

Chapter 2

Descriptive Biostatistics

Abstract In this chapter we briefly describe the main methods and techniques in descriptive biostatistics, as well as their application to the analysis of biomedical questions. With special attention to the study of cancer, this chapter provides a general understanding of the nature and relevance of descriptive biostatistical methods in medicine and biology, explains the design behind a biostatistical descriptive analysis, and stresses the paramount importance of descriptive statistics as the initial stage of any biomedical investigation.

2.1 Descriptive Statistics: The Starting Point

As explained in the former chapter, statistics can be defined as the scientific study of the numerical data that emerge from non totally predictable phenomena. Starting from this definition, we will distinguish between the two alternatives that biostatistics (and in general statistics) offers to extract conclusions and information from a set of data. The first analyzes the data without assuming any underlying structure for such data, and is called *descriptive statistics*, while the second, *inferential statistics*, operates on the basis of a given structure for the observed data.

When nothing is previously assumed for the observed data, the only possible task is to describe such data. This is why the branch of statistics pursuing this objective is called *descriptive statistics*. The descriptive stage is for obvious reasons the first step in any applied research, and must lead to a manageable numerical summary of the data concisely but accurately describing them. Indeed, the ultimate goal of descriptive statistics is not to explain the data or the event behind the data, but, on the contrary, to make a future explanation and interpretation possible.

Usually, and especially in medicine and biology, the observations originated by a phenomenon are large in number, dispersed, variable and heterogeneous, aspects that prevent the researcher from directly comprehending it. To make the understanding of the analyzed phenomenon possible, it is first necessary to present, arrange, measure, classify, describe and summarize the obtained data. These are specifically the tasks carried out by descriptive statistics, a discipline that without any doubt constitutes the entry door to any biomedical scientific investigation.

Looking at how to describe, simplify and arrange the data, descriptive (bio)statistics makes use of six main instruments:

1. Statistical tables.
2. Graphical representations.
3. Measures of central tendency.
4. Measures of dispersion.
5. Measures of form.
6. Measures of correlation.

When the data obtained from a biomedical phenomenon refer to a unique aspect or variable, only the five first instruments are susceptible to application; in addition, measures of correlation are also possible when the researcher collects data relative to several characteristics. In the following sections we will succinctly comment on these descriptive statistical tools. As is logical, our intention is not to give a condensed course on descriptive biostatistics, but to provide a general understanding of the nature and relevance of descriptive statistical methods in medicine and biology.

We will first define statistical unit, statistical population, and statistical variable. A *statistical unit* is each member—or each element, in statistical terms—of the considered set of entities under analysis. This set of studied entities is known as *statistical population*, a denomination inherited from demography, the first application field of statistics¹. Finally, a *statistical variable* is each aspect or characteristic of the statistical unit that is considered for study. These statistical variables can be qualitative or quantitative, depending on whether their nature is countable or not.

2.2 Univariate Descriptive Statistics

We will begin our discussion of the main descriptive biostatistical methods and techniques by considering the univariate case. As previously remarked, before explaining a biomedical phenomenon it must be described and characterized, descriptive analysis being the first fundamental stage in any applied research. As a result, in biology and medicine, most research programs have a set of descriptive statistical analyses as starting point.

Let us suppose that, with the ultimate goal of explaining a biomedical phenomenon, we have considered a population and measured a particular characteristic for each member of this population. Let N be the number of individuals in the population, let C denote the analyzed characteristic, and let $C_i, i = 1, 2, \dots, I$ be the different values for this characteristic. We will denote the number of individuals

¹ Indeed, traditionally, the origin of descriptive statistics dates back to the demographic work by John Graunt (1662).

Table 2.1 Univariate statistical table

Values of the characteristic	Absolute frequency	Relative frequency
C_1	n_1	$f_1 = \frac{n_1}{N}$
C_2	n_2	$f_2 = \frac{n_2}{N}$
...	...	...
C_i	n_i	$f_i = \frac{n_i}{N}$
...	...	...
C_I	n_I	$f_I = \frac{n_I}{N}$
Total	$\sum_{i=1}^I n_i = N$	$\sum_{i=1}^I f_i = 1$

presenting the value C_i as n_i . This number n_i is the *absolute frequency* of the observed value C_i , and the ratio

$$f_i = \frac{n_i}{N}$$

is known as the *relative frequency* of the observed value C_i . Note that f_i is the proportion on the total population N of individuals presenting the value C_i , and that

$$\sum_{i=1}^I n_i = N, \quad \sum_{i=1}^I f_i = 1.$$

The observed values for the characteristic and their absolute and relative frequencies are usually arranged in tables. When the number of considered characteristics is one, the table is called a *univariate statistical table*. Table 2.1 is an example of a univariate statistical table.

Usually, the information contained in this table is presented graphically under the form of histograms and cumulative frequency curves. A *histogram* is a graphical representation of the absolute or relative frequencies for each value of the characteristic. A *cumulative frequency curve* is a plot of the number or percentage of individuals falling in or below each value of the characteristic. As is obvious, the histogram shows the relative presence or weight of each value of the characteristic in the population, whilst the cumulative frequency curve shows, with respect to each value of the analyzed characteristic, the population percentages displaying equal or lower values.

