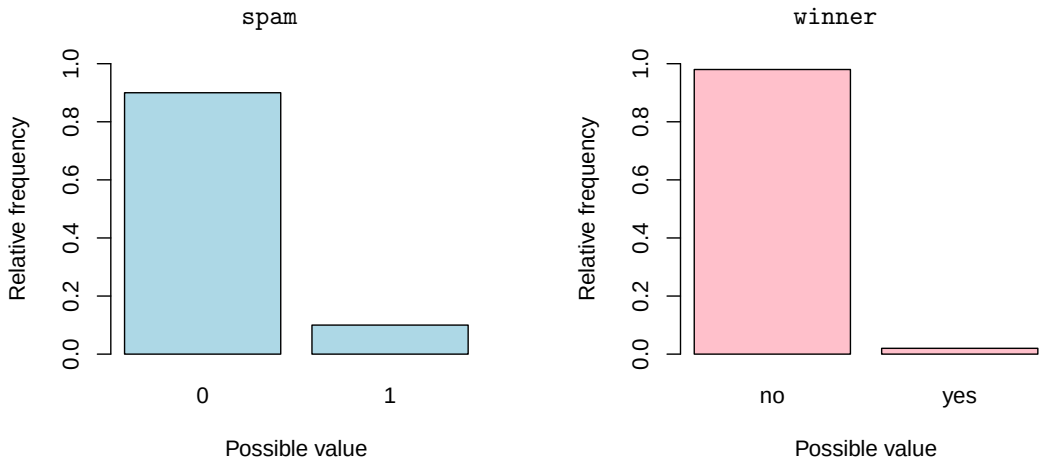


Figure 2.1: Bar plots for variables `spam` and `winner` from the *emails* data set

What about numerical variables? For discrete variables, the aggregate distribution of values in the data set can be similarly expressed with the relative frequency of each possible value. This information can in turn be visualized, for instance using *dot plots*. Figure 2.2 shows the dot plot for the variable `exclaim_mess`. Each dot represents an email, and the horizontal axis represents `exclaim_mess`, that is, the number of exclamation marks appearing in the email. For instance, we can see from the plot that, in our data set, there are 12 emails with no exclamation marks, 3 emails with 4 exclamation marks, and no email with 30 exclamation marks. From the plot, we may also notice a case with as many as 43 exclamation marks (corresponding to row 5 in Table 2.1)! Such cases whose values are exceptionally far from the rest of the data set are referred to as *outliers*, and may need our attention.

Figure 2.3 (left) shows the dot plot for the variable `line_breaks`. Note that hardly anything is visible in this plot. The reason is that the variable has a large range of values, from 0 to almost 1200, and as a result, cases with each specific value for the variable take up a tiny fraction of the cases, if at all. A remedy for this is to lower the resolution as in Figure 2.3 (right). In the latter figure, the possible values of the variable are placed into *bins* of width 50, and all the corresponding dots are piled up in one column. For instance, the leftmost column of dots in the plot corresponds to cases for which the variable has values between -25 and 24, and so on.

A much more flexible visualization can be obtained using a *histogram*. Figure 2.4 shows the histograms for the variables `line_breaks` (left) and `exclaim_mess` (right). For `line_breaks`, the range of possible values is divided into bins of size 50. The *area* of the rectangle on top of each bin represents the fraction of cases whose values fall within the bin. For example, the leftmost rectangle corresponds to the bin $[0, 50)$. It has height 0.0048 and area $50 \times 0.0048 = 0.24$. This indicates that in 24% of the cases in the data set, the value of `line_breaks` is between 0 and 49. For the histogram of `exclaim_mess`, the bins are chosen to be of width 1.

[3] Strictly speaking, the number of characters in thousands can only take values which are multiples of 0.001. However, given that the values of the variable `num_char` are here rounded to 2 decimal places, we can pretend that any (non-negative) value has been possible. A better example of a continuous numerical variable is the height of individuals in the *Hong Kong* data set discussed in Section 2.1.2 below.

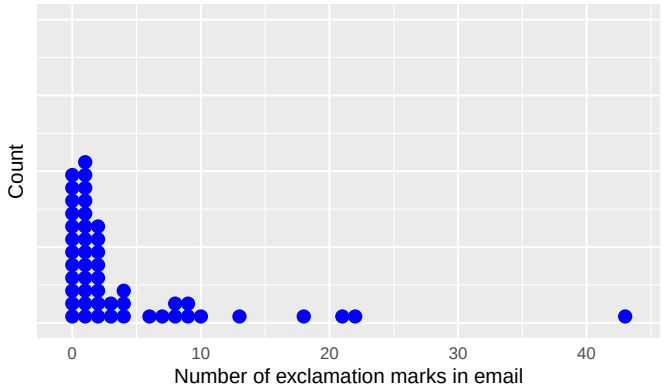
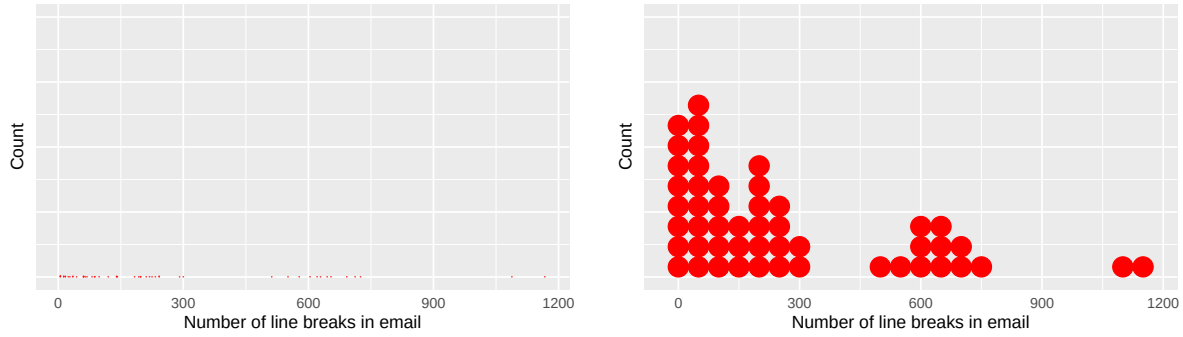


