# Chapter 2
# Methods and Principles of Statistical Analysis

**Abstract** The best way to learn statistics is by taking courses and working with data. Some books may also be helpful. A first step in applied statistics is usually to describe and summarize data using estimates and descriptive plots. The principle behind *p*-values and statistical inference in general is covered with a schematic overview of statistical tests and models.

**Keywords** Recommended textbooks • Descriptive statistics • *p*-values • Statistical models

## 2.1 Recommended Textbooks on Statistics

How does one learn statistics, epidemiology, and experimental design? The recommended approach is, of course, to take (university) courses and combine it with applied use. In the same way it takes considerable effort and time to become trained in food technology or chemistry or as a physician, learning statistics – both the mathematical theory and applied use – takes time and effort. Some courses or books that promise to teach statistics without requiring much time and that neglect all the fundamental aspects of the subject could be deceiving. Learning the technical use of statistical software without some fundamental knowledge of what these methods express and the basics of calculations may leave the statistical analysis part in a black box. Appropriate statistical analysis and a robust experimental design should be the opposite of a black box – it should shed light upon data and give clear insights. It should ideally not be Harry Potter magic!

A comprehensive introduction to statistics and experimental design goes somewhat beyond the scope of this brief text. Therefore, this section will refer the reader to several excellent textbooks on the subject available from Springer. Readers with access to a university library service should be able to obtain these texts online through

A.H. Pripp, *Statistics in Food Science and Nutrition*, SpringerBriefs in Food, Health, and Nutrition, DOI 10.1007/978-1-4614-5010-8_2,

www.springerlink.com. Readers unfamiliar with the general aspects of statistics and experimental design or who have not taken introductory courses are encouraged to study some of these textbooks. An overview of the principles of descriptive statistics, statistical inference (e.g., estimations and *p*-values), classic tests, and statistical models is given later, but it is assumed that the reader has a basic knowledge of these principles.

### *2.1.1 Applied Statistics, Epidemiology, and Experimental Design*

*Statistics for Non-Statisticians* by Madsen (2011) is an excellent introductory textbook for those new to the field. It covers the collection and presentation of data, basic statistical concepts, descriptive statistics, probability distributions (with an emphasis on the normal distribution), and statistical tests. The free spreadsheet software OpenOffice is used throughout the text. Additional material on statistical software, more comprehensive explanations on probability theory, and statistical methods and examples are provided in appendices and at the textbook's Web site. At 160 page, the textbook is not overwhelming. Readers with different interests, either in applied statistics or in mathematical-statistical concepts, are told which parts to read. Readers unfamiliar with statistics are highly encouraged to read this text or a similar introductory textbook on statistics.

*Applied Statistics Using SPSS, STATISTICA, MATLAB and R* by Marques de Sá (2007) is another recommended textbook, although it goes into somewhat more depth on mathematical-statistical principles. However, it provides a very useful introduction to using these four key statistical softwares for applied statistics. Combined with software manuals, it will give the reader an improved understanding of how to conduct descriptive statistics and tests. Both SPSS and STATISTICA have menu-based systems in addition to allowing users to write command lines (syntaxes). MATLAB and R might have a steeper learning curve and assume a more in-depth understanding of mathematical-statistical concepts, but they have many advanced functions and are used widely in statistical research. R is available for free and can be downloaded on the Internet. This is sometimes a great advantage and makes the user independent of updated licenses. Those who wish to make the effort to learn and use R will be part of a large statistical community (R Development Core Team 2012). It may, however, take some effort if one is unfamiliar with computer programming.

*Biostatistics with R: An Introduction to Statistics Through Biological Data* by Shahbaba (2012) gives a very useful step-by-step introduction to the R software platform using biological data. The many statistical methods available through so-called R packages and the (free) availability of the software makes it very attractive, but its somewhat more complicated structure compared to commercial software like SPSS, STATISTICA, or STATA might make it less relevant for those who use mostly so-called standard methods and have access to commercial software.

Various regression methods play a very important part in the analysis of biological data including food science and technology and nutrition research. *Regression Methods in Biostatistics* by Vittinghoff et al. (2012) gives an introduction to explorative and descriptive statistics and basic statistical methods. Linear, logistic, survival, and repeated measures models are covered without a too-overwhelming focus on mathematics and with applications to biological data. The software STATA that is widely used in biostatistics and epidemiology is used throughout the book.

Those who work much with nutrition, clinical trials, and epidemiology with respect to food will find very useful topics in textbooks such as *Statistics Applied to Clinical Trials* by Cleophas and Zwinderman (2012) and *A Pocket Guide to Epidemiology* by Kleinbaum et al. (2007). These books cover concepts that are statistical in nature but more related to clinical research and epidemiology. Clinical research is in many ways a scientific knowledge triangle comprised of medicine, biostatistics, and epidemiology.

