Statistical Concepts for Intelligent Data Analysis

A.J. Feelders

Utrecht University, Institute for Information & Computing Sciences,
Utrecht, The Netherlands

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 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 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 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.

In section 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 a 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 6, are rather obvious once the bias/variance decomposition is understood.

In section 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 Probability

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

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

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

Something that might or might not happen, depending on the outcome of the experiment, is called an *event*. Examples of events are “coin lands heads” or “die shows an odd number”. An event A is represented by a subset of the outcome space. For the above examples we have $A = \{\text{H}\}$ and $A = \{1,3,5\}$ respectively. Elements of the outcome space are called elementary events.

2.2 Classical definition of probability

If all outcomes in Ω are equally likely, the probability of A is the number of outcomes in A, which we denote by $M(A)$ divided by the total number of outcomes M

$$P(A) = \frac{M(A)}{M}$$

If all outcomes are equally likely, the probability of {H} in the coin tossing experiment is $\frac{1}{2}$, and the probability of {5,6} in the die rolling experiment is $\frac{1}{3}$. The assumption of equally likely outcomes limits the application of the concept of probability: what if the coin or die is not 'fair'? Nevertheless there are random experiments where this definition of probability is applicable, most importantly in the experiment of random selection of a unit from a population. This special and important kind of experiment is discussed in the section 3.

2.3 Frequency definition of probability

Recall that a random experiment may be repeated under identical conditions. When the number of trials of an experiment is increased indefinitely, the relative frequency of the occurrence of an event approaches a constant number. We denote the number of trials by m, and the number of times A occurs by $m(A)$. The frequency definition of probability states that

$$P(A) = \lim_{m \to \infty} \frac{m(A)}{m}$$

The law of large numbers states that this limit does indeed exist. For a small number of trials, the relative frequencies may show strong fluctuation as the number of trials varies. The fluctuations tend to decrease as the number of trials increases.

Figure 1 shows the relative frequencies of heads in a sequence of 1000 coin tosses as the sequence progresses. In the beginning there is quite some fluctuation, but as the sequence progresses, the relative frequency of heads settles around 0.5.

2.4 Subjective definition of probability

Because of the demand of repetition under identical circumstances, the frequency definition of probability is not applicable to every event. According to the subjective definition, the probability of an event is a measure of the *degree of belief* that the event will occur (or has occured). Degree of belief depends on the person who has the belief, so my probability for event A may be different from yours.

Consider the statement: "There is extra-terrestrial life". The degree of belief in this statement could be expressed by a number between 0 and 1. According to the subjectivist definition we may interpret this number as the probability that there is extra-terrestrial life.

The subjective view allows the expression of all uncertainty through probability. This view has important implications for statistical inference (see section 4.3).

2.5 Probability axioms

Probability is defined as a function from subsets of Ω to the real line $\mathbb{R}$, that satisfies the following axioms

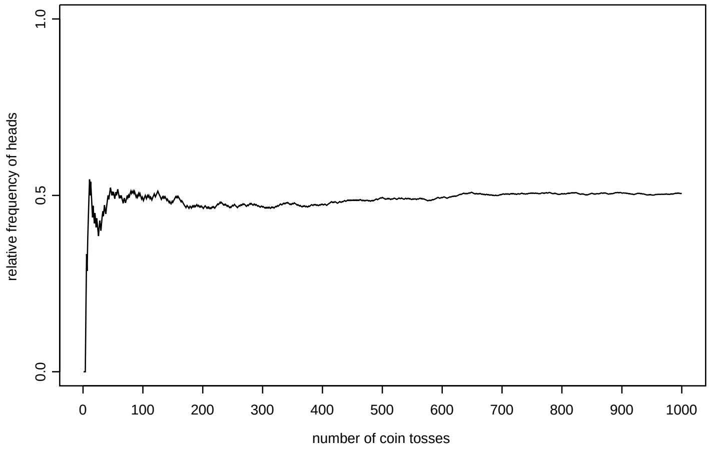


Fig. 1. Relative frequency of heads in a sequence of 1000 coin tosses

1. Non-negativity: $P(A) \geq 0$
2. Additivity: If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$
3. $P(\Omega) = 1$

The classical, frequency and subjective definitions of probability all satisfy these axioms. Therefore every property that may be deduced from these axioms holds for all three interpretations of probability.

2.6 Conditional probability and independence

The probability that event A occurs may be influenced by information concerning the occurrence of event B. The probability of event A, given that B will occur or has occurred, is called the *conditional probability* of A given B, denoted by $P(A \mid B)$. It follows from the axioms of probability that

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

for $P(B) > 0$. Intuitively we can appreciate this equality by considering that B effectively becomes the new outcome space. The events A and B are called *independent* if the occurrence of one event does not influence the probability of occurrence of the other event, i.e.

$$P(A \mid B) = P(A) \text{ , and consequently } P(B \mid A) = P(B)$$

Since independence of two events is always mutual, it is more concisely expressed by the product rule

$$P(A \cap B) = P(A)\,P(B)$$

2.7 Random variables

A random variable X is a *function* from the outcome space Ω to the real line

$$X : \Omega \rightarrow \mathbb{R}$$

Example 3. Consider the random experiment of tossing a coin twice, and observing the faces turning up. The outcome space is

$$\Omega = \{(H,T),(T,H),(H,H),(T,T)\}$$

The number of heads turning up is a random variable defined as follows

$$X((H,T)) = X((T,H)) = 1\,,\; X((H,H)) = 2\,,\; X((T,T)) = 0$$

2.8 Probability distribution

A probability function p assigns to each possible realisation x of a discrete random variable X the probability $p(x)$, i.e. $P(X = x)$. From the axioms of probability it follows that $p(x) \geq 0$, and $\sum_x p(x) = 1$.

Example 4. The number of heads turning up in two tosses of a *fair* coin is a random variable with the following probability function: $p(1) = 1/2$, $p(0) = 1/4$, $p(2) = 1/4$.

Since for continuous random variables, $P(X = x) = 0$, the concept of a probability function is useless. The probability distribution is now specified by representing probabilities as areas under a curve. The function $f : \mathbb{R} \rightarrow \mathbb{R}^+$ is called the probability density of X if for each pair $a \leq b$,

$$P(a < X \leq b) = \int_a^b f\,(x)\,dx$$

It follows from the probability axioms that $f\,(x) \geq 0$ and $\int_{-\infty}^{\infty} f\,(x)\,dx = 1$.

Example 5. Consider the random variable X with the following density function

$$f\,(x) = \begin{cases} \frac{1}{2} \text{ for } 0 \leq x \leq 2 \\ 0 \text{ otherwise} \end{cases}$$

It follows that

$$P(1/2 < X \leq 5/4) = \int_{1/2}^{5/4} 1/2dx = 1/2x|_{1/2}^{5/4} = 3/4$$

The *distribution function* is defined for both discrete and continuous random variables as the function F which gives for each $x \in \mathbb{R}$ the probability of an outcome of X at most equal to x:

$$F(x) = P(X \leq x), \quad \text{for } x \in \mathbb{R}$$

2.9 Entropy

The entropy of a random variable is the average amount of information generated by observing its value. The information provided by observing realisation $X = x$ is

$$\mathrm{H}(X = x) = \ln \frac{1}{p(x)} = -\ln p(x)$$

Example 6. Consider the random experiment of tossing a coin with probability of heads equal to 0.9, and random variable X with $X(H) = 1$ and $X(T) = 0$. What is the information generated by observing $x = 1$? $\mathrm{H}(x = 1) = -\ln 0.9 = 0.105$. The information generated by observing $x = 0$ is $\mathrm{H}(x = 0) = -\ln 0.1 = 2.303$.

Intuitively, one can appreciate that observing the outcome “heads” provides little information, since the probability of heads is 0.9, i.e. heads is almost certain to come up. Observing “tails” on the other hand provides much information, since its probability is low.

If we were to repeat this experiment very many times, how much information would be generated on average? In general

$$\mathrm{H}(X) = -\sum_i p(x_i) \ln p(x_i)$$

Example 7. The average amount of information or *entropy* generated by the previous experiment is: $-(0.9 \ln 0.9 + 0.1 \ln 0.1) = 0.325$. The entropy of tossing a fair coin is: $-(0.5 \ln 0.5 + 0.5 \ln 0.5) = 0.693$.

The information provided by the individual outcomes is weighted by their respective probabilities. Tossing a biased coin generates less information on average than tossing a fair coin, because for a biased coin, the realisation that generates much information (tails coming up in example 6) occurs less frequently.

2.10 Expectation

For a discrete random variable, the *expected value* or mean is defined as

$$\mathrm{E}(X) = \sum_x x\, p(x) \text{ , and } \mathrm{E}[h(X)] = \sum_x h(x)\, p(x)$$

for arbitrary function $h : \mathbb{R} \to \mathbb{R}$.

Example 8. Consider once more the coin tossing experiment of example 4 and corresponding probability distribution. The expected value or mean of X is

$$\mathrm{E}(X) = 1/2 \cdot 1 + 1/4 \cdot 2 + 1/4 \cdot 0 = 1$$

The definition of expectation for a continuous random variable is analogous, with summation replaced by integration.

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\, dx \text{ , and } \mathrm{E}[h(X)] = \int_{-\infty}^{\infty} h(x)\; f(x)\, dx$$

x	4	6	8	10	12
$p(x \mid C)$	1/9	2/9	1/3	2/9	1/9

Table 1. Conditional probability function $p(x \mid C)$

Example 9. (Continuation of example 5) The mean or expected value of the random variable with probability density given in example 5 is

$$\mathrm{E}(X) = \int_0^2 \frac{1}{2}\, dx = \frac{1}{2}x \Big|_0^2 = \frac{1}{2} \cdot 2 - \frac{1}{2} \cdot 0 = 1$$

The expected value $\mathrm{E}(X)$ of a random variable is usually denoted by μ. The variance σ^2 of a random variable is a measure of spread around the mean obtained by averaging the squared differences $(x - \mu)^2$, i.e.

$$\sigma^2 = \mathrm{V}(X) = \mathrm{E}(X - \mu)^2$$

The standard deviation $\sigma = \sqrt{\sigma^2}$ has the advantage that it has the same dimension as X.

2.11 Conditional probability distributions and expectation

For a discrete random variable X we define a conditional probability function as follows

$$p(x \mid C) = P(X = x \mid C) = \frac{P(\{X = x\} \cap C)}{P(C)}$$

Example 10. Two fair dice are rolled, and the numbers on the top face are noted. We define the random variable X as the sum of the numbers showing. For example $X((3,2)) = 5$. Consider now the event C : both dice show an even number. We have $P(C) = \frac{1}{4}$ and $P(\{X = 6\} \cap C) = \frac{1}{18}$ since

$$\begin{aligned} C &= \{(2,2),(2,4),(2,6),(4,2),(4,4),(4,6),(6,2),(6,4),(6,6)\} \\ \{X = 6\} \cap C &= \{(2,4),(4,2)\} \end{aligned}$$

The probability of $\{X = 6\}$ given C therefore is

$$P(X = 6 \mid C) = \frac{P(\{X = 6\} \cap C)}{P(C)} = \frac{1/18}{1/4} = \frac{2}{9}$$

The conditional probability function of X is shown in table 1. The conditional expectation of X given C is: $\mathrm{E}(X \mid C) = \sum_x x\, p(x \mid C) = 8$.

For continuous random variable X, the conditional density $f(x \mid C)$ of X given C is

$$f(x \mid C) = \begin{cases} f(x)/P(C) \text{ for } x \in C \\ \quad 0 \text{ otherwise} \end{cases}$$

2.12 Joint probability distributions and independence

The joint probability distribution of a pair of discrete random variables (X,Y) is uniquely determined by their joint probability function $p : \mathbb{R}^2 \to \mathbb{R}$

$$p(x,y) = P((X,Y) = (x,y)) = P(X = x, Y = y)$$

From the axioms of probability it follows that $p(x,y) \geq 0$ and $\sum_x \sum_y p(x,y) = 1$.

The *marginal* probability function $p_X(x)$ is easily derived from the joint distribution

$$p_X(x) = p(X = x) = \sum_y P(X = x, Y = y) = \sum_y p(x,y)$$

The conditional probability function of X given $Y = y$

$$p(x \,|\, y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p(x,y)}{p_Y(y)}$$

Definitions for continuous random variables are analogous with summation replaced by integration. The function $f : \mathbb{R}^2 \to \mathbb{R}$ is the probability density of the pair of random variables (X,Y) if for all $a \leq b$ and $c \leq d$

$$P(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f(x,y)\,dx\,dy$$

From the probability axioms it follows that

1. $f(x,y) \geq 0$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$

The marginal distribution of X is obtained from the joint distribution

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy$$

and the conditional density of X given $\{Y = y\}$ is

$$f(x \,|\, y) = \frac{f(x,y)}{f_Y(y)}$$

According to the product rule discussed in section 2.6, the events $\{X = x\}$ and $\{Y = y\}$ are independent iff

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

We now generalize the concept of independence to pairs of random variables. Discrete random variables X and Y are independent iff

$$p(x,y) = p_X(x)p_Y(y) \text{ for all } (x,y),$$

and as a consequence $p(x \,|\, y) = p_X(x)$, and $p(y \,|\, x) = p_Y(y)$. Definitions are completely analogous for continuous random variables, with probability functions replaced by probability densities.

2.13 The law of total probability

In some cases the (unconditional) probability of an event may not be calculated directly, but can be determined as a weighted average of various conditional probabilities.

Let $B_1, B_2, \ldots, B_s$ be a partition of Ω, that is $B_i \cap B_j = \emptyset$ for all $i \neq j$ and $\bigcup_{i=1}^{s} B_i = \Omega$. It follows from the axioms of probability that

$$P(A) = \sum_{i=1}^{s} P(A|B_i)P(B_i)$$

Example 11. Consider a box containing three white balls and one red ball. First we draw a ball at random, i.e. all balls are equally likely to be drawn from the box. Then a second ball is drawn at random (the first ball has not been replaced in the box). What is the probability that the second draw yields a red ball? This is most easily calculated by averaging conditional probabilities.

$$P(R_2) = P(R_2|W_1)P(W_1) + P(R_2|R_1)P(R_1) = 1/3 \cdot 3/4 + 0 \cdot 1/4 = 1/4,$$

where R_i stands for "a red ball is drawn on i-th draw" and W_i for "a white ball is drawn on i-th draw".

2.14 Bayes' rule

Bayes' rule shows how probabilities change in the light of evidence. It is a very important tool in Bayesian statistical inference (see section 4.3). Let $B_1, B_2, \ldots, B_s$ again be a partition of Ω. Bayes' rule follows from the axioms of probability

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{\sum_j P(A|B_j)P(B_j)}$$

Example 12. Consider a physician's diagnostic test for the presence or absence of some rare disease D, that only occurs in 0.1% of the population, i.e. $P(D) = .001$. It follows that $P(\bar{D}) = .999$, where $\bar{D}$ indicates that a person does not have the disease. The probability of an event before the evaluation of evidence through Bayes' rule is often called the prior probability. The prior probability that someone picked at random from the population has the disease is therefore $P(D) = .001$.

Furthermore we denote a positive test result by T^+, and a negative test result by T^-. The performance of the test is summarized in table 2.

