In the previous lectures, we were concerned with phenomena for which we were given accurate probability distributions to model them. In this lecture, we switch gears, and imagine we want to study a new phenomenon, for which we have little knowledge. An intuitive strategy is to run a large data set of observations to help to characterize the new phenomenon. The question then is: how to extract useful information from the resulting data?

We will soon consider methods based on statistical modeling and hypothesis testing. First, however, we will explore elementary methods, often visual, for data analysis. These methods fall under the category of *exploratory data analysis* (EDA).

1 Graphical methods

1.1 Sorting raw data

Consider the following results from a midterm exam worth 70 points, which 35 students took:

$$67, 47, 59, 62, 24, 66, 45, 63, 57, 10, 39, 58, 35, 23, 45, 49, 50, 51, 38, 57, 36, 40, 55, 41, 51, 53, 43, 28, 53, 19, 53, 33, 39, 50, 52$$

It is not easy to determine from this dataset how hard the test was, what proportion of the class is comfortable with the material, whether the class has a homogeneous level or not, etc. A first basic step is to sort the data in ascending order:

$$10, 19, 23, 24, 28, 33, 35, 36, 38, 39, 39, 40, 41, 43, 45, 45, 47, 49, 50, 50, 51, 51, 52, 53, 53, 53, 55, 57, 57, 58, 59, 62, 63, 66, 67$$

From this data, we can see that only 6 students got a score lower than half the total number of points, indicating that most of the students master most of the material covered in the midterm.

We can also see that the middle element, i.e. the 18^{th} score, is 49, which is significantly closer to 70 than to 0, suggesting a skewness of the data set toward the right, reinforcing the previous conclusion.

However, not much more information can be extracted by just staring at the data. To go further, it is helpful to visualize the data graphically. There are a few ways to do so.

1.2 Histograms

A well established way to represent data sets graphically is through **histograms**. The idea of the histogram is quite intuitive:

- One subdivides the range of the data (0-70 in our example) into intervals called bins, B_1, B_2,...,B_m.
- For each bin, one draws a rectangle whose height is given by the number of data points in the bin.

As an illustration, for the data set discussed above, the histogram we obtain if we choose bins of width 4 is found in Figure 1.

We see that the histogram is revealing: as a first approximation, it appears that the scores are distributed in a bell-like curve around the score 48. Upon closer look, we also see thet hint for the presence of two "modes": one group of students in the neightborhood of 36-40, and another group of students in the neighborhood of 52.

To get more precise information, we can reduce the size of the bins. Choosing bin size is part art, part science: too large bins lead to the loss of fine scale information, while too small bins can lead to too detailed data, from which it is complicated to extract insightful information. Figure 2, which shows the histograms for different bin sizes for the data set we have discussed thus far, illustrates that point. Our textbook provides some quantitative guidelines on how to determine optimal bin sizes. You may be interested in reading them. Still, it is important to understand that there does not exist a silver bullet. Different data distributions require different treatments, so experimentation is eventually always needed. Fortunately, numerical softwares for histograms have a variety of different options.

Note that although it is a common choice, bin sizes do not have to be equal throughout the range of the data.

Observe also that some people like to scale the histogram in such a way that the total area of all rectangles is 1, so that the histogram can be viewed as an approximation of a probability density function f.

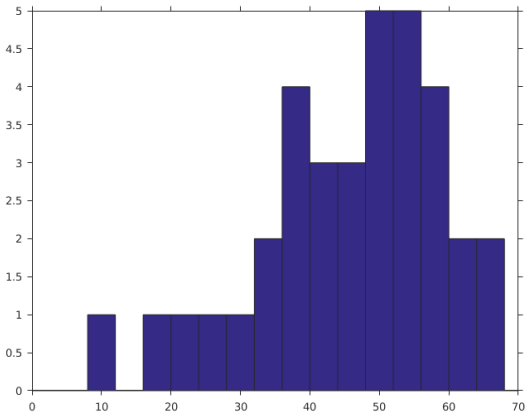

Figure 1: Histogram with bins of width 4 for the data set in Section 1.1.