Lehmann (2011) in his book *Fisher, Neyman, and the Creation of Classical Statistics* gives historical background of the scientists that laid the foundation for statistical analysis – Ronald A. Fisher, Karl Pearson, William Sealy Gosset, and Egon S. Pearson. Those with some insight into classical statistical methods and with a historic interest in the subject should derive much pleasure from reading about the discoveries that we sometimes take for granted in quantitative research. The text is not targeted at food applications but, without going into all the mathematics, provides a historical introduction to the development of statistics and experimental design. Some of the methods presented from their historical perspective might be difficult to follow if one is unfamiliar with statistics. However, a more comprehensive description of the "lady tasting tea" experiment is provided together with the many important concepts later discussed in relation to food science and technology.

### 2.1.2 *Advanced Text on the Theoretical Foundation in Statistics*

Numerous textbooks have a more theoretical approach to statistics, and many are collected in the series *Springer Texts in Statistics*. *Modern Mathematical Statistics with Applications* by Devore and Berk (2012) provides comprehensive coverage of the theoretical foundations of statistics. Another recommended text that gives an overview of the mathematics in a shorter format is the SpringerBrief *A Concise Guide to Statistics* by Kaltenbach (2011). These two and other textbooks with an emphasis on mathematical statistics are useful for exploring the fundamentals of statistical science with a more mathematical than applied approach to data analysis. However, most readers with a life science or biology-oriented background may find the formulas, notations, and equations challenging. Applied knowledge and mathematical knowledge often go hand in hand. It is usually more inspiring to learn the basic foundation if there is an applied motivation for a specific method. Many readers might therefore

wish to consult textbooks with a more mathematical approach on "a-need-to-know" basis and begin with the previously recommended texts on applied use.

## 2.2 Describing Data

Food scientists encounter many types of data. Consumers report their preferences, sensory panels give scores on flavor and taste characteristics, laboratories provide chemical and microbial data, and management sets specific targets on production costs and expected sales. Analysis of all these data begins with a basic understanding of their statistical nature.

The first step in choosing an appropriate statistical method is to recognize the type of data. From a basic statistical point of view there are two main types of data – categorical and numerical. We will discuss them thoroughly before continuing with more specific types of data like ranks, percentages, and ratios. Many data sets contain missing data and extreme observations often called outliers. They also provide information and require attention.

To illustrate the different types of data and how to describe them, we will use yogurt as an example. Yogurt is a dairy product made from pasteurized and homogenized milk, fortified to increase dry matter, and fermented with the lactic acid bacteria *Streptococcus thermophilus* and *Lactobacillus delbrueckii subspecies bulgaricus*. The lactic acid bacteria ferment lactose into lactic acid, which lowers pH and makes the milk protein form a structural network, giving the typical texture of fresh fermented milk products. It is fermented at about 45°C for 5–7 h. A large proportion of yogurts also add fruit, jam, and flavor. There are two major yogurt processing technologies – stirring and setting. Stirred yogurt is fermented to a low pH and thicker texture in a vat and then pumped into packages, while set yogurt is pumped into packages right after lactic acid bacteria have been added to the milk; the development of a low pH and the formation of a gel-like texture take part in the package. The fat content can change from 0% fat to as high as 10% for some traditional types. It is a common nutritious food item throughout the world with a balanced content of milk proteins, dairy fats, and vitamins. In some yogurt products, especially the nonfat types, food thickeners are added to improve texture and mouthfeel (for a comprehensive coverage of yogurt technology see, e.g., Tamine and Robinson 2007).

### *2.2.1 Categorical Data*

In Table 2.1 categorical and numerical data from ten yogurt samples are presented to illustrate types of data. The first three variables (flavor, added thickener, and fat content) are all derived from categorical data. Observations that can be grouped into categories are thus called categorical data. Statistically they contain less information than numerical data (to be covered later) but are often easier to interpret and

**Table 2.1** Types of data and variables given by some yogurts samples

| Type of data | Categorical | | | Numerical | |
|---|---|---|---|---|---|
| Type of variable | Nominal | Binary | Ordinal | Discrete | Continuous |
| Sample | Flavor | Added thickener | Fat content | Preference (1: low, 5: high) | pH |
| 1 | Plain | Yes | Fat free | 1 | 4.41 |
| 2 | Strawberry | Yes | Low fat | 4 | 4.21 |
| 3 | Blackberry | No | Medium | 3 | 4.35 |
| 4 | Vanilla | No | Full fat | 4 | |
| 5 | Vanilla | Yes | Full fat | 4 | 4.15 |
| 6 | | No | Low fat | 3 | 4.38 |
| 7 | Strawberry | Yes | Fat free | 2 | 4.22 |
| 8 | Vanilla | Yes | Fat free | 2 | 4.31 |
| 9 | Plain | No | Medium | 2 | 4.22 |
| 10 | Strawberry | No | Full fat | | 6.41 |

understand. Low-fat yogurt conveys more clearly the fat content to most consumers than the exact fat content. Consumers like to know the fat content relative to other varieties and not the exact amount. Categorical data are statistically divided into three groups – nominal, binary, and ordinal data. Knowing the type of data one is dealing with is essential because that dictates the type of statistical analysis and tests one will perform.

