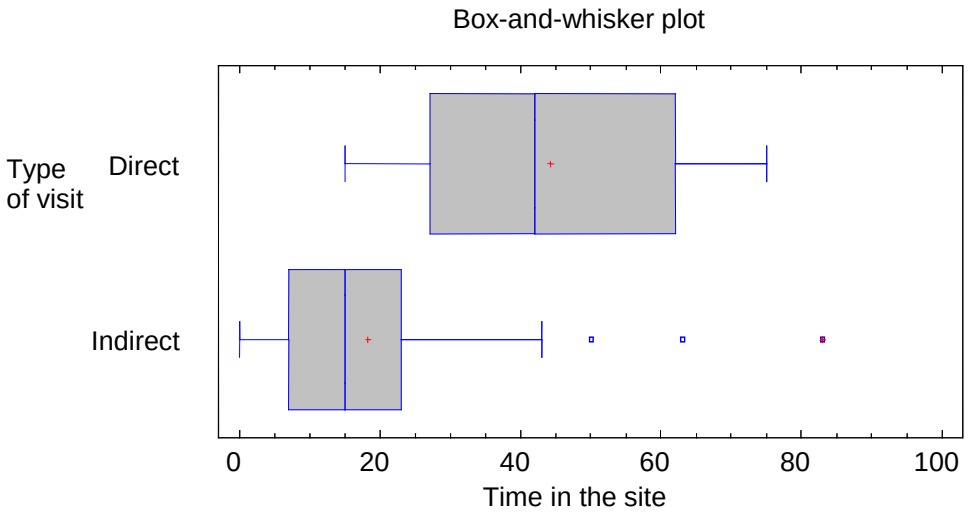
EXERCISES UD2 - DESCRIPTIVE STATISTICS

1) (1st partial 7-March-2011) One company of web design wants to study the profile of those who search in a certain web page as a function of the type of visit. One visit is direct if the user writes the web address of the site that wants to visit or accesses through the favorite markers of the navigator, while one indirect visit comes from search engines (Google, MSN Search, Yahoo!) or an external link in other web sites. The Box-and-Whiskers plot of the time registered as a function of the type of visit is the following:



According to this plot obtained, answer the following questions:

a) What would be the most suitable parameters of position and dispersion in order to characterize the variable of time, according to the type of visit?

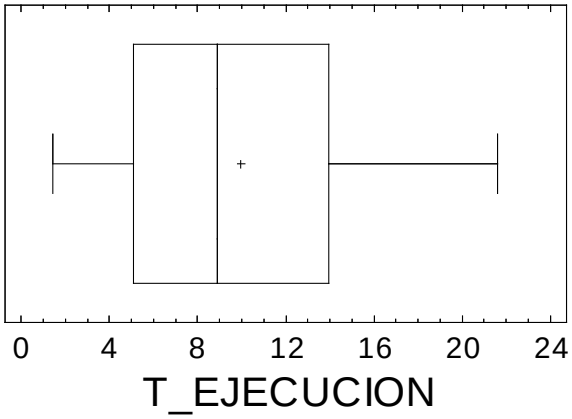
b) From a descriptive point of view, can you detect differences regarding the dispersion of the time, according to the type of visit of the user?

c) Is there any indication of asymmetry (skewness) in the times recorded according to one or the other type of visit?

d) If possible, estimate the number of data that have been taken to represent the figure.

e) Indicate if the following sentence is correct: "Approximately 50% of the persons who access by indirect visit pass less than 15 minutes in the web page while this percentage is 25% for those who access by direct visit".

2) (Final exam 14-June-2011) The following plot has been obtained with a sample of 100 data corresponding to the time of execution (in seconds) of a certain program.



According to this plot, answer the following questions justifying conveniently the reply.

a) What is the random variable under study? Indicate what is the type of this variable.

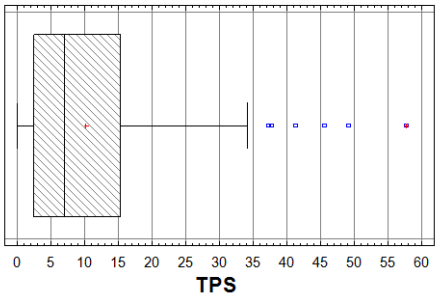
b) What is the name of this kind of plot?

c) Calculate the interquartile range. Indicate if it is a parameter of position or dispersion.

d) What parameters of position can be obtained from this plot? Which one would be the most convenient to describe this sample?

e) If the variance is calculated with the 100 data, what will be the units?

3) (1st partial 2-April-2012) In order to study the performance of an operative system, the time required for processing certain service (TPS) (expressed in ms) has been registered during one week. The following plot shows the results obtained:



Answer the following questions:

a) What is the variable under study in this case? What type is it?

