REVIEW ARTICLE

Descriptive Statistics

The Specification of Statistical Measures and Their Presentation in Tables and Graphs

Part 7 of a Series on Evaluation of Scientific Publications

Albert Spriestersbach, Bernd Röhrig, Jean-Baptist du Prel, Aslihan Gerhold-Ay, Maria Blettner

SUMMARY

Background: Descriptive statistics are an essential part of biometric analysis and a prerequisite for the understanding of further statistical evaluations, including the drawing of inferences. When data are well presented, it is usually obvious whether the author has collected and evaluated them correctly and in keeping with accepted practice in the field.

Methods: Statistical variables in medicine may be of either the metric (continuous, quantitative) or categorical (nominal, ordinal) type. Easily understandable examples are given. Basic techniques for the statistical description of collected data are presented and illustrated with examples.

Results: The goal of a scientific study must always be clearly defined. The definition of the target value or clinical endpoint determines the level of measurement of the variables in question. Nearly all variables, whatever their level of measurement, can be usefully presented graphically and numerically. The level of measurement determines what types of diagrams and statistical values are appropriate. There are also different ways of presenting combinations of two independent variables graphically and numerically.

Conclusions: The description of collected data is indispensable. If the data are of good quality, valid and important conclusions can already be drawn when they are properly described. Furthermore, data description provides a basis for inferential statistics.

Key words: statistics, data analysis, biostatistics, publication

Cite this as: Dtsch Arztebl Int 2009; 106(36): 578–83
DOI: 10.3238/arztebl.2009.0578

Institut für Medizinische Biometrie, Epidemiologie und Informatik, Johannes Gutenberg-Universität Mainz: Prof. Dr. rer. nat. Blettner, Spriestersbach, Gerhold-Ay

Zentrum Präventive Pädiatrie, Zentrum für Kinder- und Jugendmedizin, Universitätsmedizin der Johannes Gutenberg-Universität Mainz: Dr. med. du Prel, M.P.H.

MDK Rheinland-Pfalz, Referat Rehabilitation/Biometrie, Alzey: Dr. rer. nat. Röhrig

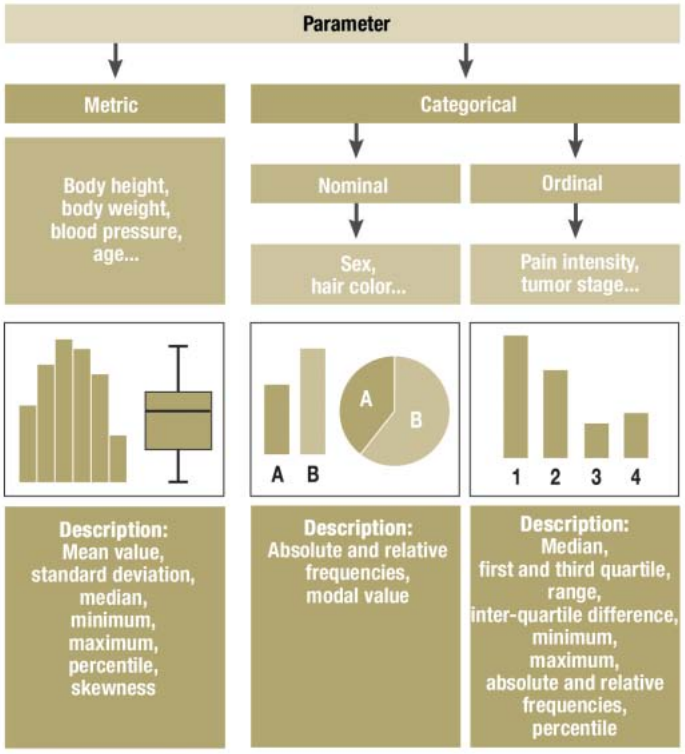
A set of medical data is based on a collection of the data of individual cases or objects, also called observation units or statistical units. Every case, for example every study participant, patient, every experimental animal, every tooth or every cell shows comparable parameters (such as body weight, gender, erosion, pH). Each of these parameters, also called variables, has a specific parameter value (gender = male, age = 30 years, weight = 70 kg) for each observation unit (for example the patient). The aim of descriptive statistics is to summarize the data, so that they can be clearly illustrated (1–3).