Data that fall under nominal variables (e.g., "flavor" in Table 2.1) are comprised of categories, but there is no clear order or rank. Perhaps one person prefers strawberry over vanilla, but from a statistical point of view there is no obvious order to yogurt flavors. Other typical examples of numerical data in food science are food group (e.g., dairy, meat, vegetables), method of conservation (e.g., canned, dried, vacuum packed), and retail (e.g., supermarket, restaurant, fast-food chain). Statistically, nominal variables contain less information than ordinal or numerical variables. Thus, statistical methods developed for nominal variables can be used on other types of data, but with lower efficiency than other more appropriate or efficient methods.

If measurements can only be grouped into two mutually exclusive groups, then the data are called binary (also called dichotomous). In Table 2.1 the variable "added thickener" contains binary data. As long as the data can be grouped into only two categories, they should be treated statistically as binary data. Binary data can always be reported in the form of *yes* or *no*. Sometimes for binary data, *yes* and *no* are coded as 1 and 0, respectively. It is not necessary, but it is convenient in certain statistical analysis, especially when using statistical software. Binary variables are statistically often associated with giving the *risk* of something. One example is the risk of foodborne disease bacteria (also called pathogenic bacteria) in a yogurt sample. Pathogenic bacteria are either detected or not. However, from a statistical point of view the risk is estimated on a scale of 0 to 1, but for individual observations the risk is either present (pathogenic bacteria detected) or not (pathogenic bacteria not detected). Thus, it could then be presented as binary data for individual observations.

Data presented by their relative order of magnitude, such as the variable "fat content" in Table 2.1, are ordinal. Fat content expressed as fat free, low fat, medium fat, or full fat has a natural order. Since it has a natural order of magnitude with more than two categories, it contains more statistical information than nominal and binary data. Ordinal data can be simplified into binary data – e.g., reduced fat (combining the categories fat free, low fat, and medium fat) or nonreduced (full fat), but with a concomitant loss of information. Statistical methods used on nominal data can also be used on ordinal data, but again with a loss of statistical information and efficiency. If ordinal data can take only two categories, e.g., thick or thin, they should be considered binary.

### *2.2.2 Numerical Data*

Observations that are measurable on a scale are numerical data. In Table 2.1, two types of numerical data are illustrated. These are discrete or continuous. In applied statistics both discrete and ordinal data are sometimes analyzed using methods developed for continuous data, even though it is not always appropriate according to statistical theory. Numerical data contain more statistical information than categorical data. Statistical methods suitable for categorical data analysis can therefore be applied to numerical data, but again with a loss of information. Therefore, it is common to apply other methods that take advantage of their additional statistical information compared with categorical data.

Participants in a sensory test may score samples on their preference using only integers like 1, 2, 3, 4, or 5. Observations that can take only integers (no decimals) are denoted discrete data. The "distance" between discrete variables is assumed to be the same. For instance the difference in preference between a score of 2 and 3 is assumed to be the same as the difference between scores 4 and 5. It is therefore possible to estimate, for example, the average and sum of discrete variables. If the "distance" cannot be assumed equal, discrete data should instead be treated as ordinal.

The pH of yogurt samples is an example of continuous data. Continuous data are measured on a scale and can be expressed with decimals. They contain more statistical information than the other types of data in Table 2.1. Thus, statistical methods applied to categorical or discrete data can be used on variables with continuous data, but not vice versa. For example, the continuous data on pH can be divided into those falling below and those falling above pH 4.6 and thereby be regarded as binary data and analyzed using methods for such data. However, if we have only information in our database about whether the yogurt sample is below or above pH 4.6, it is not possible to make such binary data continuous data. Thus, it is always useful to save the original continuous data even though they may be divided into categories for certain analysis. One may perhaps need the original data's additional statistical information at a later stage. Many advanced statistical methods like regression were first developed for continuous data as an outcome and then later expanded for use with categorical data.

### 2.2.3 Other Types of Data

Understanding the properties of categorical and numerical data serves as the foundation of quantitative and statistical analysis. However, in applied work with statistics one often encounters other specific types of data that require our attention. Some examples are missing data, ranks, ratios, and outliers. They have certain properties that one should be aware of.

Missing data are unobserved observations. Technical problems during laboratory analysis or participants' not answering all questions in a survey are typical reasons for missing data. In Table 2.1 yogurt samples 4, 6, and 10 have missing data for some of the variables. A main issue with missing data is whether there is an underlying reason why data are missing for some observations.

Statistical research on the effect of missing data is driven by medical statistics. It is a very important issue in both epidemiology and clinical studies and especially with longitudinal data (Song 2007; Ibrahim and Molenberghs 2009). What if a large proportion of those patients that do not experience any health improvement of a new drug drop out of a clinical trial? Statistical analysis could then be influenced greatly by the proportion of missing data and the biased medical conclusions that were reached. Missing data should therefore never be simply neglected or just replaced by a given value (e.g., the mean of nonmissing data) without further investigation. The issue of missing data is likewise important in food science and nutrition. We will use the terminology developed in medical statistics to understand how missing data could be approached.