What is the probability that a patient has the disease, if the test result is positive? First, notice that $D, \bar{D}$ is a partition of the outcome space. We apply Bayes' rule to obtain

$$P(D|T^+) = \frac{P(T^+|D)P(D)}{P(T^+|D)P(D) + P(T^+|\bar{D})P(\bar{D})} = \frac{.95 \cdot .001}{.95 \cdot .001 + .02 \cdot .999} = .045.$$

Only 4.5% of the people with a positive test result actually have the disease. On the other hand, the posterior probability (i.e. the probability after evaluation of evidence) is 45 times as high as the prior probability.

	T^+	T^-
D	0.95	0.05
$\bar{D}$	0.02	0.98

Table 2. Performance of diagnostic test

2.15 Some named discrete distributions

A random experiment that only distinguishes between two possible outcomes is called a *Bernoulli* experiment. The outcomes are usually referred to as *success* and *failure* respectively. We define a random variable X that denotes the number of successes in a Bernoulli experiment; X consequently has possible values 0 and 1. The probability distribution of X is completely determined by the probability of success, which we denote by π, and is: $p(X = 0) = 1 - \pi$ and $p(X = 1) = \pi$. It easily follows that $\mathrm{E}(X) = \mu = \pi$ and $\sigma^2 = \pi(1 - \pi)$.

A number of *independent*, *identical* repetitions of a Bernoulli experiment is called a *binomial* experiment. We denote the number of successes in a binomial experiment by Y which has possible values $0, 1, \ldots, m$ (where m is the number of repetitions). Any particular sequence with y successes has probability

$$\pi^y (1 - \pi)^{m-y}$$

since the trials are independent. The number of distinct ways y successes may occur in a sequence of m is

$$\binom{m}{y} = \frac{m!}{y!(m-y)!}$$

so the probability distribution of Y is

$$p(y) = \binom{m}{y} \pi^y (1 - \pi)^{m-y} \quad \text{for } y = 0, 1, \ldots, m.$$

We indicate that Y has binomial distribution with parameters m and π by writing $Y \sim B(m, \pi)$ ($\sim$ should be read "has distribution"). We can derive easily that $\mathrm{E}(Y) = \mu = m\pi$ and $\sigma^2 = m\pi(1 - \pi)$.

The multinomial distribution is a generalization of the binomial distribution to random experiments with $n \geq 2$ possible outcomes or categories. Let y_i denote the number of results in category i, and let π_i denote the probability of a result in the ith category on each trial (with $\sum_{i=1}^{n} \pi_i = 1$). The joint probability distribution of $Y_1, Y_2, \ldots, Y_n$ for a sequence of m trials is

$$P(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) = \frac{m!}{y_1!\, y_2! \ldots y_n!} \pi_1^{y_1} \pi_2^{y_2} \ldots \pi_n^{y_n}$$

The product of powers of the π_i represents the probability of any particular sequence with y_i results in category i for each $1 \leq i \leq n$, and the ratio of factorials indicates the number distinct sequences with y_i results in category i for each $1 \leq i \leq n$.

A random variable Y has Poisson distribution with parameter μ if it has probability function

$$p(y) = \frac{\mu^y}{y!}e^{-\mu} \text{ for } y = 0, 1, 2, \ldots$$

where the single parameter μ is a positive real number. One can easily show that $\mathrm{E}(Y) = \mathrm{V}(Y) = \mu$. We write $Y \sim \mathrm{Po}(\mu)$. Use of the Poisson distribution as an approximation to the binomial distribution is discussed in section 3.

2.16 Some named continuous distributions

Continuous distributions of type

$$f\,(y) = \begin{cases} \frac{1}{\beta-\alpha} \text{ for } \alpha \leq y \leq \beta \\ \quad 0 \text{ otherwise} \end{cases}$$

are called uniform distributions, denoted $U(\alpha, \beta)$. Mean and variance are respectively

$$\mu = \frac{\alpha+\beta}{2}, \text{ and } \sigma^2 = \frac{(\beta-\alpha)^2}{12}$$

Continuous distributions of type

$$f(y) = \frac{e^{-(y-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}} \qquad \text{for } y \in \mathbb{R}$$

with $\sigma > 0$ are called *normal* or *Gaussian* distributions. Mean μ and variance σ^2 are the two parameters of the normal distribution, which we denote by $\mathcal{N}(\mu, \sigma^2)$. The special case with $\mu = 0$ and $\sigma^2 = 1$, is called the standardnormal distribution. A random variable of this type is often denoted by Z, i.e. $Z \sim \mathcal{N}(0, 1)$. If the distribution of a random variable is determined by many small independent influences, it tends to be normally distributed. In the next section we discuss why the normal distribution is so important in statistical inference.

The binormal distribution is a generelization of the normal distribution to the joint distribution of pairs (X, Y) of random variables. Its parameters are μ_x, μ_y, σ_x^2, σ_y^2, and correlation coefficient ρ, with σ_x^2, $\sigma_y^2 > 0$ and $-1 \leq \rho \leq 1$. We write

$$(X, Y) \sim \mathcal{N}^2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$$

The parameter ρ is a measure for the linear dependence between X and Y (for further explanation of this parameter, the reader is referred to section 6.3). Further generelization to the joint distribution of $n \geq 2$ random variables $Y_1, Y_2, \ldots, Y_n$ yields the multivariate normal distribution. For convenience we switch to matrix notation for the parameters

$$(Y_1, Y_2, \ldots, Y_n) \sim \mathcal{N}^n(\mu, \Sigma)$$

where $\mu = (\mu_1, \mu_2, \ldots, \mu_n)$ is the vector of means and Σ is an $n \times n$ covariance matrix. The diagonal elements of Σ contain the variances $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2)$ and element (i, j) with $i \neq j$ contains the covariance between Y_i and Y_j (for an explanation of covariance, the reader is again referred to section 6.3).

A random variable T has *exponential* distribution with rate λ $(\lambda > 0)$ if T has probability density

$$f(t) = \lambda e^{-\lambda t} \quad (t \geq 0)$$

We may think of T as a random time of some kind, such as a time to failure for artifacts, or survival times for organisms. With T we associate a survival function

$$P(T > s) = \int_s^\infty f(t)dt = e^{-\lambda s}$$

representing the probability of surviving past time s. Characteristic for the exponential distribution is that it is *memoryless*, i.e.

$$P(T > t + s \,|\, T > t) = P(T > s) \quad (t \geq 0, s \geq 0)$$

Given survival to time t, the chance of surviving a further time s is the same as surviving to time s in the first place. This is obviously not a good model for survival times of systems with aging such as humans. It is however a plausible model for time to failure of some artifacts that do not wear out gradually but stop functioning suddenly and unpredictably.

A random variable Y has a Beta distribution with parameters $l > 0$ and $m > 0$ if it has probability density

$$f(y) = \frac{y^{l-1}(1-y)^{m-1}}{\int_0^1 y^{l-1}(1-y)^{m-1}dy} \quad (0 \leq y \leq 1)$$

For the special case that $l = m = 1$ this reduces to a uniform distribution over the interval $[0, 1]$. The Beta distribution is particularly useful in Bayesian inference concerning unknown probabilities, which is discussed in section 4.3.

3 Sampling and sampling distributions

The objective of sampling is to draw a sample that permits us to draw conclusions about a population of interest. We may for example draw a sample from the population of Dutch men of 18 years and older to learn something about the joint distribution of height and weight in this population.

Because we cannot draw conclusions about the population from a sample without error, it is important to know how large these errors may be, and how often incorrect conclusions may occur. An objective assessment of these errors is only possible for a *probability sample*. For a probability sample, the probability of inclusion in the sample is *known* and *positive* for each unit in the population. Drawing a probability sample of size m from a population consisting of M units, may be a quite complex random experiment. The experiment is simplified

Unit	1	2	3	4	5	6
$\mathcal{X}$	1	1	2	2	2	3

Table 3. A small population

x	1	2	3
$p_1(x) = p_2(x)$	1/3	1/2	1/6

Table 4. Probability distribution of X_1 and X_2

considerably by subdividing it into m experiments, consisting of drawing the m consecutive units. In a *simple random sample* the m consecutive units are drawn with equal probabilities from the units concerned. In random sampling *with replacement* the subexperiments (drawing of one unit) are all identical and independent: m times a unit is randomly selected from the entire population. We will see that this property simplifies the ensuing analysis considerably.

For units in the sample we observe one or more population variables. For probability samples, each draw is a random experiment. Every observation may therefore be viewed as a random variable. The observation of a population variable $\mathcal{X}$ from the unit drawn in the i^{th} trial, yields a random variable X_i. Observation of the complete sample yields m random variables $X_1, ..., X_m$. Likewise, if we observe for each unit the pair of population variables $(\mathcal{X}, \mathcal{Y})$, we obtain pairs of random variables (X_i, Y_i) with outcomes (x_i, y_i). Consider the population of size $M = 6$, displayed in table 3.

A random sample of size $m = 2$ is drawn *with replacement* from this population. For each unit drawn we observe the value of $\mathcal{X}$. This yields two random variables X_1 and X_2, with identical probability distribution as displayed in table 4. Furthermore X_1 and X_2 are independent, so their joint distribution equals the product of their individual distributions,

$$p(x_1, x_2) = \prod_{i=1}^{2} p_i(x_i) = [p(x)]^2$$

The distribution of the sample is displayed in the table 5.

Usually we are not really interested in the individual outcomes of the sample, but rather in some sample statistic. A statistic is a function of the sample observations $X_1, ..., X_m$, and therefore is itself also a random variable. Some important sample statistics are the sample mean $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$, sample variance $S^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2$, and sample fraction $Fr = \frac{1}{m}\sum_{i=1}^{m} X_i$ (for 0-1 variable $\mathcal{X}$). In table 5 we listed the values of sample statistics $\bar{x}$ and s^2, for all possible samples of size 2.

The probability distribution of a sample statistic is called its *sampling distribution*. The sampling distribution of $\bar{X}$ and S^2 is calculated easily from table 5; they are displayed in tables 6 and 7 respectively.

(x_1, x_2)	$p(x_1, x_2)$	$\bar{x}$	s^2
(1,1)	1/9	1	0
(2,2)	1/4	2	0
(3,3)	1/36	3	0
(1,2)	1/6	1.5	0.5
(1,3)	1/18	2	2
(2,1)	1/6	1.5	0.5
(2,3)	1/12	2.5	0.5
(3,1)	1/18	2	2
(3,2)	1/12	2.5	0.5

Table 5. Probability distribution of sample of size $m = 2$ by sampling with replacement from the population in table 3

$\bar{x}$	$p(\bar{x})$
1	1/9
1.5	1/3
2	13/36
2.5	1/6
3	1/36

Table 6. Sampling distribution of $\bar{X}$

s^2	$p(s^2)$
0	14/36
0.5	1/2
2	1/9

Table 7. Sampling distribution of S^2

Note that $\mathrm{E}(\bar{X}) = \frac{11}{6} = \mu$, where μ denotes the population mean, and $\mathrm{E}(S^2) = \frac{17}{36} = \sigma^2$, where σ^2 denotes the population variance.

In the above example, we were able to determine the probability distribution of the sample, and sample statistics, by complete enumeration of all possible samples. This was feasible only because the sample size and the number of distinct values of $\mathcal{X}$ was very small. When the sample is of realistic size, and $\mathcal{X}$ takes on many distinct values, complete enumeration is not possible. Nevertheless, we would like to be able to infer something about the shape of the sampling distribution of a sample statistic, from knowledge of the distribution of X. We consider here two options to make such inferences.

1. The distribution of X has some standard form that allows the mathematical derivation of the exact sampling distribution.
2. We use a limiting distribution to approximate the sampling distribution of interest. The limiting distribution may be derived from some characteristics of the distribution of X.

The exact sampling distribution of a sample statistic is often hard to derive analytically, even if the population distribution of $\mathcal{X}$ is known. As an example we consider the sample statistic $\bar{X}$. The mean and variance of the sampling distribution of $\bar{X}$ are $\mathrm{E}(\bar{X}) = \mu$ and $\mathrm{V}(\bar{X}) = \sigma^2/m$, but its exact shape can only be derived in a few special cases. For example, if the distribution of $\mathcal{X}$ is $\mathcal{N}(\mu, \sigma^2)$ then the distribution of $\bar{X}$ is $\mathcal{N}(\mu, \sigma^2/m)$. Of more practical interest is the exact sampling distribution of the sample statistic Fr, i.e. the fraction of successes in the sample, with $\mathcal{X}$ a 0-1 population variable. The number of successes in the sample has distribution $Y \sim B(m, \pi)$ where m is the sample size and π the fraction of successes in the population. We have $\mu_y = m\pi$ and $\sigma_y^2 = m\pi(1-\pi)$. Since $Fr = Y/m$, it follows that $\mu_{fr} = \pi$ and $\sigma_{fr}^2 = \pi(1-\pi)/m$. Since $P(Fr = fr) = P(Y = m fr)$, its sampling distribution is immediately derived from the sampling distribution of Y.

Example 13. Consider a sample of size 10 from a population with fraction of successes $\pi = 0.8$. What is the sampling distribution of Fr , the sample fraction of successes? The distribution is immediately derived from the distribution of the number of successes $Y \sim B(10, 0.8)$.

In practice, we often have to rely on approximations of the sampling distribution based on so called *asymptotic* results. To understand the basic idea, we have to introduce some definitions concerning the convergence of sequences of random variables. For present purposes we distinguish between convergence in probability (to a constant) and convergence in distribution (weak convergence) of a sequence of random variables. The limiting arguments below are all with respect to sample size m.

Definition 1. *A sequence $\{X_m\}$ of random variables converges in probability to a constant c if, for every positive number ε and η, there exists a positive integer $m_0 = m_0(\varepsilon, \eta)$ such that*

$$P(\,|X_m - c\,| > \varepsilon) < \eta,\ m \geq m_0$$

Example 14. Consider the sequence of random variables $\{X_m\}$ with probability distributions $P(x_m = 0) = 1 - 1/m$ and $P(x_m = m) = 1/m$. Then $\{X_m\}$ converges in probability to 0.

Definition 2. *A sequence $\{X_m\}$ of random variables converges in distribution to a random variable X with distribution function $F(X)$ if for every $\varepsilon > 0$, there exists an integer $m_0 = m_0(\varepsilon)$, such that at every point where $F(X)$ is continuous*

$$|\, F_m(x) - F(x)\,| < \varepsilon,\ m \geq m_0$$

where $F_m(x)$ denotes the distribution function of x_m.

Example 15. Consider a sequence of random variables $\{X_m\}$ with probability distributions $P(x_m = 1) = 1/2 + 1/(m+1)$ and $P(x_m = 2) = 1/2 - 1/(m+1)$, $m = 1, 2, \ldots$. As m increases without bound, the two probabilities converge to 1/2, and $P(X = 1) = 1/2$, $P(X = 2) = 1/2$ is called the *limiting distribution* of $\{X_m\}$.

