Statistics and Data Analysis
Lecture notes

Edgar Bueno

1 Introduction

There is no unique definition of *Statistics*. For the purpose of this course we simply define *Statistics* as the discipline that allows to achieve any of the following goals: **i.** describe a population of interest with respect to one or several characteristics; **ii.** make inferences regarding an unknown characteristic of a population of interest; **iii.** predict the outcome of a characteristic for a new element of the population.

Basically, working around those three goals is what we will do in this course. The first part of the course will concentrate on *descriptive statistics*. Then, at the beginning of the second part, we will cover some basic concepts in *probability* that will be needed for discussing *inferential statistics* in the second part.

The three goals in the first paragraph refer to a *population* and a *characteristic* of it, so it is adequate to begin by defining these concepts.

1.1 Populations

We begin by defining the concept of a population.

Definition 1. A *population*, denoted by U, is a set of elements of interest. □

We can classify populations with respect to two criteria. The first criterion indicates whether a population is *bounded* or *unbounded*. A *bounded* population is such that it is contained within some limits. An *unbounded* population is not contained within any limits.

The second criterion indicates whether a population is *discrete* or *continuous*. A *discrete* population is such that its elements can be counted (enumerated), whereas elements in a *continuous* population cannot be enumerated. There are, therefore, mainly four types of populations: discrete–bounded, continuous–bounded, discrete–unbounded and continuous–unbounded.

All real populations we can think of are bounded (either discrete or continuous). Some examples of discrete–bounded populations are the set of residents of a given country, the set of trees in a properly demarcated area, the set of books in a library or the set of companies of a country. Some examples of continuous–bounded populations are the set of time points in a time period of interest or the set of points within the area that demarcates a lake. Unbounded populations are merely conceptual, for example the set of boxes to be produced by a manufacturing process is a discrete population whose size is potentially non-finite.

Most of this course will concentrate on discrete populations, either bounded or unbounded. In particular, the first part (*descriptive statistics*) concerns discrete–bounded populations, also known as *finite populations*.

1.2 Study variables

Next we define a *study variable*.

Definition 2. A *study variable* is an attribute of interest associated to each element in the population. □

Consider the population of residents of a given country, examples of variables are income, sex, age, municipality of residence, height and educational level. In the case where the population of interest are the books in a library, some examples of variables are the price of each book, the number of pages or the genre.

Variables can be classified in several ways. One criterion for classification is with respect to the type of information carried by the variable. According to this criterion, variables can be classified as either *categorical*, *numerical* or *identifier*.

A *categorical* variable produces responses that belong to categories. Sex and educational level are examples of categorical variables in the population of residents of a given country, whereas genre is a categorical variable in the population of books in a library. In particular, categorical variables with only two response categories are known as binary or dichotomous variables. Sex (male/female), being a student or not, passing or failing an exam are all examples of dichotomous variables. A dichotomous variable that takes the values 0 and 1 is known as a *dummy variable*. Evidently, any dichotomous variable can be expressed as a dummy variable. For instance, the variable that indicates whether students passed or failed an exam can be written as 1 if the student passed and 0 otherwise.

A categorical variable that produces responses that can be ranked or ordered is called an *ordinal* variable. The position of drivers in a Formula 1 race is an example of an ordinal variable: being first is better than being second, which is in turn better than being third, and so on. Other classical examples of ordinal variables are satisfaction ratings (very satisfied, moderately satisfied, neutral, moderately dissatisfied, very dissatisfied) or educational level (having undergraduate studies is higher than having high school studies, but lower than having master studies).

Although the categories of an ordinal variable can be ordered, the difference or ratio between categories has no meaning. In a Formula 1 race, it makes no sense to say that the difference between the first and the second driver is the same as between fourth and fifth; or that the driver in the fourth position is twice as bad as the driver in the second position.

If the categories of the variable cannot be ranked or ordered, so that the categories are only names, we call it a *nominal* variable. For instance, sex, age color or nationality are examples of nominal variables.

As indicated by the name, *numerical* variables are variables that produce numerical responses. They are in turn subdivided into *discrete* and *continuous*. Discrete variables can take a countable number of responses, for example, the number of children a woman has, the number of students enrolled in a class or the number of university credits earned by a student at the end of a semester. Continuous variables may take on any value within an interval, for example, a person's height or weight, the time to run a race or the temperature.

