Anna Janicka

Mathematical Statistics 2018/2019
Lecture 1, Parts 1 & 2

1. Descriptive Statistics

By the term descriptive statistics we will mean the tools used for quantitative description of the properties of a *sample* (a given set of information or data, coming from a larger *population*). These tools are purely arithmetical (do not use methods based on the theory of probability), and they are aimed at summarizing or visualizing the properties of the data. The preferred tools and measures depend on the characteristics of the variables that are to be described, and will vary from case to case.

Variables studied with the use of statistical tools may be divided into two main groups: measurable and categorical. The latter group consists of variables which take on values from a limited set of values, representing categories (such as eye color, level of education, sex etc.). The first group consists of variables which take on meaningful, numerical values (that can be measured – such as height, weight etc.).[1] Measurable variables can be further decomposed into continuous (when the value may be a number from a range of real numbers if measured with infinite precision – for example velocity) and count variables (when the possible values are discrete). Typical cases of count variables are the number of children or students enrolled in a class. Some variables which at first sight resemble continuous random variables are also categorical – for example, wages can be measured only up to 1 currency unit, and not with infinite precision. Such variables, called quasi-continuous variables, are treated as continuous in all practical applications .

1.1. Visualizing the data.

We will start our presentation of descriptive statistical tools with count variables, to which one can apply basically all tools. Assume that we look at the grades from a "Mathematical Statistics" course, and that these were equal to:

$$2, 4, 3, 3, 3.5, 2, 4.5, 3, 5, 2.$$

One would need to spend some time in order to be able to say something general about the grades for this course, and we only have 10 students. It would help a little bit if we arranged the numbers in order:

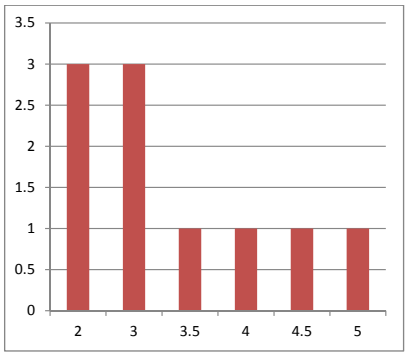
$$2, 2, 2, 3, 3, 3, 3.5, 4, 4.5, 5.$$

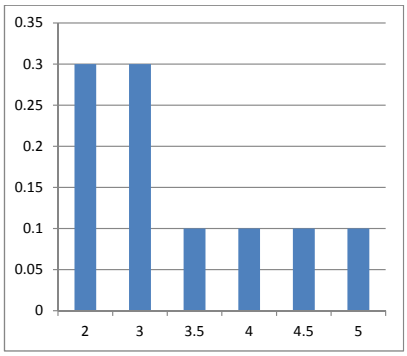
But even now, in case of larger data sets it is evident: just enumerating the values will not be enough to determine at first glance whether, say, it is hard or easy to pass this course. What we can do to better comprehend the properties of the variable under study is to group it. In the case of count variables (and also in the case of categorical variables), the easiest way of grouping is grouping by precise value of the variable. In the case of student grades, we have 6 intuitive groups, corresponding to the following possible outcomes: 2, 3, 3.5, 4, 4.5 and 5. Therefore, we could summarize the course outcome with the use of the following table:

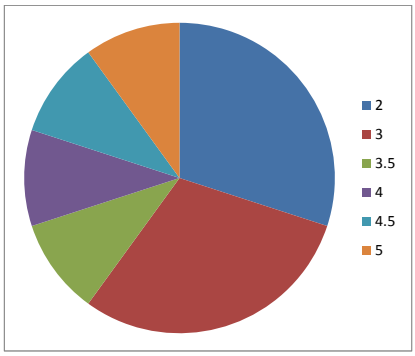
grade	Number of students	Frequency
2	3	30%
3	3	30%
3.5	1	10%
4	1	10%
4.5	1	10%
5	1	10%

[1] Numbers are also possible descriptions of the classes of categorical variables; for example, one could describe the possible outcomes of a coin toss – heads or tails – as outcome 1 and outcome 2, respectively. In this case, however, the values assigned to the two categories are not meaningful and could be changed without loss of our understanding of the phenomenon.