The property of a parameter is specified by its so-called scale of measure. Generally two types of parameters are distinguished. A variable has a metric level (= quantitative data) if it can be counted, measured or weighed in a physical unit (as in cm or kg) or at least can be recorded in whole numbers. Data with a metric scale of measure can be further classified into continuous and discrete variables. In contrast to discrete variables, continuous variables can take any value. Examples for metric continuous parameters are body height in cm, blood pressure in mmHg or the creatinine concentration in mg/L. One example for a metric discrete parameter is the number of erythrocytes per microliter of blood.

The gender of man cannot be measured, but is classified into two categories. Parameters which can be classified into two or more categories are described as categorical parameters (= qualitative data). A further classification of a categorical parameter is into nominal characteristics (unordered) and ordinal characteristics (ordered according to rank). Good basic portrayals of the descriptive statistics of medical data can be found in text books (4–9). *Figure 1* gives a review of types of parameters, as well as graphs to be used and statistical measures.

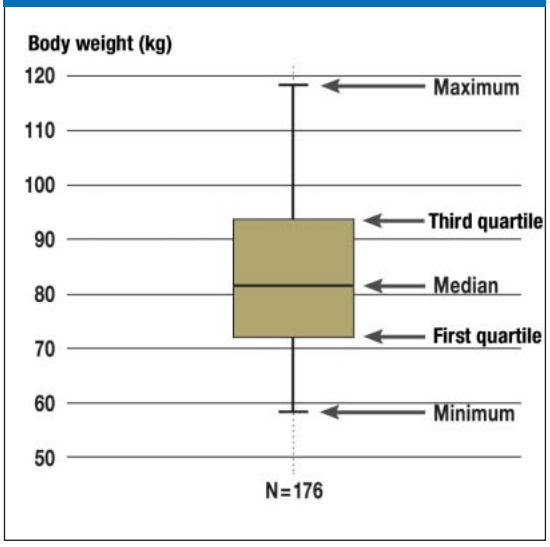
Different procedures are necessary for the statistical evaluation of metric and categorical parameters in graphic and tabular forms. The graphs used here and the evaluation tables were created with the statistics package SPSS for WINDOWS (Version 15). As example, we are using data of 176 sportsmen and women.

FIGURE 1



Sketch of parameter types and suitable statistical measures for descriptive presentation

FIGURE 2



Example for a box plot

Results

1. Description of a continuous parameter

A continuous parameter is a quantitative measure. The value is measured on a continuous scale in arbitrarily small intermediate steps (3). In this article, quantitative parameters are treated as continuous parameters. The size of the continuous parameters is presented by a measurement unit (= size unit, physical unit), for example body height in cm or body weight in kg.

Initially, graphic presentations of the distribution of a continuous variable are created in the form of box plots and histograms. Such diagrams make it possible for the researcher to get an initial visual impression of the distribution of the collected parameters.

1.1. Graphic presentation of a continuous parameter

1.1.1. The box plot diagram

Box plots offer a visual impression of the position of the first and third quartiles (25th and 75th percentile) and of the median (central value). Also minimum, maximum, and the breadth of scatter of all case values of a continuous parameter are recognizable. 50% of the values of distribution are within the box (= interquartile range). A box with a greater interquartile range indicates greater scatter of the values. *Figure 2* shows an example of the distribution of body weight in kg in 176 sportsmen and women.

1.1.2. The histogram

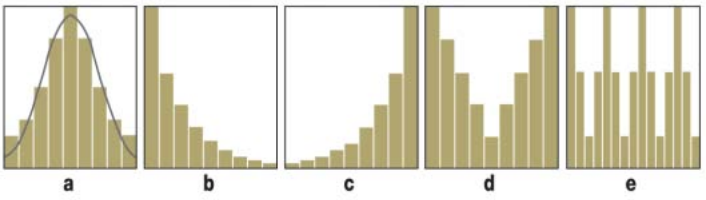
A histogram shows the distribution form of the measured values of a continuous variable. In a normal distribution, this takes the form of a "Gauss bell curve" *(Figure 3a)*. The present (measured) values are classified into an appropriate number of classes (3). If the number of classes is not "naturally" given, it is recommended to choose the number of classes as the square root of the case number N. If you had, for example, 49 cases, seven classes would be chosen for the histogram on which the measured values are distributed. Within each class, the measured values are counted and illustrated as column in the figure. *Figure 3* shows five schematic examples for distribution forms in each histogram.