It is important to note that although, strictly speaking, dummy variables are categorical, for many practical purposes they can be regarded as numerical variables. This is a property that we will use repeatedly over the course.

Identifiers are variables that allow to identify unequivocally elements of the population. The personal number (*personnummer*) assigned by the Tax Agency is an example of identifier for the Swedish citizens (and some other individuals). Identifiers take as many values as elements in the population.

We will denote variables by the last letters in the alphabet, x, y or z. When we have several study variables, we will use the notation $x_1, x_2, \cdots$.

We denote by x_i the value of the variable x associated to the ith population element. When considering several variables, we will use x_{ji} to denote the value of the variable x_j for the ith population element. In this way, if U is a population of establishments and x is the variable "number of employees", x_i denotes the number of employees of the ith establishment. If we consider more than one variable, say "number of employees" and "taxes paid during the last year" we switch to x_{1i} and x_{2i} to denote the number of employees and the taxes paid by the ith establishment, respectively.

1.3 Datasets

If we are dealing with finite populations we could, in principle, measure a set of variables on *each* element of the population of interest. This information could be arranged in a rectangular array where each row corresponds to one element of the population and each column corresponds to one variable. In this way, the information about the jth variable for the ith element will be shown in the cell $(i\,, j)$. Such arrays are known as *data tables* or *datasets*. Table 1 illustrates a dataset collecting information about J variables on the N elements of a population. When the data has been arranged into a dataset, each row is usually referred to as a *case*.

ID	**Contact**	$\mathbf{x_1}$	$\mathbf{x_2}$	$\cdots$	$\mathbf{x_J}$
ID_1	$Contact_1$	x_{11}	x_{21}	$\cdots$	x_{J1}
ID_2	$Contact_2$	x_{12}	x_{22}	$\cdots$	x_{J2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
ID_N	$Contact_N$	x_{1N}	x_{2N}	$\cdots$	x_{JN}

Table 1: Array collecting the information in a statistical study

One categorical variable x with K categories can be stored in different ways in the dataset. Probably the most common way is to store it as one column showing the category to which each element belongs. Consider, for instance, the variable $x =$ *marital status* that takes $K = 3$ categories: *single*, *married* and *divorced*. The first column in Table 2 shows the values taken by a population of ten individuals.

Marital status	Single	Married	Divorced
x	x_1	x_2	x_3
Single	1	0	0
Divorced	0	0	1
Divorced	0	0	1
Married	0	1	0
Married	0	1	0
Divorced	0	0	1
Single	1	0	0
Divorced	0	0	1
Single	1	0	0
Married	0	1	0

Table 2: Two ways of storing a categorical variable in a dataset

Alternatively, the categorical variable x can be splitted into K dummy variables, $x_1, x_2, \cdots, x_K$ as follows. Let x_i be the value of the variable x associated to the ith element ($i = 1, 2, \cdots, N$),

we define

$$x_{ki} = \begin{cases} 1 & \text{if } x_i \text{ takes the } k\text{th category} \\ 0 & \text{otherwise} \end{cases} \qquad \text{for } k = 1, \cdots, K$$

In other words, x_1 indicates whether x belongs to the first category or not, x_2 indicates whether x belongs to the second category or not, and so on. The last three columns in Table 2 show this way of storing the marital status for the ten individuals, where x_1, x_2 and x_3 indicate, respectively, if the individual is single, married or divorced. As a dummy variable can be regarded as numerical, this second way for storing a categorical variable is more convenient for analytical purposes.

By construction, exactly one value among $x_{1i}, x_{2i}, \cdots, x_{Ki}$ is equal to one. Therefore it should be noted that even when a categorical variable with K categories can be stored as K dichotomous variables, one of them is unnecessary as its values are completely specified once the remaining $K-1$ values are specified. In our example in Table 2, note that we could have removed any of the three variables x_1, x_2, x_3, as its values are known once we know the values of the remaining two variables. For instance, if we remove x_3 (which indicates whether a person is divorced or not) we know that the person is divorced if it is neither single nor married ($x_1 = 0$ and $x_2 = 0$); and the person is not divorced if it is either single or married ($x_1 = 1$ or $x_2 = 1$).

Evidently, a dataset contains data. Sometimes *a lot* of data. However raw data is often not very useful by itself as it is not easy to get any actual *information* from it. We need methods that allow to summarize and describe the main characteristics of the data. Thus, now we assume that a dataset is available and our task will be to summarize these data through measures that provide meaningful information. In the following sections we will introduce numerical and graphical tools that allow for summarizing and describing the information provided by one or two variables.