Statistical tables, histograms and cumulative frequency curves are just different ways to present the obtained data. In fact, none of these three descriptive statistical instruments imply modifications or manipulations of the data, which are simply collected and arranged. Together with these purely descriptive methods, there exist functions of the data that describe and summarize them and that are of paramount importance in descriptive biostatistics. For the univariate case we are examining,

these functions describing and summarizing the observed data are of three types: the so called *measures of central tendency*, the *measures of dispersion*, and the *measures of symmetry and form* or *moments*. They are also called *statistics*, since this term—statistics—refers not only to the science we are discussing, but is also applied to any function of the observed data. It is therefore in this latter sense—statistics as a function or modification of the obtained data—that the term statistics is used to design these measures of central tendency, dispersion and form.

The main statistics of location are the mean—which can be arithmetic, geometric or harmonic—, the median and the mode. Dispersion measures are given by four main statistics—range, standard deviation, coefficient of variation and percentiles—, and finally, symmetry and form are measured by variance, semivariance, skewness and kurtosis. We will not define or describe in detail these statistics since this is not the purpose of this book. However, two questions are worth noting. First, all these descriptive measures of the observed data set are a direct consequence of the frequency distribution—or histogram—of the data. As explained, the frequency distribution is simply an arrangement of the data showing the frequency of each value, and constitutes all the obtained information. The frequency distribution reports on the percentage weight that each observed value has on the total set of data, and makes judgements and predictions on the observations set possible. For instance, from the frequency distribution, the probability of observing a value or an interval of values in the data set can be exactly measured, and which value is the more likely to be observed in the data set can be predicted. This information is condensed by the aforementioned statistics, which provide a numerical summary of the frequency distribution. Second, among all these descriptive measures, researchers must choose the most appropriate for their purposes. For instance, if researchers want to compare the dispersion of two different populations, the coefficient of variation and not the standard deviation is the pertinent measure. In other cases, transformation of the original data can be convenient to eliminate asymmetry or to make variances independent of mean, and therefore the appropriate type of mean must be chosen. For these aspects and many others, we refer the reader to any of the excellent textbooks on biostatistics available today.

A very good example of the fundamental role that the descriptive stage plays in dealing with a biomedical question and of how descriptive biostatistics may help in extracting medical conclusions is Russo and Russo's (1987b) paper "Biological and Molecular Basis of Mammary Carcinogenesis". In this paper, the authors opened up a research avenue on breast cancer after concluding that malignant phenotypical changes in human breast epithelial cells are inversely related to the degree of glandular lobular differentiation and glandular development of the donor gland, and directly related to the proliferative activity of its epithelial cells. To arrive at these conclusions, basic for the subsequent analyses of how and when breast cancer initiates, the authors carried out an exhaustive descriptive statistical analysis of the relevant ratios and variables, that help them in a decisive way. For instance, to support one of their hypotheses, namely that the malignant phenotypical changes in breast epithelial cells are inversely related to the degree of glandular lobular differentiation and glandular development of the donor gland, the authors begin by depicting the histogram showing the decrease with aging in the number of terminal end buds for two kinds

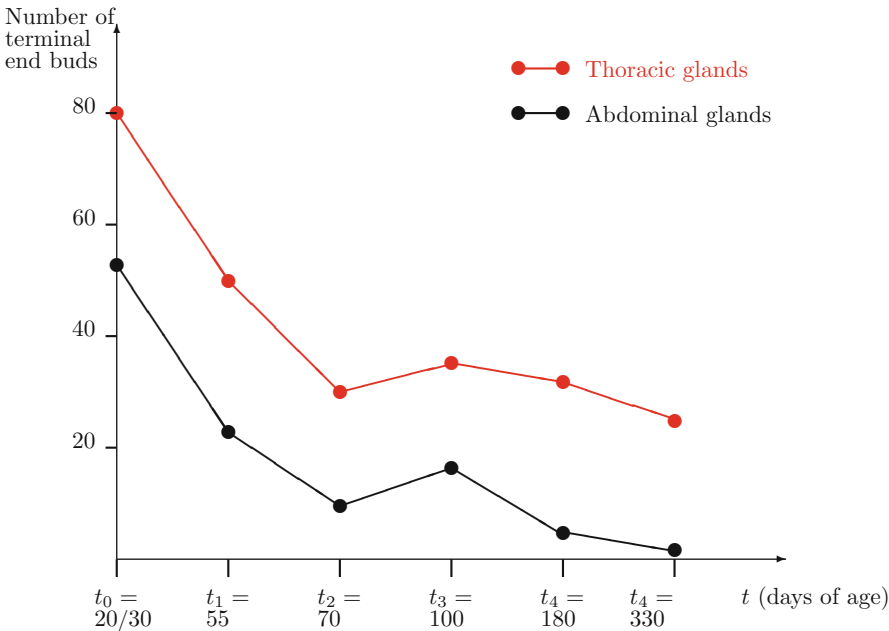


Fig. 2.1 Number of terminal end buds for thoracic and abdominal rat mammary glands. (Figure 25 in Russo and Russo (1987b))

of rat mammary glands, the thoracic glands and the abdominal glands. The evidence arising from this histogram is clear, since the number of terminal end buds for the thoracic glands always lies manifestly above the number of terminal end buds for the abdominal glands. Figure 2.1 reproduces the histogram in Russo and Russo (1987b) showing the number of terminal end buds for thoracic and abdominal rat mammary glands.