In a histogram it is recognizable whether the data are symmetrically distributed around the mean value *(Figure 3a)*. However, when the histogram has a left peak *(Figure 3b)* or a right peak *(Figure 3c)*, the values have a "skew" distribution. In some situations, it may happen that several peaks are recognizable in a histogram *(Figure 3d and e)*.

1.2. Numerical description of a continuous parameter

The distribution of a recorded continuous parameter can be numerically described with the following statistical measures: Minimum, maximum, quartiles (with median), range (difference between maximum and minimum), skewness (indicates whether the distribution is symmetric or not), arithmetic mean value, and standard deviation (= square root of the variance) (6). When the parameter "skewness" is between –1 and +1, the distribution is

FIGURE 3



Examples of the types of distribution in histograms
a) Normal distribution (symmetric)
b) left peak (= skewed to the right)
c) right peak (= skewed to the left)
d) two peaks (symmetric)
e) several peaks

symmetric, but when it is under –1 or above +1 the values are distributed with a right or a left peak *(Box)*.

2. Description of a categorical parameter

2.1. Graphic presentation of a categorical parameter

2.1.1. The pie diagram as graphic presentation of a categorical parameter

The pie diagram or circular diagram is a popular form of presentation for the distribution of characteristics of a parameter classified into groups. The number of segments in one pie diagram corresponds to the number of possible characteristics (= steps) of these variables. So, a pie diagram would have two segments for "sex." Their proportion in the total pie corresponds to their relative percentage.

BOX

Mean value	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance	$Var = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standard deviation	$s = \sqrt{Var}$
Skewness	$g = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{s^3}$
Median	$\tilde{x} = x_{(n+1)/2}$ If n is odd
	$\tilde{x} = \frac{x_{(n/2)} + x_{(1+n/2)}}{2}$ If n is even
Range	$R = x_{max} - x_{min}$

with
n = Sample size or case number
x_i = Measured value for i-th sample element or i-th case, where $i=1,\ldots,n$
$x_{(i)}$ = Describes the i-th value in the ascending order of the measured values, where $i=1,\ldots,n$

2.1.2. The bar diagram as graphic presentation of a categorical parameter

The bar diagram offers an alternative form of presentation. In contrast to the pie diagram, the frequency values are read on the y-axis. This figure can present absolute or relative frequencies. It is possible to compare the heights of the bars directly, which is not possible with the segments in the pie diagram. In contrast to the histogram for continuous parameters, there are no class intervals for the value ranges of the measurements on the x-axis of the bar diagram. On the contrary, each bar creates one unit completed to the right and to the left, according to its value. Thus, one single bar refers either to the female or to the male sex in *Figure 4*. In contrast to the histogram, the single bars or columns of a bar diagram should contain gaps.

2.2. Numerical description of a categorical parameter

Both absolute and relative frequencies of a categorical parameter can be given in a frequency table. In *Table 1*, we show our population of sportswomen and sportsmen once again: In this SPSS Version of a frequency table, you can find the absolute frequencies in the "frequency" column. If given, the number of missing values is listed here. In the columns "percentage" and "valid percentages," you can choose whether the missing values should be listed or not as a separate category. In the column "cumulated percentages," the successive cumulated relative frequencies are presented. They are only important in parameters of ordinal scale level with more than two values. The two last columns ("valid and cumulated percentages") are generally not suitable for presentation in a publication.

3. Description of correlations

Until now, we have only considered single variables, i.e., we have described our data as "univariate." No parameter was related to any other. It is also possible to describe potential correlations between two variables, for example between body weight (continuous) and body height (continuous). It is also possible to relate two categorical parameters or a metric and a categorical parameter.