Figure 2.2: Dot plot for variable `exclaim_mess` from the *emails* data set



bin width $= 1$ bin width $= 50$

Figure 2.3: Dot plots for variable `line_breaks` from the *emails* data set

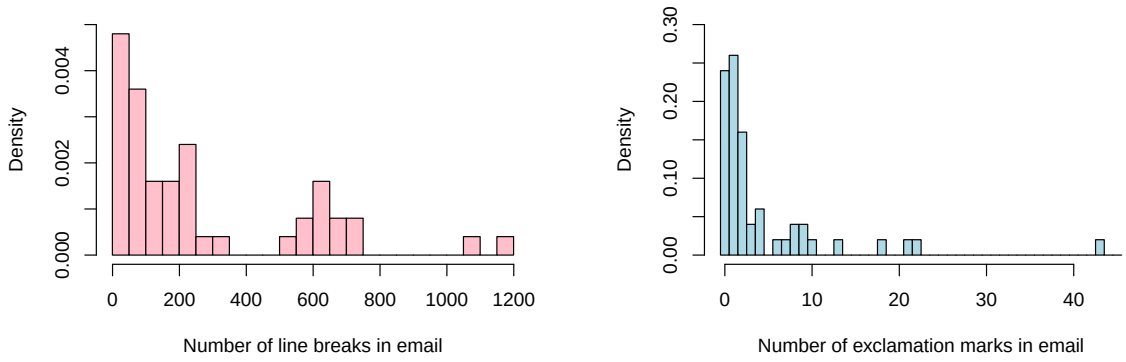


Figure 2.4: Histograms of variables `line_breaks` and `exclaim_mess` from the *emails* data set

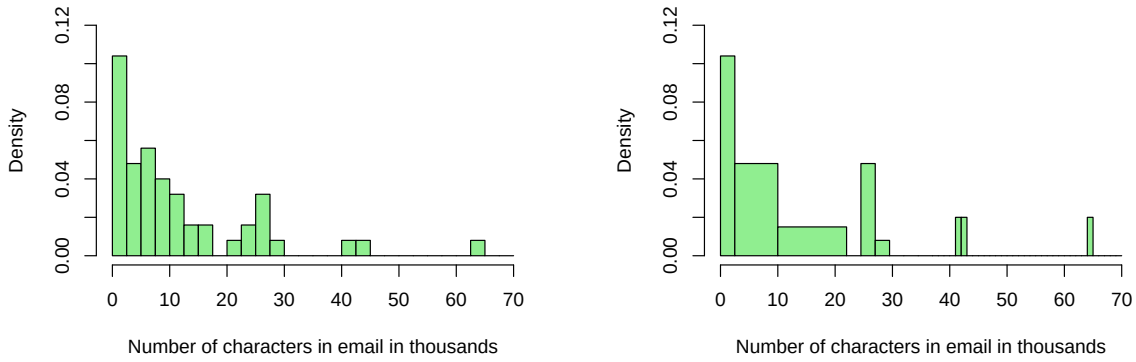
Ⓠ What is the total area of the rectangles in a histogram?

Ⓐ The total area is 1. The area of the rectangle above each bin indicates the proportion of cases that fall within that bin. Since each case falls within one and only one of the bins, the proportions must add up to 1.

One great advantage of histograms over dot plots is that histograms can be used to visualize the distribution of continuous variables as well as discrete variables. Figure 2.5 (left) depicts the histogram of the variable `num_char`.[4] Another flexibility of the histograms is that the bins do not need to have the same widths. This allows us to have different resolutions at different regions by adjusting the widths of the bins. For instance, in Figure 2.5 (right), we have placed all the values between 2.5 and 10 into a single bin (lower resolution), whereas we have chosen narrower bins for the values larger than 40 (higher resolution). The histogram still follows the same principle that the area of each rectangle indicates the fraction of cases with values in the corresponding bin.

Ⓠ Based on the histogram in Figure 2.5 (right), which of the following is larger?

[4]Recall that we are pretending that `num_char` (number of characters in thousands) is a continuous variable.



equal bin widths unequal bin widths

Figure 2.5: Histograms of variable `num_char` from the *emails* data set

– The proportion of cases for which the number of characters (in thousands) falls between 2.5 and 10.
– The proportion of cases for which the number of characters (in thousands) falls between 10 and 22.5.

A The first proportion is higher. The area of the rectangle above $[2.5, 10)$ is larger than the area of the rectangle above $[10, 22.5)$ (why?). This can also be seen from the histogram in Figure 2.5 (left), in which all bins have equal widths.

2.1.2 Typical histogram shapes

There is a remarkable statistical phenomenon that the histograms of variables from large data sets often have one of a handful of different qualitative shapes.