The following step to prove the hypothesis is to show that the thoracic glands are behind abdominal glands in development. This is again done through the appropriate histogram, which in this case represents the increment with aging in the number of alveolar buds and lobules, also for the two types of rat mammary glands. The conclusion from this histogram is also obvious, since the increments for the thoracic glands are always and evidently behind the increments for abdominal glands, as appears in Fig. 2.2.

Since (1) the terminal end buds are undifferentiated structures of the mammary gland, and (2) the increment in the number of alveolar buds and lobules is a direct sign of gland development, the final step to prove the hypothesis is to measure the tumor incidence for the two kinds of glands and to show that this incidence is greater in the thoracic glands than in the abdominal glands. This is also done by depicting a histogram showing the percentage of adenocarcinomas induced in the two types of glands, which allows the hypothesis to be proved: tumor incidence is greater in those glands located in the thoracic gland, which are less differentiated (with more terminal

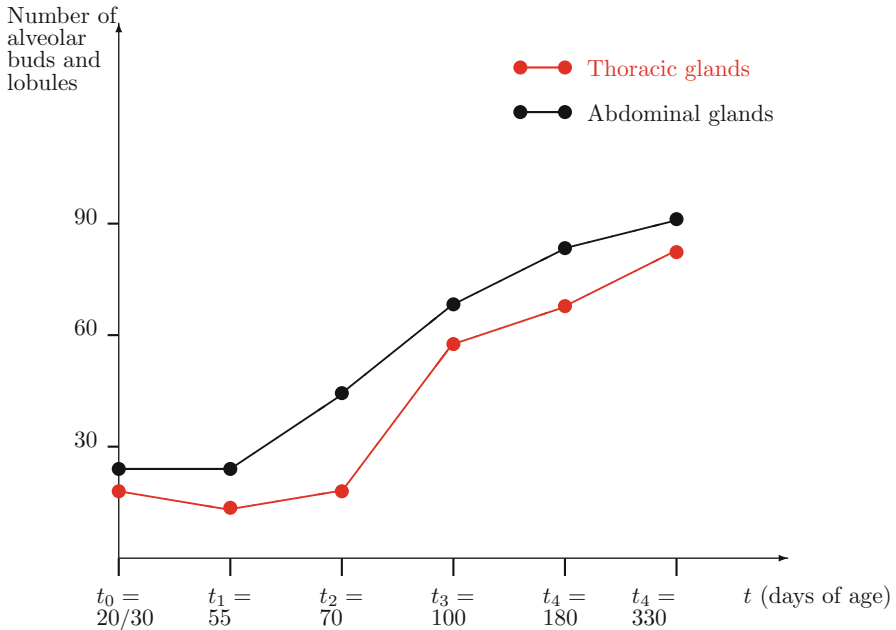


Fig. 2.2 Increment with aging in the number of alveolar buds and lobules for thoracic and abdominal rat mammary glands. (Figure 26 in Russo and Russo (1987b))

end buds) and developed (with lower increments in alveolar buds and lobules). This histogram is represented in Fig. 2.3.

As commented on above, this paper by Russo and Russo (1987b) is a very good illustration of how descriptive statistics can be a powerful tool in medical research. Indeed, although almost all the biostatistics in Russo and Russo (1987b) is descriptive statistics, the pertinent and insightful use of histograms, means, standard deviations, modes, etc., becomes an irrefutable argument to support the authors' hypotheses and conclusions. For instance, also thanks to descriptive biostatistical tools, these authors deduce that in humans, "the high risk group to be tested for cell transformation is represented by young, perimenarchal and nulliparous females between the ages of 12 and 24 years, in whom the gland has a low degree of differentiation and contains topographic areas of high proliferation".

To reach this deduction, Russo and Russo (1987b) take as the starting point their findings for rat mammary carcinomas linking malignant phenotypical changes with low glandular lobular differentiation and high proliferative activity of epithelial cells. The authors begin with a study of the human breast morphology. In this study, they grade human mammary gland development according to a classification of the lobular components of the organ. Lobules, in turn, are classified based upon determination of their size and quantification of both number and size of the alveoli that compose each lobule. As a result, the authors distinguish four different types of lobules:

