MAT 2379, Introduction to Biostatistics

Section 9.1. Random Sampling and Data Description
Part 2: Descriptive Statistics

Suppose we have a quantitative variable X whose observed values are $x_1, \ldots, x_n$. A *statistic* is a function of these observed value. *Descriptive statistics* give a useful summary of the data. The most commonly used descriptive statistics are:
a) the mean and the median
b) minimum and maximum values, the quartiles $Q1, Q_3$, and the interquartile range
c) the standard deviation

a) Measures of Center

The **sample mean** $\bar{x}$ is the average of the n observations:

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

If we put a mass of $1/n$ on each observation x_i, then $\bar{x}$ is the center of the mass.

Example 1. (weight) Let X be the weight (in kg) of a randomly chosen 12-year girl. The following observations represent the weights of $n = 10$ girls:

$x_1 = 27.7, \quad x_2 = 31.5 \quad x_3 = 30.9 \quad x_4 = 29.6 \quad x_5 = 27.0 \quad x_6 = 38.1 \quad x_7 = 32.4 \quad x_8 = 31.1 \quad x_9 = 36.7 \quad x_{10} = 28.4$

The sample mean is:

$$\bar{x} = \frac{27.7 + 31.5 + 30.9 + 29.6 + 27.0 + 38.1 + 32.4 + 31.1 + 36.7 + 28.4}{10} = 31.34$$

The difference between a data point and its mean is called a **deviation**. The deviation of the i-th observation is $x_i - \bar{x}$. Note that:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0.$$

.

Home Exercise: Calculate the deviations in the previous example. Check that their sum is equal to 0.

The **median** $\tilde{x}$ is the sample value which divides the sample into two approximately equal-sized data subsets. To obtain the median, arrange the sample values $x_1, x_2, \ldots, x_n$ in ascending order $y_1 \le y_2 \le \ldots \le y_n$. The median is given by:

$$\tilde{x} = \begin{cases} y_{\{(n+1)/2\}}, & \text{if } n \text{ is odd} \\ (y_{\{n/2\}} + y_{\{n/2+1\}})/2, & \text{if } n \text{ is even} \end{cases}$$

Note that the median is a sample value if n is odd, but not necessarily a sample value if n is even. (When is a median a sample value if n is even?)

Example 1. (continued) The sample values arranged in ascending order are:

$y_1 = 27.0, \quad y_2 = 27.7 \quad y_3 = 28.4 \quad y_4 = 29.6 \quad y_5 = 30.9 \quad y_6 = 31.1 \quad y_7 = 31.5 \quad y_8 = 32.4 \quad y_9 = 36.7 \quad y_{10} = 38.1$

Note that $n = 10$ is even. Hence, the median is calculated as the average of y_5 and y_6:

$$\tilde{x} = \frac{y_5 + y_6}{2} = \frac{30.9 + 31.1}{2} = 31$$

Example 2. (blood pressure) The following data gives the systolic blood pressure (in mmHg) of a sample of 5 persons after ten minutes of cardio exercise:

$$x_1 = 105, \quad x_2 = 120, \quad x_3 = 111, \quad x_4 = 125 \quad x_5 = 132$$

The mean is

$$\bar{x} = \frac{105 + 120 + 111 + 125 + 132}{5} = 118.6$$

To calculate the median, we arrange the data in ascending order as follows:

$$y_1 = 105, \quad y_2 = 111, \quad y_3 = 120, \quad x_4 = 125 \quad x_5 = 132$$

Since $n = 5$ is even, the median is equal to the 3rd observation, i.e.

$$\tilde{x} = y_3 = 120$$

Note: 1. The median is a "robust" statistic, i.e. it is not affected by small changes in the data set: in Example 2, if we replace $x_1 = 105$ by $x_1 = 110$, the median remains the same. The mean does not have this property.

2. The mean is an "efficient" statistic, i.e. it uses all the information in the data set. The median does not have this property.

b) Boxplots

Quartiles are values that divide the ordered data into 4 equally-sized data subsets. The three quartiles are: $q_1, Q_2 = \tilde{x}, q_3$. To calculate the quartiles, we arrange the data $x_1, x_2 \ldots, x_n$ in ascending order $y_1 \le y_2 \le \ldots \le y_n$.

(i) To find the **first quartile**, we need to calculate $(n+1)/4$. Represent this quantity as an integer part r and a fractional part $a/4$, where a can be 0, 1, 2 or 3. i.e.

$$\frac{n+1}{4} = r + \frac{a}{4}, \quad r = \text{integer}, a = 0, 1, 2 \text{ or } 3.$$

The first quartile is calculated as follows:

$$q_1 = \begin{cases} y_r, & \text{if } (n+1)/4 = r \\ (3/4)\, y_r + (1/4)\, y_{r+1}, & \text{if } (n+1)/4 = r + 1/4 \\ (1/2)\, y_r + (1/2)\, y_{r+1}, & \text{if } (n+1)/4 = r + 2/4 \\ (1/4)\, y_r + (3/4)\, y_{r+1}, & \text{if } (n+1)/4 = r + 3/4 \end{cases}$$

For example, if $n = 14$, then

$$\frac{n+1}{4} = 3.75 = 3 + \frac{3}{4} \quad \text{and} \quad q_1 = \frac{1}{4}\, y_3 + \frac{3}{4}\, y_4.$$

Note that 3.75 is between 3 and 4, but it is closer to 4. Therefore q_1 should be an weighted average between y_3 and y_4, with more weight given to y_4.

The interpretation: approximately one quarter of the data values are smaller than q_1 and three quarters of the data values are larger than q_1.