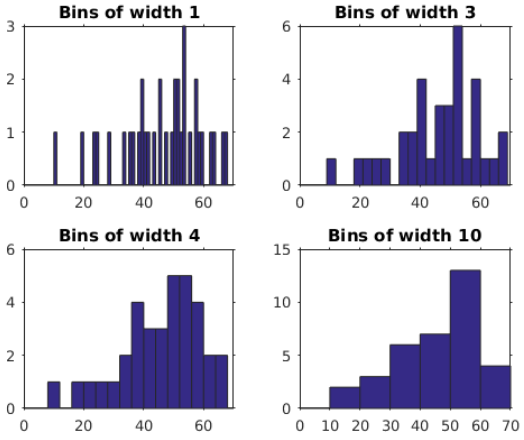

Figure 2: Histograms corresponding to 4 different bin widths for the data set in Section 1.1: bin width of 1 (top left), bin width of 3 (top right), bin width of 4 (bottom left), bin width of 10 (bottom right).

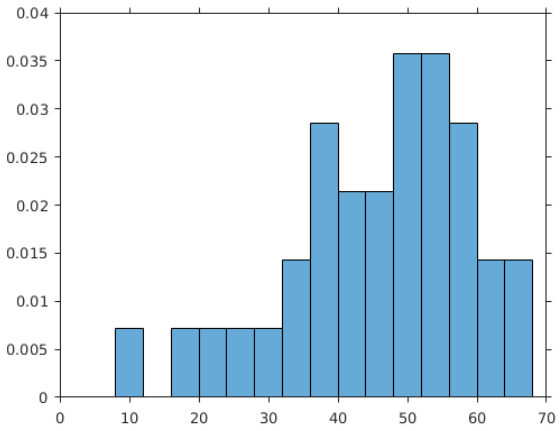

Figure 3: Histogram with bins of width 4 for the data set in Section 1.1., normalized in such a way that the total area of all rectangles is 1

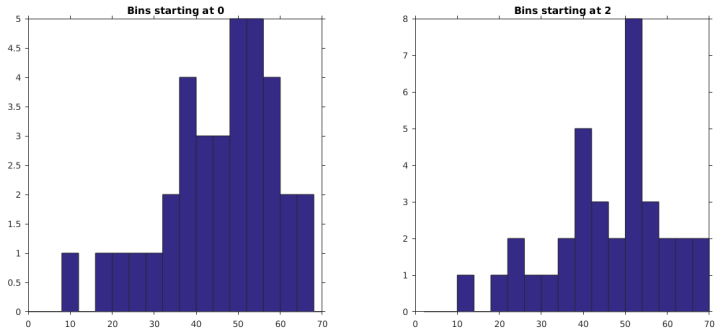

Figure 4: Histograms with bins of width 4 for the data set in Section 1.1.. For the histogram on the left, the first bin is $[0, 4]$, and for the histogram on the right, the first bin is $[2, 6]$.

To achieve this, the area of the rectangle for bin B_j must be equal to the proportion of data points in that bin:

$$\text{Area}(B_j) = \frac{\#\text{ of data points in } B_j}{\#\text{ of data points in overall}}$$

The height of that rectangle then is

$$\text{Height}(B_j) = \frac{\#\text{ of data points in } B_j}{(\#\text{ of data points in overall}) \cdot (\text{width of } B_j)}$$

A normalized version of Figure 1, constructed following that rule, is shown in Figure 3. Beyond the 'theoretical' connection with probability density functions, the advantage of normalizing histograms in this manner is that it facilitates comparison between different data sets, which may have different sample sizes, so that the unnormalized histograms may have very different heights.

Histograms are simple and intuitive, and therefore frequently used. However, they are discrete objects by nature, which leads to limitations in the method. Specifically, the key features of a given histogram can change significantly because of modest changes in the bin size or bin location. We show that in Figure 4, in which we compare two histograms for the data set corresponding to the grades on the midterm exam. Both histograms use bins with width 4, but for the first histogram, on the left, already showed in Figure 1, the first bin is $[0, 4]$, while for the second histogram, on the right, the first bin is $[2, 6]$.

This limitation is the motivation for using an alternative, *continuous* approach, called **kernel density estimation**.