The file "`hongkong.csv`" contains the heights and weights of 25000 18-year-old individuals from Hong Kong.[5] Table 2.2 shows a few rows of this data set. Each row corresponds to one individual. Figure 2.6 shows his-

	height	weight
1	167.09	51.25
2	181.65	61.91
3	176.27	69.41
4	173.27	64.56
5	172.18	65.45
⋮	⋮	⋮
25000	174.95	56.64

`height` Height of the individual (in cm)
`weight` Weight of the individual (in Kg)

Table 2.2: Few rows of the *Hong Kong* data set

tograms of the two variables from this data set. It might come as a surprise that the histograms for the height

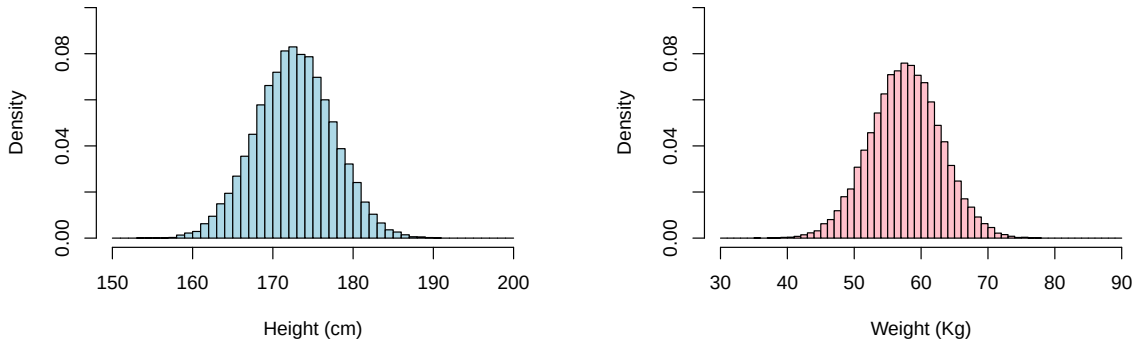


Figure 2.6: Histograms of variables `height` and `weight` from the *Hong Kong* data set

[5] This is a simulated data set. Source: Statistics Online Computational Resource (SOCR).

and the weight look so much alike: they both resemble a church bell. Similar bell-shaped histograms appear in many other completely different scenarios in statistics. In fact, bell-shaped histograms are so commonplace that their widespread appearance cannot be a coincidence. One of the principal links between statistics and probability theory is a mathematical theorem, known as the *central limit theorem*, which gives a partial explanation of the widespread occurrence of bell-shaped histograms. We will talk about the central limit theorem later in this course.

One characteristic feature of the histograms in Figure 2.6 is that each has a single peak. The histograms in the next example are different in that they have two peaks.

The file "`faithful.csv`" contains data gathered on the eruptions of the Old Faithful geyser in Yellowstone National Park in Wyoming, United States.[6] A few rows from this data set are shown in Table 2.3. Each row

	`eruptions`	`waiting`
1	3.60	79
2	1.80	54
3	3.33	74
4	2.28	62
5	4.53	85
⋮	⋮	⋮
272	4.47	74

`eruptions`	Eruption time (in mins)
`waiting`	Waiting time to next eruption (in mins)

Table 2.3: The first few rows of the *Old Faithful* data set

corresponds to one eruption of the geyser. The variable `eruptions` indicates the duration of the eruption, and the variable `waiting` indicates the waiting time till the next eruption, both measured in minutes. Figure 2.7 shows histograms of these two variables. Note that each of these two histograms has two distinct peaks. For

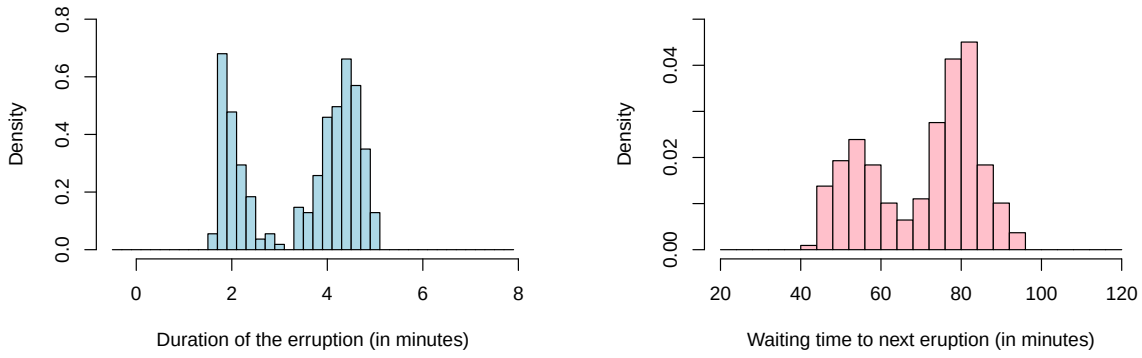


Figure 2.7: Histograms of variables `eruptions` and `waiting` from the *Old Faithful* data set

example, in the histogram for `eruptions`, the cases seemed to be roughly grouped around two peaks, one at 1.8 mins and the other at 4.4 mins. A histogram with two peaks is often called a *bi-modal* histogram. In contrast, the histograms in Figure 2.6 are *uni-modal*, meaning that each has a single peak.

Ⓠ Is the histogram of the number of line breaks in the *emails* data set (Figure 2.4, left) uni-modal, bi-modal, or neither?

Ⓐ Bi-modal. There are two distinct peaks, one at around 0 and one at around 600.

Note that, although the terms "uni-modal" and "bi-modal" have precise definitions, we often do not use them in the exact sense, but only to communicate the rough qualitative shape of the histograms.