The properties of the data under study become more apparent now: we see that approximately 30% students fail, and that another 30% of students obtain the lowest possible passing grade. We could also visualize the proportions of the particular outcomes graphically. The most commonly used graphs in such cases are the bar chart (with counts – such as the graph with red bars below – or frequencies – such as the graph with blue bars below) and pie chart.







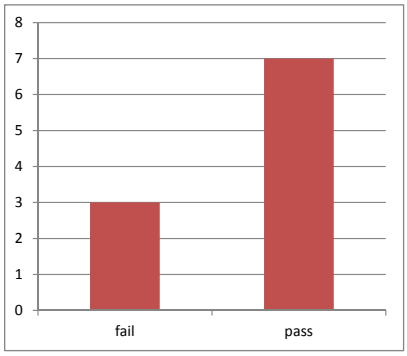
Note that we could have also grouped the data differently, for example as

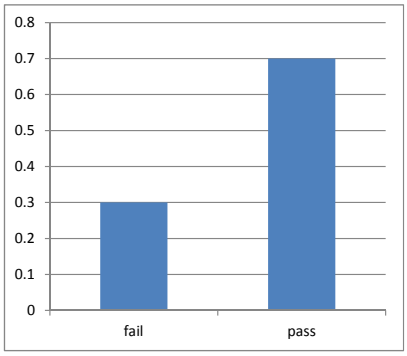
outcome	Number of students	Frequency
fail	3	30%
pass	7	70%

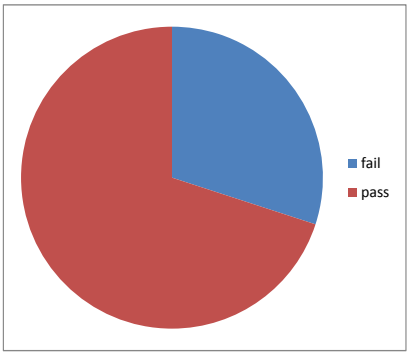
if, say, we were only interested in the failure rate for the course. Note also that this latter representation is also a grouping for the following series of categorical data:

fail, fail, fail, pass, pass, pass, pass, pass, pass, pass.

In the case of categorical data, we can also visualize it graphically by means of bar charts (for numbers or frequencies) and pie charts:







Let us now look at a different example – a continuous variable. Let us assume that we analyze the surface area of 100 apartments available for sale in a given vicinity (in square meters):
32.45, 33.21, 34.36, 35.78, 37.79, 38.54, 38.91, 38.96, 39.5, 39.67, 39.8, 41.45, 41.55, 42.27, 42.4, 42.45, 44.25, 44.5, 44.7, 44.83, 44.9, 45.1, 45.9, 46.52, 47.65, 48.1, 48.55, 48.9, 49, 49.24, 49.55, 49.65, 49.7, 49.9, 50.9, 51.4, 51.5, 51.65, 51.7, 51.8, 51.98, 52, 52.1, 52.3, 53.65, 53.89, 53.9, 54, 54.1, 55.2, 55.3, 55.56, 55.62, 56, 56.7, 56.8, 56.9, 56.95, 57.13, 57.45, 57.7, 57.9, 58, 58.5, 58.67, 58.8, 59.23, 63.4, 63.7, 64.2, 64.3, 64.6, 65, 66.29, 66.78, 67.8, 68.9, 69, 69.5, 73.2, 76.8, 77.1, 77.8, 78.9, 79.5, 82.7, 83.4, 84.5, 84.9, 85, 86, 89.1, 89.6, 93, 96.7, 98.78, 103, 107.9, 112.7, 118.9.

