01 What descriptive statistics does

Descriptive statistics treats mathematically finite numerical sequences [1] and quantities determined by them. Therefore, from the mathematical point of view, descriptive statistics is a collection of skills based on transformations of sequences. This insight is irrelevant to the way this lecture goes: descriptive statistics is presented via sequences and their transformations.
Elements forming these sequences are obtained in experiments (by direct observation and measurement), in polls (by completing questionnaires, in inquiring, by acquiring answers in surveying), are taken from reports (in particular, statistical yearbooks). This makes that elements of considered sequences are referred to as **observations**. In most cases (and, in practical applications of statistics, in all cases) observations concern a part of the whole population, and this validates to refer to sequences as **samples** (acquired from the total) [2].

Having a sample, there can be set various questions (up to the coincidence of the sample at hand with some pattern). The descriptive statistics limits to the questions which can be answered without the probability. The most simple questions treated in descriptive statistics are that for

0) the size of a sample (how big is a sample ?), minimal, maximal, and the most frequent value a sample is composed of,
1) the average value.

Answers to these questions are classified as characteristics of degree 0 and that of the 1st degree, resp. There are various other problems treated by descriptive statistics, and among them there are questions, ahich are formally classified as that of degree 2, 3 and 4; they concern

2) the variability,
3) the asymmetry,
4) the peakedness,

respectively.

---

[1] A sequence is called numerical if its elements are numbers. In general, sequences must not be numerical ones; for instance, there are sequences of vectors, of functions (both are of mathematical nature), or qualitative sequences. The last mentioned are composed of words expressing degree of some feature. A typical example is composed of adjectives ‘excellent’, ‘very good’, ‘good’, almost good’, ‘sufficient’, ‘bad’; they can be replaced by (or identified with) values 5.0, 4.5, 4.0, 3.5, 3.0 and 2.0, resp., forming the scale of marks in Polish university education. They correspond to the international assessments A, B, C, D, E and F, resp.

[2] For the integrity of the lecture we allow to repeat a little of what was just said. In particular, we list once again the places where observations are gathered.

Parameters of the degrees 0 and 1 are obvious, we all are familiarized to them, we apply them naturally (it means: without any special training, directly from the fact we live), we find them easily. All we need to know the size, the minimal, maximal and most frequent value in a sample is to count and compare. Producing the average of a number sequence is a skill we dominate early: we sum all values a sample at hand is composed of and we divide this sum by the size number of these values (and this quantity is called a size of a sample).

Characteristics of higher degrees came be noticed also from a common observation, but they are not so direct. It appeared that the descriptive statistics worked out a mathematical treatment of these quantities, provided formulae to express more exactly (it means: not only in words such as a lot, almost, very much, but in figures – they do not depend on subjective opinions, they are objective measures assigned to things the concrete formula describes) such notions as the lack of symmetry (how much asymmetric a sample is ?) and the peakedness (how peaky a sample is ?).

Moreover, basic characteristics of degree r (r is any natural number, but we will be interested with $r = 0, 1, 2, 3,$ and 4) can be expressed via quantities called r-th moments of a sample at hand. We will familiarize the concept of a moment of a sample. We will also pay attention that there exist characteristics of, let's say, the asymmetry which are different from the characteristic based on the third moment.

In this book, to keep an order and clearness, sometimes we will give definitions of notions everybody knows (e.g., what an average is). In particular, we do it because of reasons we already mentioned: several phrases are specific to statistics, several notions in statistics have a slightly (or, even, significant) meaning that in they have in mathematics. Let's recall two basic ones: in statistics the terms 'sequence' and 'series" can be finite (while in mathematics every series is always infinite). When properly handled, these differences do not mesh, they are commonly adapted also by mathematicians, at last they understand that the statistics is commonly applied by people (in numerous cases eminent scholars and practitioners) in whose activity the mathematical subtlety is not necessary (what does not want to say that it can be totally neglected).

Many of the parameters are said to be **empirical statistics**, i.e., they are **statistics** (here a 'statistic' means a parameter used to describe a phenomena in the language of the science called 'statistics') and they are determined by information saved as a sample and acquired by observation or experimentation (this word is derived from Greek ἐμπειρία = *empiria*, and Latin *experientia*, direct translation is 'an experience' and it is related to 'an experiment').

Most of parameters of a sample do not depend on the ordering, but there are some of them (in particular, so is the median) which do. In many cases, an arbitrary sample is first made non-decreasing, this pre-processing makes

a sample to be analyzed in clearer way (i.e., the analysis of a non-decreasing sequence is easier, is more readable for a common researcher, some formulas become simpler when concern non-decreasing sequences than in case of non-ordered sequences) and that's why in most cases we perform this pre-processing. The non-decreasing sample obtained from a given sample y by the ordering of its elements is called an **ordered sample** of y, a **non-decreasing sample** of y, or an **ordered version** of y, an **ordeence** (ordered sequence) of the sequence y.

In lectures there will be introduced and considered several versions of sequences, several ways to jot down sequences. These versions, these records have to be fully identified, and this requirement makes some identifiers having relatively long names. In aim to shorten them, as well as to make the text more readable, there are proposed some acronyms composed of parts of appropriate words [3], namely (just mentioned) ordeence (ordered sequence), valuence (value sequence), multence (multiplicity sequence), freqence (frequency sequence).