Another characteristic feature of the histograms in Figure 2.6 is their *symmetry*: in each of the two histograms, the cases seem to be distributed around the center of the histogram in a more-or-less symmetric fashion. Figure 2.8 shows three idealized histogram shapes. While all three are uni-modal, there is a clear distinction between them: one is symmetric (the same bell-shaped curve), one is *skewed to the right* and one *skewed to the left*.

Ⓠ Is the histogram of the number of exclamation marks in the *emails* data set (Figure 2.4, right) left-skewed, symmetric or right-skewed?

[6] Source: The R Documentation.

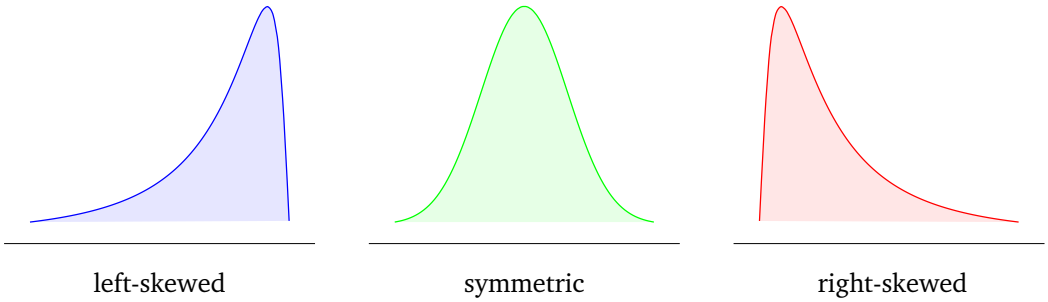


Figure 2.8: Three types of unimodal histograms

A Right-skewed.

Again, the term "skew" is often used in a rough qualitative sense rather than a precise mathematical sense.

2.2 Measures of location

In addition to visualizing the data, one often wishes to summarize the values in a data set with a few quantities. For instance, we may want to quantify what is the center of the histogram of a variable, and how widely or narrowly are the values spread around the center. We may also wish to quantify the co-dependence between two variables. Such quantities are referred to as *summary statistics*. It is remarkable that a few well-chosen summary statistics can carry a considerable amount of statistical information about the variables.

Q What single quantity is a good representative of the values of a numerical variable in a data set?

As an example, consider again the variable `height` in the *Hong Kong* data set, and its histogram in Figure 2.6 (left). All the values seem to be distributed symmetrically around ~ 173 cm, with highest concentration at the center and lower concentrations as we go farther from the center in either direction. Wouldn't ~ 173 cm be a good representative for `height`? What is the defining characteristic of this value?

A1 The *mean* (or *average*) of all the values of the variable in the data set is a good representative of the data values. For instance, from Table 2.2, the mean of the variable `height` is

$$\mathsf{mean}(\texttt{height}) = \frac{167.09 + 181.65 + \cdots + 174.95}{25000} = 172.7025 \,,$$

which is consistent with the intuitive center of the histogram. The mean of a numerical variable x is often denoted by $\mathsf{mean}(x)$ or $\overline{x}$. If x has values $x_1, x_2, \ldots, x_n$ in a data set, then

$$\mathsf{mean}(x) \coloneqq \overline{x} \coloneqq \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{k=1}^{n} x_k \,.$$

A2 Another good representative is the *median*. The median of a numerical variable is a value which divides its histogram into halves with equal areas: one half on the left, and the other half on the right. In other words, the median of a variable x is a value which is

- larger than or equal to the value of x in at least half the cases,
- less than or equal to the value of x in at least half the cases.

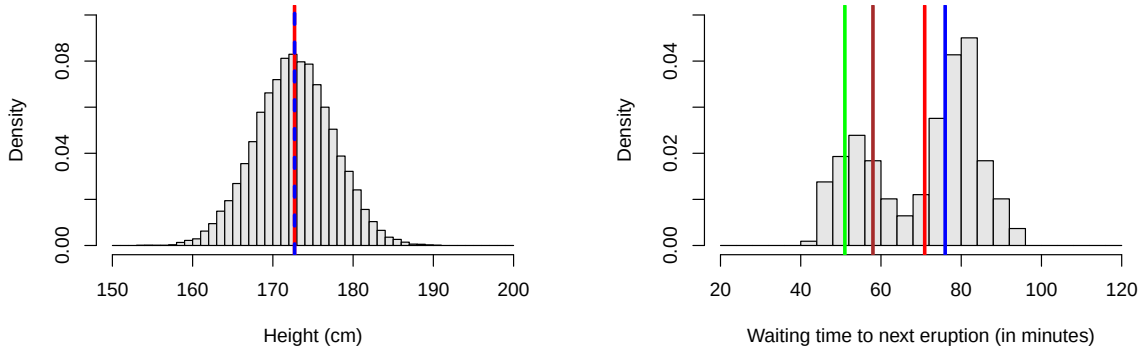
This definition is somewhat ambiguous. When the number of cases in the data set is odd, the median is simply the middle value once we sort all the values from smallest to largest. However, when the number of cases is even, any value between the two middle values is consistent with the above definition. In this case, the convention is to take the average of the two middle values as the median. The median of a variable x is often denoted by $\mathsf{median}(x)$ or $\widetilde{x}$.

From Table 2.2, the median of the variable `height` can be calculated to be

$$\mathsf{mean}(\texttt{height}) = 172.7091 \,,$$

which is again consistent with our intuitive notion of the center of the histogram.

