Chapter 2
Statistical Concepts

Ad J. Feelders
Utrecht University, The Netherlands

2.1. Introduction

Statistics is the science of collecting, organizing and drawing conclusions from data. How to properly produce and collect data is studied in experimental design and sampling theory. Organisation and description of data is the subject area of descriptive statistics, and how to draw conclusions from data is the subject of statistical inference. In this chapter the emphasis is on the basic concepts of statistical inference, and the other topics are discussed only inasfar as they are required to understand these basic concepts.

In Section 2.2 we discuss the basic ideas of probability theory, because it is the primary tool of statistical inference. Important concepts such as random experiment, probability, random variable and probability distribution are explained in this section.

In Section 2.3 we discuss a particularly important kind of random experiment, namely random sampling, and a particularly important kind of probability distribution, namely the sampling distribution of a sample statistic. Random sampling and sampling distributions provide the link between probability theory and drawing conclusions from data, i.e. statistical inference.

The basic ideas of statistical inference are discussed in Section 2.4. Inference procedures such as point estimation, interval estimation (confidence intervals) and hypothesis testing are explained in this section. Next to the frequentist approach to inference we also provide a short discussion of likelihood inference and the Bayesian approach to statistical inference. The interest in the latter approach seems to be increasing rapidly, particularly in the scientific community. Therefore a separate chapter of this volume is entirely dedicated to this topic (see Chapter 4).

In Section 2.5 we turn to the topic of prediction. Once a model has been estimated from the available data, it is often used to predict the value of some variable of interest. We look at the different sources of error in prediction in order to gain an understanding of why particular statistical methods tend to work well on one type of dataset (in terms of the dimensions of the dataset, i.e. the number of observations and number of variables) but less so on others. The emphasis in this section is on the decomposition of total prediction error into an irreducible and reducible part, and in turn the decomposition of the reducible part into a bias and variance component. Flexible techniques such as classification and regression trees, and neural networks tend to have low bias and high variance whereas the more inflexible "conventional" statitical methods such as linear regression and linear discriminant analysis tend to have more bias and less variance than their "modern" counterparts. The well-known danger of overfitting, and ideas of model averaging presented in Section 2.6, are rather obvious once the bias/variance decomposition is understood.

In Section 2.6, we address computer-intensive statistical methods based on resampling. We discuss important techniques such as cross-validation and bootstrapping. We conclude this section with two model averaging techniques based on resampling the available data, called bagging and arcing. Their well-documented success in reducing prediction error is primarily due to reduction of the variance component of error.

We close off this chapter with some concluding remarks.

2.2. Probability

The most important tool in statistical inference is probability theory. This section provides a short review of the important concepts.

2.2.1 Random Experiments

A *random experiment* is an experiment that satisfies the following conditions

1. all possible distinct outcomes are known in advance,
2. in any particular trial, the outcome is not known in advance, and
3. the experiment can be repeated under identical conditions.

The *outcome space* Ω of an experiment is the set of all possible outcomes of the experiment.

Example 2.1. Tossing a coin is a random experiment with outcome space $\Omega = \{\text{H,T}\}$

Example 2.2. Rolling a die is a random experiment with outcome space $\Omega = \{1,2,3,4,5,6\}$