Convergence in distribution is a particularly important concept in statistical inference, because the limiting distributions of sample statistics may be used as an approximation in case the exact sampling distribution cannot be (or is prohibitively cumbersome) to derive. A crucial result in this respect is the *central limit theorem* : If $(x_1, ..., x_m)$ is a random sample from any probability distribution with finite mean μ and finite variance σ^2, and $\bar{x} = 1/m \sum x_i$ then

$$\frac{(\bar{x} - \mu)}{\sigma/\sqrt{m}} \xrightarrow{D} \mathcal{N}(0,1)$$

regardless of the form of the parent distribution. In this expression, $\xrightarrow{D}$ denotes convergence in distribution. This property explains the importance of the normal distribution in statistical inference. Note that this theorem doesn't say anything however about the rate of convergence to the normal distribution. In general, the more the population distribution resembles a normal distribution, the faster the convergence. For extremely skewed distributions $m = 100$ may be required for the normal approximation to be acceptable.

A well-known application of the central limit theorem is the approximation of the distribution of the sample proportion of successes Fr by a normal distribution. Since a success is coded as 1, and failure as 0, the fraction of successes is indeed a mean. This means the central limit theorem is applicable and as a rule of thumb $Fr \approx \mathcal{N}(\pi, \pi(1-\pi)/m)$ if $m\pi \geq 5$ and $m(1-\pi) \geq 5$. Even though the exact sampling distribution can be determined in this case, as m becomes large it becomes prohibitively time-consuming to actually calculate this distribution.

If π is close to 0 or 1, quite a large sample is required for the normal approximation to be acceptable. In that case we may use the following covergence property of the binomial distribution

$$\binom{m}{y} \pi^y (1-\pi)^{m-y} \xrightarrow{D} \frac{(m\pi)^y}{y!} e^{-m\pi}$$

In words, the binomial distribution with parameters m and π converges to a Poisson distribution with parameter $\mu = m\pi$ as m gets larger and larger. Moreover, it can be shown that this approximation is quite good for $\pi \leq 0.1$, regardless of the value of m. This explains the use of the Poisson rather than the normal approximation to the binomial distribution when π is close to 0 or 1.

4 Statistical Inference

The relation between sample data and population may be used for reasoning in two directions: from known population to yet to be observed sample data (as discussed in section 3), and from observed data to (partially) unknown population. This last direction of reasoning is of inductive nature and is addressed in statistical inference. It is the form of reasoning most relevant to data analysis, since one typically has available one set of sample data from which one intends to draw conclusions about the unknown population.

4.1 Frequentist Inference

According to frequentists, inference procedures should be interpreted and evaluated in terms of their behavior in hypothetical repetitions under the same conditions. To quote David S. Moore, the frequentist consistently asks "What would happen if we did this many times?"[15]. To answer this question, the sampling distribution of a statistic is of crucial importance. The two basic types of frequentist inference are estimation and testing. In estimation one wants to come up with a plausible value or range of plausible values for an unknown population parameter. In testing one wants to decide whether a hypothesis concerning the value of an unknown population parameter should be accepted or rejected in the light of sample data.

Point Estimation In point estimation one tries to provide an estimate for an unknown population parameter, denoted by θ, with *one number*: the point estimate. If G denotes the estimator of θ, then the estimation error is a random variable $G - \theta$, which should preferably be close to zero.

An important quality measure from a frequentist point of view is the bias of an estimator

$$\mathrm{B}_\theta = \mathrm{E}_\theta(G - \theta) = \mathrm{E}_\theta(G) - \theta,$$

where expectation is taken with respect to repeated samples from the population. If $\mathrm{E}_\theta(G) = \theta$, i.e. the expected value of the estimator is equal to the value of the population parameter, then the estimator G is called *unbiased*.

Example 16. If π is the proportion of successes in some population and Fr is the proportion of successes in a random sample from this population, then $\mathrm{E}_\pi(\mathit{Fr}) = \pi$, so Fr is an unbiased estimator of π.

Another important quality measure of an estimator is its variance

$$\mathrm{V}_\theta(G) = \mathrm{E}_\theta(G - \mathrm{E}_\theta(G))^2$$

which measures how much individual estimates g tend to differ from $\mathrm{E}_\theta(G)$, the average value of g over a large number of samples.

An overall quality measure that combines bias and variance is the *mean squared error*

$$\mathrm{M}_\theta(G) = \mathrm{E}_\theta(G - \theta)^2$$

where low values indicate a good estimator. After some algebraic manipulation, we can decompose mean squared error into

$$\mathrm{M}_\theta(G) = \mathrm{B}^2_\theta(G) + \mathrm{V}_\theta(G)$$

that is mean squared error equals squared bias plus variance. It follows that if an estimator is unbiased, then its mean squared error equals its variance.

Example 17. For the unbiased estimator Fr of π we have $\mathrm{M}_\pi(Fr) = \mathrm{V}_\pi(Fr) = \pi(1-\pi)/m$.

The so-called "plug-in" principle provides a simple and intuitively plausible method of constructing estimators. The plug-in estimate of a parameter $\theta = t(F)$ is defined to be $\hat{\theta} = t(\hat{F})$. Here F denotes the population distribution function and $\hat{F}$ its estimate, based on the sample. For example, to estimate the population mean μ use its sample analogue $\bar{x} = 1/m \sum x_i$, and to estimate population variance σ^2 use its sample analogue $s^2 = 1/m \sum (x_i - \bar{x})^2$. Another well-known method for finding point estimates is the method of least squares. The least squares estimate of population mean μ is the number g for which the sum of squared errors $(x_i - g)^2$ is at a minimum. If we take the derivative of this sum with respect to g, we obtain

$$\frac{\partial}{\partial g} \sum_{i=1}^{m} (x_i - g)^2 = \sum_{i=1}^{m} (x_i - g)(-2) = -2m(\bar{x} - g)$$

When we equate this expression to zero, and solve for g we obtain $g = \bar{x}$. So $\bar{x}$ is the least squares estimate of μ. A third important method of estimation is *maximum likelihood estimation*, which is discussed in section 4.2.

Interval Estimation An interval estimator for population parameter θ is an interval of type (G_L, G_U). Two important quality measures for interval estimates are:

$$\mathrm{E}_\theta(G_U - G_L),$$

i.e. the expected width of the interval, and

$$P_\theta(G_L < \theta < G_U),$$

i.e. the probability that the interval contains the true parameter value. Clearly there is a trade-off between these quality measures. If we require a high probability that the interval contains the true parameter value, the interval itself has to become wider. It is customary to choose a confidence level $(1-\alpha)$ and use an interval estimator such that

$$P_\theta(G_L < \theta < G_U) \geq 1 - \alpha$$

for all possible values of θ. A realisation (g_L, g_U) of such an interval estimator is called a $100(1-\alpha)\%$ *confidence interval.*

The form of reasoning used in confidence intervals is most clearly reflected in the estimation of the mean of a normal population with variance σ^2 known, i.e. $X \sim \mathcal{N}(\mu, \sigma^2)$. The distribution of the sample mean for random samples of size m from this population is known to be $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/m)$. First $\bar{X}$ is standardized to obtain

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{m}} \sim \mathcal{N}(0,1)$$

which allows us to use a table for the standardnormal distribution $Z \sim \mathcal{N}(0,1)$ to find the relevant probabilities. The probability that $\bar{X}$ is more than one standard error (standard deviation of the sampling distribution) larger than unknown μ is

$$P(\bar{X} > \mu + \frac{\sigma}{\sqrt{m}}) = P(\frac{\bar{X} - \mu}{\sigma/\sqrt{m}} > 1) = P(Z > 1) = 0.1587$$

But we can *reverse* this reasoning by observing that

$$P(\bar{X} - \frac{\sigma}{\sqrt{m}} < \mu) = 1 - 0.1587 = 0.8413$$

because $\bar{X} - \frac{\sigma}{\sqrt{m}} < \mu$ holds unless $\bar{X} > \mu + \frac{\sigma}{\sqrt{m}}$. Therefore, the probability that the interval $(\bar{X} - \sigma/\sqrt{m}, \infty)$ will contain the true value of μ equals 0.8413. This is called a left-sided confidence interval because it only states a lower bound for μ. In general a $100(1-\alpha)\%$ left-sided confidence interval for μ reads $(\bar{x} - z_\alpha \frac{\sigma}{\sqrt{m}}, \infty)$, where $P(Z > z_\alpha) = \alpha$. Likewise, we may construct a right-sided confidence interval $(-\infty, \bar{x} + z_\alpha \frac{\sigma}{\sqrt{m}})$ and a two-sided confidence interval

$$(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{m}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{m}}).$$

If the distribution of X is unknown, i.e. $X \sim \mu, \sigma^2$, then for sufficiently large m we may invoke the central limit theorem and use $\bar{X} \approx \mathcal{N}(\mu, \sigma^2/m)$, and proceed as above.

In most practical estimation problems we don't know the value of σ^2, and we have to estimate it from the data as well. A rather obvious estimator is the sample variance

$$S^2 = \frac{1}{m-1} \sum_{i=1}^{m} (x_i - \bar{x})$$

Now we may use

$$\frac{\bar{X}-\mu}{S/\sqrt{m}} \sim t_{m-1}$$

where t_{m-1} denotes the t-distribution with $m-1$ degrees of freedom. This distribution has a higher variance than the standardnormal distribution, leading to somewhat wider confidence intervals. This is the price we pay for the fact that we don't know the value of σ^2, but have to estimate it from the data. On the other hand we have $t_\nu \approx \mathcal{N}(0,1)$ for $\nu \geq 100$, so if m is large enough we may use the standardnormal distribution for all practical purposes.

Hypothesis Testing A test is a statistical procedure to make a choice between two hypotheses concerning the value of a population parameter θ. One of these, called the *null hypothesis* and denoted by H_0, gets the "benefit of the doubt". The two possible conclusions are to reject or not to reject H_0. H_0 is only rejected if the sample data contains strong evidence that it is not true. The null hypothesis is rejected iff realisation g of test statistic G is in the *critical region* denoted by C. In doing so we can make two kinds of errors

Type I error: Reject H_0 when it is true.
Type II error: Accept H_0 when it is false.

Type I errors are considered to be more serious than Type II errors. Test statistic G is usually a point estimator for θ, e.g. if we test a hypothesis concerning the value of population mean μ, then $\bar{X}$ is an obvious choice of test statistic. As an example we look at hypothesis test

$$H_0 : \theta \geq \theta_0 \ , \ H_a : \theta < \theta_0$$

The highest value of G that leads to the rejection of H_0 is called the critical value c_u, it is the upper bound of the so-called critical region $C = (-\infty, c_u]$. All values of G to the left of c_u lead to the rejection of H_0, so this is called a left one-sided test. An overall quality measure for a test is its power β

$$\beta(\theta) = P_\theta(\text{Reject } H_0) = P_\theta(G \in C)$$

Because we would like a low probability of Type I and Type II errors, we like to have $\beta(\theta)$ small for $\theta \in H_0$ and $\beta(\theta)$ large for $\theta \in H_a$. It is common practice in hypothesis testing to restrict the probability of a Type I error to a maximum called the *significance level* α of the test, i.e.

$$\text{max}_{\theta \in H_0} \beta(\theta) \leq \alpha$$

Since the maximum is reached for $\theta = \theta_0$ this reduces to the restriction $\beta(\theta_0) \leq \alpha$. If possible the test is performed in such a way that $\beta(\theta_0) = \alpha$ (This may not be possible for discrete sampling distributions). Common levels for α are 0.1, 0.05 and 0.01. If in a specific application of the test, the conclusion is that H_0 should be rejected, then the result is called *significant*.

Consider a left one-sided test on population mean μ with $X \sim \mathcal{N}(\mu, \sigma^2)$ and the value of σ^2 known. That is

$$H_0 : \mu \geq \mu_0 \ , \ H_a : \mu < \mu_0$$

We determine the sampling distribution of the test statistic $\bar{X}$ under the assumption that the $\mu = \mu_0$, i.e. $\bar{X} \sim \mathcal{N}(\mu_0, \sigma^2/m)$. Now

$$\alpha = P_{\mu_0}(\bar{X} \leq c_u) = P(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{m}} \leq \frac{c_u - \mu_0}{\sigma/\sqrt{m}}) = P(Z \leq \frac{c_u - \mu_0}{\sigma/\sqrt{m}})$$

and since $P(Z \leq -z_\alpha) = \alpha$, we obtain

$$\frac{c_u - \mu_0}{\sigma/\sqrt{m}} = -z_\alpha, \text{ and therefore } c_u = \mu_0 - z_\alpha \frac{\sigma}{\sqrt{m}}$$

Example 18. Consider a random sample of size $m = 25$ from a normal population with known $\sigma = 5.4$ and unknown mean μ. The observed sample mean is $\bar{x} = 128$. We want to test the hypothesis

$$H_0 : \mu \geq 130, \text{ against } \ H_a : \mu < 130$$

i.e. $\mu_0 = 130$. The significance level of the test is set to $\alpha = 0.05$. We compute the critical value

$$c_u = \mu_0 - z_{0.05} \frac{\sigma}{\sqrt{m}} = 130 - 1.645 \frac{5.4}{\sqrt{25}} = 128.22$$

where $z_{0.05} = 1.645$ was determined using a statistical package (many books on statistics contain tables that can be used to determine the value of z_α). So the critical region is $(-\infty, 128.22]$ and since $\bar{x} = 128$ is in the critical region, we reject H_0.

Similarly, if

$$H_0 : \theta \leq \theta_0 \ , \ H_a : \theta > \theta_0$$

the critical region is $[c_l, \infty)$, and for a two-sided test

$$H_0 : \theta = \theta_0 \ , \ H_a : \theta \neq \theta_0$$

it has the form $(-\infty, c_u] \cup [c_l, \infty)$.

As with the construction of a confidence interval for the mean, for a hypothesis test concerning the mean we may invoke the central limit theorem if $X \sim \mu, \sigma^2$ and m is large. Furthermore, if σ^2 is unknown, we have to estimate it from the data and use a t_{m-1} distribution rather than the standardnormal distribution to determine the critical region.

Sometimes one doesn't want to specify the significance level α of the test in advance. In that case it us customary to report so-called p-values, indicating the *observed significance*.

Example 19. Consider the test of example 18. The p-value of the observed outcome $\bar{x} = 128$ is

$$P_{\mu_0}(\bar{X} \leq 128) = P(Z \leq \frac{128 - \mu_0}{\sigma/\sqrt{m}}) = P(Z \leq -1.852) = 0.0322$$

Since the p-value is 0.0322, we would reject H_0 at $\alpha = 0.05$, but we would accept H_0 at $\alpha = 0.01$.

4.2 Likelihood

The deductive nature of probability theory versus the inductive nature of statistical inference is perhaps most clearly reflected in the “dual” concepts of (joint) probability distribution and likelihood.

Given a particular probability model and corresponding parameter values, we may calculate the probability of observing different samples. Consider the experiment of 10 coin flips with probability of heads $\pi = 0.6$. The probability distribution of random variable “number of times heads comes up” is now the following function of the data

$$P(y) = \binom{10}{y} 0.6^y \, 0.4^{10-y}$$

We may for example compute that the probability of observing $y = 7$ is

$$\binom{10}{7} 0.6^7 0.4^3 \approx 0.215$$

In statistical inference however, we typically have one data set and want to say something about the relative likelihood of different values of some population parameter. Say we observed 7 heads in a sequence of ten coin flips. The likelihood is now a function of the unknown parameter π

$$L(\pi \,|\, y = 7) = \binom{10}{7} \pi^7 (1 - \pi)^3$$

where the constant term is actually arbitrary, since we are not interested in absolute values of the likelihood, but rather in ratios of likelihoods for different values of π.