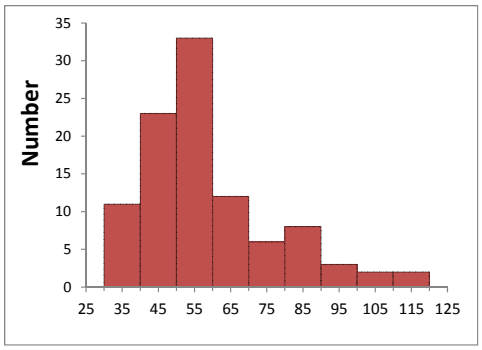
This time, a sensible "visual" analysis of the raw series is impossible to conduct. Constructing a simple frequency table for single values would not lead to better understanding of the phenomenon, as there are no repeated values in the series. In this case, in order to better see the properties, we need to group the series into class intervals. The choice of intervals for grouping is often not an easy one. First of all, it is optimal if the intervals are meaningful: if the government subsidizes the acquisition of apartments not exceeding 75 square meters, it would be preferable that the value of 75 be a bound to an interval, etc. Second, it is better if the interval ranges are of similar length (say, $10m^2$ each). Third, in some cases it is better

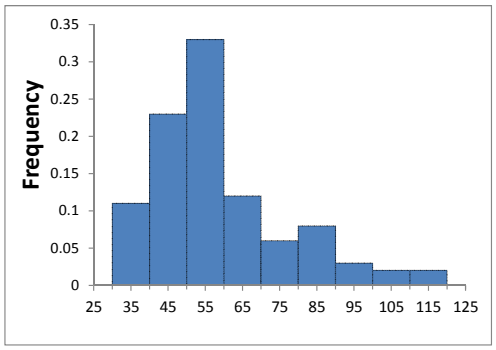
(for computational reasons) if the frequencies in the particular classes are balanced (i.e. we should avoid groupings such that one class has 95 elements and the other has 5 elements, etc.). We should also avoid groupings which are too detailed or not detailed enough. And, last but not least, both for computational ease and visual clarity, it is preferable that the classes have "neat" values.

In the case of our apartment size example, it seems reasonable to group the series into intervals of width equal to 10, starting from 30. In this case, all intervals have the same length and all have "neat" centers (or class marks).

Interval	Class mark $\bar{c}_i$	Number of apartments n_i	Frequency f_i	Cumulative number cn_i	Cumulative frequency cf_i
(30,40]	35	11	0,11	11	0,11
(40,50]	45	23	0,23	34	0,34
(50,60]	55	33	0,33	67	0,67
(60,70]	65	12	0,12	79	0,79
(70,80]	75	6	0,06	85	0,85
(80,90]	85	8	0,08	93	0,93
(90,100]	95	3	0,03	96	0,96
(100,110]	105	2	0,02	98	0,98
(110,120]	115	2	0,02	100	1
Total		100	1		

Based on the grouping, we can now visualize the data with the use of a histogram (the difference between a histogram and the bar chart lies in the horizontal axis – the bars are adjacent to each other, unlike in the previous case, where the categories were separate). One can construct both a histogram of numbers (counts), and frequencies.





Note that in case of continuous variables it is preferable use histograms instead of simple bar charts, and pie charts are usually not a good choice (unless we categorize the variable). One can also visualize the distribution by means of a cumulative frequency histogram or the empirical CDF.

1.2. **Characteristics of the data.** In case of categorical variables, there is not much more that we can do in terms of descriptive statistics to visualize the data. In case of measurable variables, we have a whole array of arithmetical tools that can be used to describe the properties of the data set.

There are two basic distinctions for the characteristics describing the studied variable. The first differentiation is based on the feature of the distribution that we want to describe – whether it is the overall magnitude (how large, on average, are the values of the variable), the variability, the asymmetry etc. The second distinction is based on the values that we will be using for description – whether we will be using different moments of the distribution (in which case we will be talking about classical measures) or measures of position (such as the minimum, maximum, median etc.).

1.2.1. *Measures of central tendency.* Measures of central tendency are those which tell us where – on average – is the "middle" of the distribution located. The basic measures are: the arithmetic mean (the average, a classical measure) or the median and the mode (positional measures). Other measures of position (not necessarily talking about the "middle" of the distribution) include other quantiles (such as quartiles, deciles, percentiles etc.).

