Presenting data in tables and charts*

Rodrigo Pereira Duquia[1] João Luiz Bastos[2] Renan Rangel Bonamigo[1]
David Alejandro González-Chica[2] Jeovany Martínez-Mesa[3]

DOI: http://dx.doi.org/10.1590/abd1806-4841.20143388

Abstract: The present paper aims to provide basic guidelines to present epidemiological data using tables and graphs in Dermatology. Although simple, the preparation of tables and graphs should follow basic recommendations, which make it much easier to understand the data under analysis and to promote accurate communication in science. Additionally, this paper deals with other basic concepts in epidemiology, such as variable, observation, and data, which are useful both in the exchange of information between researchers and in the planning and conception of a research project.

Keywords: Epidemiology; Epidemiology, descriptive; Tables

INTRODUCTION

Among the essential stages of epidemiological research, one of the most important is the identification of data with which the researcher is working, as well as a clear and synthetic description of these data using graphs and tables. The identification of the type of data has an impact on the different stages of the research process, encompassing the research planning and the production/publication of its results. For example, the use of a certain type of data impacts the amount of time it will take to collect the desired information (throughout the field work) and the selection of the most appropriate statistical tests for data analysis.

On the other hand, the preparation of tables and graphs is a crucial tool in the analysis and production/publication of results, given that it organizes the collected information in a clear and summarized fashion. The correct preparation of tables allows researchers to present information about tens or hundreds of individuals efficiently and with significant visual appeal, making the results more easily understandable and thus more attractive to the users of the produced information. Therefore, it is very important for the authors of scientific articles to master the preparation of tables and graphs, which requires previous knowledge of data characteristics and the ability of identifying which type of table or graph is the most appropriate for the situation of interest.

BASIC CONCEPTS

Before evaluating the different types of data that permeate an epidemiological study, it is worth discussing about some key concepts (herein named data, variables and observations):

Data – during field work, researchers collect information by means of questions, systematic observations, and imaging or laboratory tests. All this gathered information represents the data of the research. For example, it is possible to determine the color of an individual's skin according to Fitzpatrick classification or quantify the number of times a person uses sunscreen during summer.[1,2] All the information collected during research is generically named "data." A set of individual data makes it possible to perform statistical analysis. If the quality of data is good, i.e., if the way information was gathered was appropriate, the next stages of database preparation, which will set the ground for analysis and presentation of results, will be properly conducted.

Observations – are measurements carried out in one or more individuals, based on one or more variables. For instance, if one is working with the variable "sex" in a sample of 20 individuals and knows the exact amount of men and women in this sample (10 for each group), it can be said that this variable has 20 observations.

Received on 02.12.2013.
Approved by the Advisory Board and accepted for publication on 02.02.2014.
* Work performed at the Dermatology service, Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Departamento de Saúde Pública e Departamento de Nutrição da UFSC.
Financial Support: None
Conflict of Interest: None

[1] Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA) – Porto Alegre (RS), Brazil.
[2] Universidade Federal de Santa Catarina (UFSC) – Florianópolis (SC) Brazil.
[3] Latin American Cooperative Oncology Group (LACOG) – Porto Alegre (RS) Brazil.

Variables – are constituted by data. For instance, an individual may be male or female. In this case, there are 10 observations for each sex, but "sex" is the variable that is referred to as a whole. Another example of variable is "age" in complete years, in which observations are the values 1 year, 2 years, 3 years, and so forth. In other words, variables are characteristics or attributes that can be measured, assuming different values, such as sex, skin type, eye color, age of the individuals under study, laboratory results, or the presence of a given lesion/disease. Variables are specifically divided into two large groups: **(a)** the group of categorical or qualitative variables, which is subdivided into dichotomous, nominal and ordinal variables; and **(b)** the group of numerical or quantitative variables, which is subdivided into continuous and discrete variables.

Categorical variables

a) Dichotomous variables, also known as binary variables: are those that have only two categories, i.e., only two response options. Typical examples of this type of variable are sex (male and female) and presence of skin cancer (yes or no).

b) Ordinal variables: are those that have three or more categories with an obvious ordering of the categories (whether in an ascending or descending order). For example, Fitzpatrick skin classification into types I, II, III, IV and V.[1]

c) Nominal variables: are those that have three or more categories with no apparent ordering of the categories. Example: blood types A, B, AB, and O, or brown, blue or green eye colors.