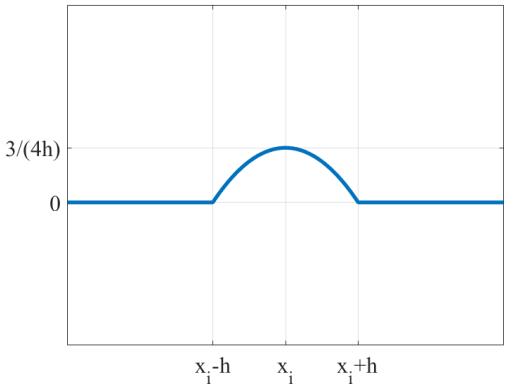

Figure 5: Local parabola as defined in Eq.(1)

1.3 Kernel density estimation

The idea of kernel density estimation is to associate a local, symmetric, compactly supported probability density function K_i to each data point x_i, and to approximate the probability density function f of the entire data set by summing the K_i (and dividing by n, the number of data points, to make sure the area under f is 1).

Specifically, for data point x_i, consider the local parabola

$$K_i(x) = \begin{cases} \frac{3}{4h}\left[1 - \left(\frac{x-x_i}{h}\right)^2\right] & \text{if } -h \leq x - x_i \leq 0 \\ 0 \quad \text{otherwise} \end{cases} \tag{1}$$

This local parabola is plotted in Figure 5. For a given data set, we obtain the entire approximation $f_{n,h}(x)$ by the formula

$$f_{n,h}(x) = \frac{1}{n}\sum_{i=1}^{n} K_i(x)$$

The approximation $f_{35,h}$ of the data set corresponding to the grades on the midterm exam we have considered in the previous sections is shown in Figure 6. Specifically, we produced 4 different approximations $f_{35,h}$, for 4 different values of h, and therefore 4 different values of the support of the local parabola we introduced in Eq.(5). The analogy often made for this approach is that one puts a "pile of sand" around each data point. The sand will then pile up where the data points are bunched up.

We observe that the kernel density estimation addresses two limitations of histograms:

- The fact that histograms are not smooth
- The strong dependence of the histograms on the location of the end points of the bins

However, the method has the same issue as histograms regarding the size of bins. In the context of kernel density estimation, h plays the role of the size of bins, and is often called the bandwidth. A small bandwidth leads to an $f_{n,h}$ with many features, from which insight may be hard to extract. A large bandwidth leads to a smooth $f_{n,h}$, which may not be accurate enough to highlight key features. This can be clearly seen by comparing the different panels in Figure 6. Here too, the textbook also provides some quantitative estimates for how to best choose h.

Note that there also is freedom in how one may choose the kernels K_i. In the previous page, we explicitly treated the case in which K_i is a section of a parabola, but K_i could also be a hat function, as pictured in Figure 7. Many other similar possibilities are discussed in the textbook. Fortunately, as the textbook shows, the kernel density estimate does not often sensitively depend on the choice of the particular form of K_i.

1.4 The empirical cumulative distribution function

Another intuitive way to represent a data set and extract valuable information is to construct the **empirical cumulative distribution function** F_n of the data.

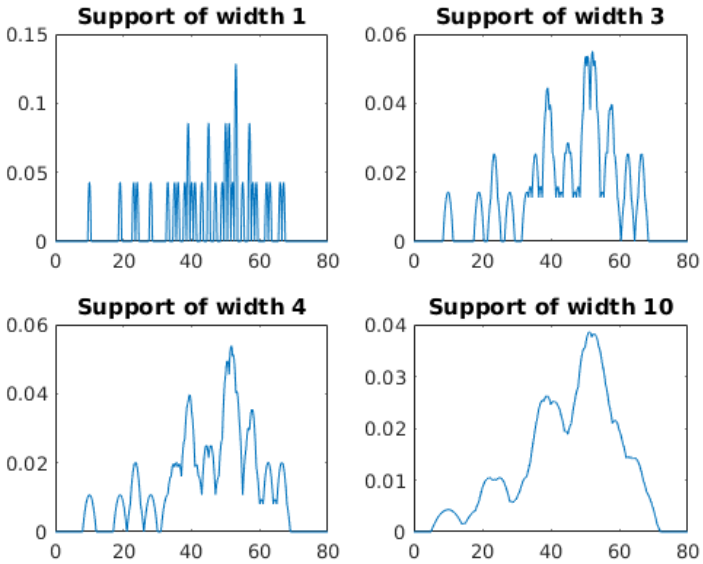

Figure 6: Kernel density estimation $f_{35,h}$ using local parabolas for the data set in Section 1.1.. Each panel corresponds to a kernel density estimation with a different value of the parameter h in Eq.(1): $h = \frac{1}{2}$ (top left), $h = \frac{3}{2}$ (top right), $h = 2$ (bottom left), $h = 5$ (bottom right).