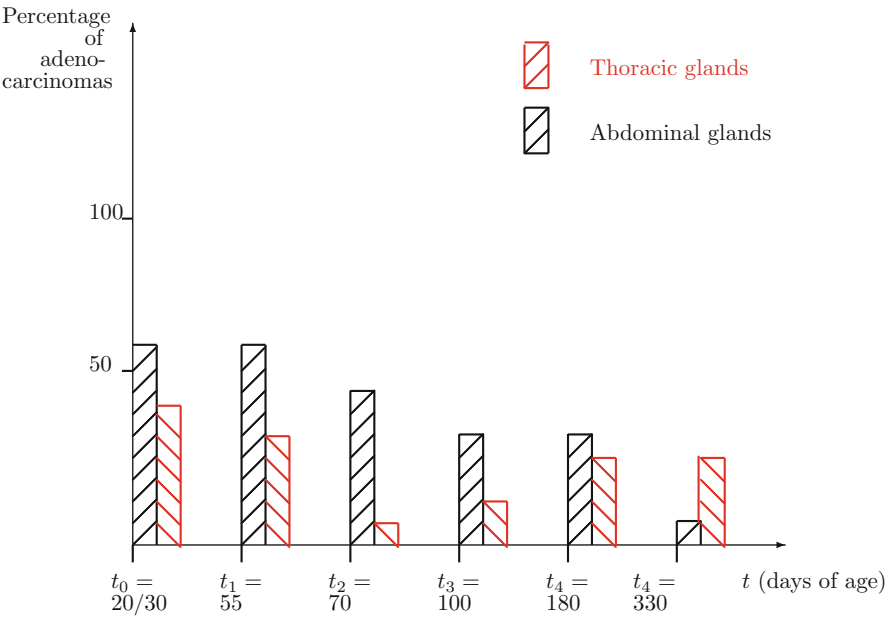


Fig. 2.3 Percentage of adenocarcinomas induced in thoracic and abdominal rat mammary glands. (Figure 29 in Russo and Russo (1987b))

Type 1 lobules are composed of approximately 5 or 6 alveoli; type 2 of approximately 47 alveoli; type 3 of about 81 alveoli; and type 4 lobules, observed only during pregnancy and not discussed, of approximately 180 alveoli. This morphological analysis of the human mammary gland allows its degree of glandular lobular differentiation and glandular development to be established. To grade the proliferative activity of cells, the second aspect to consider in order to determine the risk of cell malignant transformation, the researchers analyze DNA synthesis measured by [3H]thymidine incorporation—or DNA labeling index, DNA-LI—by using the organ culture technique.

The results of this examination of the human mammary gland showing the degree of morphological differentiation and the level of cell proliferation are those in Table 2.2.

As appears in Table 2.2, type 1 lobules present the lowest differentiation and the highest proliferation level. Since breast tissues composed predominantly of type 1 lobules are more frequently found on young, perimenarchal and nulliparous women, Russo and Russo (1987b) conclude that this group of females constitutes the high risk group to be tested for malignant cell transformation. As the authors remark, the elevated incidence of breast cancer due to physical agents such as ionizing radiations acting on the mammary gland of young women supports the validity of their hypothesis.

Table 2.2 Identification profile of the different compartments of human breast. (Russo and Russo (1987b))

Structure	Area (mm ²)	Number of alveoli	DNA-LI
Lobule 1	0.048 ± 0.044	11.20 ± 6.34	5.45 ± 2.50
Lobule 2	0.060 ± 0.026	47.00 ± 11.70	0.99 ± 1.24
Lobule 3	0.129 ± 0.049	81.00 ± 16.60	0.25 ± 0.33

Values are expressed as the mean ± standard deviation

It is worth noting again that, to reach this conclusion, Russo and Russo (1987b) make almost exclusive use of descriptive statistical tools. In the specific case of Table 2.2, the researchers consider mean values and standard deviations of the obtained data to describe and summarize both the mammary gland degree of development and differentiation and the proliferative activity of cells. As commented above, these technically simple analyses in terms of descriptive statistics opened up a new research avenue on breast cancer, and perfectly illustrate how descriptive statistics is indispensable for any further work in medicine or biology, the reason why it is present in almost any research article on biomedicine. This presence varies from the pure presentation and discussion of data to the use of descriptive statistics and histograms for presenting, confirming or rejecting hypotheses of several natures. For instance, in Myers and Gloeckler (1989), the authors carry out an analysis of the cancer patient survival rates that only involves descriptive biostatistical arguments. The main conclusion is that as a patient survives for a number of years after the diagnosis of cancer, the probability of surviving cancer each subsequent year increases and approaches that of the general population. To arrive at this conclusion, the authors simply present, at five and ten years, the survival percentages of several original groups of patients diagnosed with cancer—which represent the chance of avoiding any cause of death—, and the relative to cancer survival rates for the same original group of patients, that is, the percentages of patients who escaped death due to cancer. Then, by considering the ratio between relative to cancer survival rates and survival rates, given that this ratio is greater for the ten years temporal horizon, it can be concluded that the chance of surviving cancer increases over time. In addition, the authors also compute the ratio between the observed survival rate for the cancer patients who were alive five years after diagnosis—the observed five-to-ten survival rate—and the general population survival percentage. As this ratio approaches 100, the survival rates for the cancer patients after five years approaches that of the general population, and since this relative ratio is manifestly greater than the survival percentage for the five years temporal horizon, it can be concluded that the probability of surviving cancer each subsequent year approaches that of the general population. The detailed analysis by types of cancer, sex and races that the authors carried out in this paper allows several significant scenarios to be identified concerning medical surveillance, interestingly obtained with the sole application of descriptive statistics. As an example, we reproduce the Table 2.3 from Myers and Gloeckler (1989), corresponding to the case *patients diagnosed in 1973 to 1975, SEER program, all races, males and females*, in Table 2.3.

