Descriptive Statistics (Part I)

In a study or an experiment, one has statistical estimates (which can be obtained from an experiment, such as in biology, or through observation, such as in astronomy). We describe these statistical estimates with variables. A variable can be classified as being qualitative or quantitative. A qualitative variable takes values from a finite set of possible categories or classes. For example, a person's favorite ice cream flavor among those available at some ice cream store could be one of five 5 categories (blueberry, chocolate, vanilla, Oreo, strawberry). Conformity ("yes" or "no") is another example of a qualitative variable. A quantitative variable, on the other hand, is a variable whose values are numbers. Examples of quantitative variables include the age in years, distance in meters, or volume in cubic centimeters. A discrete variable (such as one that takes values in $0, 1, 2, 3, \ldots$) is a also a quantitative variable.

In these notes, we will describe the distribution of a quantitative variable.

Quantitative Summary

Suppose that we have n data points and we want to describe each of these data points with a quantitative variable x. We will denote the n values by $x_1, x_2, \ldots, x_n$. We say that these n values make up a **random sample** of size n.

Example 39 : The $n = 9$ observations that follow are the oven temperatures recorded for successive batches of silicon wafers produced (in °F) :
953, 950, 948, 955, 951, 949, 957, 954, 955

So, $x_1 = 953$, $x_2 = 950, \ldots, x_9 = 955$.

Note : We will describe the central tendancies of this variable with *sample mean*, *sample variance*, and *median* (we will define the sample variance later).

The sample mean is

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{953 + 950 + \cdots + 955}{9} = \frac{8572}{9} = 952.44°F.$$

Remark : Before defining the median, we will define the notion of percentile.

Percentile : A percentile is one of 99 values that divide the random samples into 100 equal parts such that each part represents 1/100 of the sample of size n.

Remark : There are many ways to compute a percentile. Here is one way to do so

Calculation of the k'th percentile : We need to arange our n values in increasing order :

$$y_1 \leq y_2 \leq \ldots \leq y_n.$$

For our temperature example (Example 39), we obtain :
$y_1 = 948$, $y_2 = 949$, $y_3 = 950$, $y_4 = 951$, $y_5 = 953$, $y_6 = 954$, $y_7 = 955$, $y_8 = 955$, $y_9 = 957$

Consider : With 2 numbers, we can divide the real line into 3 intervals. With 3 numbers, we have 4 intervals, and so on. We want to know in which interval our percentile lies :

The **rank** of the k'th percentile is $(n+1) \times k/100 = m + s$, where m is the whole part and s is the remainder, i.e., $0 \leq s < 1$.

The k'th percentile is

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

Remarks :

— If the rank is 7.5, then $m = 7$ and $s = 0.5$.
— If the rank is 9.25, then $m = 9$ and $s = 0.25$.
— If the rank is 10.75, then $m = 10$ and $s = 0.75$.
— This particular method of calculating the percentile is called *linear interpolation*.

The **median** is the 50'th percentile. If $n = 9$, the rank of the median is $(n + 1) \times 50/100 = 5$. Since the fractional part is zero for this example, the median is $y_5 = 953\,°F$.

Remark : Why are we interested in both the mean and the median ? We oftentimes interpret the median as a "typical value". For example, consider the following $n = 10$ temperatures :

$$700, 953, 950, 948, 955, 951, 949, 957, 954, 955.$$

The mean temperature is $\overline{x} = 927.2$ and the median temperature is 952. In this example, the mean temperature is therefore smaller than all the values in the sample except for the smallest value, so the mean does not give us a "typical" value in this case. The median, on the other hand, is always in the middle of the sample. So in this example it is better to use the median as a "typical value".

Example 40 : The concentration of ingredients in a liquid detergent can be affected by the catalyst used in a fabrication process. The values of the concentration were recorded for two types of catalysts :

$$\text{catalyst 1 : } 57.9,\ 58.0,\ 62.6,\ 63.7,\ 65.2,\ 65.2,\ 65.4,\ 66.2,\ 67.2,\ 67.6$$

$$\text{catalyst 2 : } 64.8,\ 65.3,\ 65.3,\ 68.6,\ 68.8,\ 69.3,\ 69.4,\ 69.6$$

For each of the two catalysts, give the sample size, the mean concentration, and the median concentration. **solution (catalyst 1) :** Sample size for catalyst 1 is : $n = 10$ mean is $\mu = 63.9$

To find the median (i.e., the 50'th percentile), we must first compute the rank of the $k = 50$'th percentile :

$$\text{rank of 50'th percentile } = (n{+}1){\times}k/100 = (10{+}1){\times}50/100 = 5.5 = 5{+}0.5 = m{+}s$$

where $m = 5$ is the whole part of the rank, and $s = 0.5$ is the remainder.