Let us assume that we are conducting a survey on a new yogurt brand. We want to examine how fat content influences sensory preferences. A randomly selected group of 500 consumers is asked to complete a questionnaire about food consumption habits including their consumption of different yogurts. However, only 300 questionnaires are returned. Thus, we have 200 missing observations in our data set. According to statistical theory on missing data, these 200 missing observations can be classified as missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR). This terminology is, unfortunately, not self-explanatory and somewhat confusing. However, one may say generally that it concerns the probability that an observation is missing.

*Missing completely at random (MCAR)*: It is assumed here that the probability of missing data is unrelated to the possible value of a given missing observation (given that the observation was not missing and was actually made) or any other observations in one's data set. For instance, if the 200 missing observations were randomly lost, then it is unlikely that the probability to be missing is related to the preference scores of yogurt or any selected demographic data. Perhaps the box with the last 200 questionnaires was accidentally thrown away! For MCAR any piece of data is just as likely to be missing as any other piece of data. The nice feature is that the statistical estimates and resulting conclusions are not biased by the missing data. Fewer observations give increased uncertainty (i.e., reduced statistical power or conse-

quently broader confidence intervals), but what we find is unbiased. They may just remain missing in your data set in further statistical analyses. All statistical analyses with MCAR give unbiased information on what influences yogurt preferences.

*Missing at random (MAR)*: It is also assumed that the probability of missing data is unrelated to the possible value of a given missing observation (given that the observation was not missing and was actually made) but related to some other observed data in the data set. For example, if younger participants are less likely to complete the questionnaire than older ones, the overall analysis will be biased with more answers from older participants. However, separate analysis of young and old participants will be unbiased. A simple analysis to detect possible MAR in the data set entails examining the proportion of missing data between key baseline characteristics. Such characteristics in a survey could be the age, gender, and occupation of the participants.

*Missing not at random (MNAR)*: Missing data known as MNAR present a more serious problem! It is assumed here that the probability of a missing observation is related to the possible value of a given missing observation (given that the observation was not missing and was actually made) or other unobserved or missing data. Thus, it is very difficult to say how missing data could influence one's statistical analysis. If participants who prefer low-fat yogurt are less likely to complete the questionnaire, then the results will be biased, but the information that it is due to their low preference for low-fat yogurt is lacking! The overall results will be biased and incorrect conclusions could be reached.

Whenever there are missing data, one needs to determine if there is a pattern in the missingness and try to explain why the data are missing. In Table 2.1 data on preference are missing for yogurt sample 10. However, the pH is exceptionally high. Perhaps something went wrong during the manufacture of the yogurt and the lactic acid bacteria did not ferment the lactose into lactic acid and so did not lower the pH. That could explain why preference was not examined for this sample. Therefore, always try to gather information to explain why data are missing. The best strategy is always to design a study in a way that minimizes the risk for missing data.

Especially in consumer surveys and sensory analysis, it is common to rank food samples. Ranking represents a relationship between a set of items such that, for any two items, the first is either ranked higher than, lower than, or equal to the second. For example, a consumer might be asked to rank five yogurt samples based on preference. This is an alternative to just giving a preference score for each sample. If there is no defined universal scale for the measurements, it is also feasible to use ranking for comparison of samples. Statistically speaking, data based on ranking have a lot in common with ordinal data, but they may be better analyzed using methods that take into account the ranks given by each consumer. It is therefore important to recognize rankings from other types of data.

A ratio is a relationship between two numbers of the same kind. We might estimate the ratio of calorie intake from dairy products compared with that from vegetables. Percentage is closely related as it is expressed as a fraction of 100. In Latin *per cent* means *per hundred*. Both ratios and percentages are sometimes treated as continuous data in statistical analysis, but this should be done with great caution. The statistical properties might be different around the extremes of 0 or 100%. Therefore, it is important

to examine ratio and percentage data to assess how they should be treated statistically. Sometimes ratios and percentages are divided into ordinal categories if they cannot be properly analyzed with methods for continuous data.

Take a closer look at the data for sample 10 in Table 2.1. All the other samples have pH measurements around 4.5, but the pH of sample 10 is 6.41. It is numerically very distant from the other pH data. Thus, it might be statistically defined as an outlier, but it is not without scientific information. Since it deviates considerably from the other samples, the sample is likely not comparable with the other ones. This could be a sample without proper growth of the lactic acid bacteria that produce the acid to lower the pH during fermentation. Outliers need to be examined closely (just like missing data) and be treated with caution. With the unnatural high pH value of sample 10 compared with the other samples, the average pH of all ten samples would not be a good description of the typical pH value among the samples. Therefore, it might be excluded – or assessed separately – in further statistical analysis.

