

MAT 2377 3X (Spring 2010)

Section 6-1 : Numerical Summaries

With a random experiment comes data. In these notes, we learn techniques to describe the data.

Data : We will denote the n observations of the random variable X as x_1, x_2, \dots, x_n .

Example 1 : The following $n = 8$ measurements (in millimeters) were made on the inside diameter of forged piston rings used in an automobile engine.

74.001, 74.003, 74.015, 74.000, 74.005, 74.002, 74.005, 74.004.

Hence, $x_1 = 74.001, x_2 = 74.003, \dots, x_8 = 74.004$.

Note : We will describe the central tendencies or the location of the distribution of these measurements with the sample mean and the sample median.

The sample mean is defined as the average of the n values and it is denoted \bar{x} . In our example, we get

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{74.001 + 74.003 + \dots + 74.004}{8} = \frac{592.035}{8} = 74.004375.$$

Percentile : A percentile is a value among 99 values that divide the sample into 100 (approximately) equal parts in the sense that about 1% of the values are in the class.

To compute the k th percentile : We order the values in ascending order :

$$y_1 \leq y_2 \leq \dots \leq y_n.$$

In our example, we get :

74, 74.001, 74.002, 74.003, 74.004, 74.005, 74.005, 74.015.

We consider the n values in the sample as splitting the number line into $n + 1$ parts. The position of the k th percentile is $(n + 1) \times k/100 = m + p$, where m is the whole part and p is the fractional part, that is $0 \leq p < 1$.

The k th percentile is

$$\begin{cases} y_m, & \text{if } p = 0 \\ y_m + p(y_{m+1} - y_m), & \text{if } 0 < p < 1. \end{cases}$$

The median is the 50th percentile. If $n = 8$, the position of the median is $(n + 1) \times 50/100 = 4.5$. Thus, the median is

$$\text{median} = y_4 + 0.5(y_5 - y_4) = 74.003 + 0.5(74.004 - 74.003) = 74.0035.$$

Mesures of Dispersion (or Variability) : We will use the variance, the standard deviation, the range and the inter-quartile range to describe the dispersion of the distribution.

The **sample variance** is

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} = \frac{(\sum_{i=1}^n x_i^2) - (\sum_{i=1}^n x_i)^2/n}{n - 1} \\ &= \frac{(74.001^2 + 74.003^2 + \dots + 74.004^2) - (74.001 + 74.003 + \dots + 74.004)^2/8}{8 - 1} \\ &= 2.1696 \times 10^{-5} \end{aligned}$$

and the **sample standard deviation** is $s = \sqrt{s^2} = 0.004658$.

The **sample range** is

$$r = \max(x_i) - \min(x_i) = y_n - y_1 = 74.015 - 74 = 0.015.$$

Quartile : The first quartile Q_1 , the second quartile Q_2 and the third quartile Q_3 , are respectively the 25th, the 50th and the 75th percentile.

For our piston rings example of $n = 8$ observations, the position of Q_1 is $(n + 1) \times 25/100 = 25\%(9) = 2.25$. Thus,

$$Q_1 = y_2 + .25(y_3 - y_2) = 74.001 + 0.25(74.002 - 74.001) = 74.00125.$$

The position of Q_3 is $75\%(n + 1) = 6.75$. Alors

$$Q_3 = y_6 + .75(y_7 - y_6) = 74.005 + 0.75(74.005 - 74.005) = 74.005.$$

The interquartile distance is

$$\text{IQR} = Q_3 - Q_1 = 74.005 - 74.00125 = 0.00375.$$

Remark : The IQR represents the 50% middle. It it becomes larger then the values become more dispersed and vice-versa. Hence it can be used as a measure of dispersion.

Output from R commander : Using **statistics -> summaries -> numerical summaries** in R commander, we obtain the following output.

| mean | sd | 0% | 25% | 50% | 75% | 100% | n |
|----------|-------------|----|----------|---------|--------|--------|---|
| 74.00437 | 0.004657943 | 74 | 74.00175 | 74.0035 | 74.005 | 74.015 | 8 |

Remarks :

- There are no an accepted way to obtain sample percentiles. Different software will give different values. But they all can be interpreted in the same way. Here $Q_1 = 74.00175$ but we computed $Q_1 = 74.00125$. In both cases, about 25% of the sampled values are at most equal to Q_1 . In this example, there are $2/8=25\%$ of the values that are at most either of these values. For samples that are not too small, the values will usually be similar.
- From the R output we can compute :

$$\text{range} = 74.015 - 74 = 0.015 \quad \text{and} \quad \text{IQR} = 74.005 - 74.00175 = 0.00325.$$

Histograms - Section 6-3

A histogram is a visual tool that can be used to describe the **shape** of the distribution of a numerical variable. We can either build a **frequency**, **relative frequency** or **probability density** histogram.