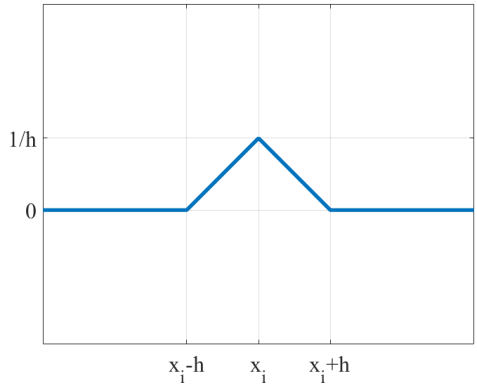

Figure 7: Local hat function, which could also be used as an appropriate kernel for kernel density estimation.

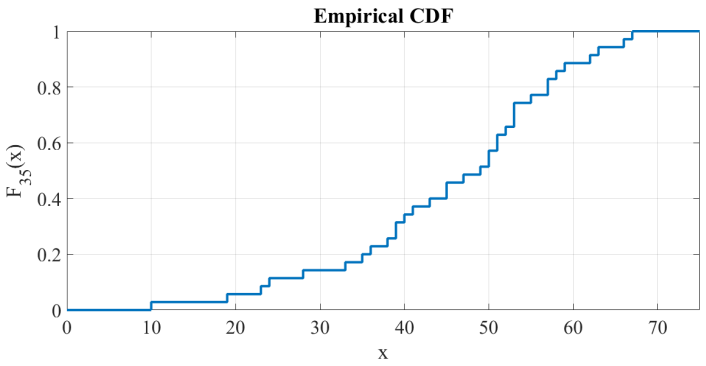

Figure 8: Empirical cumulative distribution function for the data set in Section 1.1..

It is defined in a way that is reminiscent of the cumulative distribution function F_X of a random variable X:

$$\boxed{F_n(x) = \frac{\#\text{ of data points} \leq x}{n}}$$

Since the data points are discrete, F_n is discontinuous, and looks like a staircase. The staircase is such that:

- $F_n(x) = 0$ for all x smaller than the smallest data point.
- $F_n(x) = 1$ for all x larger than the largest data point.
- The more bunched up the data points are, the steeper the staircase.

The empirical cumulative distribution F_{35} for the data set of the grades for the midterm taken as example all along this lecture is shown in Figure 8.

1.5 Scatterplot

Thus far, we have focused on data sets corresponding to measurements or observations of one particular quantity. As we have often seen in this course, however, we also sometimes want to investigate the relationship between two or more quantities.

The most basic idea to visualize a possible relationship is to plot all points corresponding to measurements of pairs (x_i, y_i) of observations in the Cartesian plane. This is called a **scatterplot** (because the data is likely to be scattered in a somewhat disorganized way).

As an example, imagine an electrical engineering student measuring in her lab the electric power P dissipated in an ohmic resistor as a function of the current I running through the resistor. Her measurements are as shown in Figure 9.

From this scatterplot, the student can conclude that the higher the current through the resistor, the higher the dissipated power seems to be. It also seems like a quadratic relationship (parabola) could make sense, where P would be proportional to I^2.

We will discuss methods for constructing such a quadratic fit later in this course.

2 Numerical summaries

In past lectures, we learned several mathematical objects used to characterize a random variable, from quite coarse, such as the expected value, to very detailed, such as the cumulative distribution function.

Many of these concepts have an empirical counterpart for data sets. The empirical cumulative distribution function is an example we just encountered. We will now discuss a few more.

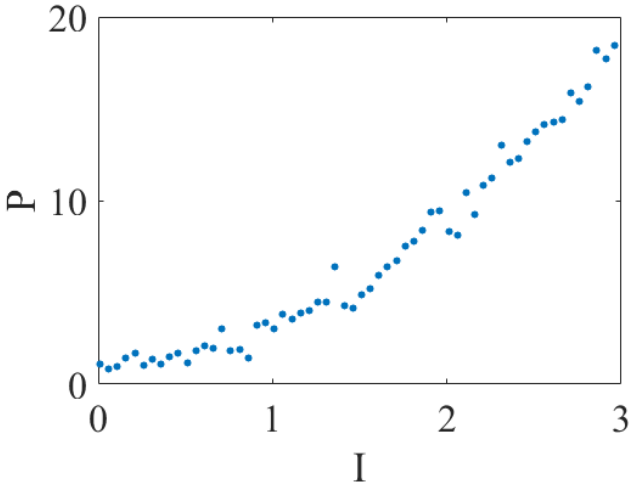

Figure 9: Measurements of the electric power P in an ohmic resistor for different values of the electric current I running through the resistor.