## 2.3 Summarizing Data

### *2.3.1 Contingency Tables (Cross Tabs) for Categorical Data*

A contingency table is very useful for describing and comparing categorical variables. Table 2.2 is a contingency table with exemplified data to illustrate a comparison of preferences for low- or full-fat yogurt between men and women. The number of men and women in these data is different, so it is very useful to provide the percentage distribution in addition to the actual numbers. It makes the results much easier to read and interpret. Statistically, it does not matter which categorical variable is presented in rows or columns. However, it is rather common to have the variable defining the outcome of interest (preferred type of yogurt in our example) in columns and the explanation (gender of survey participants) in rows (Agresti 2002). In these illustrative data, women seem on average to prefer low-fat yogurt, and men seem to prefer full-fat yogurt. Perhaps this is a coincidence just for these 100 people, or is it a sign of a general difference in preference for yogurt types among men and women? Formal statistical tests and models are needed to evaluate this.

### *2.3.2 The Most Representative Value of Continuous Data*

Let us examine again the pH measurements of our ten yogurt samples. Remember, we have missing data for sample 4; therefore, we have only nine data observations. Reordering the pH data in ascending yields 4.15, 4.21, 4.22, 4.22, 4.31, 4.35, 4.38,

**Table 2.2** Comparison of two categorical variables

| | Low-fat yogurt | Full-fat yogurt | Total |
|---|---|---|---|
| Men | 12 (30%) | 28 (70%) | 40 (100%) |
| Women | 45 (75%) | 15 (25%) | 60 (100%) |
| Total | 60 (60%) | 40 (40%) | 100 (100%) |

4.41, and 6.41. What single number represents the most typical value in this data set? For continuous data, the most "typical" value, or what is referred to in statistics as the central location, is usually given as either the mean or median. The mean is the sum of values divided by the number of values (the mean is also known as the "standard" average). It is defined for a given variable $X$ with $n$ observations as

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and is estimated in our example as

$$\text{mean} = \frac{4.15 + 4.21 + 4.22 + 4.22 + \mathbf{4.31} + 4.35 + 4.38 + 4.41 + 6.41}{9} = 4.52.$$

The single outlier measurement of pH 6.41 has a relatively large influence on the estimated mean. An alternative to the mean could be to use the median. The median is the numeric values separating the upper half of the sample or, in other words, the value in the middle of our data set. The median is found by ranking all the observations from lowest to highest value and then picking the middle one. If there is an even number of observations and thus no single middle value, then the median is defined as the mean of the two middle values. In our example the middle value is 4.31 (indicated by bold typeface in the equation estimating the mean). A rather informal approach to deciding whether to use the mean or median for continuous data is to estimate them both. If the median is close to the mean, then one can usually use the mean, but if they are substantially different, then the median is usually the better choice.

### *2.3.3 Spread and Variation of Continuous Data*

Describing the central location or the most typical value is telling only half the story. One also needs to describe the spread or variation in the data. For continuous data it is common to use the standard deviation or simply the maximum and minimum values. These might not be so intuitive as the mean and median. If the data set has no extreme outliers or a so-called skewed distribution (many very high or low

values compared with the rest of the data), it is common to use the standard deviation. It can be estimated for a given variable $X$ with $n$ observations as

$$\text{Standard deviation (SD)} = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}{n-1}}$$

If we exclude the extreme pH value in sample 10 (regarded as an outlier), then the new mean of our remaining eight data points on pH is estimated to be 4.28 and the standard deviation is estimated as

$$\text{SD} = \sqrt{\frac{\left(4.41-4.28\right)^2 + \left(4.21-4.28\right)^2 + \ldots + \left(4.22-4.28\right)^2}{8-1}} = 0.09$$

Fortunately, most spreadsheets such as Excel or OpenOffice or statistical software can estimate standard deviations and other statistics efficiently and lessen the need to know the exact estimation formulas and computing techniques. If we assume that our data are more or less normally distributed, then a distance of one standard deviation from the mean will contain approximately 65% of our data. Two standard deviations from the mean will contain approximately 95% of our data. This is the main reason why continuous data are often described using the mean and standard deviation.

If a data set has a skewed distribution or contains many outliers or extreme values, it is more common to describe the data as the median, with the spread represented by the minimum and maximum values. To reduce the effect of extreme values, the so-called interquartile range is an alternative measure of spread in data. It is equal to the difference between the third and first quartiles. It can be found by ranking all the observations in ascending order. For the sake of simplicity, let us assume one has 100 observations. The lower boundary of the interquartile range is at the border of the first 25% of observations – in this example observation 25 if they are ranked in ascending order. The higher boundary of the interquartile range is at the border of the first 75% of observations – in this example observation 75 if they are ranked in ascending order.