2 Describing one variable

In this section we introduce several measures and tools that allow for summarizing and describing the information provided by one variable measured on the elements of a finite population of size N. The general setting is this: A variable x has been measured on the elements of a population of size N, so we have the observations $x_1, x_2, \cdots, x_N$. Our task is to summarize the main characteristics of the observations through a few meaningful measures.

2.1 Numerical description

Formally, any numerical measure that we calculate from a population is known as a *parameter*.

Definition 3. A *parameter* is a characteristic of interest from the population. □

We will consider three types of parameters. We begin by introducing some parameters that are useful for describing where the observations are located.

2.1.1 Measures of location

In this subsection we introduce several measures that are useful for summarizing where the observations of a given variable are located. The first parameter we introduce is the arithmetic mean or average.

The average or mean:

Definition 4. The *average* or *arithmetic mean* or simply, the mean, of a numerical variable x in the population U is defined as

$$\bar{x}_U \equiv \frac{1}{N}\sum_U x_i \tag{1}$$

i.e. the sum of all values of x in the population divided by the size of the population.

In particular, the mean of a dummy variable x is known as a *proportion* and is denoted by P_x. □

Example 5. Let U be the population of $N = 10$ students taking a Master course in statistics, let x_i be the number of points the ith student got in the final exam and y_i be a dummy variable indicating the sex of the student (male=0; female=1). Table 3 shows the observed values:

i	1	2	3	4	5	6	7	8	9	10
x_i	8	15	5	36	40	30	9	21	32	27
y_i	1	1	1	1	0	1	0	0	1	0

Table 3: Points of ten students in an exam in Statistics

Although graphical tools for describing data will be the topic of Subsection 2.2, an exception will be made and we will introduce at this point one type of chart, namely, the *dotplot*. In a dotplot the observations are represented as dots (or other symbols, like line segments) over a number line. If one value occurs multiple times, we just stack them over each other. Figure 1 shows a dotplot of the ten exam points in Example 5. (The function `stripchart()` allows to create dotplots in R.)

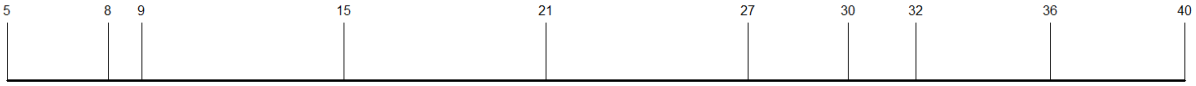

Figure 1: Dot plot of the exam points of the students in Example 5.

The mean number of points received by the students in the course is

$$\bar{x}_U = \frac{1}{N}\sum_U x_i = \frac{1}{10}\left(8 + 15 + \cdots + 27\right) = \frac{1}{10}223 = 22.3.$$

The proportion of female students (i.e. the mean of y) is

$$P_y = \bar{y}_U = \frac{1}{N}\sum_U y_i = \frac{1}{10}\left(1 + 1 + \cdots + 0\right) = \frac{1}{10}6 = 0.6. \qquad \square$$

The mean can be interpreted as the center of gravity of the observations of the variable. Imagine the different observations as weights over a bar. This bar will get balanced at the mean. At any other point, the bar will not find balance. This is represented on Figure 2 using the number of points in the exam for the population of students in Example 5.

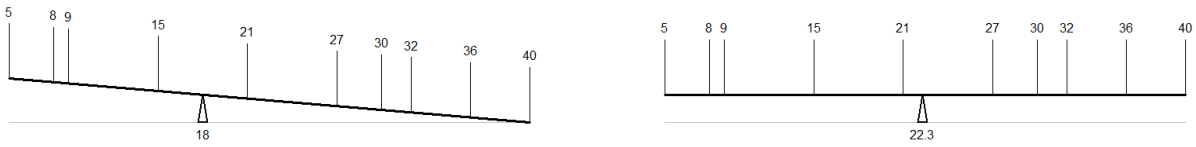

Figure 2: The mean as the center of gravity of the observations.

The mean is only defined for numerical (quantitative) variables. However, it is not uncommon to use the mean with ordinal variables. For instance, if an athlete has finished his last three races in the tenth, fourth and twentieth position, respectively, one may say that, on average, he has finished in the eleventh position. Strictly speaking, this use of the mean is not correct.