If $X_1, X_2, \ldots, X_n$ are the sample values of the variable under study (for example, the raw data for the surface areas of apartments), then the **arithmetic mean** can be calculated as

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

The average value of the surface area of apartments, for the data presented above, would be equal to

$$\bar{X} = \frac{1}{100} (32.45 + 33.21 + 34.36 + \cdots + 112.7 + 118.9) = 59.58.$$

If we are dealing with grouped data (for example, like in the case of student grades above), we can simplify the calculations from the above formula by avoiding summing identical values and using multiplication instead:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{k} X_i n_i,$$

where k is the number of groups into which we have divided our data, X_i are the values in the groups, and n_i are the counts of these groups.

In our example of student grades, we could calculate the average as

$$\bar{X} = \frac{1}{10} (3 \cdot 2 + 3 \cdot 3 + 1 \cdot 3.5 + 1 \cdot 4 + 1 \cdot 4.5 + 1 \cdot 5) = 3.2.$$

In both of these examples, the arithmetic mean was calculated precisely. In some cases, however, we may be faced with a situation where we do not have exact data at our disposal, but only some approximations – as is the case if we are provided with data aggregated into class intervals. In such situations, we will not be able to calculate the true value of the mean, but we will be able to calculate an approximation of the average. This is achieved by treating all observations from a given class as being equal to the middle of the class interval (the so-called class mark, see the headers in the table above), and applying the formula for grouped data, i.e.

$$\bar{X} \cong \frac{1}{n} \sum_{i=1}^{k} \bar{c}_i \cdot n_i.$$

In the apartment surface area example, the approximation of the mean, calculated based on class interval data, would be equal to

$$\bar{X} \cong \frac{1}{100} (35 \cdot 11 + 45 \cdot 23 + 55 \cdot 33 + 65 \cdot 12 + 75 \cdot 6 + 85 \cdot 8 + 95 \cdot 3 + 105 \cdot 2 + 115 \cdot 2) = 58.7.$$

This result is different than the exact amount, which was equal to 59.58. Obviously, the narrower the class intervals used for grouping the data, the less information is lost and the more accurate is the approximation. If raw data are available, they should always be used for calculating characteristics in order to avoid precision losses.

The mean is a good measure to approximate the center of the distribution governing the data, provided that the distribution has an expected value (and, as we know from probability calculus, not all distributions do). Also, if there are outliers (very high or very low values, or erroneous observations) in the data, the average will be affected by these observations. A measure of central tendency which does not have these flaws is the **median**, or the middle observation: (any) number such that at least half of the observations are less than or equal to it and at least half of the observations are greater than or equal to it.

In order to calculate the median (as well as any other measure based on rank), we will need to rearrange our observations in ascending order. We will adopt the following notation: $X_{i:n}$ will be the i-th smallest value of the n element sample (**the i-th order statistic**). In this notation, $X_{1:n}$ is the smallest value (minimum) in the sample, and $X_{n:n}$ is the largest value in the sample (maximum). As for the median, the calculations will depend on whether the sample size is odd or even. If it is odd, then there exists a single "middle" observation; if it is even, there exist two observations "in the middle", and we will take an average of these two as the median value. Therefore,

$$\text{Med} = \begin{cases} X_{\frac{n+1}{2}:n} & \text{if } n \text{ is odd,} \\ \frac{1}{2}\left(X_{\frac{n}{2}:n} + X_{\frac{n}{2}+1:n}\right) & \text{if } n \text{ is even.} \end{cases}$$

Going back to our examples, we can see that:

- In the case of surface apartment area raw data, we have $n = 100$ which is even, so the median will be the average of the 50th and 51st observations:
$$\text{Med} = \frac{1}{2}(55.2 + 55.3) = 55.25.$$
- In the case of grades (grouped data), we need to find the class with the fifth and sixth observations; in our case, it is going to be the class of the 3 grade; therefore, the
$$\text{Med} = \frac{1}{2}(3 + 3) = 3.$$