In table 8, each column specifies the probability distribution of Y for a different value of π. Each column sums to 1, since it represents a probability distribution. Each row, on the other hand, specifies a likelihood function, or rather: it specifies the value of the likelihood function for 9 values of π. So for example, in the third row we can read off the probability of observing 2 successes in a sequence of 10 coin flips for different values of π.

In general, if $\mathbf{y} = (y_1, \ldots, y_m)$ are independent observations from a probability density $f(y \,|\, \theta)$, where θ is the parameter vector we wish to estimate, then

$$L(\theta \,|\, \mathbf{y}) \propto \prod_{i=1}^{m} f(y_i | \, \theta)$$

y	π								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0	.349	.107	.028	.006	.001				
1	.387	.269	.121	.04	.01	.002			
2	.194	.302	.234	.121	.044	.01	.002		
3	.057	.201	.267	.215	.117	.043	.009	.001	
4	.011	.088	.2	.251	.205	.111	.036	.005	
5	.002	.027	.103	.201	.246	.201	.103	.027	.002
6		.005	.036	.111	.205	.251	.2	.088	.011
7		.001	.009	.043	.117	.215	.267	.201	.057
8			.002	.01	.044	.121	.234	.302	.194
9				.002	.01	.04	.121	.269	.387
10					.001	.006	.028	.107	.349
	1	1	1	1	1	1	1	1	1

Table 8. Probability distributions (columns) and likelihood functions (rows) for $Y \sim B(10, \pi)$

The *likelihood function* then measures the relative likelihood that different θ have given rise to the observed $\mathbf{y}$. We can thus try to find that particular $\hat{\theta}$ which maximizes L, i.e. that $\hat{\theta}$ such that the observed $\mathbf{y}$ are more likely to have come from $f(y \,|\, \hat{\theta})$ than from $f(y \,|\, \theta)$ for any other value of θ.

For many parameter estimation problems one can tackle this maximization by differentiating L with respect to the components of θ and equating the derivatives to zero to give the *normal equations*

$$\frac{\partial L}{\partial \theta_j} = 0$$

These are then solved for the θ_j and the second order derivatives are examined to verify that it is indeed a maximum which has been achieved and not some other stationary point.

Maximizing the likelihood function L is equivalent to maximizing the (natural) log of L, which is computationally easier. Taking the natural log, we obtain the log-likelihood function

$$l(\theta \,|\, \mathbf{y}) = \ln(L(\theta \,|\, \mathbf{y})) = \ln(\prod_{i=1}^{m} f(y_i \,|\, \theta)) = \sum_{i=1}^{m} \ln f(y_i \,|\, \theta)$$

since $\ln ab = \ln a + \ln b$.

Example 20. In a coin flipping experiment we define the random variable Y with $y = 1$ if heads comes up, and $y = 0$ when tails comes up. Then we have the following probability distribution for one coin flip

$$f(y) = \pi^y (1 - \pi)^{1-y}$$

For a sequence of m coin flips, we obtain the joint probability distribution

$$f(\mathbf{y}) = f(y_1, y_2, ..., y_m) = \prod_{i=1}^{m} \pi^{y_i}(1-\pi)^{1-y_i}$$

which defines the likelihood when viewed as a function of π. The log-likelihood consequently becomes

$$l(\pi \,|\, \mathbf{y}) = \sum_{i=1}^{m} y_i \ln(\pi) + (1-y_i)\ln(1-\pi)$$

In a sequence of 10 coin flips with seven times heads coming up, we obtain

$$l(\pi) = \ln(\pi^7(1-\pi)^3) = 7\ln\pi + 3\ln(1-\pi)$$

To determine the maximum we take the derivative

$$\frac{dl}{d\pi} = \frac{7}{\pi} - \frac{3}{1-\pi} = 0$$

which yields maximum likelihood estimate $\hat{\pi} = 0.7$.

The reader may notice that the maximum likelihood estimate in this case is simply the fraction of heads coming up in the sample, and we could have spared ourselves the trouble of maximizing the likelihood function to obtain the required estimate. Matters become more interesting (and complicated) however, when we make π a function of data *and* parameters. Suppose that for each y_i in our sample, we observe a corresponding measure x_i which we assume is a continuous variable. We could write $\pi_i = g(x_i)$, where g is some function. In so-called Probit analysis [10] we assume

$$\pi_i = \Phi(\alpha + \beta x_i)$$

where Φ denotes the standard normal distribution function. The parameters of the model are now α and β, and we can write the log-likelihood function as

$$l(\alpha, \beta) = \sum_{i=1}^{m} y_i \ln(\Phi(\alpha + \beta x_i)) + (1-y_i)\ln(1 - \Phi(\alpha + \beta x_i))$$

This is the expression of the log-likelihood for the Probit model. By maximizing with respect to α and β, we obtain maximum likelihood estimates for these parameters.

Example 21. Consider a random sample $\mathbf{y} = (y_1, ..., y_m)$ from a normal distribution with unknown mean μ and variance σ^2. Then we have likelihood

$$L((\mu, \sigma^2)' \,|\, \mathbf{y}) = \prod_{i=1}^{m} \frac{e^{-(y_i-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}} = \frac{1}{\sigma^m (2\pi)^{m/2}} \exp\left[-\frac{1}{2}\sum_{i=1}^{m}\left(\frac{y_i-\mu}{\sigma}\right)^2\right]$$

The natural log of this expression is

$$l = \ln(L) = -m \ln \sigma - \left(\frac{m}{2}\right) \ln 2\pi - \frac{1}{2\sigma^2} \sum_{i=1}^{m} (y_i - \mu)^2$$

To determine the maximum likelihood estimates of μ and σ, we take the partial derivative of l with respect to these parameters, and equate them to zero

$$\frac{\partial l}{\partial \mu} = \frac{m}{\sigma^2} (\bar{y} - \mu) = 0$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{m}{2\sigma^2} + \frac{m}{2\sigma^4} (s^2 + (\bar{y} - \mu)^2) = 0$$

Solving these equations for μ and σ, we obtain maximum likelihood estimates $\hat{\mu} = \bar{y}$ and $\hat{\sigma}^2 = s^2$, where $s^2 = 1/m \sum (y_i - \mu)^2$.

Another important aspect of the log-likelihood function is its shape in the region near the maximum. If it is rather flat, one could say that the likelihood contains little information in the sense that there are many values of θ with log-likelihood near that of $\hat{\theta}$. If, on the other hand, it is rather steep, one could say that the log-likelihood contains much information about $\hat{\theta}$. The log-likelihood of any other value of θ is approximately given by the Taylor expansion

$$l(\theta) = l(\hat{\theta}) + (\theta - \hat{\theta}) \frac{dl}{d\theta} + \frac{1}{2} (\theta - \hat{\theta})^2 \frac{d^2 l}{d\theta^2} + ...$$

where the differential coefficients are evaluated at $\theta = \hat{\theta}$. At this point, $\frac{dl}{d\theta}$ is zero, so approximately

$$l(\theta) = l(\hat{\theta}) + \frac{1}{2} (\theta - \hat{\theta})^2 \frac{d^2 l}{d\theta^2}.$$

Minus the second derivative of the log-likelihood function is known as the (Fisher) *information*. When evaluated at $\hat{\theta}$ (the maximum likelihood estimate of θ) it is called the *observed information*.

Note 1. This concept of information should not be confused with the one discussed in section 2.9.

Some authors take the view that all statistical inference should be based on the likelihood function rather than the sampling distribution used in frequentist inference (see [5, 17]). In this sense likelihood inference differs from frequentist inference.

Example 22. Figure 2 displays the likelihood function for π corresponding to 7 successes in a series of 10 coin flips. The horizontal line indicates the range of values of π for which the ratio of $L(\pi)$ to the maximum $L(0.7)$ is greater than 1/8. The 1/8 likelihood interval is approximately $(0.38, 0.92)$. Such an interval is similar in spirit to a confidence interval in the sense that it intends to provide a range of "plausible values" for π based on the sample data. A confidence interval for π is based however on the sampling distribution of some sample statistic (the sample proportion of successes is the most obvious choice) whereas a likelihood interval is based, as the name suggests, on the likelihood function.

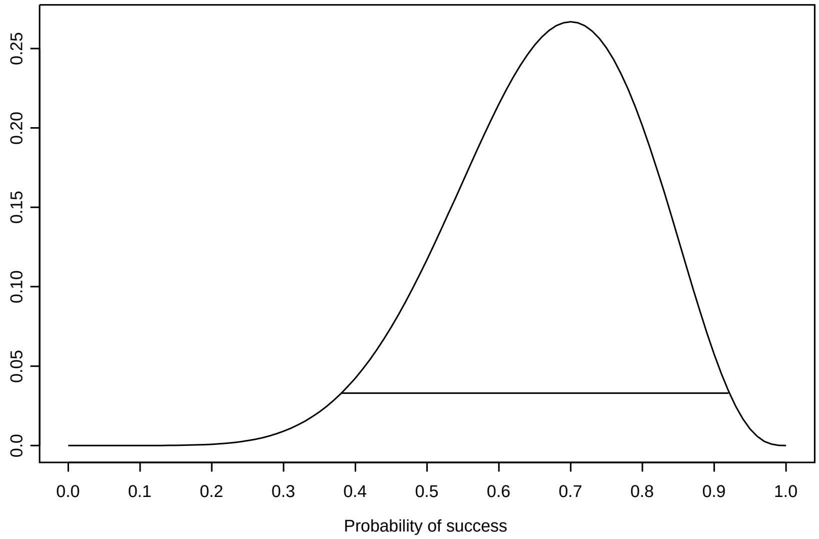


Fig. 2. Likelihood function $L(\pi \mid y = 7) = 120\pi^7(1-\pi)^3$.

On the other hand, maximum likelihood estimation may be used and evaluated from a frequentist perspective. This motivates the study of the sampling distribution of maximum likelihood estimates. If we know the true value of $\theta = \theta^*$, we can determine the *expected* log-likelihood, i.e. the mean value of the log-likelihood conditional on $\theta = \theta^*$ (still expressed as a function of θ). The expected log-likelihood has a maximum at $\theta = \theta^*$. Minus the second derivative of the expected log-likelihood evaluated at $\theta = \theta^*$, is called the *expected information*. Assuming parameter vector θ with several components the expected information matrix is defined as

$$I(\theta) = -\left\{ E\left(\frac{\partial^2 l}{\partial\theta_j \partial\theta_k}\right)_{\theta^*} \right\}$$

In large samples, the maximum likelihood estimate $\hat{\theta}$ is approximately normally distributed with mean θ^*, and covariance matrix $I(\theta)^{-1}$. Unfortunately, we cannot in practice determine $I(\theta)$, since θ^* is unknown. It is therefore set equal to $\hat{\theta}$ so that $I(\theta)$ can be calculated. An alternative estimate for the covariance matrix is the observed information matrix

$$-\left(\frac{\partial^2 l}{\partial\theta_j \partial\theta_k}\right)_{\hat{\theta}}$$

which is easier to compute since it does not involve an expectation. For the exponential family of distributions these two estimates are equivalent.

Example 23. Consider a sequence of m coin tosses, with heads coming up y times. We are interested in the probability of heads π. We have seen that

$$l(\pi) = y \ln(\pi) + (m - y) \ln(1 - \pi)$$

Setting the first derivative to zero and solving for π yields $\hat{\pi} = y/m$. The information is

$$-\frac{d^2 l}{d\pi^2} = \frac{y}{\pi^2} + \frac{(m - y)}{(1 - \pi)^2}$$

Evaluating this expression at $\hat{\pi} = y/m$ we obtain the observed information

$$\frac{m}{\hat{\pi}(1 - \hat{\pi})}.$$

In large samples, $\hat{\pi}$ is approximately normally distributed with mean π^* and variance $\pi^*(1 - \pi^*)/m$, i.e. the reciprocal of the expected information. The estimated variance of $\hat{\pi}$ is equal to the reciprocal of the observed information, i.e. $\hat{\pi}(1 - \hat{\pi})/m$.

4.3 Bayesian Inference

In this section we briefly consider the principal idea of Bayesian inference. In [1, 9], Bayesian inference is discussed in greater detail.

Consider again the coin tossing experiment. We stated that the probability of heads, denoted by π, is a fixed yet unknown quantity. From a relative frequency viewpoint, it makes no sense to talk about the probability distribution of π since it is not a random variable. In Bayesian inference one departs from this strict interpretation of probability. We may express prior, yet incomplete, knowledge concerning the value of π through the construction of a *prior distribution*. This prior distribution is then combined with sample data (using Bayes rule, see section 2.14) to obtain a posterior distribution. The posterior distribution expresses the new state of knowledge, in light of the sample data. We reproduce Bayes' rule using symbols that are more indicative for the way it is used in Bayesian inference:

$$P(M_i|D) = \frac{P(D|M_i)P(M_i)}{\sum_j P(D|M_j)P(M_j)}$$

	Prior $P(M_i)$	Likelihood $P(y = 5 \mid M_i)$	Posterior $P(M_i \mid y = 5)$
M_1: $\pi = 0.8$	0.7	0.027	0.239
M_2: $\pi = 0.4$	0.3	0.201	0.761

Table 9. Prior and posterior probabilities of M_1 and M_2

where the M_i specify different models for the data, i.e. hypotheses concerning the parameter value(s) of the probability distribution from which the data were drawn. Note that in doing so, we actually assume that this probability distribution is known up to a fixed number of parameter values.

Example 24. Consider the somewhat artificial situation where two hypotheses concerning the probability of heads of a particular coin are entertained, namely M_1: $\pi = 0.8$ and M_2: $\pi = 0.4$ (see table 9). Prior knowledge concerning these models is expressed through a prior distribution as specified in the first column of table 9. Next we observe 5 times heads in a sequence of 10 coin flips, i.e. $y = 5$. The likelihood of this outcome under the different models is specified in the second column of table 9 (the reader can also find them in table 8). The posterior distribution is obtained via Bayes' rule, and is specified in the last column of table 9. Since the data are more likely to occur under M_2, the posterior distribution has clearly shifted towards this model.

In general, the probability distribution of interest is indexed by a number of continuous valued parameters, which we denote by parameter vector θ. Replacing probabilities by probability densities and summation by integration, we obtain the probability density version of Bayes' rule

$$f(\theta \,|\, \mathbf{y}) = \frac{f(\mathbf{y} \,|\, \theta)\; f(\theta)}{\int_\Omega f(\mathbf{y} \,|\, \theta)\; f(\theta)\; d\theta}$$

where $\mathbf{y}$ denotes the observed data and Ω denotes the parameter space, i.e. the space of possible values of θ.