Numerical variables

a) Discrete variables: are observations that can only take certain numerical values. An example of this type of variable is subjects' age, when assessed in complete years of life (1 year, 2 years, 3 years, 4 years, etc.) and the number of times a set of patients visited the dermatologist in a year.

b) Continuous variables: are those measured on a continuous scale, i.e., which have as many decimal places as the measuring instrument can record. For instance: blood pressure, birth weight, height, or even age, when measured on a continuous scale.

It is important to point out that, depending on the objectives of the study, data may be collected as discrete or continuous variables and be subsequently transformed into categorical variables to suit the purpose of the research and/or make interpretation easier. However, it is important to emphasize that variables measured on a numerical scale (whether discrete or continuous) are richer in information and should be preferred for statistical analyses. Figure 1 shows a diagram that makes it easier to understand, identify and classify the abovementioned variables.

DATA PRESENTATION IN TABLES AND GRAPHS

Firstly, it is worth emphasizing that every table or graph should be self-explanatory, i.e., should be understandable without the need to read the text that refers to it refers.

Presentation of categorical variables

In order to analyze the distribution of a variable, data should be organized according to the occurrence of different results in each category. As for categorical variables, frequency distributions may be presented in a table or a graph, including bar charts and pie or sector charts. The term *frequency distribution* has a specific meaning, referring to the the way observations of a given variable behave in terms of its absolute, relative or cumulative frequencies.

In order to synthesize information contained in a categorical variable using a table, it is important to count the number of observations in each category of the variable, thus obtaining its absolute frequencies. However, in addition to absolute frequencies, it is worth presenting its percentage values, also known as relative frequencies. For example, table 1 expresses, in absolute and relative terms, the frequency of acne scars in 18-year-old youngsters from a population-based study conducted in the city of Pelotas, Southern Brazil, in 2010.[3]

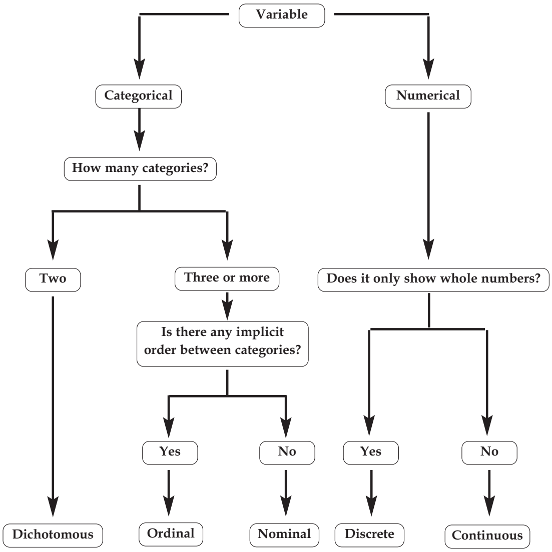

FIGURE 1: Types of variables

The same information from table 1 may be presented as a bar or a pie chart, which can be prepared considering the absolute or relative frequency of the categories. Figures 2 and 3 illustrate the same information shown in table 1, but present it as a bar chart and a pie chart, respectively. It can be observed that, regardless of the form of presentation, the total number of observations must be mentioned, whether in the title or as part of the table or figure. Additionally, appropriate legends should always be included, allowing for the proper identification of each of the categories of the variable and including the type of information provided (absolute and/or relative frequency).

Presentation of numerical variables

Frequency distributions of numerical variables can be displayed in a table, a histogram chart, or a frequency polygon chart. With regard to discrete variables, it is possible to present the number of observations according to the different values found in the study, as illustrated in table 2. This type of table may provide a wide range of information on the collected data.

Table 2 shows the distribution of educational levels among 18-year-old youngsters from Pelotas, Southern Brazil, with absolute, relative, and cumulative relative frequencies. In this case, absolute and relative frequencies correspond to the absolute number and the percentage of individuals according to their distribution for this variable, respectively, based on complete years of education. It should be noticed that there are 450 adolescents with 8 years of education, which corresponds to 20.5% of the subjects. Tables may also present the cumulative relative frequency of the variable. In this case, it was found that 50.6% of study subjects have up to 8 years of education. It is important to point that, although the same data were used, each form of presentation (absolute, relative or cumulative frequency) provides different information and may be used to understand frequency distribution from different perspectives.