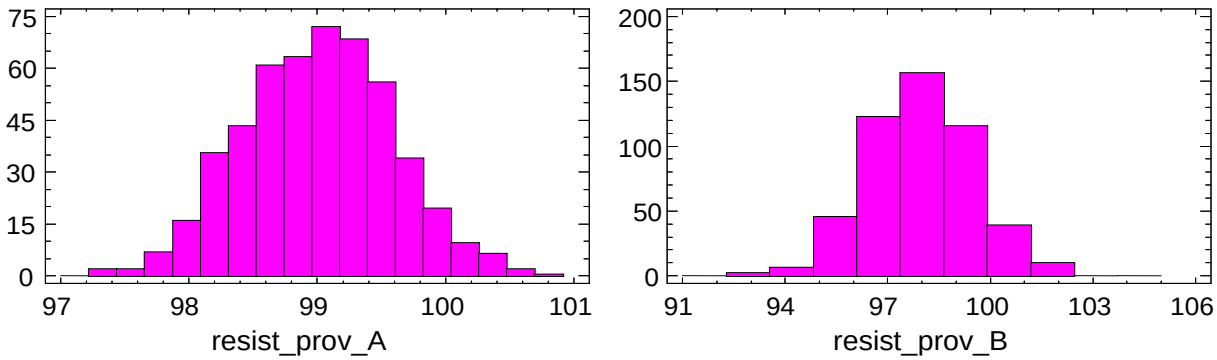
b) What is the name of this type of plot? For what types of variables is its use recommended?

c) According to the information provided by the plot shown above, what can we say about the distribution of the variable analyzed?

d) What parameter would provide reliable information about the position of the data analyzed? Justify your answer and calculate, if possible, the approximate value of that parameter.

e) What parameter would provide reliable information about the dispersion of the data analyzed? Justify your answer and calculate, if possible, the approximate value of this parameter.

4) (final exam 15-June-2012) One company that manufactures computer equipment uses certain electronic component with a resistance of 100 ohms, which can be purchased from two different suppliers (A or B). In order to study the differences in the resistance of components sold by each supplier, a sample of 500 components is taken from supplier A and another sample of 500 components from B. After measuring the resistance of these components, the following histograms were obtained:



According to these histograms, answer the following questions justifying conveniently the replies.

a) What does the vertical scale indicate? Why is it so different in the two cases?

b) In which of the two suppliers is there a higher dispersion in the values of resistance? Why?

c) Do you think that a histogram is an adequate technique to detect outliers? What other techniques would you use?

d) Indicate what parameters are the most appropriate in this case to quantify the data position of resistance. Calculate approximately such parameters for each supplier.

SOLUTIONS - PROBLEMS OF DESCRIPTIVE STATISTICS

Question 3:

3a) The variable under study is the time required for processing certain service (TPS), expressed in ms. It is a one-dimensional quantitative continuous variable, because the values of time can be measured with decimals.

3b) Box-and-whisker plot. It is recommended for continuous variables or for discrete variables with a high number of different values.

3c) The distribution is characterized by: position, dispersion and shape parameters. TPS ranges from 0 to 57 with an average of 10 and a median of 7. Fifty percent of the values are comprised between 7.5 and 15.5. The distribution is positively skewed (because the right whisker is much longer than the left one) and there is no clear evidence about outliers. The data might follow an exponential distribution because most values are close to zero.

3d) As the distribution is skewed the median is a better parameter of position than the average because it is not affected by extreme values. Median = 7 (line inside the box).

3e) As the distribution is skewed, the interquartile range (IQR) is a better parameter of dispersion than the variance or the standard deviation, because it is not affected by extreme values.
IQR = Q3 - Q1 = 15.5 - 2.5 = **13**

Question 4:

a) The vertical scale is absolute frequency, i.e., number of data contained in each interval of the hystogram. This scale is much bigger in the histogram of supplier B because it contains mess intervals (i.e., a lower number of bars). Taking into account that both histograms were built with 500 data, when dividing the range of variation of resistance by a lower number of intervals, it turns out a higher number of data in each interval, and as a result the absolute frequency becomes higher.

b) Range of A ≈ 101 – 97 = 4 ohms Range of B ≈ 102.5 – 92.5 = 10 ohms
Given that the ranges are so different and taking into account that in both cases data follow approximately a Normal distribution (as the hystogram resembles a Gauss function), supplier B presents a bigger variability than supplier A (i.e., bigger standard deviation, variance and interquartile range).

c) Generally speaking, a histogram is not a convenient tool to detect outliers because a single extreme value will result in a bar with a small height that can easily be unnoticed except if the total number of values is rather small. In order to detect outliers it would be more convenient to use a box-whisker plot or a normal probability plot.

d) Both histograms are rather symmetric and resemble a Gauss function, which suggests a Normal distribution of the data. In such cases, the axis of symmetry of the histogram corresponds to the average and the median, which are the most convenient parameters of position. Thus, it can be deduced from the plot that the average of supplier A is approximately 99 ohms, and 98 ohms in the case of B.