Consider the case where we have no prior knowledge whatsoever concerning the probability of heads π. How should this be reflected in the prior distribution? One way of reasoning is to say that all values of π are considered equally likely, which can be expressed by a uniform distribution over $\Omega = [0, 1]$: the range of possible values of π. Let's consider the form of the posterior distribution in this special case.

$$f(\pi \,|\, \mathbf{y}) = \frac{f(\mathbf{y} \,|\, \pi) f(\pi)}{\int_0^1 f(\mathbf{y} \,|\, \pi) f(\pi) d\pi}$$

If we observe once again 7 times heads in a sequence of 10 coin flips, then $f(\mathbf{y} \,|\, \pi) = \pi^7(1-\pi)^3$. Since $f(\pi) = 1$, the denominator of the above fraction becomes

$$\int_0^1 \pi^7(1-\pi)^3 d\pi = \frac{1}{1320}$$

and so the posterior density becomes

$$f(\pi \,|\, \mathbf{y}) = 1320\,\pi^7(1-\pi)^3$$

It is reassuring to see that in case of prior ignorance the posterior distribution is proportional to the likelihood function of the observed sample. Note that the constant of proportionality merely acts to make the integral of the expression in the numerator equal to one, as we would expect of a probability density!

In general, the computationally most difficult part of obtaining the posterior distribution is the evaluation of the (multiple) integral in the denominator of the expression. For this reason, a particular class of priors, called conjugate priors, have received special attention in Bayesian statistics. Assume our prior knowledge concerning the value of π may be expressed by a Beta(4,6) distribution (see section 2.16), i.e.

$$f(\pi) = \frac{\pi^3(1-\pi)^5}{\int_0^1 \pi^3(1-\pi)^5 d\pi}$$

Since $\int_0^1 \pi^3(1-\pi)^5 d\pi = \frac{1}{504}$, we get $f(\pi) = 504\,\pi^3(1-\pi)^5$.

Multiplied with the likelihood this results in $504\,\pi^3(1-\pi)^5\pi^7(1-\pi)^3 = 504\pi^{10}(1-\pi)^8$, so the denominator becomes

$$\int_0^1 504\,\pi^{10}(1-\pi)^8 = \frac{28}{46189}$$

and the posterior density becomes

$$f(\pi \,|\, \mathbf{y}) = 831402\,\pi^{10}(1-\pi)^8$$

which is a Beta(11,9) distribution. In general, when we have a binomial sample of size m with r successes, and we combine that with a Beta(l, k) prior distribution, then the posterior distribution is Beta$(l + r, k + m - r)$. Loosely speaking, conjugate priors allow for simple rules to update the prior with sample data to arrive at the posterior distribution. Furthermore, the posterior distribution belongs to the same family as the prior distribution. Since the uniform distribution over the interval $[0, 1]$ is the same as a Beta(1,1) distribution (see section 2.16), we could have used this simple update rule in the "prior ignorance" case as well: combining a Beta(1,1) prior with a binomial sample of size 10 with 7 successes yields a Beta(8,4) posterior distribution.

Once we have calculated the posterior distribution, we can extract all kinds of information from it. We may for example determine the mode of the posterior distribution which represents the value of π for which the posterior density is maximal. When asked to give a point estimate for π, it makes sense to report this value. When asked for a range of plausible values for π we may use the posterior distribution to determine a so-called $100(1-\alpha)\%$ probability interval, which is an interval $[g_l, g_u]$ such that $P(\pi < g_l) = \alpha/2$ and $P(\pi > g_u) = \alpha/2$ where the relevant probabilities are based on the posterior distribution for π.

5 Prediction and prediction error

The value of a random variable Y depends on the outcome of a random experiment (see section 2.7). Before the experiment is performed, the value of the random variable is unknown. In many cases, we would like to *predict* the value of a yet to be observed random variable. The usual assumption is that the distribution of Y depends in some way on some random vector $\mathbf{X}=(X_1,...,X_n)$. For example, in the simple linear regression model, the assumption is that $Y_i \sim \mathcal{N}(\beta_0+\beta_1 x_i, \sigma_\epsilon^2)$, i.e. Y is normally distributed with mean a linear function of x, and constant variance σ_ϵ^2. If Y is a 0-1 variable, a common assumption is $Y_i \sim B(1, \Phi(\beta_0+\beta_1 x_i))$, i.e. the Y_i are Bernoulli random variables. The probability of success is the area under the standardnormal distribution to the left of a value that depends on x_i: the so-called Probit model [10]. Replacing the standardnormal distribution by the logistic distribution leads to the very popular logistic regression model [13].

Often, the goal of data analysis is to gain information from a training sample $T=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),...,(\mathbf{x}_m,y_m)\}$ in order to estimate the parameters of such a model. Once the model is estimated, we can use it to make predictions about random variable Y. If Y is numeric, we speak about *regression*, and if Y takes its values from a discrete unordered set we speak about *classification*.

A fundamental choice is which assumptions are made concerning *how* the distribution of Y (or some aspect of it: typically the expected value of Y), depends on $\mathbf{x}$. Such assumptions are often called *inductive bias* in the machine learning literature [14] and are part of what is called *model specification* in the econometrics literature [10]. The assumptions of the linear regression model, for example, are quite restrictive. If the true relationship is not linear, the estimated model will produce some prediction error due to the unwarranted assumption of linearity. One might therefore argue that no assumptions at all should be made, but let the data "speak entirely for itself". However if no assumptions whatsoever are made, there is no rational basis to generalize beyond the data observed [14]. If we make very mild assumptions, basically allowing the relation between $\mathrm{E}(Y)$ and $\mathbf{x}$ to be chosen from a very large class of functions, the estimated function is capable to adapt very well to the data. In fact so well that it will also model the peculiarities that are due to random variations. The estimate becomes highly sensitive to the particular sample drawn, that is upon repeated sampling it will exhibit high variance.

We consider a simple regression example to illustrate these ideas. Suppose that $Y_i \sim \mathcal{N}(\mu = 2.0 + 0.5x_i, \sigma_\varepsilon^2 = 1)$, i.e. the true relation between $\mathrm{E}(Y)$ and x is

$$\mathrm{E}(Y) = 2.0 + 0.5x.$$

We have a sample T of ten (x,y) observations, which is displayed in the scatterplot of Fig. 3(a). Note that x is not a random variable but its values are chosen by us to be $1,2,\ldots,10$. Although $\mathrm{E}(Y)$ is a linear function of x, the observations do not lie on a straight line due to the inherent variability of Y. We pretend we don't know the relation between x and y, but only have T at our disposal, as would be the case in most data analysis settings. We consider three classes of models to describe the relation between x and y

Linear Model: $\mathrm{E}(Y) = f_1(x) = \beta_0 + \beta_1 x$
Quadratic Model: $\mathrm{E}(Y) = f_2(x) = \beta_0 + \beta_1 x + \beta_2 x^2$
Cubic Model: $\mathrm{E}(Y) = f_3(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$

Note 2. In this section, the symbol f does not denote a probability density. Probability densities and probability functions are henceforth both denoted by the symbol p.

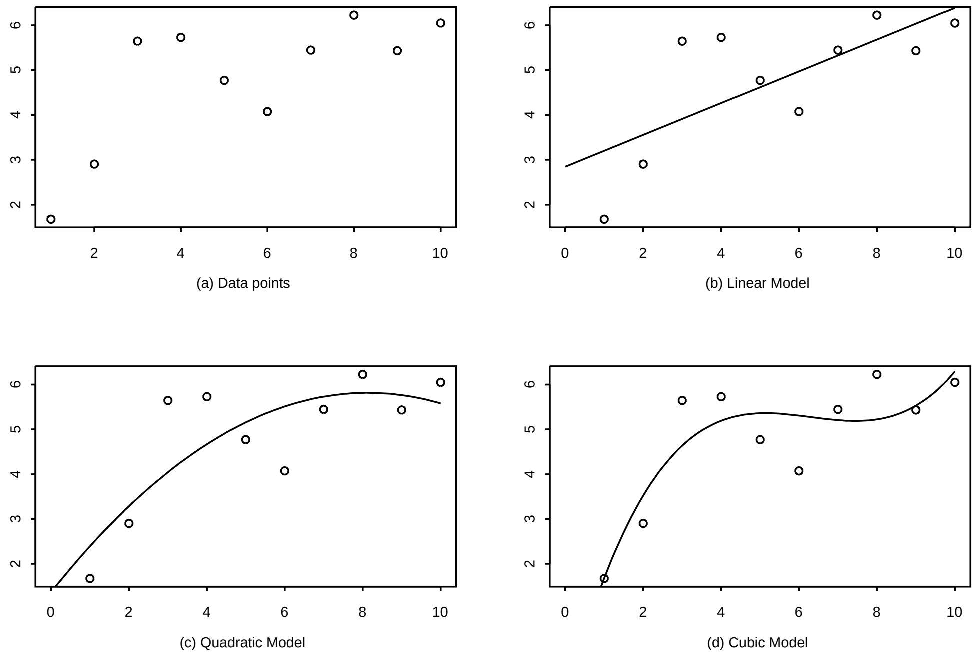
Note that (2) encompasses (1) in the sense that if $\beta_2 = 0$, (2) reduces to the linear function (1). Likewise, (3) encompasses (2), and consequently also (1). The β_j are the parameters of the model, whose estimates are chosen in such a way that the sum of squared *vertical* distances from the points (x_i, y_i) to the fitted equation is minimized. For example, for the simple linear regression model we choose the estimates of β_0 and β_1 such that

$$\sum_{i=1}^{m} [y_i - (\hat{\beta}_0 + \hat{\beta}_1\, x_i)]^2$$

is minimal. The expression $\hat{\beta}_0 + \hat{\beta}_1\, x_i$ denotes the predicted value for y_i, so one effectively minimizes the sum of squared differences between predicted values and realisations of y. The estimates $\hat{\beta}_j$ of the β_j thus obtained are called the *least squares* estimates. The choice for minimizing the squared differences rather than using some other criterion (e.g. minimize the absolute value of the differences) was historically for a large part justified by analytical convenience. Under the usual assumption of the linear regression model concerning the normal distribution of Y (conditional on x), the least squares estimates and maximum likelihood estimates coincide.

The equations obtained by least squares estimation for the respective models are displayed in Fig. 3 (b) to (d). Without performing the actual calculations, one can easily see that the linear model gives the worst fit, even though the true (population) relation is linear. The quadratic model gives a somewhat better fit, and the cubic model gives the best fit of the three. In general, the more parameters the model has, the better it is able to adjust to the data in T. Does this mean that the cubic model gives better predictions than the linear model? It does on T, but how about on data that were not used to fit the equation? We drew a second random sample, denoted by T', and looked at the fit of the equations to T' (see Fig. 4). The fit of the cubic model is clearly worse than that of the linear model. The reason is that the cubic model has adjusted itself to the random variations in T, leading on average to bad predictive performance on new samples. This phenomenon is called *overfitting*.

In the next section we discuss the decomposition of prediction error into its components to gain a further understanding of the phenomenon illustrated by the above example.



(a) Data points

(b) Linear Model

(c) Quadratic Model

(d) Cubic Model

Fig. 3. Equations fitted by least squares to the data in T

5.1 Prediction error in regression

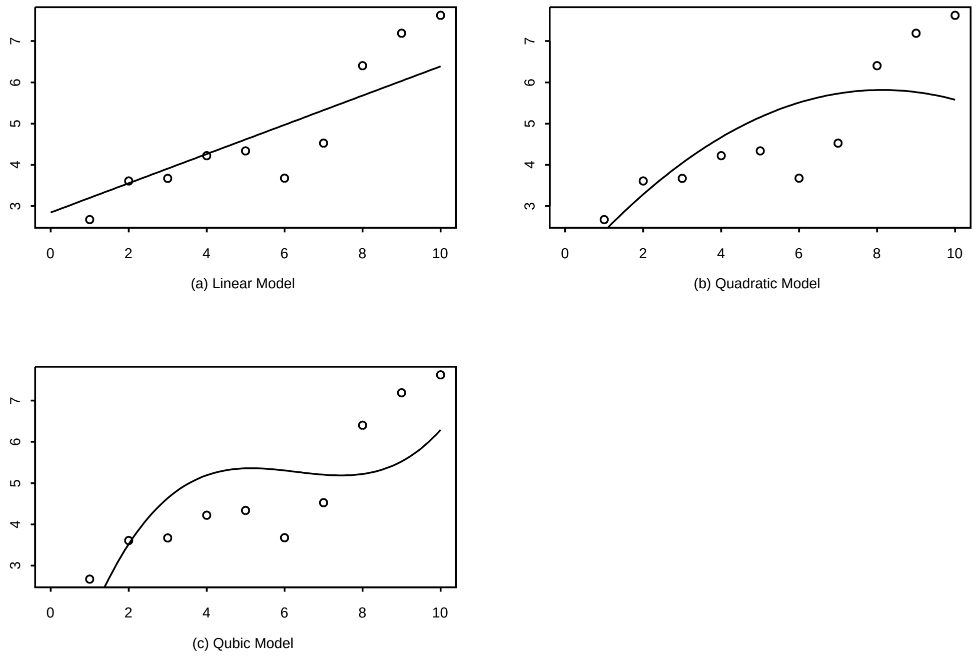
Once we have obtained estimates $\hat{\beta}_j$ by estimation from some training set T, we may use the resulting function to make predictions of y when we know the corresponding value of $\mathbf{x}$. Henceforth we denote this prediction by $\hat{f}(\mathbf{x})$. The difference between prediction $\hat{f}(\mathbf{x})$ and realisation y is called prediction error. It should preferrably take values close to zero. A natural quality measure of $\hat{f}$ as a predictor of Y is the mean squared error. For fixed T and $\mathbf{x}$

$$\mathrm{E}\,[(Y - \hat{f}(\mathbf{x}\,|\,T))^2]$$

where the expectation is taken with respect to $p(Y\,|\,\mathbf{x})$, the probability distribution of Y at $\mathbf{x}$. We may decompose this overall error into a *reducible* part, and an *irreducible* part that is due to the variability of Y at $\mathbf{x}$, as follows

$$\mathrm{E}\,[(Y - \hat{f}(\mathbf{x}\,|\,T))^2] = [f(\mathbf{x}) - \hat{f}(\mathbf{x}\,|\,T)]^2 + \mathrm{E}[(y - f(\mathbf{x}))^2]$$

where $f(\mathbf{x}) \equiv \mathrm{E}[Y\,|\,\mathbf{x}]$. The last term in this expression is the mean square error of the best possible (in the mean squared error sense) prediction $\mathrm{E}[Y\,|\,\mathbf{x}]$. Since



(a) Linear Model

(b) Quadratic Model

(c) Qubic Model

Fig. 4. Fit of equations to new sample T'

we can't do much about it, we focus our attention on the other source of error $[f(\mathbf{x}) - \hat{f}(\mathbf{x} \mid T)]^2$. This tells us something about the quality of the estimate $\hat{f}(\mathbf{x} \mid T)$ for a particular realisation of T. To say something about the quality of the estimator $\hat{f}$, we take its expectation over all possible training samples (of fixed size) and decompose it into its bias and variance components as discussed in section 4.1:

$$\mathrm{E}_T[(f(\mathbf{x}) - \hat{f}(\mathbf{x} \mid T))^2] = (f(\mathbf{x}) - \mathrm{E}_T[\hat{f}(\mathbf{x} \mid T)])^2 + \mathrm{E}_T[(\hat{f}(\mathbf{x} \mid T) - \mathrm{E}_T[\hat{f}(\mathbf{x} \mid T)])^2]$$

The first component is the squared bias, where bias is the difference between the best prediction $f(\mathbf{x})$ and its average estimate over all possible samples of fixed size. The second component, variance, is the expected squared difference between an estimate obtained for a single training sample and the average estimate obtained over all possible samples.