Table 2.3 Long-term survival rates for patients diagnosed in 1973 to 1975, SEER program, all races, males and females. (Myers and Gloeckler (1989))

Primary site	5-Year rate			10-Year rate			5–10 year relative rate for 5 year survivors
	Observed	Relative	Ratio	Observed	Relative	Ratio	
All sites	40	48	120	28	41	146	84
Oral cavity and pharynx	43	51	119	28	41	146	78
Esophagus	4	4	100	2	3	150	67
Stomach	11	14	127	7	11	157	79
Colon	38	48	126	26	44	169	89
Rectum	37	46	124	25	40	160	84
Liver	3	3	100	2	2	100	61
Gallbladder	6	8	133	2	4	200	58
Pancreas	2	3	150	2	2	100	79
Larynx	55	64	116	36	51	142	79
Lung and bronchus	10	11	110	6	8	133	70
Bone	48	52	108	40	46	115	88
Soft tissue	52	59	113	41	53	129	89
Melanoma	69	76	110	56	68	121	89
Breast	65	73	112	47	60	128	82
Urinary bladder	55	71	129	37	63	170	86
Kidney	42	49	117	30	42	140	84
Brain	16	18	113	12	14	117	73
Thyroid gland	86	91	106	81	90	111	98
Hodgkin's disease	62	66	106	50	57	114	83
Non-Hodgkin's lymphomas	38	45	118	23	33	143	90
Multiple myeloma	19	23	121	5	9	180	36
Leukemias	27	33	122	13	20	154	61

2.3 Multivariate Descriptive Statistics

Unlike univariate descriptive statistics, in which the number of analyzed characteristics in the population reduces to one, multivariate descriptive statistics considers several characteristics of the individuals in the population. This is indeed the key distinctive feature of all multivariate statistics, namely the simultaneous examination of two or more characters of the statistical entities.

To fix concepts, let us consider a population with N individuals, simultaneously described according to two characteristics A and B . Let $A_1, \dots, A_i, \dots, A_I$ be the I observed values for the A characteristic, and let $B_1, \dots, B_j, \dots, B_J$ be the J observed values for the B characteristic. We will denote the number of entities presenting simultaneously the values A_i and B_j by n_{ij} . Analogously, we will denote the proportion—or percentage—of entities presenting simultaneously the values A_i

and B_j by $f_{ij} = \frac{n_{ij}}{N}$. Obviously,

$$\sum_{i=1}^I \sum_{j=1}^J n_{ij} = N, \quad \sum_{i=1}^I \sum_{j=1}^J f_{ij} = 1.$$

The numbers n_{ij} and f_{ij} are known as *absolute frequency* and *relative frequency*, respectively, of the values A_i and B_j , and both refer to the concurrent occurrence of the two values A_i and B_j , one of each considered characteristic.

In addition, we can also focus on the incidence of exclusively one character. This is done through the analysis of the *marginal distributions*. When we accumulate on the index i , relative to the A characteristic, we obtain the *marginal absolute frequency* $n_{.j}$ and the *marginal relative frequency* $f_{.j}$ for the value B_j of the B characteristic. More specifically, $n_{.j}$ and $f_{.j}$ are given by the expressions

$$n_{.j} = \sum_{i=1}^I n_{ij}, \quad f_{.j} = \sum_{i=1}^I f_{ij} = \frac{n_{.j}}{N}, \quad j = 1, 2, \dots, J.$$

Analogously, by totalizing on the index j , that is on the B character, we get the the *marginal absolute frequency* $n_{i.}$ and the *marginal relative frequency* $f_{i.}$ for the value A_i of the A characteristic:

$$n_{i.} = \sum_{j=1}^J n_{ij}, \quad f_{i.} = \sum_{j=1}^J f_{ij} = \frac{n_{i.}}{N}, \quad i = 1, 2, \dots, I.$$

As is logical,

$$\begin{aligned} \sum_{j=1}^J n_{.j} &= N, & \sum_{j=1}^J f_{.j} &= 1, \\ \sum_{i=1}^I n_{i.} &= N, & \sum_{i=1}^I f_{i.} &= 1. \end{aligned}$$

Unlike the original bivariate distribution described by n_{ij} and f_{ij} , reporting on the simultaneous occurrence of the different values of the two characters, the marginal distributions are univariate distributions informing on the manifestation of only one characteristic, A or B , and must be interpreted as a standard univariate distribution defined for the entire population of N entities. In Table 2.4 we represent a bivariate statistical table with the bivariate and the two marginal distributions.