TABLE 1: Absolute and relative frequencies of acne scar in 18-year-old adolescents (n = 2.414). Pelotas, Brazil, 2010

Prevalence	Absolute frequency (n)	Relative frequency (%)
No	1.855	76.84
Yes	559	23.16
Total	2.414	100.00

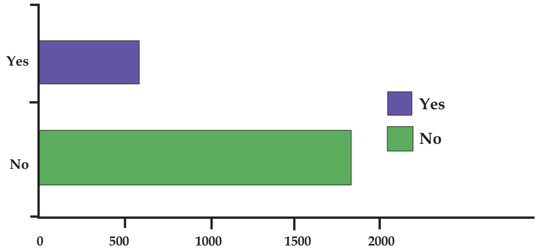

FIGURE 2: Absolute frequencies of acne scar in 18-year-old adolescents (n = 2.414). Pelotas, Brazil, 2010

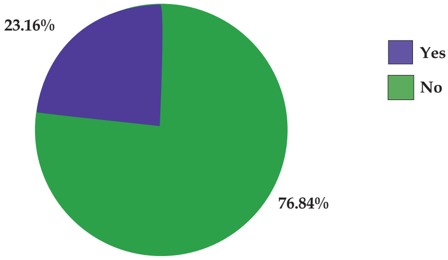

FIGURE 3: Relative frequencies of acne scar in 18-year-old adolescents (n = 2.414). Pelotas, Brazil, 2010

When one wants to evaluate the frequency distribution of continuous variables using tables or graphs, it is necessary to transform the variable into categories, preferably creating categories with the same size (or the same amplitude). However, in addition to this general recommendation, other basic guidelines should be followed, such as: (1) subtracting the highest from the lowest value for the variable of interest; (2) dividing the result of this subtraction by the number of categories to be created (usually from three to ten); and (3) defining category intervals based on this last result.

For example, in order to categorize height (in meters) of a set of individuals, the first step is to identify the tallest and the shortest individual of the sample. Let us assume that the tallest individual is 1.85m tall and the shortest, 1.55m tall, with a difference of 0.3m between these values. The next step is to divide this difference by the number of categories to be created, e.g., five. Thus, 0.3m divided by five equals 0.06m, which means that categories will have exactly this range and will be numerically represented by the following range of values: 1st category – 1.55m to 1.60m; 2nd category – 1.61m to 1.66m; 3rd category – 1.67m to 1.72m; 4th category – 1.73m to 1.78m; 5th category – 1.79m to 1.85m.

Table 3 illustrates weight values at 18 years of age in kg (continuous numerical variable) obtained in a study with youngsters from Pelotas, Southern Brazil.[4,5] Figure 4 shows a histogram with the variable weight categorized into 20-kg intervals. Therefore, it is possible to observe that data from continuous numerical variables may be presented in tables or graphs.

Table 2: Educational level of 18-year-old adolescents (n = 2,199). Pelotas, Brazil, 2010

Educational level	Absolute frequency	Relative frequency	Cumulative relative frequency
(in years of education)	(n)	(%)	(%)
0	1	0.05	0.05
1	2	0.09	0.14
2	2	0.09	0.23
3	11	0.50	0.73
4	100	4.55	5.28
5	156	7.09	12.37
6	169	7.69	20.05
7	221	10.05	30.10
8	450	20.46	50.57
9	251	11.41	61.98
10	320	14.55	76.53
11	479	21.78	98.32
12	31	1.41	99.73
13	6	0.27	100.00
Total	**2.199**	**100.00**	**-**

Table 3: Weight distribution among 18-year-old young male sex (n = 2.194). Pelotas, Brazil, 2010

Weight at 18 years of age (in kg)	Absolute frequency(n)	Relative frequency (%)
40.5 to 59.9	554	25.25
60.0 to 65.8	543	24.75
65.9 to 74.6	551	25.11
74.7 to 147.8	546	24.89
Total	**2.194**	**100.00**