We illustrate these concepts by a simple simulation study using the models introduced in the previous section. The expectations in the above decomposition are taken over *all possible* training samples, but this is a little bit to much to compute. Instead we use the computer to draw a large number of random samples

x	$f(x)$	$\mathrm{E}(\hat{f}_1)$	$\mathrm{E}(\hat{f}_2)$	$\mathrm{E}(\hat{f}_3)$	$\mathrm{V}(\hat{f}_1)$	$\mathrm{V}(\hat{f}_2)$	$\mathrm{V}(\hat{f}_3)$
1	2.50	2.48	2.48	2.49	0.34	0.61	0.84
2	3.00	2.99	2.98	2.98	0.25	0.27	0.29
3	3.50	3.49	3.49	3.48	0.18	0.18	0.33
4	4.00	3.99	4.00	3.99	0.13	0.20	0.32
5	4.50	4.50	4.50	4.50	0.10	0.23	0.25
6	5.00	5.00	5.00	5.01	0.10	0.22	0.23
7	5.50	5.50	5.51	5.52	0.13	0.19	0.28
8	6.00	6.01	6.01	6.02	0.17	0.18	0.31
9	6.50	6.51	6.51	6.51	0.24	0.28	0.30
10	7.00	7.01	7.01	7.00	0.33	0.62	0.80

Table 10. Expected value and variance of $\hat{f}_j$

to obtain an estimate of the desired quantities. In the simulation we sampled 1000 times from

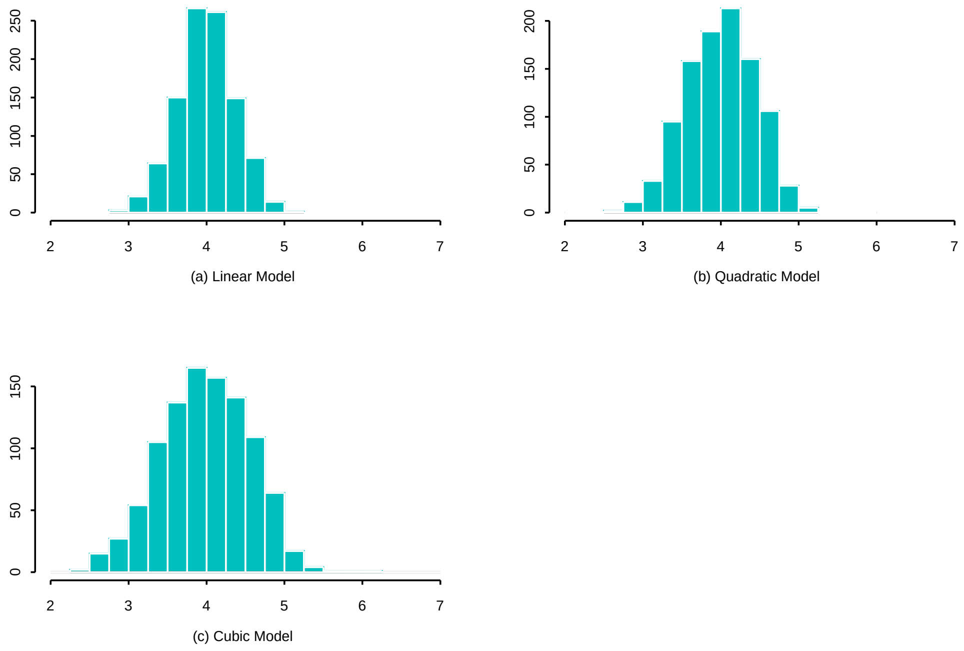
$$Y_i \sim \mathcal{N}(\mu = 2 + 0.5x_i, \sigma_\varepsilon^2 = 1)$$

with $x_i = 1, 2, \ldots, 10$. In other words we generated 1000 random samples, $T_1, T_2, \ldots, T_{1000}$ each consisting of 10 (x, y) pairs. For each T_i, the least squares parameter estimates for the three models were computed. Using the estimated models we computed the predicted values $\hat{f}(x)$. From the 1000 predicted values we computed the mean to estimate the expected value $\mathrm{E}(\hat{f}(x))$ and variance to estimate $\mathrm{V}(\hat{f}(x))$. The results of this simulation study are summarized in Table 10. Consider the fourth row of this table for the moment. It contains the simulation results of the predictions of the different models for $x = 4$. The expected value is $f(4) = \mathrm{E}(Y|x = 4)$ is $2 + 0.5 \cdot 4 = 4$. From the first three columns we conclude that all models have no or negligable bias; in fact we can prove mathematically they are unbiased since all three models encompass the correct model. But now look at the last three columns of table 10. We see that the linear model has lowest variance, the cubic model has highest variance, and the quadratic model is somewhere inbetween. This is also illustrated by the histograms displayed in Fig. 5. We clearly see the larger spread of the cubic model compared to the linear model. Although all three models yield unbiased estimates, the linear model tends to have a lower prediction error because its variance is smaller than that of the quadratic and cubic model.

The so-called bias/variance dillema lies in the fact that there is a tradeoff between the bias and variance components of error. Incorrect models lead to high bias, but highly flexible models suffer from high variance. For a fixed bias, the variance tends to decrease when the training sample gets larger and larger. Consequently, for very large training samples, bias tends to be the most important source of prediction error.

This phenomenon is illustrated by a second simulation. We generated the training sets by drawing from

$$Y_i \sim \mathcal{N}(\mu = 2 + 0.5x_i + 0.02x_i^2, \sigma_\varepsilon^2 = 1).$$



(a) Linear Model

(b) Quadratic Model

(c) Cubic Model

Fig. 5. Histograms of $\hat{f}_j(4)$ based on 1000 samples

The true model is quadratic, so the linear model is biased whereas the quadratic and cubic model are unbiased. We generated 1000 training samples of size 10, 100 and 1000 respectively. The first three columns of Table 11 summarize the estimated squared bias for the different models and sample sizes.

The results confirm that the linear model is biased, and furthermore indicate that the bias component of error does not decrease with sample size. Now consider the variance estimates shown in the middle three columns of Table 11. Looking at the rows, we observe that variance does decrease with the size of the sample. Taking these two phenomena together results in the summary of mean square error given in the last three columns of Table 11. The linear model outperforms the other models for small sample size, despite its bias. Because variance is a substantial component of overall error the linear model profits from its smaller variance. As the sample size gets larger, variance becomes only a small part of total error, and the linear model becomes worse due to its bias.

These phenomena explain the recent interest in increasingly flexible techniques that reduce estimation bias, since large data sets are quite commonly analyzed in business nowadays, using so-called *data mining* techniques.

	Squared bias			Variance			Mean square error		
m	10	100	1000	10	100	1000	10	100	1000
Linear (f_1)	.021	.022	.022	.197	.022	.002	.218	.043	.024
Quadratic (f_2)	.000	.000	.000	.299	.037	.004	.299	.037	.004
Cubic (f_3)	.001	.000	.000	.401	.054	.006	.401	.054	.006

Table 11. Bias, variance and mean squared estimation error for different sample sizes

5.2 Prediction error in classification

In classification y assumes values on an unordered discrete set $y \in \{y_1, ..., y_L\}$. For ease of exposition we assume $L = 2$, and $y \in \{0, 1\}$. We now have

$$\mathrm{E}[y \,|\, \mathbf{x}] \equiv f(\mathbf{x}) = P(y = 1 | \, \mathbf{x}) = 1 - P(y = 0 \,|\, \mathbf{x})$$

The role of a classification procedure is to construct a rule that makes a prediction $\hat{y}(\mathbf{x}) \in \{0, 1\}$ for the class label y at every input point $\mathbf{x}$.

For classification it is customary to measure prediction error in terms of the error rate $P(\hat{y} \neq y)$ [11, 13]. The best possible allocation rule, in terms of error rate, is the so called Bayes rule:

$$y_B(\mathbf{x}) = I(f(\mathbf{x}) \geq 1/2)$$

where $I(\cdot)$ is an indicator function of the truth of its argument. This rule simply states that $\mathbf{x}$ should be allocated to the class that has highest probability at $\mathbf{x}$. The probability of error $P(y_B \neq y)$ of the Bayes rule represents the irreducible error rate, analogous to $\mathrm{E}[(y - f(\mathbf{x}))^2]$ in regression.

Training data T are again used to form an estimate $\hat{f}(\mathbf{x} \,|\, T)$ of $f(\mathbf{x})$ to construct an allocation rule:

$$\hat{y}(\mathbf{x} \,|\, T) = I(\hat{f}(\mathbf{x} \,|\, T) \geq 1/2)$$

We discuss two alternative decompositions of the prediction error of a classification method. The first one is due to Friedman [8], and aims to keep a close analogy to the decomposition for regression problems. The second one is an additive decomposition due to Breiman [3], and is aimed at explaining the success of model aggregation techniques such as bagging (see section 6.4) and arcing (see section 6.5).

Friedman's bias-variance decomposition for classifiers Friedman [8] proposes the following decomposition of the error rate of an allocation rule into reducible and irreducible error

$$P(\hat{y}(\mathbf{x}) \neq y(\mathbf{x})) = |2f(\mathbf{x}) - 1| P(\hat{y}(\mathbf{x}) \neq y_B(\mathbf{x})) + P(y_B(\mathbf{x}) \neq y(\mathbf{x}))$$

The reader should note that the reducible error is expressed as a *product* of two terms, rather than a *sum*, in the above equation. The first term of the product denotes the increase in error if $\hat{f}$ is on the wrong side of 1/2, and the second term of the product denotes the probability of this event. We consider a simple example to illustrate this decomposition.

Example 25. Suppose $P(y = 1|\mathbf{x}) = f(\mathbf{x}) = 0.8$, and $\mathrm{E}(\hat{f}(\mathbf{x})) = 0.78$. Since $f(\mathbf{x}) \geq 1/2$, Bayes rule allocates $\mathbf{x}$ to class 1, that is $y_B(\mathbf{x}) = 1$. The error rate of Bayes rule is $P(y = 0|\mathbf{x}) = 1 - 0.8 = 0.2$. Now $\hat{y}(\mathbf{x})$ differs from Bayes rule if it is below 1/2. Assume the distribution $p(\hat{f}(\mathbf{x}))$ of $\hat{f}(\mathbf{x})$ is such that $P(\hat{f}(\mathbf{x}) < 1/2) = 0.1$, and consequently $P(\hat{y}(\mathbf{x}) \neq y_B(\mathbf{x})) = 0.1$. We can now compute the error rate as follows

$$\begin{aligned} P(\hat{y}(\mathbf{x}) \neq y(\mathbf{x})) &= P(\hat{f}(\mathbf{x}) \geq 1/2)P(y = 0|\mathbf{x}) + P(\hat{f}(\mathbf{x}) < 1/2)P(y = 1|\mathbf{x}) \\ &= 0.9 \cdot 0.2 + 0.1 \cdot 0.8 = 0.26 \end{aligned}$$

The decomposition of Friedman shows how this error rate may be split into reducible and irreducible parts as follows. We saw the irreducibe error $P(y_B(\mathbf{x}) \neq y(\mathbf{x})) = 0.2$. We assumed $P(\hat{y}(\mathbf{x}) \neq y_B(\mathbf{x})) = 0.1$. If $\hat{f}(\mathbf{x}) < 1/2$ the error rate increases from 0.2 to 0.8, an increase of 0.6 that is represented by the term

$$|2f(\mathbf{x}) - 1| = 2 \cdot 0.8 - 1 = 0.6$$

Summarizing Friedman's decomposition gives

$$P(\hat{y}(\mathbf{x}) \neq y(\mathbf{x})) = 0.6 \cdot 0.1 + 0.2 = 0.26.$$

The important difference with regression is that a wrong estimate $\hat{f}(\mathbf{x})$ of $f(\mathbf{x})$ does not necessarily lead to a different allocation as Bayes rule, as long as they are on the same side of 1/2 [8]. In Fig. 6 we show a simple representation of the decision boundary when $\mathbf{x}$ only has two components, i.e. $\mathbf{x} = (x_1, x_2)$. The picture shows the optimal decision boundary corresponding to the Bayes allocation rule as the curved solid line $f(x_1, x_2) = 1/2$. The dotted line denotes the average decision boundary $\mathrm{E}[\hat{f}(x_1, x_2)]$ for some unspecified classifier. The areas in Fig. 6 labeled "pbb" correspond to the regions where the classifier is on average on the wrong side of 1/2; these regions are called regions of positive boundary bias (hence "pbb") by Friedman [8]. In the remaining regions the classifier is either unbiased, or its bias is of the "harmless" kind in the sense that the classifier is on average on the right side of 1/2. In the latter case we speak of negative boundary bias.

When $\mathrm{E}[\hat{f}]$ and f are on the same side of 1/2 (negative boundary bias), then

1. the classification error decreases with increasing $|\,\mathrm{E}[\hat{f}] - 1/2\,|$ irrespective of the amount of bias $f - \mathrm{E}[\hat{f}]$, and
2. one can reduce the classification error towards its minimal value by reducing the variance of $\hat{f}$ alone.