2.1 Center of a data set

2.1.1 Sample mean

An intuitive method to identify the center of a data set is to decide that it is well approximated by the **sample mean**, defined as usual:

$$\boxed{\overline{x_n} = \frac{x_1 + x_2 + \ldots + x_n}{n}}$$

As an example, the sample mean for the midterm exam results discussed throughout this lecture is $\overline{x_n} = 45.46$, which indeed is a fairly good approximation of the center of our data set.

2.1.2 Sample median

An alternative way to identify the center of the data set is to compute the **sample median**, written $\mathbf{Med}_n$, which is the element in a data set organized in ascending order which is such that half of the elements are smaller than it, and half of the elements are greater than it.

If there is an even number of elements in the data set, Med_n is the average of the two middle elements.

As an illustration, in our midterm data set, the 18^{th} largest grade is 49, so the median is 49.

2.1.3 Sample mean vs sample median

The sample mean is often favored because it is the analog of the expected value for a random variable. However, the sample mean tends to be more sensitive to outliers than the sample median, which can be a weakness.

2.2 Variability of a data set

As was the case for random variables, one often wants to go beyond simply identifying the center of a data set. The next natural step is to quantify the variability of the data set. We will see two common approaches below.

2.2.1 Sample variance and standard deviation

The **sample variance** if a data set is defined in a way that is reminiscent of the variance of a random variable:

$$\boxed{s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x_n})^2 \qquad \text{Sample variance}}$$

The careful reader may at first think that the $\frac{1}{n-1}$ prefactor is incorrect, and should be replaced by $\frac{1}{n}$. $\frac{1}{n-1}$ is in fact a better suited prefactor, as we will explain in a future lecture.

The **sample standard deviation** s_n has the same units as the data itself, and is given by

$$\boxed{s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x_n})^2} \qquad \text{Sample standard deviation}}$$

For the data set corresponding to the midterm exam, we find $s_n^2 \approx 184.49$, and $s_n \approx 13.58$

The sample variance is unfortunately sensitive to outliers in much the same way the sample mean is. To address this limitation, an alternative measure of variability is sometimes used, called the **median of absolute deviations** (MAD).

2.2.2 Median of absolute deviations (MAD)

The MAD is constructed as follows.

Consider the set

$$\{|x_1 - \text{Med}_n|, |x_2 - \text{Med}_n|, \ldots, |x_n - \text{Med}_n|\}$$

of absolute deviation of every element with respect to the sample median. Order this set in ascending order, and find the median of the resulting set. This is the MAD:

$$\text{MAD}(x_1, x_2, \ldots, x_n) = \text{Med}(|x_1 - \text{Med}_n|, |x_2 - \text{Med}_n|, \ldots, |x_n - \text{Med}_n|)$$

For the data set of the midterm exam, we find MAD = 9.

2.3 Empirical quantiles and Inter-Quantile Range (IQR)

A more refined understanding of a data set, beyond just the center and variability, but short of the full empirical cumulative distribution function, is provided by the calculation of **empirical quantiles**, constructed as follows.

Imagine the data set $\{x_1, x_2, \ldots, x_n\}$ has been sorted in ascending order:

$$\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\} \qquad \text{with } x_{(1)} < x_{(2)} < \ldots < x_{(n)}$$

The $x_{(i)}$ are called **order statistics** of our data set.

The quantile q_p of a data set is the number q_p such that a proportion p of the values are less than or equal to q_p. q_p is obtained by **taking the rank** $k = (n+1)p^{\text{th}}$ **order statistic**: $x_{(n+1)p}$.

Of course, for an arbitrarily chosen $p \in [0, 1]$, $(n+1)p$ is *not* an integer, so we cannot immediately pick an order statistic for q_p. In such cases, we obtain q_p by *linear interpolation* of the order statistics $x_{(k)}$ and $x_{(k+1)}$, with k such that $k < (n+1)p < k+1$. The formula is:

$$\boxed{q_p = x_{(k)} + (p(n+1) - k)(x_{(k+1)} - x_{(k)})}$$

Note that q_p can be read from the empirical cumulative distribution function by looking at the x-value corresponding to the $y - value\ p$ of the distribution function.