Construction of a histogram :

1. Dived the horizontal axis into sub-intervals (preferably of equal length). Each sub-interval represents a range values for the random variable. It is often suggested to use between 5 to 20 classes. Often # of classes = \sqrt{n} works well.
2. Different statistical packages use different techniques to determine the number of subintervals. However often the default works well.
3. **Terminology** : Often a subinterval is called a **bin**.
4. For each bin, erect a rectangle whose height is equal to either the frequency, the relative frequency or the density.
5. If you use the density, that is

$$\text{density} = \text{relative frequency}/\text{length of bin},$$

then the area of the bin, which is density \times length of bin, is equal to the relative frequency (i.e. probability).

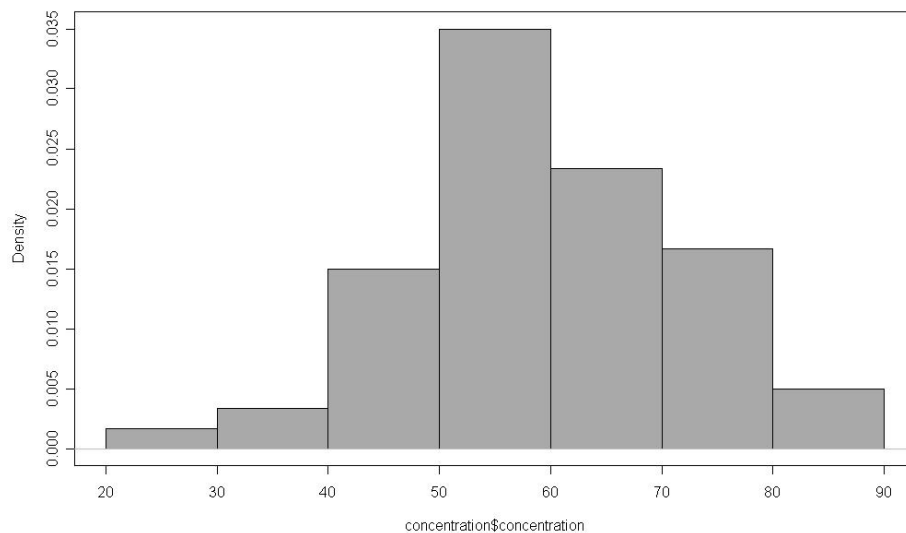
Example 2 : Consider the 60 observations in the file *concentration.txt*. We start by ordering them in ascending order :

29.8, 31.6, 39.3, ..., 81.3, 89.8.

We will construct a probability density histogram. We will use 7 bins of length 10.

| | | | | |
|-------------|------------------|------------------|-----|---------------------|
| classe | $20 \leq x < 30$ | $30 \leq x < 40$ | ... | $80 \leq x \leq 90$ |
| frequency | 1 | 2 | ... | 3 |
| probability | 0.0167 | 0.0333 | ... | 0.05 |
| density | 0.00167 | 0.00333 | ... | 0.005 |

Using R commander : **Graphs-> Histogram**, we produced the following histogram.



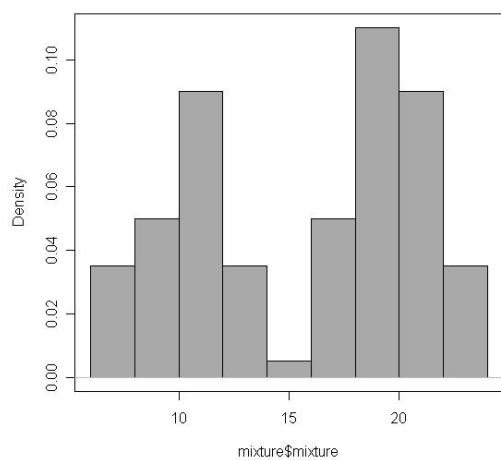
Describe the shape of the distribution.

Remark :

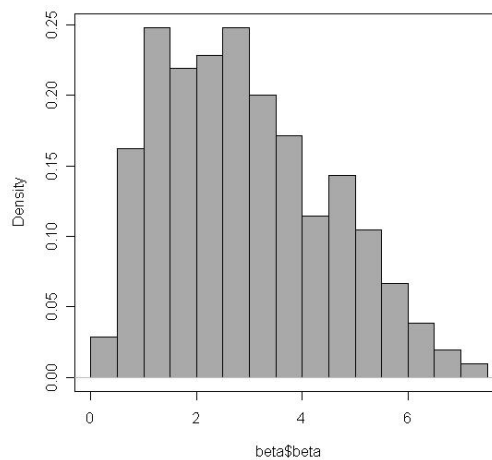
- R commander allows us to change the number of bins manually.
- It should be noted that if the sample size is small it usually is difficult to describe the shape of the distribution with a histogram.