Consider the distributions of $\hat{f}(\mathbf{x})$ displayed in Fig. 7 to get an intuitive appreciation for these phenomena. Assume $f(\mathbf{x}) = 0.6$, so in the top row $\mathrm{E}[\hat{f}(\mathbf{x})]$

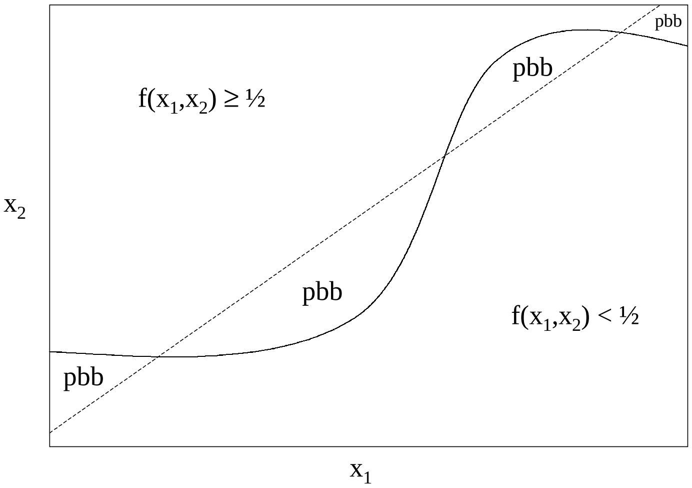
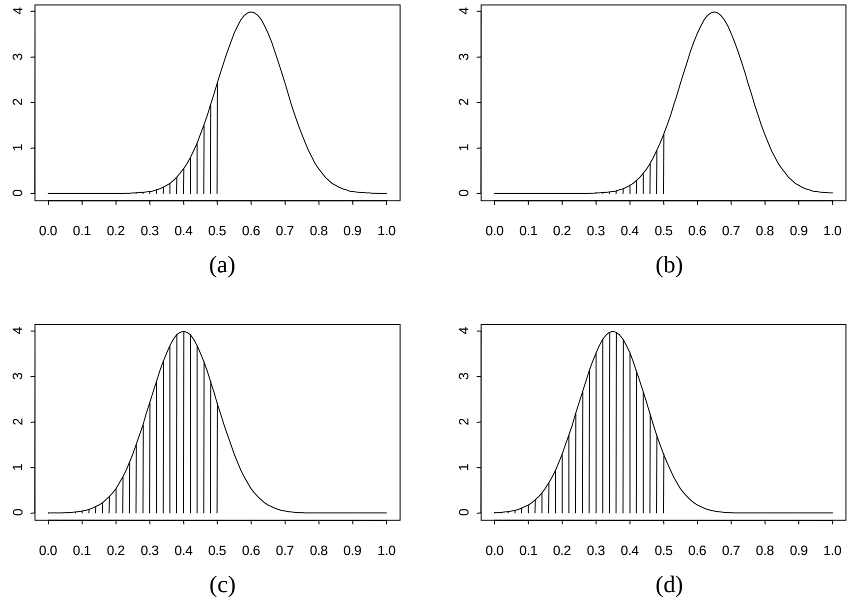


Fig. 6. Example regions of positive boundary bias: solid line denotes $f(x_1, x_2) = 1/2$ and dotted line denotes $\mathrm{E}[\hat{f}(x_1, x_2)] = 1/2$

is on the correct side of 1/2 (negative boundary bias). The shaded areas indicate the probability that $\hat{y}(\mathbf{x})$ differs from $y_B(\mathbf{x})$. In Fig. 7 (a) $\mathrm{E}[\hat{f}(\mathbf{x})] = 0.6$, so $\hat{f}(\mathbf{x})$ is unbiased in the mean squared error sense. In Fig. 7 (b) $\mathrm{E}[\hat{f}(\mathbf{x})] = 0.65$, so $\hat{f}(\mathbf{x})$ is biased in the mean squared error sense; $f - \mathrm{E}[\hat{f}] = -0.05$. Since the distributions have the same variance, the biased estimator actually has a lower error rate! An increase in bias (with constant variance) actually leads to a decrease in error rate, something that could never occur with mean squared error since that is simply the sum of squared bias and variance. In the bottom row of Fig. 7 $\mathrm{E}[\hat{f}(\mathbf{x})]$ is on the wrong side of 1/2 (positive boundary bias). In this case an increase in bias increases the error rate, as indicated by the size of the shaded areas in panel (c) and (d).

Next consider the influence of variance on the error rate, assuming constant bias. Again the qualitative behaviour differs, depending on whether there is negative boundary bias (top row of Fig. 8) or positive boundary bias (bottom row of Fig. 8). In the first case an increase in variance leads to an increase of the error rate, as witnessed by the increase of the shaded area. In the latter case however, an increase in variance leads to a decrease in the error rate.

We may summarize these findings as follows. The effect of bias and variance (as defined for regression models) on prediction error in classification involves a rather complex interaction, unlike the nice additive decomposition obtained



(a)

(b)

(c)

(d)

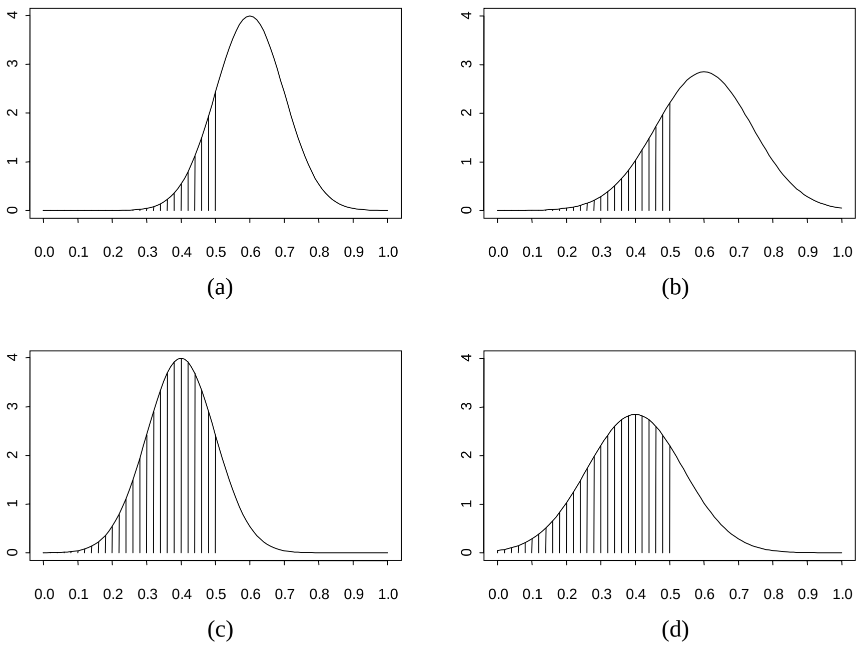
Fig. 7. Distributions $p(\hat{f}(\mathbf{x}))$; $f(\mathbf{x}) = 0.6$.

for prediction error in regression. This interaction explains the success of some obviously biased classification procedures. As long as boundary bias is predominantly negative the error rate can be made arbitrarily small by reducing variance. As indicated in [8] classification procedures such as naive Bayes and k-nearest neighbour tend to produce negative boundary bias. This explains their relative success in comparison to more flexible classifiers, even on data sets of moderate size [12].

Breiman's bias-variance decomposition for classifiers In [3] Breiman proposes an additive decomposition of the prediction error of classifiers aimed at explaining the success of particular aggregation techniques such as bagging and arcing (see sections 6.4 and 6.5). To this end an aggregate classifier is definded as follows

$$y_A(\mathbf{x}) = I(\mathrm{E}\hat{f}(\mathbf{x}) \geq 1/2)$$

where the subscript A denotes aggregation.



(a)

(b)

(c)

(d)

Fig. 8. Distributions $p(\hat{f}(\mathbf{x}))$; $f(\mathbf{x}) = 0.6$.

Note 3. In fact, Breiman defines the aggregate classifier somewhat differently as follows

$$f_A(\mathbf{x}) = \mathrm{E}(I(\hat{f}(\mathbf{x}) \geq 1/2))$$

and then

$$y_A(\mathbf{x}) = I(f_A(\mathbf{x}) \geq 1/2)$$

In the first definition the class probabilities of the individual classifiers are averaged and the aggregate classifier assigns $\mathbf{x}$ to the class with largest average probability. In the second definition the aggregate classifier assigns $\mathbf{x}$ to the class to which the individual classifiers allocate $\mathbf{x}$ most often (majority voting). If $p(\hat{f})$ is symmetric and unimodal, the two definitions are equivalent. Henceforth we stick to the first definition.

We already saw that the reducible error at $\mathbf{x}$ is

$$r(\mathbf{x}) = P(\hat{y}(\mathbf{x}) \neq y(\mathbf{x})) - P(y_B(\mathbf{x}) \neq y(\mathbf{x}))$$

The bias at $\mathbf{x}$ is now defined as follows

$$\mathrm{bias}(\mathbf{x}) = I(y_A(\mathbf{x}) \neq y_B(\mathbf{x}))\, r(\mathbf{x})$$

and likewise

$$\mathrm{var}(\mathbf{x}) = I(y_A(\mathbf{x}) = y_B(\mathbf{x}))\, r(\mathbf{x})$$

By definition $r(\mathbf{x}) = \mathrm{bias}(\mathbf{x}) + \mathrm{var}(\mathbf{x})$, so

$$P(\hat{y}(\mathbf{x}) \neq y(\mathbf{x})) = \mathrm{bias}(\mathbf{x}) + \mathrm{var}(\mathbf{x}) + P(y_B(\mathbf{x}) \neq y(\mathbf{x}))$$

Example 26. (Continuation of example 25) Recall that $\mathrm{E}[\hat{f}(\mathbf{x})] = 0.78$. The aggregate classifier allocates $\mathbf{x}$ to class

$$y_A(\mathbf{x}) = I(\mathrm{E}[\hat{f}(\mathbf{x})] \geq 1/2) = I(0.78 \geq 0.5) = 1$$

The reducible error $r(\mathbf{x}) = 0.26 - 0.2 = 0.06$. Since the aggregate classifier allocates $\mathbf{x}$ to the same class as Bayes rule, i.e. $y_A(\mathbf{x}) = y_B(\mathbf{x})$, we have

$$\mathrm{bias}(\mathbf{x}) = I(y_A(\mathbf{x}) \neq y_B(\mathbf{x}))\, r(\mathbf{x}) = 0 \cdot 0.06 = 0$$

and

$$\mathrm{var}(\mathbf{x}) = I(y_A(\mathbf{x}) = y_B(\mathbf{x}))\, r(\mathbf{x}) = 1 \cdot 0.06 = 0.06$$

In comparison to Friedman's decomposition, the reducible error is called bias in regions of positive boundary bias, and variance in regions of negative boundary bias. At any given point $\mathbf{x}$, the classifier has either bias or variance, depending upon whether or not the aggregated classifier disagrees with the Bayes rule there.

6 Resampling

6.1 Introduction

Resampling techniques are computationally expensive techniques that reuse the available sample to make statistical inferences. Because of their computational requirements these techniques were infeasible at the time that most of "classical" statistics was developed. With the availability of ever faster and cheaper computers, their popularity has grown very fast in the last decade. In this section we provide a brief introduction to some important resampling techniques.

6.2 Cross-Validation

Cross-Validation is a resampling technique that is often used for model selection and estimation of the prediction error of a classification- or regression function. We have seen already that squared error is a natural measure of prediction error for regression functions:

$$\mathrm{PE} = \mathrm{E}(y - \hat{f})^2$$

Estimating prediction error on the same data used for model estimation tends to give downward biased estimates, because the parameter estimates are "fine-tuned" to the peculiarities of the sample. For very flexible methods, e.g. neural networks or tree-based models, the error on the training sample can usually be

	in-sample	leave-one-out
linear	150.72	167.63
quadratic	16.98	19.89
cubic	16.66	20.66

Table 12. Mean square error of candidate models: in-sample and leave-one-out

made close to zero. The true error of such a model will usually be much higher however: the model has been "overfitted" to the training sample.

An alternative is to divide the available data into a training sample and a test sample, and to estimate the prediction error on the test sample. If the available sample is rather small, this method is not preferred because the test sample may not be used for model estimation in this scenario. Cross-validation accomplishes that all data points are used for training as well as testing. The general K-fold cross-validation procedure works as follows

1. Split the data into K roughly equal-sized parts.
2. For the kth part, estimate the model on the other $K-1$ parts, and calculate its prediction error on the kth part of the data.
3. Do the above for $k = 1, 2, \ldots, K$ and combine the K estimates of prediction error.

If $K = m$, we have the so-called *leave-one-out* cross-validation: one observation is left out at a time, and $\hat{f}$ is computed on the remaining $m-1$ observations.

Now let $k(i)$ be the part containing observation i. Denote by $\hat{f}_i^{-k(i)}$ the value predicted for observation i by the model estimated from the data with the $k(i)$th part removed. The cross-validation estimate of mean squared error is now

$$\widehat{\mathrm{PE}}_{cv} = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{f}_i^{-k(i)})^2$$

We consider a simple application of model selection using cross-validation, involving the linear, quadratic and cubic model introduced in section 5. In a simulation study we draw 50 (x, y) observations from the probability distributions

$$X \sim U(0, 10) \quad \text{and} \quad Y \sim \mathcal{N}(\mu = 2 + 3x + 1.5x^2, \sigma_\varepsilon = 5),$$

i.e. $\mathrm{E}(Y)$ is a quadratic function of x. For the purposes of this example, we pretend we don't know the true relation between x and y, as would usually be the case in a practical data analysis setting. We consider a linear, quadratic and cubic model as the possible candidates to be selected as the model with lowest prediction error, and we use leave-one-out cross validation to compare the three candidates.

The first column of Table 12 contains the "in-sample" estimate of the mean square error of all three models. Based on the in-sample comparison one would select the cubic model as the best model since it has the lowest prediction error.

We already noted however that this estimate tends to be too optimistic, and the more flexible the model the more severe the optimism tends to be. In the second column the cross-validation estimates of prediction error are listed. As one would expect they are higher than their in-sample counterparts. Furthermore, we see that the quadratic model (the true model) has the lowest cross-validation prediction error of the three. The lower in-sample prediction error of the cubic was apparently due to a modest amount of overfitting.

6.3 Bootstrapping

In section 3 we saw that in some special cases we can derive mathematically the exact distribution of a sample statistic, and in some other cases we can rely on limiting distributions as an approximation to the sampling distribution for a finite sample. For many statistics that may be of interest to the analyst, such exact or limiting distributions cannot be derived analytically. In yet other cases the asymptotic approximation may not provide a good fit for a finite sample. In such cases an alternative approximation to the sampling distribution of a statistic $t(\mathbf{x})$ may be obtained using just the data at hand, by a technique called *bootstrapping* [6,4]. To explain the basic idea of the *non-parametric* bootstrap, we first introduce the *empirical distribution function*

$$\hat{F}(z) = \frac{1}{m}\sum_{i=1}^{m} I\left(x_i \leq z\right) \quad -\infty < z < \infty$$

where I denotes the indicator function and $\mathbf{x} = (x_1, x_2, ..., x_m)$ is a random sample from population distribution function F. We now approximate the sampling distribution of $t(\mathbf{x})$ by repeated sampling from $\hat{F}$. This is achieved by drawing samples $\mathbf{x}^{(r)}$ of size m by sampling independently *with replacement* from $(x_1, x_2, ..., x_m)$. If all observations are distinct, there are $\binom{2m-1}{m}$ distinct samples in

$$\mathcal{B} = \{\mathbf{x}^{(r)}, r = 1, ..., \binom{2m-1}{m}\}$$

with respective multinomial probabilities (see section 2.15)

$$P(\mathbf{x}^{(r)}) = \frac{m!}{j_1^{(r)}!\, j_2^{(r)}! \ldots j_m^{(r)}!}(\frac{1}{m})^m$$

where $j_i^{(r)}$ is the number of copies of x_i in $\mathbf{x}^{(r)}$. The bootstrap distribution of $t(\mathbf{x})$ is derived by calculating the realisation $t(\mathbf{x}^{(r)})$ for each of the resamples and assigning each one probability $P(\mathbf{x}^{(r)})$. As $m \to \infty$, the empirical distribution $\hat{F}$ converges to the underlying distribution F, so it is intuitively plausible that the bootstrap distribution is an asymptotically valid approximation to the sampling distribution of a statistic.

We can in principle compute all $\binom{2m-1}{m}$ values of the statistic to obtain its “ideal” bootstrap distribution, but this is computationally infeasible even for moderate m. For $m = 15$ there are already 77558760 distinct samples. The usual alternative is to use Monte-Carlo simulation, by drawing a number B of samples and using them to approximate the bootstrap distribution.

If a *parametric* form is adopted for the underlying distribution, where θ denotes the vector of unknown parameters, then the parametric bootstrap uses an estimate $\hat{\theta}$ formed from $\mathbf{x}$ in place of θ. If we write F_θ to signify its dependence on θ, then bootstrap samples are generated from $\hat{F} = F_{\hat{\theta}}$.

The non-parametric bootstrap makes it unneccesary to make parametric assumptions about the form of the underlying distribution. The parametric bootstrap may still provide more accurate answers than those provided by limiting distributions, and makes inference possible when no exact or limiting distributions can be derived for a sample statistic.