3.1. Description of the correlation of two continuous parameters

One of the variables is allocated to each of the two axes in the scatter plot (point cloud diagram). If there is a correlation, it is expressed in the trend of the point cloud to form an ellipse. If there were a 100 percent linear correlation between two parameters, all points would lie on a straight line. In our example (10), the measured values of bone density on the proximal site of measurement (variable: 'spa_prox') and the values of bone density on the distal site of measurement (variable: 'spa_dist') are correlated with each other *(Figure 5)*. If there were no correlation, the area of the diagram would be unstructured and covered with points.

The degree of correlation can be numerically expressed as linear correlation with a coefficient of correlation. This index can have values between –1 and +1. If both distributions are symmetric, the coefficient of correlation

can be calculated according to Pearson. Otherwise, the coefficient of correlation according to Spearman is appropriate. This is not directly calculated from the values but from their ranks. In linear distribution, the coefficient of correlation should be calculated according to Pearson. The Pearson coefficient of correlation is calculated as 0.886 for the measurement of bone density. This is a very strong correlation.

3.2. Description of the correlation of two categorical parameters

A grouped bar diagram is suitable for the graphical presentation of the correlation between two categorical parameters. The standard form uses the x-axis for entry of the single categories of the first variable and portrays within these categories the absolute (number) or relative frequencies (percentage) of the second variable through different colored bars. In *Figure 6*, a possible correlation between smokers and (smokers') cough is to be studied. The bars represent relative frequencies in percentages. They are more suitable for comparison than the absolute frequencies, as the groups are often not of the same size. In this figure, a bar diagram is presented and it is recognizable that smokers (with 71%) more often have cough than non-smokers (29%).

The numerical description of information about absolute and relative frequencies of the two correlated categorical variables is summarized in a cross table. We differentiate between line variables (smokers and non-smokers) and column variables (cough). It is possible to give the absolute and relative frequencies, i.e. line and column percentages, within the cells *(Table 2)*.

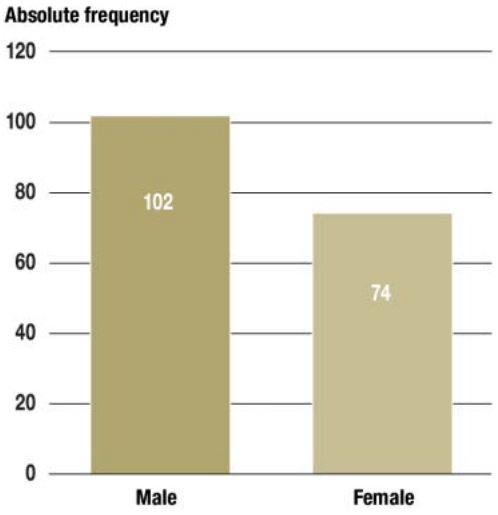
You can see from this presentation that 21/29 = 72% of non-smokers do not cough but 8/29 = 28% cough. The cumulative percentage for all lines is 100% and the cumulative percentage for all columns is also 100%. The four field table helps the reader to understand the reference values and the results.

To simplify the interpretation, it should be made clear which variable is the target parameter. In our example, it was important to know whether smoking influences the occurrence of cough. Thus, cough is the (dependent) target parameter and smoking the (independent) possible factor. In the interest of clarity of the cross table, it is recommended to write the factor in the cells and the target parameter in the columns. Then you can dispense with the percentages in the columns. In this way, the cross table is clearer and easier to understand. In practice, medical expertise is required to be able to formulate a sensible question. This results in the arrangement of the variables. Further parameters which can be calculated from the four field table (relative risk, odds ratio) will be discussed in a later publication.