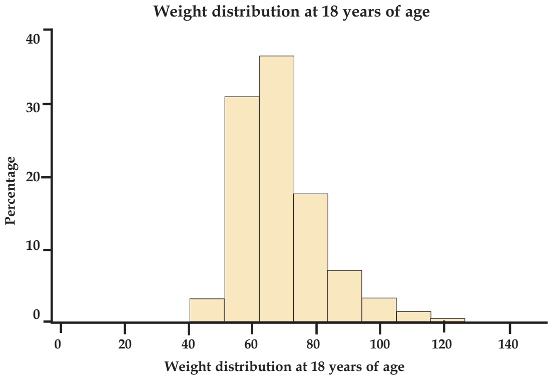

Figure 4: Weight distribution at 18 years of age among youngsters from the city of Pelotas. Pelotas (n = 2.194), Brazil, 2010

Assessing the relationship between two variables

The forms of data presentation that have been described up to this point illustrated the distribution of a given variable, whether categorical or numerical. In addition, it is possible to present the relationship between two variables of interest, either categorical or numerical.

The relationship between categorical variables may be investigated using a contingency table, which has the purpose of analyzing the association between two or more variables. The lines of this type of table usually display the exposure variable (independent variable), and the columns, the outcome variable (dependent variable). For example, in order to study the effect of sun exposure (exposure variable) on the development of skin cancer (outcome variable), it is possible to place the variable sun exposure on the lines and the variable skin cancer on the columns of a contingency table. Tables may be easier to understand by including total values in lines and columns. These values should agree with the sum of the lines and/or columns, as appropriate, whereas relative values should be in accordance with the exposure variable, i.e., the sum of the values mentioned in the lines should total 100%.

It is such a display of percentage values that will make it possible for risk or exposure groups to be compared with each other, in order to investigate whether individuals exposed to a given risk factor show higher frequency of the disease of interest. Thus, table 4 shows that 75.0%, 9.0%, and 0.3% of individuals in the study sample who had been working exposed to the sun for 20 years or more, for less than 20 years, and had never been working exposed to the sun, respectively, developed non-melanoma skin cancer. Another way of interpreting this table is observing that 25.0%, 91%,.0%, and 99.7% of individuals who had been working exposed to the sun for 20 years of more, for less than 20 years, and had never been working exposed to the sun did not develop non-melanoma skin cancer. This form of presentation is one of the most used in the literature and makes the table easier to read.

The relationship between two numerical variables or between one numerical variable and one categorical variable may be assessed using a scatter diagram, also known as dispersion diagram. In this diagram, each pair of values is represented by a symbol or a dot, whose horizontal and vertical positions are determined by the value of the first and second variables, respectively. By convention, vertical and horizontal axes should correspond to outcome and exposure variables, respectively. Figure 5 shows the relationship between weight and height among 18-year-old youngsters from Pelotas, Southern Brazil, in 2010.[3,4] The diagram presented in figure 5 should be interpreted as follows: the increase in subjects' height is accompanied by an increase in their weight.

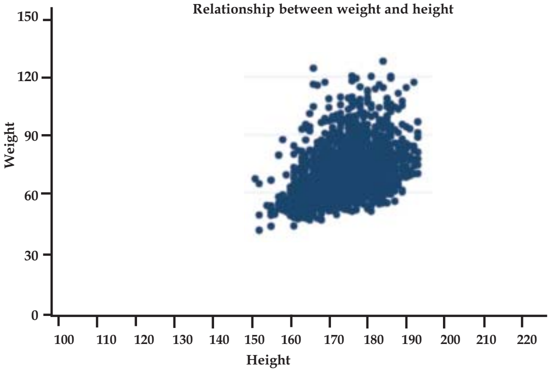

FIGURE 5: Point diagram for the relationship between weight (kg) and height (cm) among 18-year-old youngsters from the city of Pelotas (n = 2.194). Pelotas, Brazil, 2010.

BASIC RULES FOR THE PREPARATION OF TABLES AND GRAPHS

Ideally, every table should:

- Be self-explanatory;
- Present values with the same number of decimal places in all its cells (standardization);
- Include a title informing what is being described and where, as well as the number of observations (N) and when data were collected;
- Have a structure formed by three horizontal lines, defining table heading and the end of the table at its lower border;
- Not have vertical lines at its lateral borders;
- Provide additional information in table footer, when needed;
- Be inserted into a document only after being mentioned in the text; and
- Be numbered by Arabic numerals.