## 2.4 Descriptive Plots

The adage “a picture is worth a thousand words” refers to the idea that a complex idea can be conveyed with just a single still image. Actually, some attribute this quote to the Emperor Napoleon Bonaparte, who allegedly said, “Un bon croquis vaut mieux qu’un long discours” (a good sketch is better than a long speech). We might venture to rephrase Napoleon to describe data – “A good plot is worth more

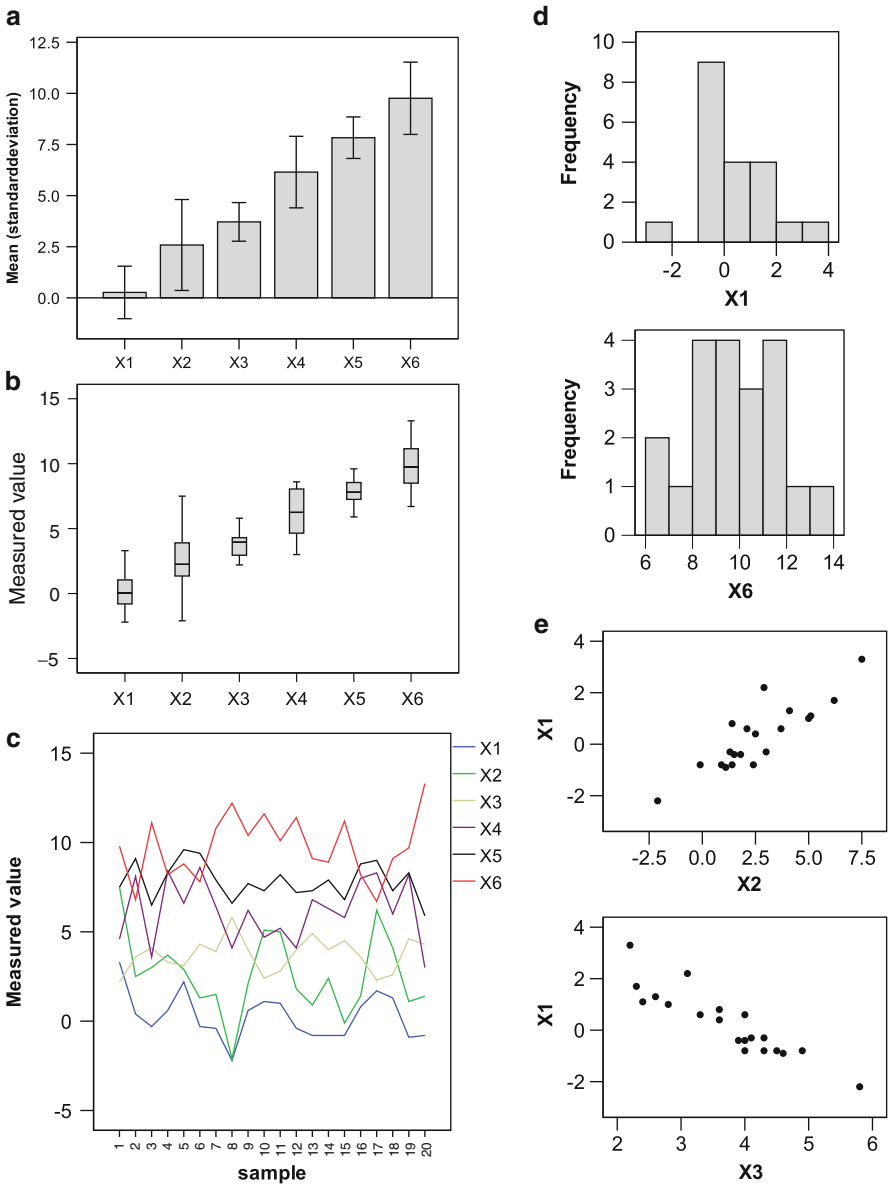


**Fig. 2.1** Typically used descriptive plots. The plots are (**a**) bar chart, (**b**) box plot, (**c**) line plot, (**d**) histogram, and (**e**) scatterplot. All plots were made using data in Box 5.1 in Chap. 5

than a thousand data." Plots are very useful for describing the properties of data. It is recommended that these be explored before further formal statistical analysis is conducted. Some examples of descriptive plots are given in Fig. 2.1

### 2.4.1 Bar Chart

A bar chart or bar graph is a chart with rectangular bars with lengths proportional to the values they represent. They can be plotted vertically or horizontally. For categorical data the length of the bar is usually the number of observations or the percentage distribution, and for discrete or continuous data the length of the bar is usually the mean or median with (error) lines sometimes representing the variation expressed as, for example, the standard deviation or minimum and maximum values. Bar charts are very useful for presenting data in a comprehensible way to a nonstatistical audience. Bar charts are therefore often used in the mass media to describe data.

### 2.4.2 Histograms

Sometimes it is useful to know more about the exact spread and distribution of a data set. Are there many outliers, or is the data distribution equally spread out? To know more about this, one could make a histogram, which is a simple graphical way of presenting a complete set of observation in which the number (or percentage frequency) of observations is plotted for intervals of values.