Together with the marginal distributions, the multivariate analysis allows other interesting univariate distributions to be derived, the so called *conditional distributions*. These conditional distributions describe the behavior of a particular character not for the entire population, as the marginal distributions do, but for the subset of

Table 2.4 Bivariate and marginal distributions

Values of character A	Values of character B					Marginal distribution of A
	B_1	$\dots$	B_j	$\dots$	B_J	
A_1	n_{11}	$\dots$	n_{1j}	$\dots$	n_{1J}	$n_{1.} = \sum_{j=1}^J n_{1j}$
	$f_{11} = \frac{n_{11}}{N}$	$\dots$	$f_{1j} = \frac{n_{1j}}{N}$	$\dots$	$f_{1J} = \frac{n_{1J}}{N}$	$f_{1.} = \frac{n_{1.}}{N}$
A_2	n_{21}	$\dots$	n_{2j}	$\dots$	n_{2J}	$n_{2.} = \sum_{j=1}^J n_{2j}$
	$f_{21} = \frac{n_{21}}{N}$	$\dots$	$f_{2j} = \frac{n_{2j}}{N}$	$\dots$	$f_{2J} = \frac{n_{2J}}{N}$	$f_{2.} = \frac{n_{2.}}{N}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
A_i	n_{i1}	$\dots$	n_{ij}	$\dots$	n_{iJ}	$n_{i.} = \sum_{j=1}^J n_{ij}$
	$f_{i1} = \frac{n_{i1}}{N}$	$\dots$	$f_{ij} = \frac{n_{ij}}{N}$	$\dots$	$f_{iJ} = \frac{n_{iJ}}{N}$	$f_{i.} = \frac{n_{i.}}{N}$
$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$	$\dots$
A_I	n_{I1}	$\dots$	n_{Ij}	$\dots$	n_{IJ}	$n_{I.} = \sum_{j=1}^J n_{Ij}$
	$f_{I1} = \frac{n_{I1}}{N}$	$\dots$	$f_{Ij} = \frac{n_{Ij}}{N}$	$\dots$	$f_{IJ} = \frac{n_{IJ}}{N}$	$f_{I.} = \frac{n_{I.}}{N}$
Marginal distribution of B	$n_{.1} = \sum_{i=1}^I n_{i1}$	$\dots$	$n_{.j} = \sum_{i=1}^I n_{ij}$	$\dots$	$n_{.J} = \sum_{i=1}^I n_{iJ}$	
	$f_{.1} = \frac{n_{.1}}{N}$	$\dots$	$f_{.j} = \frac{n_{.j}}{N}$	$\dots$	$f_{.J} = \frac{n_{.J}}{N}$	
Total	$\sum_{i=1}^I n_{i.} = N, \sum_{j=1}^J n_{.j} = N, \sum_{i=1}^I f_{i.} = 1, \sum_{j=1}^J f_{.j} = 1$					

the population that presents a specific value for the other characteristic. For instance, let us consider the $n_{.j}$ entities in the population with the value B_j for the B characteristic. Among these $n_{.j}$ individuals, n_{ij} also present the value A_i for the A character, and then

$$f_i^j = \frac{n_{ij}}{n_{.j}},$$

Table 2.5 Conditional distributions. A characteristic

Values of character A	Values of character B				
	B_1 ($n_{.1}$ entities)	...	B_j ($n_{.j}$ entities)	...	B_J ($n_{.J}$ entities)
A_1	$f_1^1 = \frac{n_{11}}{n_{.1}}$	...	$f_1^j = \frac{n_{1j}}{n_{.j}}$	...	$f_1^J = \frac{n_{1J}}{n_{.J}}$
A_2	$f_2^1 = \frac{n_{21}}{n_{.1}}$	...	$f_2^j = \frac{n_{2j}}{n_{.j}}$	...	$f_2^J = \frac{n_{2J}}{n_{.J}}$
...	...	...	...	...	...
A_i	$f_i^1 = \frac{n_{i1}}{n_{.1}}$	...	$f_i^j = \frac{n_{ij}}{n_{.j}}$	...	$f_i^J = \frac{n_{iJ}}{n_{.J}}$
...	...	...	...	...	...
A_I	$f_I^1 = \frac{n_{I1}}{n_{.1}}$	...	$f_I^j = \frac{n_{Ij}}{n_{.j}}$	...	$f_I^J = \frac{n_{IJ}}{n_{.J}}$
Total	$\sum_{i=1}^I f_i^1 = 1$	...	$\sum_{i=1}^I f_i^j = 1$	...	$\sum_{i=1}^I f_i^J = 1$

known as the relative frequency of A_i conditional to B_j , represents the proportion on the $n_{.j}$ individuals displaying the value B_j for the B characteristic that also present A_i for the A character. Table 2.5 collects the conditional distributions for the A characteristic. To derive the conditional distributions for the B characteristic, the process is totally analogous.