In the case of grouped class interval data, the situation becomes more complicated. We will not be able to provide a specific value, but only the range into which the median should fall into (this is the interval where the cumulative frequency reaches 0.5 for the first time). If we are interested in a single value, rather than an interval, we will provide an approximation. In order to derive the formula, please note the following. If we know how many observations there are in the sample in the classes before the class of the median, we will know how much "additional" observations from the median class we should take in order to reach the middle observation. Then, knowing how many observations are actually in the median class, and assuming observations are uniformly spaced in the class they fall into, we should have that the median value is proportionally as far in the class interval as the ratio of the number of observations we need to reach it to the number of observations we have in the class. Therefore, we will use the following approximation:

$$\text{Med} \cong c_L + b \cdot \frac{\frac{n}{2} - \sum_{i=1}^{M-1} n_i}{n_M},$$

where M is the number of the class interval of the median, c_L is the lower bound of the class interval of the median, b is the length of the class interval of the median and n_i is the number of observations in the i-th class interval.

In our surface area example, we would have the following: the class in which the 0.5 threshold is reached, is the 3rd class, ie. the (50,60] class. This is going to be the class interval of the median (i.e., $M = 3$). The lower bound of this class is 50, so $c_L = 50$. In the classes before the third class, we have $11 + 23 = 34$ observations, so we need the $\frac{100}{2} - 34 = 16$-th observation from the 3rd class (this is going to be our median). Since the length of the class is $b = 10$, and there are $n_M = 33$ observations overall in this class, the median can be approximated as:

$$\text{Med} \cong 50 + 10 \cdot \frac{50 - 34}{33} \approx 54.85.$$

Please note that this number, again, differs from the true value of 55.25, which means that the approximate formula should be used only in cases where raw data is not available.

In some cases, we may be interested in describing "the middle" of the distribution with the most frequent observation in the sample. This is not always possible – there are distributions which do not have the property of a single most frequent value (for example, if the histogram has several "peaks"). Therefore, the most frequent value – called the **mode** – is usually

only calculated if the data come from a distribution with a "standard" shape, i.e. when the histogram has a single local maximum. The mode is then equal to this single maximum (the most frequent observation in the sample). In the student grades example, there are two equally frequent groups. We would not define a mode in this case.

Please note that for continuous variables, it is not possible to calculate the sample mode unless we group the data – because for continuous distributions, we will not see two observations which are equal to each other, and thus each observation has the same frequency. But in such cases, it is possible to approximate the mode if we group the data first. If we are interested in calculating the mode, we should make sure that the intervals (at least in the "middle" of the distribution) have equal lengths (otherwise, the results of the calculations would be biased – please note that wider intervals will naturally have more observations). Once we have intervals of equal length, we can calculate the mode using the following formula:

$$\mathrm{Mo} \cong c_L + b \cdot \frac{n_{Mo} - n_{Mo-1}}{(n_{Mo} - n_{Mo-1}) + (n_{Mo} - n_{Mo+1})},$$

where, similarly to the formula for the median, we take c_L, the lower bound of the class of the mode, and add to it the appropriate fraction of the length of the class of the mode (b). In this case, the appropriate fraction is calculated as the ratio of the difference between the counts of the class of the mode (n_{Mo}) and the class adjacent to the left (with a count of n_{Mo-1}), to the sum of the differences between the count of the class of the mode and the classes adjacent to the left and to the right (which has a count equal to n_{Mo+1}). This means that if the classes adjacent to the mode are equally less frequent than the class of the mode, we should take the midpoint of the interval as the approximation of the mode. If the distribution is shifted to the left, i.e. smaller observations are more frequent, then the mode should not be in the middle of the interval but more to the left; if the distribution is shifted to the right, i.e. larger observations are more frequent, then the mode should be shifted to the right.