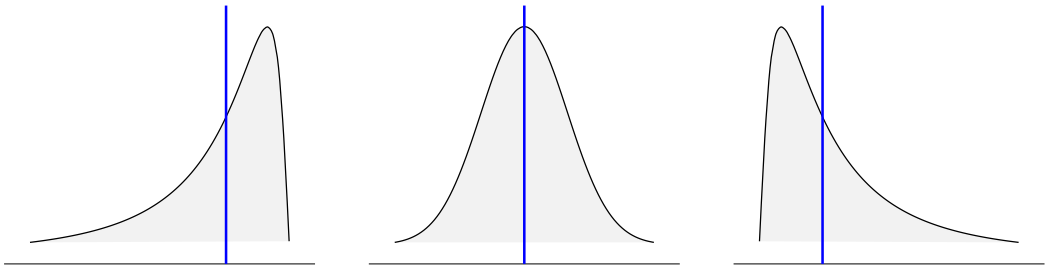
The mean and the median are both good representatives of the values of a variable in a data set. They both indicate the "center" of the histogram, but for different interpretations of the word "center". They carry somewhat different information about the distribution of the variable. In Figure 2.9, the mean and the median of the two variables `height` (from the *Hong Kong* data set) and `waiting` (from the *Old Faithful* data set) are indicated on the corresponding histograms. While in the symmetric example, the mean and the median almost coincide, the two quantities in the non-symmetric example are quite different.



Hong Kong *Old Faithful*

Figure 2.9: Mean and median of two variables `height` (from the *Hong Kong* data set) and `waiting` (from the *Old Faithful* data set). Mean is depicted in red and median in blue. For `waiting`, the 10th percentile (green) and the 25th percentile (brown) of the data values are also depicted.

As an exercise, consider the three idealized uni-modal histograms in Figure 2.10. The blue line shows the position of the median in each case.



left-skewed symmetric right-skewed

Figure 2.10: Three unimodal histograms. The blue line indicates the position of the median.

Q What is the area under each histogram to the left of the blue line?

A $1/2$. The proportion of the cases for which the the value of the variable is smaller than the median is $1/2$.

Q In each of the three histograms in Figure 2.10, can you guess the position of the mean relative to the median? Is the mean smaller than the median, equal to the median, or larger than the median?
Hint: Imagine the median is zero to make things simpler.

A The means are indicated in Figure 2.11.

In the symmetric case, the mean and the median are equal. Let us consider the the right-skewed histogram. For simplicity, imagine that the median is zero. This means that half the values in the data set are positive and half are negative. However, the negative values are concentrated around smaller negative values whereas the positive values are spread over a larger range of positive values. In other words, the positive values typically have larger magnitudes compared to the negative values. Therefore, the contribution of the positive values is larger, and as a result, the mean is positive, that is, larger than the median. (As a small example, you could consider a variable whose values in seven cases are $-1, -1, 0, 0, 0, 1, 2$. Draw the histogram with bins of width 1 around $-1, 0, 1, 2$ and note that it is skewed to the right. What are the mean and the median in this case?)

The case of left-skewed histogram can be understood similarly.

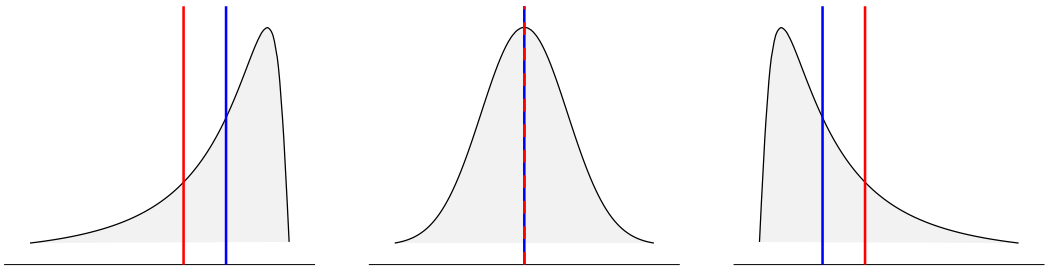


Figure 2.11: Three unimodal histograms. The blue line indicates the position of the median. The red lines indicate the position of the mean.

In Figure 2.9 (right), two additional lines are drawn. The green line indicates the 0.1-*quantile* ($= 10$*th percentile*) of the histogram, that is, the value which partitions the histogram into two unequal parts: 10% on the left, and 90% on the right. Similarly, the brown line indicates the 0.25-*quantile* ($= 25$*th percentile* $=$ *first quartile*). In general, when $0 < p < 1$, the p-*quantile* of a variable x in a data set refers to a value q for which[7]

- the proportion of cases for which $x \leq q$ is at least p,
- the proportion of cases for which $x \geq q$ is at least $1 - p$.

Observe that the median is the same as the 0.5-quantile ($= 50$th percentile $=$ second quartile).

2.3 Measures of spread

Figure 2.12 shows the histograms of the observed values for two variables, both of which are measured in the same units, say centimeters. The red lines indicate the position of the means. While the qualitative shape of the

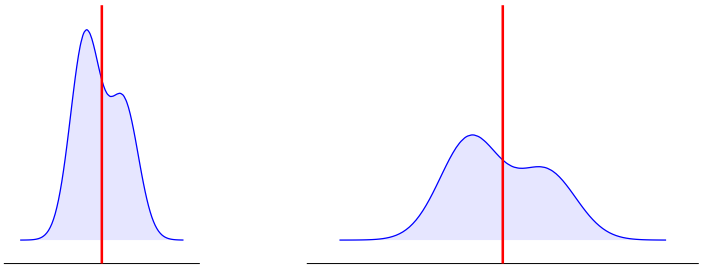


Figure 2.12: Two histograms with different spreads

two histograms are very much the same, there is major distinction between the two: in the left histogram, the values seem to be more concentrated around the mean, whereas in the right histogram, the values are spread farther from the mean. We would like to make this distinction quantitative.