Example 1. (continued) In this example, $n = 10$ and

$$\frac{n+1}{4} = \frac{11}{4} = 2.75 = 2 + \frac{3}{4} \text{ (between 2 and 3, closer to 3).}$$

Therefore,

$$q_1 = \frac{1}{4}\, y_2 + \frac{3}{4}\, y_3 = (0.25)(27.7) + (0.75)(28.4) = 28.225$$

(ii) To find the **third quartile**, we need to calculate $3(n+1)/4$. Represent this quantity as an integer part r and a fractional part $a/4$, where a can be 0, 1, 2 or 3. i.e.

$$\frac{3(n+1)}{4} = r + \frac{a}{4}, \quad r = \text{integer}, a = 0, 1, 2 \text{ or } 3.$$

The third quartile is calculated as follows:

$$q_3 = \begin{cases} y_r, & \text{if } 3(n+1)/4 = r \\ (3/4)\, y_r + (1/4)\, y_{r+1}, & \text{if } 3(n+1)/4 = r + 1/4 \\ (1/2)\, y_r + (1/2)\, y_{r+1}, & \text{if } 3(n+1)/4 = r + 2/4 \\ (1/4)\, y_r + (3/4)\, y_{r+1}, & \text{if } 3(n+1)/4 = r + 3/4 \end{cases}$$

For example, if $n = 14$, then

$$\frac{3(n+1)}{4} = 11.25 = 11 + \frac{1}{4} \quad \text{and} \quad q_3 = \frac{3}{4}\, y_{11} + \frac{1}{4}\, y_{12}.$$

Note that 11.25 is between 11 and 12, but it is closer to 11. Therefore q_3 should be an weighted average between y_{11} and y_{12}, with more weight given to y_{11}.

The interpretation is: approximately three quarters of the data values are smaller than Q_3 and one quarter of the data values are larger than q_3.

Example 1. (continued) In this example, $n = 10$ and

$$\frac{3(n+1)}{4} = \frac{33}{4} = 8.25 = 8 + \frac{1}{4} \text{ (between 8 and 9, closer to 8).}$$

Therefore,

$$q_3 = \frac{3}{4}\, y_8 + \frac{1}{4}\, y_9 = (0.75)(32.4) + (0.25)(36.7) = 33.475$$

(iii) The **interquartile range** (IQR) is the difference between the third and first quartiles:

$$IQR = q_3 - q_1.$$

The interpretation is: the IQR gives the range of the middle half of the data values.

Example 1. (continued) The interquartile range is:

$$IQR = Q_3 - Q_1 = 33.475 - 28.225 = 5.25$$

(iv) The **sample range** (R) is the difference between the maximum and the minimum values in the sample:

$$R = y_n - y_1$$

The interpretation is: R gives the range of all the data values.

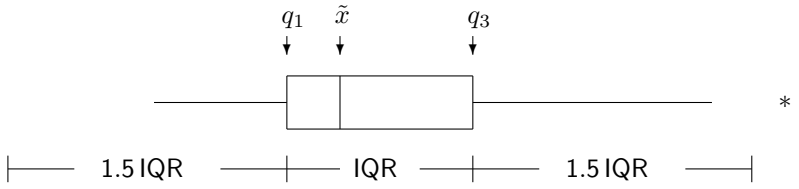
Example 1. (continued) The minimum value is $y_1 = 27.0$. The maximum value is $y_{10} = 38.1$. The range is:

$$R = y_{10} - y_1 = 38.1 - 27.0 = 11.1.$$

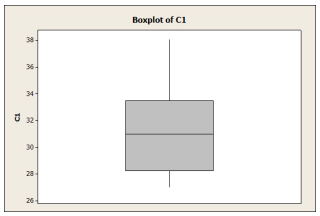
(v) The **5-number summary** consists of: $y_1, q_1, \tilde{x}, q_3, y_n$. A **boxplot** is a graphical display of the 5-number summary, which is constructed as follows:

- Extend the box from the first quartile to the third quartile. (The box displays the interquartile range.)
- Within the box, display a line at the median.
- Imaginary fences are placed at a distance of $1.5\, IQR$ above the third quartile and below the first quartile.

- Whiskers extend from the ends of the box to the smallest and the largest value in the sample, *within* the imaginary fences.
- Values outside the fences are called **outliers**. Each outlier is displayed as a point. Sometimes a different symbol is used for extreme outliers that are at least $3\,IQR$ above q_3 or below q_1.



Example 1. (continued) Here is the boxplot for this data set:



c) Measures of Dispersion

The **sample standard deviation** is denoted by s and is the most commonly used measure for dispersion. Its square is called the **sample variance** and is calculated by the formulas:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)$$

Example 2. (continued) Recall the data:

$$x_1 = 105, \quad x_2 = 120, \quad x_3 = 111, \quad x_4 = 125 \quad x_5 = 132$$

The mean is $\bar{x} = 118.6$. The sample variance is:

$$s^2 = \frac{1}{4}[(105-118.6)^2 + (120-118.6)^2 + (111-118.6)^2 + (125-118.6)^2 + (132-118.6)^2] = 116.3$$

The sample standard deviation is

$$s = \sqrt{116.3} = 10.78$$

Consider now the following data set:

$$x_1 = 109, \quad x_2 = 116, \quad x_3 = 120, \quad x_4 = 122 \quad x_5 = 126$$

We see that this sample has the same mean $\bar{x} = 118.6$ as the first sample, but a sample variance $s^2 = 41.8$ and a sample standard deviation $s = 6.47$. There is more dispersion in the first sample.