$q_{0.25}$ is called the **lower quartile**, and $q_{0.75}$ is called the **upper quartile**. $q_{0.5}$ is clearly the median.

A good summary of a data set is given by the following 5 numbers:

$$\min\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}, q_{0.25}, q_{0.5}, q_{0.75}, \max\{x_{(1)}, x_{(2)}, \ldots, x_{(n)}\}$$

From this summary, we can in particular compute

$$\frac{q_{0.5} - q_{0.25}}{q_{0.75} - q_{0.5}}$$

which gives an idea of the *skewness* of the data. We can also compute the **interquartile range (IQR)**, defined by

$$\boxed{IQR = q_{0.75} - q_{0.25}}$$

The IQR is a measure of the variability of the data set.

As an example, for the midterm exam, we have:

$$q_{0.25} = 38 \quad , \quad q_{0.5} = 49 \quad , \quad q_{0.75} = 55 \quad , \quad \frac{q_{0.5} - q_{0.25}}{q_{0.75} - q_{0.5}} = \frac{49 - 38}{55 - 49} = \frac{11}{6}$$

which indicates that the data distribution is skewed toward the right, which we had already graphically observed with the histograms and kernal density estimations.

The IQR is

$$55 - 38 = 17$$

2.4 The box-and-whisker plot

We have just seen that the minimum of the data set, its maximum, its median, and its lower and upper quartiles provide a valuable five-number summary of the data set. This summary is often visualized graphically, with a **boxplot**.

A boxplot is constructed as follows. It is a box of arbitrary width, and whose **height is the IQR**. The **base** of the box is located at the **lower quartile**, so the **top of the box is at the upper quartile**. Inside the box, one draws a **horizontal line at the median of the data set**.

In addition, one measures vertical distances of length $1.5IQR$ above the upper quartile, and $1.5IQR$ below the lower quartile. One draws a horizontal line of length the width of the box at the highest data point still within $1.5IQR$ above the upper quartile, and a horizontal line of length the width of the box at the lowest data point still within $1.5IQR$ below the lower quartile.

One often draws dashed vertical lines between the upper quartile and the last data point with $1.5IQR$ above it, and between the lower quartile and the last data point within $1.5IQR$ below it. These vertical lines are called **whiskers**.

Finally, data points beyond the whisker lines are viewed as outliers, and plotted individually.

Two typical boxplots are shown in Figure 10 to illustrate this construction.

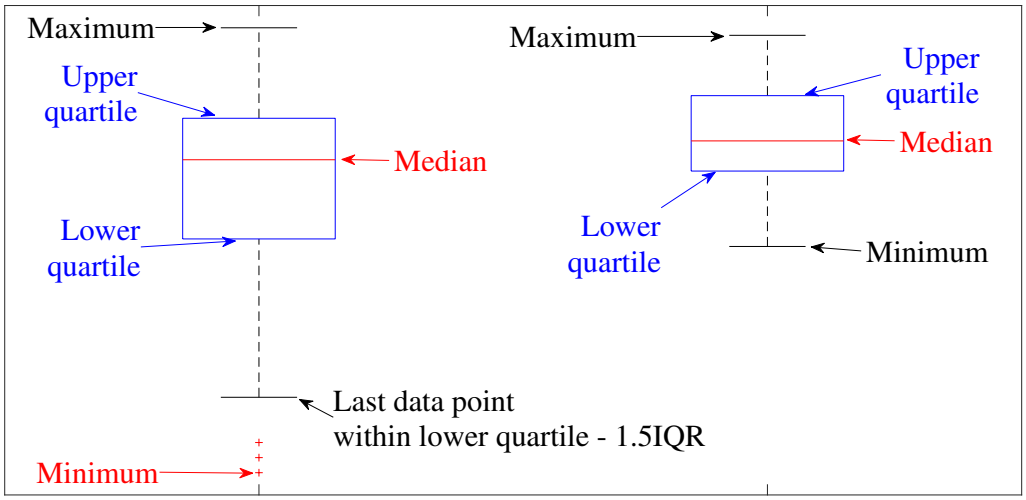

Figure 10: Boxplots for two unspecified data sets, to illustrate the construction of box plots, and how to read this information they contain.

Boxplots provide less information than histograms or kernel density estimates, for a given data set. However, they provide a convenient way to compare two or more data sets, as shown in Figure 10 above.