Ⓠ What single quantity is a good measure of the spread (or dispersion) of the values of a numerical variable in a data set?

For the two examples in Figure 2.12, the measure of spread must be larger for the histogram on the right and smaller for the histogram on the left.

A1 In a more dispersed histogram, the values are typically farther from the center of the histogram. Hence, a reasonable measure of dispersion is the *average distance from the mean*. More specifically, suppose that a variable x has values $x_1, x_2, \ldots, x_n$ in a data set. The distance between the k'th value and the mean is simply $|x_k - \overline{x}|$, hence the average distance from the mean is

$$\langle \text{average distance from mean for } x \rangle := \frac{|x_1 - \overline{x}| + |x_2 - \overline{x}| + \cdots + |x_n - \overline{x}|}{n} = \frac{1}{n} \sum_{k=1}^{n} |x_k - \overline{x}| \ .$$

[7] As in the case of the median, this definition is ambiguous. A similar convention can be used to make the value unique.

Note that we cannot drop the absolute value from $|x_k - \overline{x}|$, for otherwise, some values would turn out negative and some positive, and the negative and positive values would cancel each other.

Exercise. Verify that

$$\langle \text{average } \underline{\text{difference}} \text{ from mean for } x \rangle \coloneqq \frac{(x_1 - \overline{x}) + (x_2 - \overline{x}) + \cdots + (x_n - \overline{x})}{n} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x}) = 0 \ ,$$

irrespective of the values of x.

A2 An alternative measure of dispersion is the average *squared* distance from the mean. Namely, if a variable x has values $x_1, x_2, \ldots, x_n$ in a data set, then

$$\langle \text{average } \underline{\text{squared}} \text{ distance from mean for } x \rangle = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n} = \frac{1}{n}\sum_{k=1}^{n}(x_k - \overline{x})^2 \ .$$

Note that since the square of a number is always non-negative, the problem of cancellation mentioned above does not appear here.

For a reason which we will learn about later in the course, one often uses a variant of the average squared distance from the mean, where instead of dividing the sum by n (the number of cases in the data set), we divide it by $n - 1$. This is what we call the *variance* of x

$$\mathsf{var}(x) \coloneqq s_x^2 \coloneqq \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n-1} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2 \ .$$

The variance of a numerical variable x is often denoted by $\mathsf{var}(x)$ or s_x^2. Note that if n is large, it does not make much difference whether we divide by n or by $n - 1$. However, when n is small (say 5 or 10), the two values will be significantly different.[8]

One drawback of the variance over the average distance from the mean is that the variance of a variable x has a different unit from x. For instance, if x refers to the height of individuals measured in cm, then the the variance of x is measured in cm^2. For a measure of spread, this is not desirable. As an example, if two variables x and y are connected via the relation $y = 2x$, then it is natural to think that the values of y are twice as dispersed as the values of x, but the variance of y is 4 times the variance of x. A simple remedy for this defect is to use the square root of the variance.

A3 The *standard deviation* of a variable x in a data set is the square root of the variance of x. It is often denoted by $\mathsf{sd}(x)$ or s_x. More specifically, if x has values $x_1, x_2, \ldots, x_n$ in the data set, then

$$\mathsf{sd}(x) \coloneqq s_x \coloneqq \sqrt{s_x^2} = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2} \ .$$

Compared to the variance, the standard deviation has the advantage that it is measured in the same units as the variable x itself.

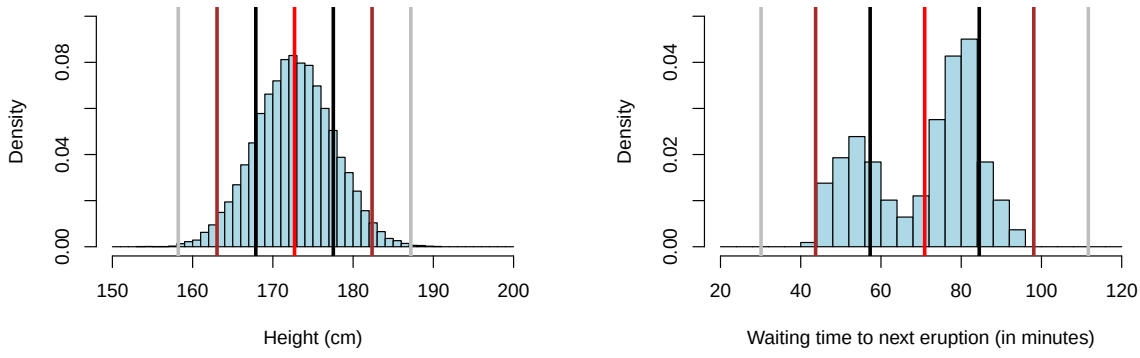
For mathematical reasons, it is more convenient to work with the standard deviation (or variance) than with the average distance from the mean. Standard deviation and variance are algebraically easier to work with (no absolute values!) and have more elegant mathematical properties.