In the case of our surface area example, all class intervals have equal length ($b = 10$), so we can calculate the mode. The class with the largest amount of observations is the third class, so we have $c_L = 50$, $n_{Mo} = n_3 = 33$, $n_{Mo-1} = n_2 = 23$, $n_{Mo+1} = n_4 = 12$, and

$$\mathrm{Mo} \cong 50 + 10 \cdot \frac{33 - 23}{(33 - 23) + (33 - 12)} \approx 53.23.$$

1.2.2. *Other measures of "location".* Additional characteristics, which may be calculated in order to show where the values of the distribution (not necessarily the center of the distribution) are located, include quantiles other than the median. For example, if we calculate the **first and the third quartiles** (i.e., values such that they divide the sample into subsamples counting at least $\frac{1}{4}$ and $\frac{3}{4}$ observations), we will know in what range the "middle" 50% observations are to be found. In order to calculate the quartiles, we will use the same method as for calculating the median; the only difference is that we will be looking for observations which are ranked not (approximately) $\frac{1}{2}n$ out of n, but (approximately) $\frac{n}{4}$ and $\frac{3}{4}n$ out of n.

In particular, in cases where raw data are available, the quartiles will be calculated using the general formula for quantiles of rank p, applied to $p = \frac{1}{4}$ and $p = \frac{3}{4}$. This general formula states that

$$Q_p = \begin{cases} X_{\lfloor np \rfloor + 1:n} & \text{if } np \notin \mathbf{Z} \\ \frac{1}{2}\left(X_{np:n} + X_{np+1:n}\right) & \text{if } np \in \mathbf{Z}. \end{cases}$$

For our apartment surface area example, $\frac{n}{4}$ and $\frac{3}{4}n$ are integer values, so we would take the average of observations numbered 25 and 26 as Q_1, and the average of observations ranked 75 and 76 as Q_3, i.e.

$$Q_1 = \frac{47.65 + 48.1}{2} \approx 47.88, \text{ and } Q_3 = \frac{66.78 + 67.8}{2} \approx 67.29.$$

For grouped class interval data, we will use the same mechanism of determining the approximate value from the appropriate interval that we used for the median, albeit with adjusted counts, i.e.:

$$Q_1 \cong c_L + b \cdot \frac{\frac{n}{4} - \sum_{i=1}^{M-1} n_i}{n_M},$$

and

$$Q_3 \cong c_L + b \cdot \frac{\frac{3}{4}n - \sum_{i=1}^{M-1} n_i}{n_M},$$

where the values of M, c_L, b and n_i are defined analogously as in the case of the median (but for the first and third quartiles, respectively).

For example, if we wanted to calculate the first and third quartiles of the distribution of apartment surface areas, we would search for the observations for which the cumulated frequency reaches 0.25 and 0.75, respectively; we would therefore have that the first quartile is located in the interval $(40, 50]$, while the third quartile is located in the interval $(60, 70]$, and we would have:

$$Q_1 \cong 40 + 10 \cdot \frac{25 - 11}{23} \approx 46.09,$$

and

$$Q_3 \cong 60 + 10 \cdot \frac{75 - 67}{12} \approx 66.67.$$

The two values calculated on the base of grouped data are, again, only approximations of the true values (calculated above).

1.2.3. *Measures of variability.* Once we know where the values of the variable under study are located (more or less), we may wish to determine whether they are concentrated around the center of the distribution, or dispersed. In order to do so, we will use measures of variability. These, too, can be calculated based on moments of the empirical distribution, or on order statistics.

We will start with the latter group. The most simple measure of the variability of a random variable is the **range**, i.e. the difference between the smallest and the largest values observed in the data. In case of grouped class interval data, we take the difference between the lower bound of the lowest interval, and the upper bound of the highest interval. This measure, although simple, has many drawbacks; the most important one is that it is very susceptible to outliers (atypical observations). Therefore, in many cases, instead of this range, we look at the spread between the first and the third quartiles:

$$IQR = Q_3 - Q_1$$

i.e. the **interquartile range** (also called midspread or middle fifty). This measure is much more robust, as it covers the middle 50% observations only. Based on this measure – which depends on the scale of the variable under study – we can calculate coefficients of variation:

$$V_Q = \frac{Q}{\text{Med}}, \quad V_{Q_1Q_3} = \frac{IQR}{Q_3 + Q_1}$$

(where $Q = IQR/2$ is the **quartile deviation**). These coefficients allow us to compare dispersion of different variables.