The mean is sensitive to extreme observations (either too large or too small), for instance, if a new student scores 100 points in the exam in Example 5, the mean becomes 29.4. A change of more than seven units! For this reason, we say that the mean is not a *robust* parameter. This is illustrated in Figure 3.

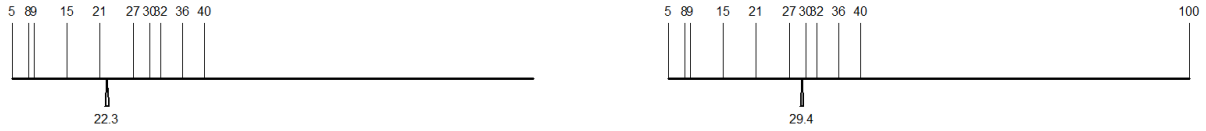

Figure 3: The mean as a parameter sensitive to extreme observations.

Some properties of the mean or average:

- if a constant is added to each observation, the mean of the resulting variable is simply the mean of the original variable plus the constant. More formally, let $y_i = x_i + c$ for $i = 1, 2, \cdots, N$, then $\bar{y}_U = \bar{x}_U + c$;
- if each observation is multiplied by a constant, the mean of the resulting variable is simply the mean of the original variable times the constant. More formally, let $y_i = cx_i$ for $i = 1, 2, \cdots, N$, then $\bar{y}_U = c\bar{x}_U$;
- the mean of the sum of two variables is the sum of the means. More formally, let x and y be two variables with means given by $\bar{x}_U$ and $\bar{y}_U$, respectively. Let $z_i = x_i + y_i$, then $\bar{z}_U = \bar{x}_U + \bar{y}_U$;
- putting all together we get the following result. Let $\bar{x}_U$ and $\bar{y}_U$ be, respectively, the means of two variables x and y,

$$\text{if } z_i = a + b\,x_i + c\,y_i \qquad \text{then} \qquad \bar{z}_U = a + b\,\bar{x}_U + c\,\bar{y}_U.$$

Or more generally, let $\bar{x}_{1U}, \bar{x}_{2U}, \cdots, \bar{x}_{JU}$ be the means of J variables $x_1, x_2, \cdots, x_J$,

$$\text{if } z_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + \cdots + a_J x_{Ji} \qquad \text{then} \qquad \bar{z}_U = a_0 + a_1 \bar{x}_{1,U} + a_2 \bar{x}_{2,U} + \cdots + a_J \bar{x}_{J,U}.$$

The function `mean()` can be used in R for computing the mean of a variable.

The median

Now we turn our attention to a second parameter: the median. Before defining the median we need to define the x-ordered population. If x is an ordinal or a numerical variable, we denote by $x_{(1)}$ the smallest x-value, by $x_{(2)}$ the second to smallest x-value and so on. In this way, $x_{(N)}$ is the largest x-value. The x-ordered population is

$$x_{(1)}, x_{(2)}, \cdots, x_{(N)}.$$

In other words, the x-ordered population are the x-values arranged from smallest to largest.

$x_{(1)}$ and $x_{(N)}$ are known as the minimum and the maximum of x, respectively.

Definition 6. Let x be a variable that is at least ordinal, the *median* of x, $\breve{x}_U$, is the value that divides the population in two halves, in such a way that (at least) half of the x-values are smaller or equal than $\breve{x}_U$ and (at least) half of the x-values are larger or equal than $\breve{x}_U$.

$$\breve{x}_U \equiv \begin{cases} x_{\left(\frac{N+1}{2}\right)} & \text{if } N \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}\right) & \text{if } N \text{ is even} \end{cases} \qquad \square \tag{2}$$

Example 7. Consider the population of ten students in Example 5, the x-ordered population is shown in Table 4.

i	1	2	3	4	5	6	7	8	9	10
$x_{(i)}$	5	8	9	15	21	27	30	32	36	40

Table 4: Ordered points of ten students in an exam in Statistics

As $N = 10$ is even, the median of x is

$$\breve{x}_U = \frac{1}{2}\left(x_{\left(\frac{N}{2}\right)} + x_{\left(\frac{N}{2}+1\right)}\right) = \frac{1}{2}\left(x_{(5)} + x_{(6)}\right) = \frac{1}{2}\left(21 + 27\right) = 24.$$

Being a nominal variable, the median of y is not defined. $\square$

Unlike the mean, the median is not very sensitive to extreme observations, so we say that it is a robust parameter. For instance, if a new student scores 100 points in the exam in Examples 5 and 7, the median becomes 27. A change of three units, which is way smaller than the change in the mean. This is represented in Figure 4.