Example 3 : Consider the following histograms. Describe the shape of the distribution. Is it a unimodal distribution? If so, is it symmetric, skewed to the right, or skewed to the left?

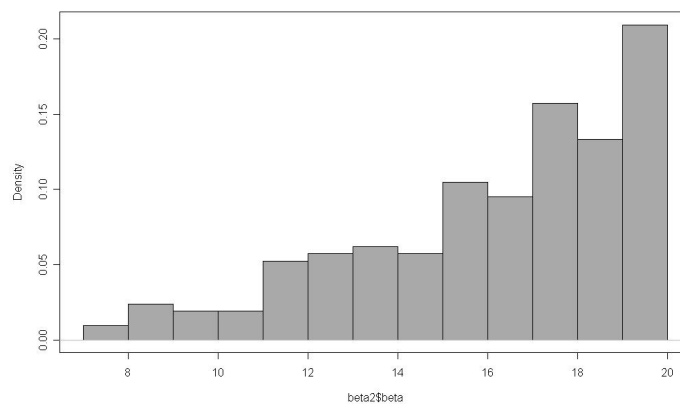
(a)



(b)



(c)



Remarks :

- When describing skewness, it is in the direction of the atypical values. A right skewed distribution as a few atypical large values on the right.
- A common shape for biological data is unimodal and skewed to the right.
- Left skewed distributions are less common.
- A bimodal distribution is often an indication that the observational units are heterogeneous, that is we have data from different subpopulations.

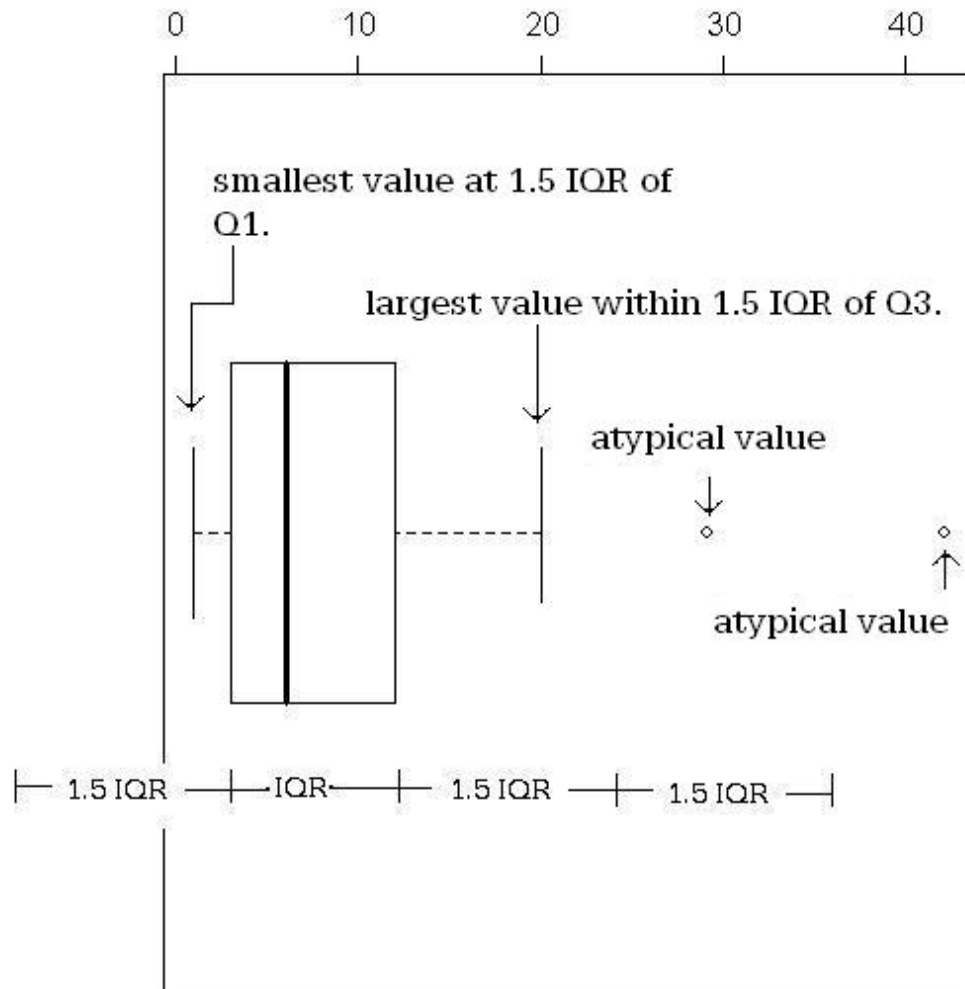
Box Plot :

We will discuss another useful visual tool invented by John Tukey in 1977 to describe the distribution a numerical variable. The diagram will us to identify the central tendency and the dispersion of the distribution. It will also help us to describe the asymmetry of the distribution and we will also be able to identify atypical values.