These $I + J$ conditional distributions— J conditional distributions for the A character, one for each value of the B characteristic, and I conditional distributions for the B character, one for each value of the A characteristic—are univariate distributions, susceptible to the same exploitation and analysis as the marginal univariate distributions. The only difference lies in the reference population: the marginal distributions are defined on the entire population, that is on the N entities, whilst the conditional distributions are specified on the subset of the population displaying the conditioning value. For instance, the conditional distribution of B conditional to A_i is defined on the $n_{i.}$ entities presenting the value A_i .

As univariate distributions, these marginal and conditional distributions can be graphically represented by histograms and cumulative curves and described by the statistics we have discussed. In addition, the consideration of the original bivariate distribution opens up new graphical and analytical possibilities, namely, the joint representation and study of the observed values for the two characteristics A and B . From the graphic perspective, the basic idea is to plot the observations by representing the values for the A characteristic on the X axis and the values for the B characteristic on the Y axis. The represented values can be those directly observed for the two characters, or convenient transformations of the measured data. In any

Table 2.6 Bivariate and marginal distributions. Number of microvessels (*A*) and presence/absence of metastases (*B*)

Number of microvessels	Presence/absence of metastases		Marginal distribution of the number of microvessels
	Present	Absent	
0–33	$n_{11} = 1$ $f_{11} = \frac{1}{49}$	$n_{12} = 6$ $f_{12} = \frac{6}{49}$	$n_{1.} = 7$ $f_{1.} = \frac{7}{49}$
34–67	$n_{21} = 9$ $f_{21} = \frac{9}{49}$	$n_{22} = 11$ $f_{22} = \frac{11}{49}$	$n_{2.} = 20$ $f_{2.} = \frac{20}{49}$
68–100	$n_{31} = 5$ $f_{31} = \frac{5}{49}$	$n_{32} = 2$ $f_{32} = \frac{2}{49}$	$n_{3.} = 7$ $f_{3.} = \frac{7}{49}$
100	$n_{41} = 15$ $f_{41} = \frac{15}{49}$	$n_{42} = 0$ $f_{42} = \frac{0}{49}$	$n_{4.} = 15$ $f_{4.} = \frac{15}{49}$
Marginal distribution of presence/absence of metastases	$n_{.1} = \sum_{i=1}^4 n_{i1} = 30$ $f_{.1} = \frac{30}{49}$	$n_{.2} = \sum_{i=1}^4 n_{i2} = 19$ $f_{.2} = \frac{19}{49}$	
Total	$\sum_{i=1}^4 n_{i.} = 49, \sum_{j=1}^2 n_{.j} = 49, \sum_{i=1}^4 f_{i.} = 1, \sum_{j=1}^2 2f_{.j} = 1$		

case, the objective is to extract conclusions from the shape and form of the cloud of points that emerges from the joint representation of the two characteristics in the entire population, and that informs on the relationship between *A* and *B*. The same idea, mathematically expressed, leads to the four basic measures of the association existing between the two characters, namely, the regression curve, the covariance, the correlation ratio, and the correlation coefficient.

Since the purpose of this book is not to give a course on biostatistics but to illustrate its applicability, and given that all these concepts and methods are not excessively difficult to understand, we will not describe in detail these mathematical measures of association. We will rather resort to a practical example of medical research to clarify the use of bivariate descriptive analysis. In this respect, the paper “Tumor angiogenesis and metastases correlation in invasive breast carcinoma”, by Weidner et al. (1991), constitutes a good application of the methods and techniques of bivariate descriptive statistics. In this paper, the authors investigate how tumor angiogenesis is related with metastases for the particular case of invasive breast carcinoma. To carry out this research, Weidner et al. (1991) counted the number of microvessels within the initial invasive carcinomas for 49 breast cancer patients, evaluated the presence

Table 2.7 Conditional distributions for the number of microvessels

Number of microvessels	Metastases	
	Present ($n_{.1} = 30$)	Absent ($n_{.2} = 19$)
0–33	$f_1^1 = \frac{1}{30}$	$f_1^2 = \frac{6}{19}$
34–67	$f_2^1 = \frac{9}{30}$	$f_2^2 = \frac{6}{19}$
68–100	$f_3^1 = \frac{5}{30}$	$f_3^2 = \frac{2}{19}$
100	$f_4^1 = \frac{15}{30}$	$f_4^2 = \frac{0}{19}$
Total	$\frac{30}{30}$	$\frac{19}{19}$

Table 2.8 Conditional distributions for the presence/absence of metastases

Presence/absence of metastases	Number of microvessels			
	0–33 ($n_{1.} = 7$)	34–67 ($n_{2.} = 20$)	68–100 ($n_{3.} = 7$)	100 ($n_{4.} = 15$)
Present	$f_1^1 = \frac{1}{7}$	$f_1^2 = \frac{9}{20}$	$f_1^3 = \frac{5}{7}$	$f_1^4 = \frac{15}{15}$
Absent	$f_2^1 = \frac{6}{7}$	$f_2^2 = \frac{11}{20}$	$f_2^3 = \frac{2}{7}$	$f_2^4 = \frac{0}{15}$
Total	$\frac{7}{7}$	$\frac{20}{20}$	$\frac{7}{7}$	$\frac{15}{15}$

or absence of metastases, and developed a bivariate biostatistical analysis. As is logical, the starting point of their study is a descriptive biostatistical examination of the data. In this case, the population consists of 49 breast cancer patients, and the considered characteristics are the presence/absence of metastasis, and the number of microvessels within the primary tumor. With these data, Weidner et al. (1991) perform the bivariate descriptive statistical analysis summarized in Table 2.6.