3.3. Correlation between a continuous and a categorical parameter

Several box plots can be presented in one diagram with the help of a statistics program like SPSS, STATISTICA, or SAS. Thus, it is possible to compare groups for which for example the continuous parameter "weight in kg"

FIGURE 4



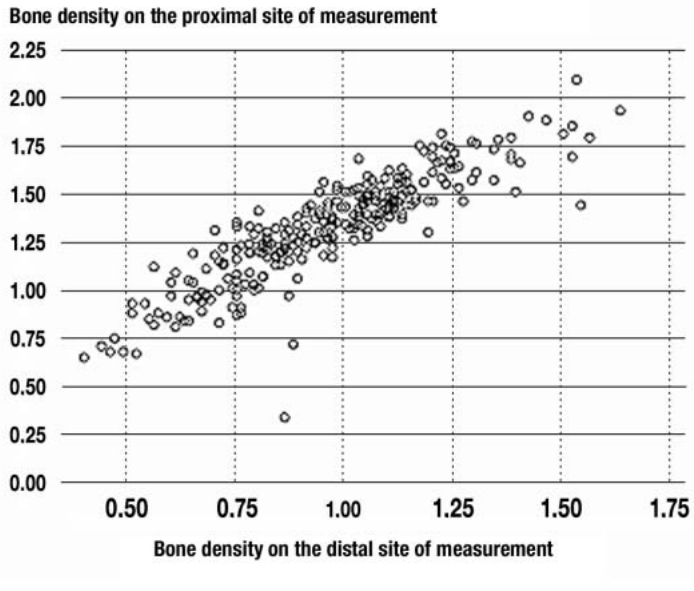
Example of a bar diagram of two groups

TABLE 1

Sex

		Frequency	Percentage	Valid percentages	Cumulated percentages
Valid	Male	102	58.0	58.0	58.0
	Female	74	42.0	42.0	100.0
	Total	176	100.0	100.0	

FIGURE 5



Example of a scatter diagram

Example of a grouped bar diagram

FIGURE 6

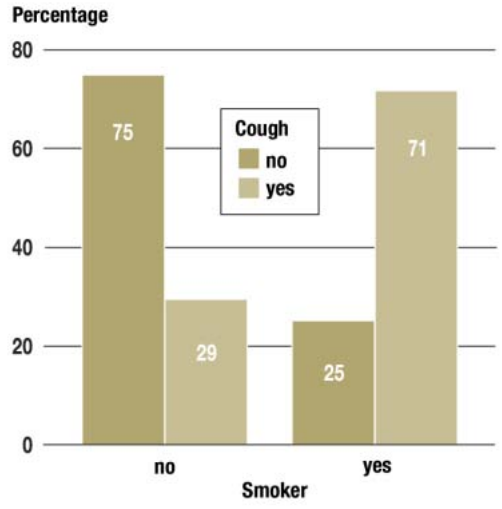


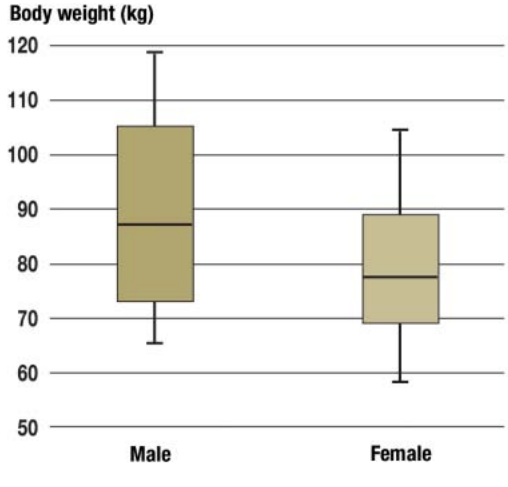
TABLE 2

Distribution of the parameters smoking and cough

			Cough		Total
			No	Yes	
Smokers	No	Number	21	8	29
		% of smokers	72%	28%	100%
		% with cough	75%	29%	52%
	Yes	Number	7	20	27
		% of smokers	26%	74%	100%
		% with cough	25%	71%	48%
Total		Number	28	28	56
		% of smokers	50%	50%	100%
		% with cough	100%	100%	100%

Example of grouped box plots

FIGURE 7



was recorded. In *Figure 7*, the distribution of the weight of 74 sportswomen was compared with the distribution of the weight of 102 sportsmen. When comparing the two groups, a trend is recognizable: Men tend to have higher weights than women (consideration of medians). The values are more scattered with men than with women, as the box plot for men is more stretched than for women.