Similarly to tables, graphs should:

- Include, below the figure, a title providing all relevant information;
- Be referred to as figures in the text;

TABLE 4: Sun exposure during work and non-melanoma skin cancer (hypothetical data).

Work exposed to the sun	Non-melanoma skin cancer				Total	
	Yes		No			
	N	%	N	%	N	%
20 or more years	30	75.0	10	25.0	40	100
<20 years	9	9.0	90	91.0	99	100
Never	1	0.3	300	99.7	301	100
Total	40	9.0	400	91.0	440	100

- Identify figure axes by the variables under analysis;
- Quote the source which provided the data, if required;
- Demonstrate the scale being used; and
- Be self-explanatory.

The graph's vertical axis should always start with zero. A usual type of distortion is starting this axis with values higher than zero. Whenever it happens, differences between variables are overestimated, as can been seen in figure 6.

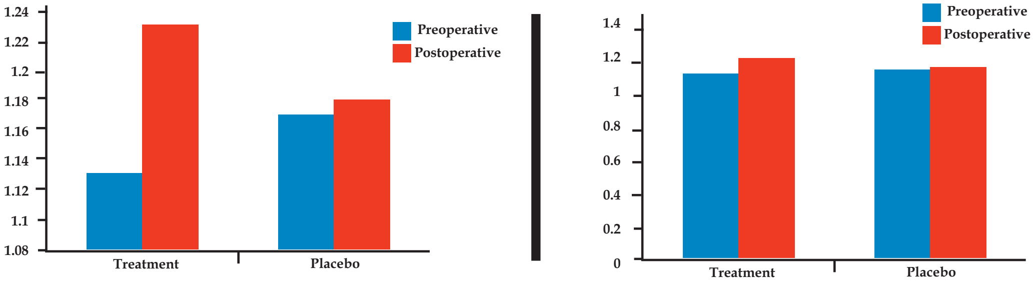

Figure 6: Figure showing how graphs in which the Y-axis does not start with zero tend to overestimate the differences under analysis. On the left there is a graph whose Y axis does not start with zero and on the right a graph reproducing the same data but with the Y axis starting with zero.

CONCLUSION

Understanding how to classify the different types of variables and how to present them in tables or graphs is an essential stage for epidemiological research in all areas of knowledge, including Dermatology. Mastering this topic collaborates to synthesize research results and prevents the misuse or overuse of tables and figures in scientific papers. ❑

REFERENCES

1. Walker SL, Hawk JLM, Young AR. Acute and chronic effects. In: Freedberg IM, Eisen AZ, Wolff K, Austen KF, Goldsmith LA, Katz SI, editors. Fitzpatrick's Dermatology in General Medicine 8th ed. p. 1275-81.
2. Duquia RP, Baptista Menezes AM, Reichert FF, de Almeida HL Jr. Prevalence and associated factors with sunscreen use in Southern Brazil: A population-based study. J Am Acad Dermatol. 2007;57:73-80.
3. Duquia RP, de Almeida HL Jr, Breunig JA, Souzat PR, Göellner CD. Most common patterns of acne in male adolescents: a population-based study. Int J Dermatol. 2013;52:550-3.
4. Breunig Jde A, de Almeida HL, Jr., Duquia RP, Souza PR, Staub HL. Scalp seborrheic dermatitis: prevalence and associated factors in male adolescents. Int J Dermatol. 2012;51:46-9.
5. Almeida H, Jr., Cecconi J, Duquia RP, Souza PR, Breunig J. Sensitivity and specificity of self-reported acne in 18-year-old adolescent males. Int J Dermatol. 2013;52:946-8.

Mailing address:
Rodrigo Pereira Duquia
R. Independência, 172 - sala 902
90035-070 - Independência - RS
Brazil
E-mail: rodrigoduquia@gmail.com

How to cite this article: Duquia RP, Bastos JL, Bonamigo RR, González-Chica DA, Martínez-Mesa J. Presenting data in tables and charts. An Bras Dermatol. 2014;89(2):280-5.