Examples: In our student grade example, calculating the range does not tell us much – it is equal to $5 - 2 = 3$ and actually does not depend on the distribution of the grades (i.e., on whether the subject is easy or hard to pass). In the surface area example, the range, calculated for raw data is equal to $118.9 - 32.45 = 86.45$, while for grouped class interval data it is equal to $120 - 30 = 90$, and is obviously always biased upwards (the wider the intervals, the more so).

On the other hand, if we calculate the interquartile range, it is equal to $66.67-46.09 = 20.58$. This value is telling, in that it shows that the middle 50% observations are quite concentrated, i.e. half of the the surface areas of apartments are relatively close to the median value.

Turning to the classical measures of dispersion, we will start with the **variance**, which, for raw data, is calculated as

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2 - (\bar{X})^2,$$

for grouped data as

$$\hat{S}^2 = \frac{1}{n}\sum_{i=1}^{k}n_i(X_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{k}n_iX_i^2 - (\bar{X})^2,$$

and for grouped class interval data is approximated as:

$$\hat{S}^2 \cong \frac{1}{n}\sum_{i=1}^{k}n_i(\bar{c}_i - \bar{X})^2 = \frac{1}{n}\sum_{i=1}^{k}n_i\bar{c}_i^2 - (\bar{X})^2.$$

The last formula gives unbiased results if the variable is distributed uniformly – i.e., we expect that the observations classified in intervals are distributed symmetrically around the centers of these intervals. If this assumption is not true – as it is, for example, if the data come from a normal distribution, where values further from the center of the distribution are less common, and we therefore expect values in intervals to be located more to the side – the approximate formula for the variance is going to systematically overestimate the variability in the data. In case of the normal distributions, the following formula for a correction (the so called **Sheppard's correction**) has been proposed:

$$\bar{S}^2 = \hat{S}^2 - \frac{1}{12n}\sum_{i=1}^{k}n_i(c_i - c_{i-1})^2,$$

where c_is denote the bounds of the class intervals. If all the intervals are of equal length, the value of the correction reduces to $\frac{c^2}{12}$, where c is the length of the interval.

Now, if we calculated the variance for the surface area example based on raw data and using the value of the mean also calculated for raw data (i.e. 59.58), we would find that the variance is equal to 333.85. If, on the other hand, we were to use the approximate formula for grouped class interval data, and assuming that the mean was also calculated based on grouped data (and equal to 58.7, approximately), we would have that the approximation of the variance is equal to

$$\begin{aligned}\hat{S}^2 \cong \frac{1}{100}\cdot\Big((35-58.7)^2\cdot 11 + (45-58.7)^2\cdot 23 + (55-58.7)^2\cdot 33 + (65-58.7)^2\cdot 12\\ +(75-58.7)^2\cdot 6 + (85-58.7)^2\cdot 8 - (95-58.7)^2\cdot 3 + (105-58.7)^2\cdot 2 + (115-58.7)^2\cdot 2\Big)\\ \approx 331.31\end{aligned}$$

Please note that this approximation is already smaller than the true value of the variance, and therefore subtracting the Sheppard's correction (equal to $\frac{10^2}{12}$) would just introduce additional error. This is because although the distribution of surface areas is not uniform, it isn't normal, either; also, the sample size may be too small (the errors resulting from small sample size may be larger than the errors arising from class grouping) to use the correction.

Please also note that the variance is a measure expressed in squares of the units of the variable under study. If we wished to have a measure of variability expressed in the same units, we would take the square root of the variance and calculate the **standard deviation**:

$$\hat{S} = \sqrt{\hat{S}^2}, \text{ or } \bar{S} = \sqrt{\bar{S}^2}.$$

In the surface area example, we would have

$$\hat{S} = 18.27$$

Now, if we were interested in comparing the dispersion of different variables for the same population, or the same variable for different populations, we would need a measure of variability which would be invariant to scaling (and units) of the variables. We can construct such a measure, called the **coefficient of variation**, by taking the ratio of the standard deviation and the mean of the variable under study:

$$V_S = \frac{\hat{S}}{\bar{X}}.$$

1.2.4. *Measures of asymmetry.*