### 2.4.3 Box Plots

A box plot (also known as a box-and-whisker diagram) is a very efficient way of describing numerical data. It is often used in applied statistical analysis but is not as intuitive for nonstatistical readers. The plot is based on a five-number summary of a data set: the smallest observation (minimum), the lower quartile (cutoff value of the lowest 25% of observations if ranked in ascending order), the median, the upper quartile (cutoff value of the first 75% of observations if ranked in ascending order), and the highest observation (maximum). Often the "whiskers" may indicate the 2.5% and 97.5% values with outliers and extreme values indicated by individual dots. Box plots provide more information about the distribution than bar charts. If the line indicating the median is not in the middle of the box, then this is usually a sign of a skewed distribution.

### 2.4.4 Scatterplots

Scatterplots are very useful for displaying the relationship between two numerical variables. These plots are also sometimes called XY-scatter or XY-plots in certain software. A scatterplot is a simple graph in which the values of one variable are

plotted against those of the other. These plots are often the first step in the statistical analysis of the correlation between variables and subsequent regression analysis.

### *2.4.5 Line Plots*

A line plot or graph displays information as a series of data points connected by lines. Depending on what is to be illustrated, the data points can be single observations or statistical estimates as, for example the mean, median, or sum. As with the bar chart, vertical lines representing data variation, for example standard deviation, may then be used. Line plots are often used if one is dealing with repeated measurements over a given time span.

## 2.5 Statistical Inference (the *p*-Value Stuff)

Descriptive statistics are used to present and summarize findings. This may form the basis for decision making and conclusions in, for example, scientific and academic reports, recommendations to governmental agencies, or advice for industrial production and food development. However, what if the findings were just due to a coincidence? If the experiment were repeated and new data collected, a different conclusion might be reached. With statistical methods it is necessary to assess whether findings are due to randomness and coincidence or are representative of the "true" or underlying effect. One set of tools is called statistical tests (or inference) and form the basis of *p*-values and confidence intervals.

The basis is a hypothesis that could be rejected in relation to an alternative hypothesis given certain conditions. In statistical sciences these hypotheses are known as the null hypothesis (typically a conservative hypothesis of no "real" difference between samples, no correlation, etc.) and the alternative hypothesis (i.e., that the null hypothesis is not "in reality" true). The principle is to assume that the null hypothesis is true. Methods based on mathematical statistics have been developed to estimate the probability of outcomes that are at least as "rare" as the observed outcomes, given the assumption that the null hypothesis is true. This probability is the well-known *p*-value. If this probability is small (typical less than 5%), then the null hypothesis is typically rejected in favor of the alternative hypothesis. The level of this probability before the null hypothesis is rejected is called the significance level (often denoted $\alpha$).

The relationship between the (unknown) reality if the null hypothesis is true or not and the decision to accept or reject the null hypothesis is shown in Table 2.3. Two types of error can be made – Type I and Type II errors. The significance level – $\alpha$ – is typically set low (e.g., 5%) to avoid Type I errors that from a methodological point of view are regarded as being more "serious" than Type II errors. The null hypothesis is usually very conservative and assumes, for example, no difference between groups or no correlation. The Type II error is denoted by $\beta$. The statistical

**Table 2.3** Two types of statistical errors: Types I and II errors and their relationship to significance level $\alpha$ and the statistical power (1-$\beta$)

| | Null hypothesis ($H_0$) is true | Alternative hypothesis ($H_1$) is true |
|---|---|---|
| Accept null hypothesis | Correct decision | Type II error: $\beta$ |
| Reject null hypothesis | Type I error: $\alpha$ | Correct decision |

power is the ability of a test to detect a true effect, i.e., reject the null hypothesis if the alternative hypothesis is true. Thus, this is the opposite of a Type II error and consequently equal to 1-$\beta$.

## 2.6 Overview of Classical Statistical Tests

Classical statistical tests are pervasive in research literature. More complex and general statistical models can often express the same information as these tests. Table 2.4 presents a list of some common statistical tests. It goes beyond the scope of this brief text to explain the statistical and mathematical foundations of these tests, but they are covered in several of the recommended textbooks. Modern software often has menu-based dialogs to help one determine the correct test. However, a basic understanding of their properties is still important.

## 2.7 Overview of Statistical Models

Generally speaking, so-called linear statistical models state that your outcome of interest (or a mathematical transformation of it) can be predicted by a linear combination of explanatory variables, each of which is multiplied by a parameter (sometimes called a coefficient and often denoted $\beta$). To avoid having the outcome be estimated as zero if all explanatory variables are zero, a constant intercept (often denoted $\beta_0$) is included. The outcome variable of interest is often called the dependent variable, while the explanatory variables that can predict the outcome are called independent variables.