From this bivariate statistical table it is immediate to deduce the conditional distributions. In this respect, the conditional distributions of microvessel number (characteristic A) and metastases presence/absence (characteristic B) are those in Tables 2.7 and 2.8, respectively. The conclusion is clear: number of microvessels and presence of metastases are correlated in the sense explained in the two following chapter, since these two characteristics are associated.

2.4 Descriptive Statistics in Biostatistical and Biomathematical Models

As we have seen, descriptive statistics can be used to present and confirm a medical hypothesis, as in the already mentioned paper by Russo and Russo (1987b), or to merely present observed evidence, as in Myers and Gloeckler (1989). In addition,

descriptive biostatistics can also be applied to validate biostatistical or biomathematical assumptions and to test the capability of biostatistical and biomathematical models to explain and predict biomedical behaviors. Since we will explain in detail how biomathematical and biostatistical models are designed and applied in medical research in the following chapters, we will only comment here on the role played in these models by descriptive biostatistics.

For instance, in Iwata et al. (2000), descriptive statistic techniques are used to test the fitting of a mathematical model in predicting the growth and size of multiple metastatic tumors. The procedure followed by these authors is easy to explain. Firstly, they design a mathematical model that theoretically describes the growth and size of multiple metastatic tumors. To do this, they use a system of equations that incorporates both the colonization by metastasis and the growth of each colony. Secondly, from these equations, they obtain a specific dynamics for the growth and size of the multiple metastatic tumors, that must be understood as the theoretical predictions of their mathematical model. More specifically, they obtain a prediction for the cumulative number of tumors as a function of the colony size and a prognosis for the tumor growth over time. Thirdly, they use descriptive statistics techniques to test the validity of their model. To do so, the authors survey multiple metastatic tumors in a liver with a hepatocellular carcinoma as a primary tumor, measuring the cumulative number of metastases for each tumor colony size on successive dates, and also the growth of the primary tumor over time. In descriptive statistical terms, they obtain, first, the observed cumulative distribution of metastases as a function of the measured colony size, and, second, the observed values for the tumor sizes. Note that, in order to derive the observed cumulative distribution of metastases as a function of the measured colony size, the authors must carry out a bivariate descriptive statistical analysis such as that described in the former section. Finally, the authors compare the theoretical behaviors, given by the mathematical model, with the observed behaviors, characterized by descriptive statistics. Since the predictions agree well with the clinical data, the mathematical model becomes a useful instrument to predict the future behavior of metastases in size and number, and also the time of origin of metastases.

While in Iwata et al. (2000) descriptive statistics is used to validate a biomathematical model, in Boucher et al. (1998) descriptive statistics helps to analyze the validity of a biostatistical model. The design of the two papers is from the methodological point of view very similar. Indeed, the main methodological difference is that Boucher et al. (1998) draw up not a mathematical model as in Iwata et al. (2000), but a stochastic model. The two kinds of model differ in the way they predict the biological phenomena: the mathematical model seeks to exactly describe the phenomena, whilst the stochastic model provides the probability of observing such phenomena². Despite this fact, the arguments followed are quite similar. Firstly, Boucher et al. (1998) use a set of equations that probabilistically describes multiple tumorigenesis induced by chemical carcinogens when cell death is incorporated. In medical terms, they consider that urethane has two kinds of effect on cells, one carcinogenic and the

² This difference will be explained in detail in the following sections.

other a toxic effect that leads to the cell death. Secondly, from these equations, they obtain a specific probabilistic behavior for the number of tumors for different carcinogenic doses and time exposures, which constitutes their theoretical prediction. Thirdly, they use descriptive statistical techniques to test the validity of their model, surveying the published data on multiple tumors induced by the considered carcinogenic. Finally, the authors compare the theoretical behaviors, given by the stochastic model, with the observed behaviors, characterized by descriptive statistics. Since the clinical data are very close to the expected values predicted by the stochastic model, it can be concluded that this probabilistic model helps to predict the carcinogenic potency of the analyzed compound.

Further Readings Given the paramount importance of descriptive biostatistics as the starting point of any biomedical research, textbooks in biostatistics usually incorporate chapters explaining this discipline. We refer the interested reader to the sections and chapters on descriptive biostatistics in Fisher (1925, 1956), Sokal and Rohlf (1987, 1995) or Glantz (2005). In addition, we remit the reader to Calot (1973), a classic text on descriptive statistics, and to Gaddis and Gaddis (1990a,b), where the basic concepts and tools of descriptive biostatistics are explained in detail.

The Evolution of the Use of Mathematics in Cancer
Research

Gutiérrez Diez, P.J.; Russo, I.H.; Russo, J.

2012, XII, 404 p., Hardcover

ISBN: 978-1-4614-2396-6