Since the samples are already ordered from smallest to largest for us, we have

$$y_1 \leq y_2 \leq \ldots \leq y_n$$

, where

$$y_1 = 57.9, y_2 = 58.0, y_3 = 62.6, y_4 = 63.7, y_5 = 65.2, y_6 = 65.2, y_7 = 65.4,$$

$$y_8 = 66.2, y_9 = 67.2, y_{10} = 67.6$$

so the median is

$$\text{median} = 50\text{'th percentile} = y_m + s\,(y_{m+1} - y_m) = y_5 + 0.5\,(y_6 - y_5)$$

$$= 65.2 + 0.5\,(65.2 - 65.2) = 65.2$$

Remark : Not all the values will be in the center, so we should also describe the "dispersion" of the samples.

Measures of dispersion : We will use the standard deviation, the range and the interquartile distance to describe the dispersion of the data. Refer to the temperatures in Example 39 :

The **sample variance** is

$$s^2 = \frac{s_{xx}}{n-1},$$

where

$$s_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n\,\overline{x}^2 = \left(\sum_{i=1}^{n} x_i^2\right) - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}.$$

So,

$$\begin{aligned} s^2 &= \frac{\left(\sum_{i=1}^{n} x_i^2\right) - n\,\overline{x}^2}{n-1} \\ &= \frac{(953^2 + 950^2 + \cdots + 955^2) - 9\,(8572/9)^2}{9-1} \\ &= \frac{8\,164\,430 - 9\,(8572/9)^2}{9-1} \\ &= 9.528\,(^\circ F)^2 \end{aligned}$$

and the **sample standard deviation** is $s = \sqrt{s^2} = 3.086\,^\circ F$.

The **range** of the samples is

$$r = \max(x_i) - \min(x_i) = y_n - y_1 = 957 - 948 = 9\,^\circ F.$$

Quartile : The first quartile q_1, the second quartile q_2 and the third quartile q_3, are, respectively, the 25'th, the 50'th and the 75'th percentiles.

For our example of $n = 9$ temperatures, the rank of q_1 is $(n+1)\times 25/100 = 2.5$. So

$$q_1 = y_2 + 0.5(y_3 - y_2) = 949 + 0.5(950 - 949) = 949.5\,^\circ F$$

and the rank of q_3 is $(n+1)75/100 = 7.5$. So

$$q_3 = y_7 + 0.5(y_8 - y_7) = 955 + 0.5(955 - 955) = 955\,^\circ F.$$

The **interquartile distance** (IQD) is

$$\mathsf{IQD} = q_3 - q_1 = 955 - 949.5 = 5.5\,^\circ F.$$

Example 40 [Part II] : The concentration of active ingredients in a liquid detergent can be affected by the catalyst used in the fabrication process. The concentration was measured when each of the two catalysts was used (the measurements are listed in order of increasing concentration)

catalyst 1 : 57.9, 58.0, 62.6, 63.7, 65.2, 65.2, 65.4, 66.2, 67.2, 67.6

catalyst 2 : 64.8, 65.3, 65.3, 68.6, 68.8, 69.3, 69.4, 69.6

For each of the catalysts, give the interquartile distance (IQD) of the concentration.

Solution (for catalyst 1) :

rank of 25'th percentile $= (n{+}1){\times}25/100 = (10{+}1){\times}25/100 = 2.75 = 2{+}0.75 = m{+}s$

where $m = 2$ and $s = 0.75$, so

$$q_1 = 25\text{'th percentile} = y_m + s\,(y_{m+1} - y_m) = y_2 + 0.75\,(y_3 - y_2)$$
$$= 58.0 + 0.75\,(62.6 - 58.0) = 61.45$$

Also,

rank of 75'th percentile $= (n{+}1){\times}75/100 = (10{+}1){\times}75/100 = 8.25 = 8{+}0.25 = m{+}s$

where $m = 8$ and $s = 0.25$, so

$$q_3 = 75\text{'th percentile} = y_m + s\,(y_{m+1} - y_m) = y_8 + 0.25\,(y_9 - y_8)$$
$$= 66.2 + 0.25\,(67.2 - 66.2) = 66.45$$

so the interquartile distance is

$$\mathsf{IQD} = q_3 - q_1 = 66.45 - 61.45 = 5$$