Figure 4: The median as a parameter that is not very sensitive to extreme observations.

As indicated before the median is the value that splits the population in two halves. Figure 4 also allows to see this interpretation of the median. If the population size is even (as in the left panel), the median is the average of the two observations in the middle of the population. If the population size is odd (as in the right panel), the median is simply the observation in the middle of the population.

The function `median()` can be used in R for computing the median of a variable.

The mode

A third parameter of location is the *mode*. Most authors consider the mode as one more parameter of location, De Veaux et al. (2021), however, consider it to be a measure of the shape of the observations. This distinction is irrelevant for practical purposes and should not cause any confusion.

Definition 8. The *mode* of a variable x in the population U, $\dot{x}_U$, is defined as the most frequently occurring value. $\square$

If all x values have the same frequency, we say that the variable has no mode; variables with only one mode are called *unimodal*; variables with two modes are called *bimodal*; and variables with more than two modes are called *multimodal*.

Example 9. In the example of the points of the ten students (variable x in Examples 5 and 7), all values occur exactly one time, therefore there is no mode.

Consider the following variables (whose dotplots are shown in Figure 5):

$$\begin{aligned} x_1 &= \{0, 1, 2, 4, 6, 7, 8, 9, 9, 9\} \\ x_2 &= \{1, 2, 6, 7, 8, 8, 8, 9, 9, 9\} \\ x_3 &= \{6, 6, 6, 7, 8, 8, 8, 9, 9, 9\} \end{aligned}$$

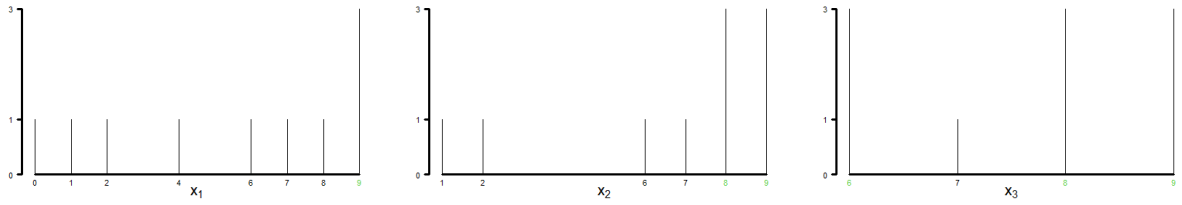

Figure 5: Dotplots of the three variables in Example 9.

The mode of x_1 is $\dot{x}_{1U} = 9$ and it is a unimodal variable; the modes of x_2 are 8 and 9, and we say that x_2 is bimodal; x_3 is multimodal, with modes 6, 8 and 9. □

Recall that the mean is defined for numerical variables only and the median is defined for all non-nominal variables. The mode is defined for any type of variable.

In principle, the interpretation of the mode is quite straightforward. However, some care is needed when interpreting the mode in real practice. Consider, for instance, the following three situations:

- a population of size $N = 10\,000$ with values 1 to 1000 occurring one time each, the value 2000 occurring 4000 times and the value 3000 occurring 5000 times. The left panel of Figure 6 shows a dotplot of such a variable. Strictly speaking one would say that the mode is 3000 and we have a unimodal variable, but this would completely ignore the fact that 2000 is also a remarkable value in terms of its frequency. It is probably more "correct" to say that we have a bimodal variable with 3000 having a larger frequency than 2000.

- a population of size 10 201 with values 1 to 50 occurring 200 times each and the value 100 occurring 201 times (see the central panel of Figure 6). Strictly speaking the mode is 100 and we have a unimodal variable, but once again, this would completely ignore the fact that almost all the observations lie in the interval [1 , 50] and although 100 is the most frequently occurring value it does not step out in terms of frequency with respect to any other value.

- a population of size 10 000 with values 1, 39 to 61 and 100 occurring 400 times each (see the right panel of Figure 6). As all values have the exact same frequency, strictly speaking, there is no mode. But this interpretation ignores the fact that almost all observations lie in the interval [39 , 61].