As described in chapter 1.2., suitable measures are numerically calculated for the comparison of the groups of the two sexes.

Discussion

The exact description of data collected in a study is sensible and important. The correct descriptive presentation of the results is the first step in evaluating and graphically presenting the results (7–9, 11). The description is the basis of the biometric evaluation and is the indispensable starting point for further methodological procedures such as statistical significance tests. The descriptive presentation of study results usually occupies most of the space in publications. The description covers the graphical and tabular presentation of the results. The exact assessment of the scale level of the parameter is important as the scale level determines the type and procedure, both in the descriptive, and in the explorative (= generating of hypothesis) and confirmatory (= biometric testing of hypothesis) evaluation. The selection of a suitable statistical test procedure for controlling the significance is determined by the scale level of the investigated parameters.

In normally distributed data, the arithmetic mean value is the same as the median. The skewness has the value zero. Unfortunately there rarely is a normal distribution in natural systems as in parameters collected in patients. It is sensible to give the arithmetic mean value, as well as the median for continuous data. A normal distribution cannot be assumed when the two values are very different. The arithmetic mean value cannot be calculated in data with a purely ordinal scale. It is often asked whether graphical or numerical presentations are preferable in data description. Graphics serve to give a first impression and to visually illustrate the situation of distribution parameters. It can be difficult to read the exact values of the median or the percentiles on the y-axis in a box plot diagram. For this reason, calculation and presentation of the exact statistical characteristic values is indispensable.

In the individual case, the information of further biometric measures is of course valuable, including measures not mentioned in the article. Examples would be effect sizes, confidence intervals, Cohen's kappa, relative risk, and cumulated values.

The use of suitable validated statistics software like SPSS or SAS is recommended for statistical evaluation of data.

Conflict of interest statement
The authors declare that there is not conflict of interest according to the guidelines of the International Committee of Medical Journal Editors.

Manuscript received on 4 February 2009, revised version accepted on 16 March 2009.

Translated from the original German by Rodney A. Yeates, M.A., Ph.D.

REFERENCES

1. Greenfield MLVH, Kuhn JE, Wojtys EM: A statistics primer: descriptive measures for continuous data. Am J Sports Med 1997; 25: 720–3.
2. McHugh ML: Descriptive statistics, part I: level of measurement. JSPN 2003; 8: 35–7.
3. Overholser BR, Sowinski KM: Biostatistics primer: part I. Nutr Clin Pract 2007; 22: 629–35.
4. SPSS Incorporated: SPSS 16.0 Schneller Einstieg. Dublin: SPSS Inc. 2007; 55–62.
5. Bortz J: Statistik für Sozialwissenschaftler. Berlin Heidelberg New York: Springer 1999; 5. Auflage: 17–47.
6. Sachs L: Angewandte Statistik: Anwendung statistischer Methoden. Berlin, Heidelberg, New York: Springer 2004; 11. Auflage: 1–177.
7. Trampisch HJ, Windeler J: Medizinische Statistik. Berlin, Heidelberg, New York: Springer 2000; 2. Auflage: 52–82.
8. Hilgers RD, Bauer P, Schreiber V: Einführung in die Medizinische Statistik. Berlin, Heidelberg, New York: Springer 2003; 3–43.
9. Altman DG: Practical statistics for medical research. Boca Raton, London, New York, Washington D.C.: Chapman & Hall/CRC 1999; 10–45.
10. Zawalski R: Messung der Hautfaltendicke am Handrücken mit Hilfe einer Mikrometerschraube [Dissertation]. Mainz: Fachbereich Medizin der Johannes Gutenberg-Universität; 1997.
11. Du Prel JB, Röhrig BBM: Kritisches Lesen wissenschaftlicher Artikel. Dtsch Arztebl Int 2009; 106: 100–5.

Coresponding author
Prof. Dr. rer. nat. Maria Blettner
Institut für Medizinische Biometrie,
Epidemiologie und Informatik
Johannes Gutenberg-Universität
55101 Mainz, Germany
sprieste@mail.uni-mainz.de