To make the lecture as easy as possible, we keep following meaning of letters

$y = (y_1, y_2, \ldots, y_N)$ – a given sequence (consisted of N elements),
$z = (z_1, z_2, \ldots, z_N)$ – the ordeence of y,
$\{ z_1, z_2, \ldots, z_N \}$ – the bag of y,
$\{ x_1, x_2, \ldots, x_n \}$ – the value set of y (it has n different elements),
$x = (x_1, x_2, \ldots, x_n)$ – the sequence of values which are present in y,
$m = (m_1, m_2, \ldots, m_n)$ – the sequence of corresponding multiplicities,
$f = (f_1, f_2, \ldots, f_n)$ – the sequence of corresponding frequencies,

all y_k, z_k, x_j, m_j, f_j, are numbers,
$m_j \in \boldsymbol{N}$,
$z_j \le z_k$ for $j < k$, (and this non-decreasing order is kept
when we list elements of corresponding sets, so $x_j \le x_k$ for $j < k$),
$0 < f_j < 1$ (or $f_j = 1$ and it can take place only in the one-point distribution).

Any departure form above convention is expressively announced.

All above structures are generated by the first listed one, namely by the sequence y, also referred to as a(n empirical) sample. Any its element, y_j ($j = 1..n$), is said to be a j-th **result of an experiment**, a j-th **outcome of an experiment**, a j-th **observation** (noticed in the experiment at hand). Values $y_1, y_2, \ldots, y_N$ are

---

[3] Examples of such acronyms are BC and AD (to identify eras: before the year 0 and after the year 0, resp.), a.m. and p.m. (before the midday, after the noon, resp.), aka (also known as), wrt (with respect to), PIN (personal identification number, originated in 1967 as an efficient way for banks to authenticate customers taking cash), Interpol (International Criminal Police Organization, founded in 1923), poset (partially ordered set, the word introduced by Garret Birkhoff in his book *Lattice theory*, 1890). We follow the portmanteau, the contraction John Tukey used when, in 1847, created the word bit (binary digit)

- obtained in an experiment (e.g., aimed to state what is the value of the thermal elongation coefficients of the specific rode),
- by inquiring (persons on a subject under investigation, e.g., their opinion on the quality of their life),
- by taking data from statistical yearbooks (e.g., on the number of cars produced in various countries, on how regions are the forested in percentage of their total area) or other documents (e.g., from a payrolls in an enterprise).

Let's keep an eye for the last mentioned example. We have an access to the payroll in a firm *We20*, where there are employed 20 people. This list is alphabetically ordered, and we read, let's say,

1. Adamski Jan – 3500,
2. Brown Victoria – 6400,
3. Dudek Henryk – 2900,

…

20. Zetowski Marek – 7000

In statistics we are concerned in this book in we pay no attention to personal issues which can be deduced form this payroll, we are not interested in answers to such questions as "Who is best paid and why ?" (we can guess, with the certainty 100%, that the best paid is the boss, and that the reason for it is because he/she is a boss). We are interested in a structure, in a distribution of wages, the questions we set are of how-many-type, for example: "How many employees get salaries between 3 and 4 thousand zlotys ?", "How many workers gain less than the average salary ?". Therefore, we forget names and what we deal with is the sequence as the following one

$$y = (y_k)_{k=1..N} = (3.5, 6.4, 2.9, 2.0, 2.1, 2.2, 3.3, 4.3, 2.2, 2.9,$$
$$2.0, 2.9, 2.9, 3.1, 3.3, 10.0, 3.3, 3.5, 3.8, 7.0),$$

where there are listed month salaries, expressed in thousand zlotys, read out the payroll the four position of which are presented above.

Identically we (can) look at, for example, data on the GDPpc (the gross domestic product per capita and year, let's say that expressed in dollars): the task we undertake is to recognize the distribution of the wealth (GDPpc is one of possible measures of it, and it is widely accepted) among various countries, so we ask, e.g., "How many countries have their GDPpc greater than 70 000 int$?", "How many have it within the range 23-25 thousand Int$?" [4]. Obviously,

[4] After data provided by the International Monetary Fund, in 2013 there five countries where the GDPpc surpassed 70 thousand international dollars (Qatar, Luxemburg, Singapore, Brunei and Kuwait, with 146, 90, 79, 74 and 71 thousand int$; they are leaders in GDPpc, the next country is Norway with 64 000 int$), and GDPpc in said range was in 6 countries: Kazakhstan, Malaysia, Hungary, Poland, Seychelles, Russia and the Bahamas (23.0, 23.2, 23.2, 23.3, 23.5, 24.3 and 24.7 thousand intS, resp.). This implies that Poland is 48th richest country (against 187 listed).

the questions such as “What are top GDPpi countries ?”, “How Poland ranges wrt GDPpi ?” are intriguing, or even of a great importance, but they do not belong to those statistics deal with; they are of political type, they be corrected answered and provide appropriate conclusions thanks to statistical sight.

In general case y is also called a sample, but it has not be empirical only. An other type of samples are so-called theoretical distributions, sequences whose elements are produced by mathematical formulas that are deduced as appropriate to describe concrete phenomena (such as results of rolling a dice, choosing an ace out form the deck). Theoretical distributions can be seen as idealized empirical distribution (e.g., theoretically a dice casted 120 times should show six points in 20 cases, but it happens so really rarely), they serve as patterns (so plays an analogue role as an ideal being a pattern to a real) and they are basic objects considered in inferential statistics.