The mean and the standard deviation together provide a remarkably efficient statistical summary of the the distribution of a variable. Consider again the histograms of the two variables `height` (from the *Hong Kong* data set) and `waiting` (from the *Old Faithful* data set) in Figure 2.13. In each histogram, the mean is indicated by a red line. The black lines indicate the mean $\pm$ 1 standard deviation. Observe that the range between the two black lines take up a considerable proportion of the cases. The brown lines indicate the mean $\pm$ 2 standard deviations. As you can observe, for the great majority of the cases, the values of the variable seem to fall between the two brown lines. The gray lines indicate the mean $\pm$ 3 standard deviations. This time, virtually all the cases fall within the interval between the two lines.

It turns out that for *any* numerical variable x in a data set,

- $\geq 75\%$ of the data values fall within 2 standard deviations from the mean, that is, within the interval $(\overline{x} - 2s_x, \overline{x} + 2s_x)$.

[8] Keep this curious modification in mind. Towards the end of the course, we will see a nice justification of why dividing by $n - 1$ is preferable to dividing by n.



Hong Kong *Old Faithful*

Figure 2.13: Concentration of the two variables `height` (from the *Hong Kong* data set) and `waiting` (from the *Old Faithful* data set) around their means. The mean is indicated by a red line. The black lines indicate mean $\pm$ sd. The brown lines indicate mean ± 2 sd. The gray lines indicate mean ± 3 sd.

- $\geq 88.9\%$ of the data values fall within 3 standard deviations from the mean, that is, within the interval $(\overline{x} - 3s_x, \overline{x} + 3s_x)$.

More generally, for any $k > 0$,

- a fraction $\geq 1 - 1/k^2$ of the data values fall within k standard deviations from the mean, that is, within the interval $(\overline{x} - ks_x, \overline{x} + ks_x)$.

This is *irrespective* of the shape of the histogram. The mathematical justification of this fact is given by the so-called *Chebyshev's inequality*, which we will talk about later in the course.

We conclude this section with yet another measure of dispersion. The first, second and third quartiles of a variable x from a data set are often denoted by $Q_1(x)$, $Q_2(x)$ and $Q_3(x)$ respectively.

A4 The *inter-quartile range* of a variable x from a data set is defines as

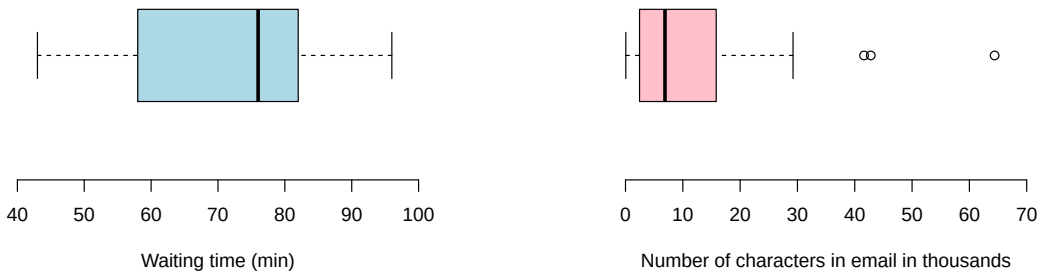
$$\mathsf{IQR}(x) := Q_3(x) - Q_1(x) \ .$$

Intuition on the relevance of the inter-quartile range will come in the next section.

2.4 Visualization of spread

2.4.1 Box plots

A *box plot* is often used to visualize the spread of the values of a variable in a data set (see Figure 2.14). Compared to a histogram, a box plot emphasizes on different features of the distribution and ignores the rest.



Old Faithful *emails*

Figure 2.14: Box plots of variables `waiting` (*Old Faithful* data set) and `num_char` (*emails* data set)

Figure 2.14 (left) shows the box plot of the variable `waiting` from the *Old Faithful* data set. The horizontal axis represents the possible values of the variable.

Ⓠ Can you guess what the thick line in the middle indicates?

Ⓐ The median.

Ⓠ Can you guess what the central box represents?

Ⓐ The left edge indicates the first quartile Q_1 and the right edge indicates the third quartile Q_3.

Ⓠ What percentage of the values of the variable fall within the box?

Ⓐ 50%. Note that 25% of the cases fall to the left of the left edge (i.e., are smaller than Q_1) and 25% of the cases fall to the right of the right edge (i.e., are large than Q_3). The remaining 50% fall within the box (i.e., between Q_1 and Q_3).

Ⓠ Can you guess what the two *whiskers* at the two ends of the plot indicate?

Ⓐ The minimum and the maximum of all the values.

Figure 2.14 (right) shows the box plot of the variable `num_char` from the *emails* data set. Compared to the previous box plot, the new one has an additional feature, namely the little circles.

Ⓠ Can you guess what the three little circles on the right stand for?

Ⓐ The *outliers*. These are the values which are exceptionally far from the rest of the values. Which values are to be judged "exceptionally far" is a matter of convention, which we will discuss in the next subsection. Note that in general, the outliers could appear on both sides. It is a coincidence that in this example all the three outliers are on the right.

Ⓠ If the small circles indicate exceptional values, the right whisker cannot possibly stand for the maximum value? What do the two whiskers represent then?

Ⓐ The minimum and the maximum once we ignore the outliers.

2.4.2 Outliers

Which values are exceptional enough to be considered as outliers is subjective, and a matter of choice. A convention suggested by John Tukey is often used in practice:[9]

- A value should be considered as an *outlier* if it is outside the range $[Q_1 - 1.5\,\mathsf{IQR}, Q_3 + 1.5\,\mathsf{IQR}]$. (In words, an outlier has distance more than $1.5\,\mathsf{IQR}$ from the main box in the box plot.)