The terminology in statistics and experimental design may sometimes be somewhat confusing. In all practical applications, models like linear regression, analysis of covariance (ANCOVA), analysis of variance (ANOVA), or general linear models (GLM) are very similar. Their different terminology is due as much to the historical tradition in statistical science as to differences in methodology. Many of these models with their different names and terminologies can be expressed within the framework of generalized linear models. It was common to develop mathematical methods to estimate parameter values and $p$-values that could be calculated manually by hand and

**Table 2.4** Proposed statistical tests or models depending on properties of the outcome and explanatory variable. Nonparametric alternative is given in *brackets* if assumptions on normal distributions are not valid. The number of mentioned tests is limited and recommendations may vary depending on the nature of the data and purpose of analysis

| Purpose with statistical analysis | Type of outcome data | | | | |
|---|---|---|---|---|---|
| | Nominal | Binary | Ordinal | Discrete | Continuous |
| Against specific null hypothesis about expected mean or proportion | Chi-squared test | Binomial test | Chi-squared test | One sample *t*-test | One sample *t*-test |
| Relationship with continuous explanatory variable | "Use a statistical model" | "Use a statistical model" | Spearman correlation | Pearson (Spearman) correlation | Pearson (Spearman) correlation |
| Difference in expected mean or proportions between two groups | Chi-squared test for cross tabs | Chi-squared test for crosstabs | Chi-squared test for crosstabs | Two-sample *t*-test (Mann–Whitney *U* test) | Two-sample *t*-test (Mann–Whitney *U* test) |
| Difference between mean or proportions between more than two groups | Chi-squared test for crosstabs | Chi-squared test for crosstabs | Chi-squared test for crosstabs | Analysis of variance (Kruskal–Wallis H test) | Analysis of variance (Kruskal–Wallis H test) |
| Analyzed as linear statistical model | Multinomial logistic regression | Binary logistic regression | Ordinal logistic regression | Linear regression/ general linear model | Linear regression/ general linear model |
| Two clustered or repeated measurements | McNemar–Bowker test | McNemar test | McNemar–Bowker test | Paired sample *t*-test (Wilcoxon signed-rank test) | Paired sample *t*-test (Wilcoxon signed-rank test) |
| Statistical model for clustered or repeated measurements | Mixed multinomial logistic regression or GEE | Mixed binary logistic regression or GEE | Mixed ordinal logistic regression or GEE | Linear mixed model or GEE | Linear mixed model or GEE |

*GEE* generalized estimating equations

with the help of statistical tables. Most graduates in statistics are familiar with such methods for simple regression and ANOVA methods. However, recent innovations in mathematical statistics, and not least computers and software, have in an applied sense replaced such "manual" methods. These computer-assisted methods are usually based on the theory of so-called likelihood functions and involve finding their maximum values by using iterations. In other words, these are methods where computer software is needed for most applied circumstances. The theory behind maximum-likelihood estimations is covered in several of the more advanced recommended textbooks.

Linear statistical models are often described within the framework of generalized linear models. The type of model is determined by the properties of the outcome variable. A dependent variable with continuous data is usually expressed with an identity link and is often referred to by more traditional terms such as linear regression or analysis of variance. If the dependent variable is binary, then it is usually expressed by a logit link and is often referred to by the more traditional term logistic regression. Count data use a log link and the statistical model is traditionally referred to as Poisson regression (e.g., Dobsen and Barnett 2008).

## References

Agresti A (2002) Categorical data analysis, 2nd edn. Wiley, Hoboken

Cleophas TJ, Zwinderman AH (2012) Statistics applied to clinical studies, 5th edn. Springer, Dordrecht

Devore JL, Kenneth N (2012) Modern mathematical statistics with applications, 2nd edn. Springer, New York

Dobsen AJ, Barnett A (2008) An introduction to generalized linear models, 3rd edn. CRC Press, London

Ibrahim JG, Molenberghs G (2009) Missing data methods in longitudinal studies: a review. Test 18:1–43. doi:10.1007/s11749-009–0138-x

Kaltenbach HM (2011) A concise guide to statistics. Springer, New York

Kleinbaum DG, Sullivan K, Barker N (2007) A pocket guide to epidemiology. Springer, New York

Lehmann EL (2011) Fisher, Neyman, and the creation of classical statistics. Springer, New York

Madsen B (2011) Statistics for non-statisticians. Springer, Heidelberg

Marques de Sá JP (2007) Applied statistics using SPSS, STATISTICA, MATLAB and R, 2nd edn. Springer, Berlin

Shahbaba R (2012) Biostatistics with R: an introduction to statistics through biological data. Springer, New York

Song PXK (2007) Missing data in longitudinal studies. In: Correlated data analysis: modeling, analytics, and applications. Springer, New York

Tamine AY, Robinson RK (2007) Tamine and Robinson's yoghurt science and technology, 3rd edn. CRC Press, Cambridge

R Development Core Team (2012) The R project for statistical computing. http://www.r-project.org. Accessed 30 Apr 2012

Vittinghoff E, Glidden DV, Shiboski SC, McCulloch CE (2012) Regression methods in biostatistics: linear, logistic, survival and repeated measures models, 2nd edn. Springer, New York

http://www.springer.com/978-1-4614-5009-2

Statistics in Food Science and Nutrition

Pripp, A.H.

2013, VIII, 66 p. 15 illus., 3 illus. in color., Softcover
ISBN: 978-1-4614-5009-2