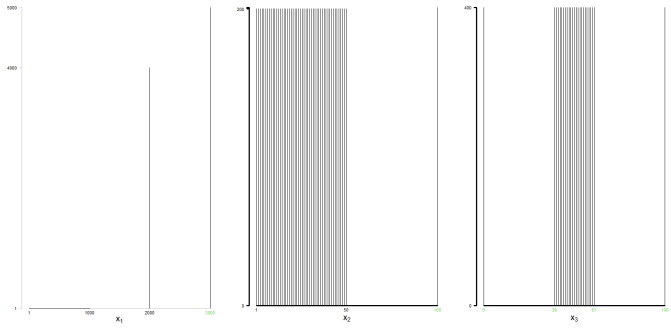

Figure 6: Three situations in which the mode should be read with care.

Although the situations described above are a bit too dramatic, it is not uncommon to have measurements in real life that behave pretty much like that. Especially when the population size is small.

There are (at least) two different views on how to proceed in situations like the ones described above. Some statisticians (including myself) would say that the mode is not an adequate parameter for describing those variables, so it should not be presented. Other statisticians (like the authors of the book) would say that some "flexibility" is needed when interpreting the mode. What matters is that all statisticians agree in the fact that the mode shall not be reported blindly. We must make sure that it is telling something useful and meaningful about the data we are analyzing.

For some strange reason, by default, R does not include a built-in function for calculating the mode of a variable. However, there are several packages that provide functions for this task. For instance, the function `Mode()` from the package `DescTools`.

Percentiles

We close this section by introducing *percentiles* and *quartiles*.

For p in the interval $(0\,,1)$, the $100p$th *percentile* of x, $\breve{x}_{p,U}$, is the value that divides the population in two parts, in such a way that (at least) $100p\%$ of the x-values are smaller or equal than $\breve{x}_{p,U}$ and (at least) $100(1-p)\%$ of the x-values are larger or equal than $\breve{x}_{p,U}$.

The description above defines informally what a percentile is. However, formally there is no unique definition of a percentile. For instance, R includes nine different definitions of a percentile.

Although for small populations the different definitions may yield different results, they all tend to coincide for large populations. The following definition will be used throughout the course.

Definition 10. Let x be a variable that is at least ordinal. Let $c = (N-1)p + 1$, a be the integer part of c and b be the decimal part of c, the pth percentile is

$$\breve{x}_{p,U} \equiv (1-b)x_{(a)} + bx_{(a+1)}, \tag{3}$$

where $x_{(a)}$ and $x_{(a+1)}$ are, respectively, the ath and $(a+1)$th observations in the x-ordered population. □

Let us illustrate the concept of percentiles with an example.

Example 11. Consider the population of ten students in Examples 5. Let us calculate the 90th percentile, $\breve{x}_{0.9,U}$, of the exam points. The x-ordered population is shown in Table 4. As we are looking for the 90th percentile, we have that $p = 0.9$ then

$$c = (N-1)p + 1 = (10-1)0.9 + 1 = 9.1,$$

therefore $a = 9$ and $b = 0.1$, and we get

$$\breve{x}_{0.9,U} = (1-b)x_{(a)} + bx_{(a+1)} = (1-0.1)x_{(9)} + 0.1x_{(9+1)} = 0.9 \cdot 36 + 0.1 \cdot 40 = 36.4.$$

Which means that 90% of the students got less than 36.4 in the exam as illustrated in Figure 7. $\square$

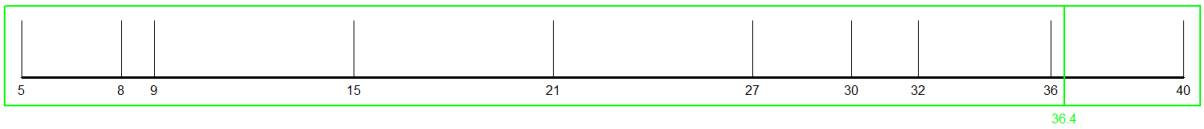

Figure 7: The 90th percentile of the exam points of the students.

Figure 7 gives some intuition about why there are several different definitions of percentiles. Note that any value in the interval $[36\,,40)$ leaves 90% of the observations under it, so any value in this interval can be taken as the 90th percentile.

The function `quantile()` can be used in R for computing the percentiles.