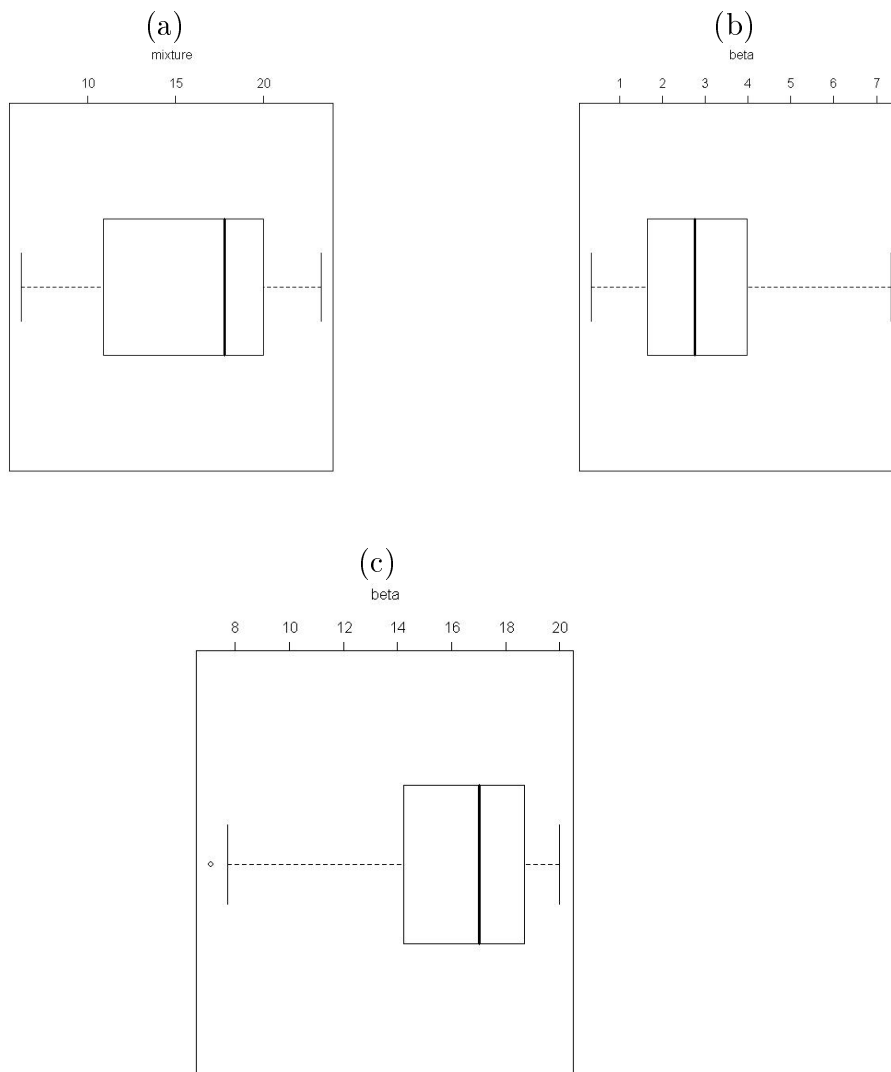
- We draw a box from the first quartile Q_1 to the third quartile Q_3 which is cut at the median.
- The range from Q_1 to Q_3 is the inter-quartile range which is measure of dispersion. The median is a mesure of central tendency.
- At the left of the box, we draw a stem up to the smallest value that is within 1.5 times the interquartile range of Q_1 .
- At the right, we draw a stem up to the largest value that is within 1.5 times the interquartile range of the 3rd quartile.
- For values that are past a distance of $1.5IQR$ to the right or left of the box, then we put a point for each of these values. We consider these values as **outlying** or **atypical** points.
- We should investigate these outlying points. They could be atypical because of an input error that should be corrected.
- Sometimes values that are a distance of $3IQR$ to the left of Q_1 or to the right of Q_3 are called **extreme** outliers.

Example 4 : Here is a box plot of the following data.

1 2 2 3 3 3 4 5
7 7 9 10 14 20 29 42



Example 5 : Here are the box plots for the distributions from Example 3. Discuss the asymmetry of the distributions and identify any outlying values (if there are any).



As we see in the following example, side-by-side box plots are useful to compare the central tendencies and the dispersions of many groups.

Example 5 : Based on the following diagrams, compare the central tendencies and within sample variabilities of the following samples :

sample a :

66.1, 64, 64.4, 60.65.3, 66.9, 61.5, 63.5, 61.6, 62.3

sample b :

66.3, 68.5, 68.4, 68.5, 68.3, 67.4, 66.1, 67.3, 69.2, 68.7