Tukey further classified the outliers into mild and extreme ones:

- A value should be considered as an *extreme outlier* if it is outside the range $[Q_1 - 3\,\mathsf{IQR}, Q_3 + 3\,\mathsf{IQR}]$. (In words, an extreme outlier has distance more than $3\,\mathsf{IQR}$ from the main box in the box plot.)
- A non-extreme outlier is referred to as a *mild outlier*.

Ⓠ Where do outliers in a data set come from?

Ⓐ The cause of the occurrence of outliers could be different from case to case:

  - The exceptional values could be due to error in measurement or data collection (either due to a faulty measurement device or due to human error).
  - The exceptional values could be due to unknown factors which are not taken into account. For instance, there could be an underlying division of the data points into categories which we have overlooked, and the exceptional cases could belong to a category with characteristics which are distinct from the rest of the data points.
  - Exceptional values could occasionally appear due to chance. In fact, in a large data set (such as the *Hong Kong* data set), some values are bound to fall off the $1.5\,\mathsf{IQR}$ range (or even the $3\,\mathsf{IQR}$ range) because of sheer random variations. See Figure 2.15.

[9] In particular, Tukey's convention is used in the box plots of Figures 2.14 and 2.15.

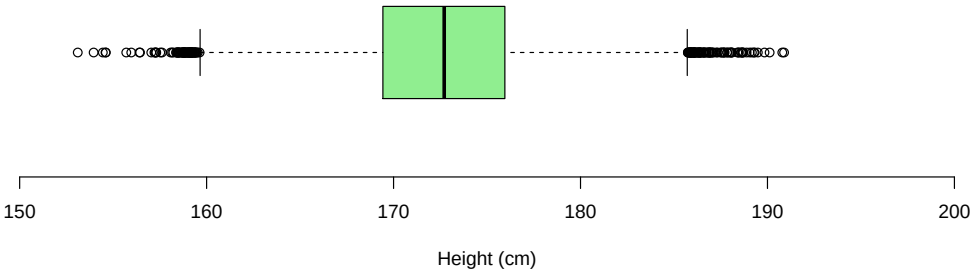


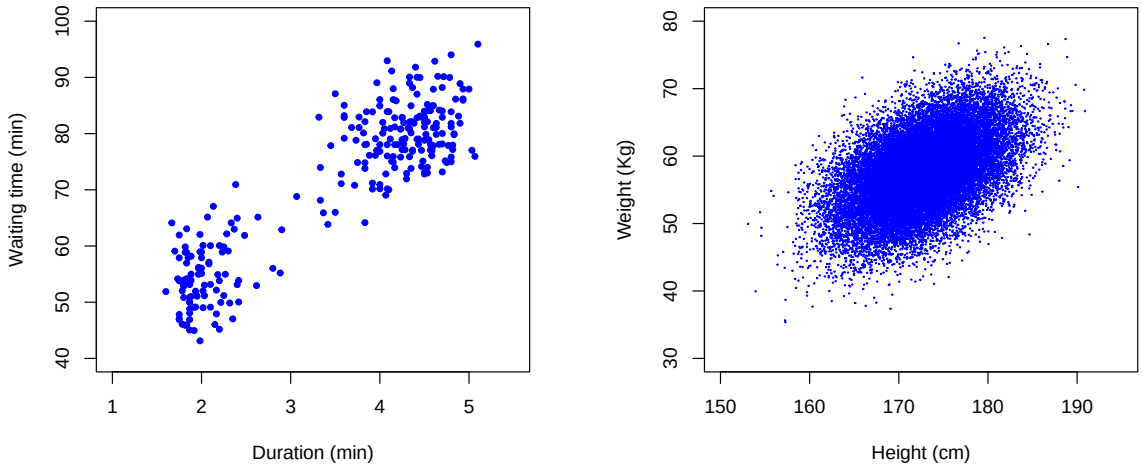
Figure 2.15: Box plots of variables `height` from the *Hong Kong* data set

2.5 Co-dependence between variables

Let us go back to the *Old Faithful* data set, and the histograms of `eruptions` and `waiting` in Figure 2.7. It is curious that both histograms are bi-modal.

Ⓠ Do the two peaks in the histogram of `eruptions` correspond to the two peaks in the histogram of `waiting`?

While histograms are quite convenient for visualizing the distribution of individual variables, they do not carry information on how different variables are related to one another. To visualize the relationship between two numerical variables, one can use a *scatter plot*. Figure 2.16 (left) shows the scatter plot of `waiting` against `eruptions`. Each dot indicates an eruption recorded in the data set; the horizontal coordinate of the dot is the duration of the eruption and the vertical coordinate of the dot is the waiting time to the next eruption for that eruption.



Old Faithful *Hong Kong*

Figure 2.16: Scatter plots of `waiting against erruptions` (*Old Faithful* data set) and `weight against height` (*Hong Kong* data set)

From the scatter plot, one can see that the eruptions recorded in the data set are roughly divided into two clusters. In one cluster, we have short eruptions with short waiting times and in the other cluster, we have long eruptions with long waiting times. The first cluster corresponds to the left peaks in the histograms of Figure 2.7, while the second cluster corresponds to the right peaks in the two histograms.

As another example, Figure 2.16 (right) depicts the scatter plot of `weight` against `height` in the *Hong Kong* data set. Again, each dot represents one case, that is, an 18-year-old individual. The horizontal and vertical coordinates of the dot are, respectively, the height and the weight of the individual. This scatter plot seems to suggest that there is a direct association between the height and the weight of the individuals surveyed in this data set: the individuals who are shorter tend to be lighter, and those who are taller tend to be heavier. However, this is only a *statistical* association, as there are random variations from case to case.