We present an elementary example to illustrate the parametric and non-parametric bootstrap. The population parameter of interest is the correlation coefficient, denoted by ρ. We first discuss this parameter before we show how to use bootstrapping to make inferences about it.

The linear dependence between population variables $\mathcal{X}$ and $\mathcal{Y}$ is measured by the covariance

$$\sigma_{xy} = \frac{1}{M} \sum_{i=1}^{M} (x_i - \mu_x)(y_i - \mu_y)$$

A term $(x_i - \mu_x)(y_i - \mu_y)$ from this sum is positive if both factors are positive or both are negative, i.e. if x_i and y_i are both above or both below their mean. Such a term is negative if x_i and y_i are on opposite sides of their mean. The dimension of σ_{xy} is the product of the dimensions of X and Y; division by both σ_x and σ_y yields a dimensionless number called the correlation coefficient, i.e.

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Evidently ρ has the same sign as σ_{xy}, and always lies between -1 and $+1$. If $\rho = 0$ there is no linear dependence: the two variables are uncorrelated. The linear dependence increases as $|\rho|$ gets closer to 1. If all pairs (x, y) are on a straight line with positive slope, then $\rho = 1$; if all pairs are on a straight line with negative slope then $\rho = -1$.

To make inferences about ρ we use the sample correlation coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m} (x_i - \bar{x})^2 \sum_{i=1}^{m} (y_i - \bar{y})^2}}$$

The sampling distribution of this statistic can't be mathematically derived in general, in fact there is no general expression for the expected value of r_{xy}. Therefore it makes sense to use the bootstrap to make inferences concerning ρ.

In our study, we draw 30 (x, y) pairs from a standard binormal distribution with $\rho = 0.7$, i.e.

$$(X, Y) \sim \mathcal{N}^2(\mu_x = 0, \mu_y = 0, \sigma_x^2 = 1, \sigma_y^2 = 1, \rho = 0.7)$$

Based on this dataset, bootstrapping proceeds as follows

Non-parametric: Draw samples of 30 (x, y) pairs (with replacement) from the data. For each bootstrap sample, compute r, to obtain an empirical sampling distribution.

Parametric: Make appropriate assumptions about the joint distribution of X and Y. In our study we assume

$$(X, Y) \sim \mathcal{N}^2(\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$$

which happens to be correct. In a practical data analysis situation we would evidently not know that, and it would usually be hard to ascertain that our assumptions are appropriate. We build an empirical sampling distribution by drawing samples of size 30 from

$$\mathcal{N}^2(\bar{x}, \bar{y}, s_x^2, s_y^2, r)$$

In both cases we draw 1000 samples to generate the empirical sampling distribution of r. To construct $100(1-\alpha)\%$ confidence intervals for ρ, we simply take the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of this distribution.

In order to determine whether the bootstrap provides reliable confidence intervals with the right coverage, we repeated the following procedure 100 times

1. Draw a sample of size 30 from the population.
2. Build a bootstrap distribution for r, and construct 90% confidence intervals for ρ. (both parametric and non-parametric)
3. Determine whether the true value of ρ is inside the confidence interval.

Like any conventional method for constructing confidence intervals, the bootstrap will sometimes miss the true value of the population parameter. This happens when the data is not representative for the population. For example, in 1 of the 100 samples the sample correlation coefficient was 0.36. This is highly unlikely to occur when sampling from a population with $\rho = 0.7$ but it will occur occasionally. In such a case the bootstrap distribution of r is bound to be way off as well. In Fig. 9 the non-parametric bootstrap distribution for this particular sample is displayed. The 90% confidence interval computed from this distribution is $(0.064, 0.610)$. Not surprisingly it does not contain the true value of ρ.

On average, one would expect a 90% confidence interval to miss the true value in 10% of the cases; that's why it's called a 90% confidence interval. Furthermore the narrower the confidence intervals, the more informative they are. Both the parametric and non-parametric bootstrap missed the true value of ρ 13 times out of 100, where one would expect 10 misses. Now we may test whether the bootstrap confidence intervals have the right coverage:

$$H_0 : \alpha = 0.1 \quad \text{against} \quad H_a : \alpha \neq 0.1$$

We observed 13 misses out of 100, so the observed value of our test statistic is $a = 0.13$. The distribution of $\hat{\alpha}$ under H_0 (the null-hypothesis) may be approximated by

$$\hat{\alpha} \approx_{H_0} \mathcal{N}(\mu = \alpha, \sigma^2 = \alpha(1-\alpha)/m)$$

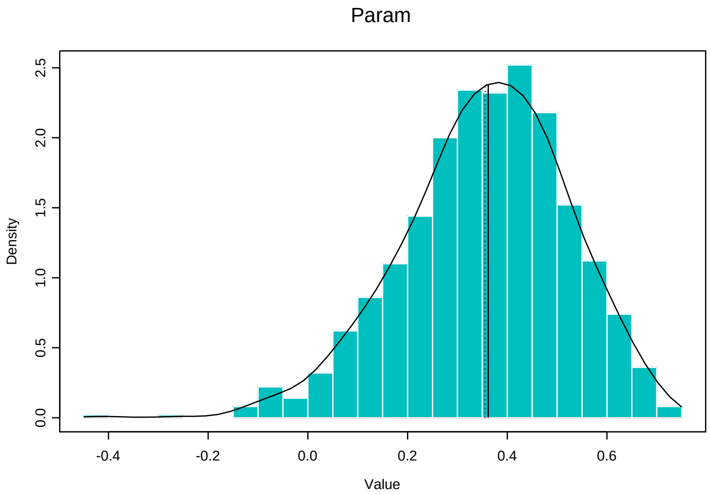


Fig. 9. Bootstrap distribution for r. Observed value of r is 0.36.

which yields $\hat{\alpha} \approx \mathcal{N}(0.1, 0.0009)$. We may now compute the p-value of the observed value under the null-hypothesis as follows

$$P_{H_0}(\hat{\alpha} > a) = P_{H_0}(\hat{\alpha} > 0.13) = P(Z > \frac{0.13 - 0.1}{\sqrt{0.0009}}) = P(Z > 1) = 0.1587$$

where the value 0.1587 was looked-up in a table for the standardnormal distribution. Since we are performing a two-sided test this probability should be doubled, so we obtain a p-value of $2 \times 0.1587 = 0.3174$. This means we would not reject the null-hypothesis under any conventional significance level. The probability under the null-hypothesis of obtaining a result at least as far from $\alpha_0 = 0.1$ (to either side) as the one we observed is "pretty high".

The mean length of the confidence intervals is 0.31 for the non-parametric bootstrap, and 0.32 for the parametric bootstrap. Even though the assumptions of the parametric bootstrap were correct it did not give shorter confidence intervals on average.

6.4 Bagging Predictors

Bagging [2] is an acronym for **b**ootstrap **agg**regat**ing**, and is used to construct aggregated predictors with the intention of reducing the variance component of prediction error (see section 5). The rationale of this idea uses a form of "reasoning by analogy" from population to sample that is characteristic for bootstrap procedures.

If we were to have access to the entire population of interest we could in principle compute aggregated regression and classification functions as follows.

If y is numerical, we replace $\hat{f}(\mathbf{x} \,|\, T)$ by the average of $\hat{f}(\mathbf{x} \,|\, T)$ over all possible samples T, that is by

$$f_A(\mathbf{x}) = \mathrm{E}_T \, \hat{f}(\mathbf{x} \,|\, T)$$

where the A in f_A denotes aggregation. This aggregation reduces mean prediction error, since the bias of $f_A(\mathbf{x})$ is the same as that of $\hat{f}(\mathbf{x} \,|\, T)$, but its variance is 0. The degree of this reduction depends on the relative importance of variance $V(\hat{f}(\mathbf{x} \,|\, T))$ as compared to squared bias $B^2(\hat{f}(\mathbf{x} \,|\, T))$. In general high variance predictors such as neural networks and regression trees tend to benefit substantially from aggregation.

If y is a class label, say $y \in \{0, 1\}$, recall from section 5.2 that the aggregated classifier may defined as

$$y_A(\mathbf{x}) = I(f_A \geq 1/2) = I(\mathrm{E}_T \hat{f}(\mathbf{x} \,|\, T) \geq 1/2)$$

where again the A in y_A denotes aggregation.

Now of course in practice we only have a single sample, so we cannot compute f_A. Here's where the bootstrap "reasoning by analogy" comes in. Instead of sampling repeatedly from the population, we *resample* the single sample T we have, and estimate f_A by

$$f_B(\mathbf{x}) = \hat{f}_A(\mathbf{x}) = \mathrm{E}_{T^{(r)}} \hat{f}(\mathbf{x} \,|\, T^{(r)})$$

where $T^{(r)}$ are bootstrap samples from T. Now there are two factors influencing the performance of f_B. If the predictor has high variance , it can give improvement through aggregation. If the predictor has low variance, then the unaggregated predictor $\hat{f}(\mathbf{x} \,|\, T)$ tends to be close to $\mathrm{E}_T \, \hat{f}(\mathbf{x} \,|\, T)$, and so $f_B = \mathrm{E}_{T^{(r)}} \hat{f}(\mathbf{x} \,|\, T^{(r)})$ may be less accurate then $\hat{f}(\mathbf{x} \,|\, T)$.

A simple simulation example serves to illustrate the idea and potential benefit of bagging. In this example we use a neural network as the regression method. We do not consider the details of neural networks here, as this is not required to understand the idea of bagging. Neural networks can be used as flexible regression methods and can approximate complex non-linear functions. In general, neural networks tend to have a low bias and high variance, so they may benefit from aggregation methods such as bagging.

The following steps were performed in the simulation study. We generated a simulated data set of size 200 with ten independent predictor variables $x_1, \ldots, x_{10}$ with $x_j \sim U(0, 1)$, $j = 1, \ldots, 10$. The response is given by

$$y = 10 \sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \varepsilon$$

with $\varepsilon \sim \mathcal{N}(0, 1)$. This data generating mechanism was used by Friedman in a paper on a flexible regression technique called MARS [7]. On this dataset we performed the following computations. First, we estimated ("trained") a single neural network on the dataset. This represented the "unaggregated" regression function. We computed the predictions of this network on a test sample of 1000 observations not used in training. Next, we drew 25 bootstrap samples from the data set at hand, and trained a neural network on each bootstrap sample. The

												unaggregated			bagged		
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	$f(\mathbf{x})$	$\mathrm{E}(\hat{f})$	B^2	V	$\mathrm{E}(\hat{f})$	B^2	V
1	.17	.29	.33	.11	.70	.69	.86	.83	.11	.78	6.79	5.62	1.36	98.58	6.75	0.00	23.59
2	.67	.25	.37	.51	.73	.99	.94	.35	.64	.36	14.14	14.27	0.02	1.81	14.50	0.13	0.55
3	.91	.34	.58	.17	.91	.26	.73	.40	.31	.72	14.60	14.31	0.08	2.72	13.93	0.45	5.29
4	.80	.09	.25	.08	.40	.70	.46	.36	.35	.09	6.31	7.85	2.37	77.48	7.80	2.20	1.28
5	.88	.43	.70	.83	.36	.55	.30	.90	.68	.05	20.19	19.80	0.15	1.29	19.24	0.90	1.24
6	.91	.53	.12	.04	.42	.88	.81	.87	.35	.33	15.36	15.07	0.09	83.13	13.42	3.77	2.03
7	.01	.53	.98	.09	.32	.91	.26	.06	.09	.67	7.22	5.99	1.52	105.79	5.83	1.94	36.42
8	.54	.33	.20	.67	.93	.42	.47	.14	.60	.06	18.53	18.36	0.03	14.62	18.35	0.03	1.57
9	.32	.42	.64	.07	.04	.12	.39	.14	.75	.07	5.40	5.38	0.00	1.70	5.77	0.14	7.17
10	.54	.93	.71	.79	.84	.18	.31	.33	.84	.06	23.05	22.04	1.01	12.86	22.12	0.85	2.64

Table 13. Bias and variance of single and aggregated neural network

predictions of the 25 networks were averaged to obtain averaged predictions for the test sample.

This whole procedure was repeated 100 times to obtain 100 predictions for each point in the test sample. We may now decompose the mean square estimation error into its bias and variance components. We would expect the aggregated neural network to show smaller variance and comparable bias, on average leading to smaller prediction error than the unaggregated version.

Averaged over the 1000 test observations, the variance of the unaggregated neural network is 47.64, whereas the variance of its aggregated counterpart is about 10 times smaller, namely 4.66. The bias components of error are about equal: 1.51 for the unaggregated neural network against 1.97 for the aggregated version. This example illustrates that bagging is a variance reduction technique; the bias component of error is not reduced by bagging. Table 13 displays the bias-variance decomposition for the unaggregated and bagged neural network for 10 of the 1000 observations in the test set.

In [2] and [3], elaborate experiments with bagging are presented on simulated as well as "real-life" data sets. The overall conclusion is that bagging tends to lead to a substantial reduction in prediction error for regression as well as classification methods. Since it is a variance-reduction technique, it tends to work well for methods with high variance (such as neural networks and tree-based methods) but does not improve the performance of methods with low variance such as linear regression and linear discriminant analysis.

6.5 Arcing classifiers

Arcing [3] is an acronym for **a**daptive **r**esampling and **c**ombin**ing,** and has proved to be a succesfull technique in reducing the prediction error of classifiers. Like bagging, arcing is an aggregation method based on resampling the available data. The main difference with bagging is the *adaptive* resampling scheme. In bagging bootstrap samples are generated by sampling the original data where each data point has probability $1/m$ (where m is the number of elements in the original sample). In the arcing algorithm these inclusion probabilities may change.

At the start of each construction, there is a probability distribution $\{p(i)\}, i = 1, ..., m$ on the cases in the training set. A bootstrap sample $T^{(r)}$ is constructed by sampling m times (with replacement) from this distribution. Then the probabilities are updated in such a way that the probability of point $\mathbf{x}$ is increased if it is misclassified by the classifier constructed from $T^{(r)}$ and remains the same otherwise (of course the revised "probabilities" are rescaled to sum to one). After R classifiers have been constructed in this manner, a (possibly weighted) voting for the class allocation is performed.

The intuitive idea of arcing is that the points most likely to be selected for the replicate data sets are those most likely to be misclassified. Since these are the troublesome points, focussing on them using the adaptive resampling scheme of arcing may do better than the "neutral" bagging approach.

We perform a simple simulation study to illustrate the idea. First we specify the exact arcing algorithm used in this simulation study. It is essentially the arcx4 algorithm proposed by Breiman [3] except that we don't aggregate the classifiers by voting, but rather by averaging the predicted class probabilities (see section 5.2). The algorithm works as follows

1) At the r^{th} step, sample with replacement from T to obtain training set $T^{(r)}$, using inclusion probabilities $\{p^{(r)}(i)\}$. Use $T^{(r)}$ to construct (estimate) $\hat{f}_r$.

2) Predict the class of each point in T using $\hat{f}_r$, i.e. determine $\hat{y}_r = I(\hat{f}_r \geq 1/2)$ for each point in T. Now let $k(i)$ be the number of misclassifications of the i^{th} point by $\hat{y}_1, \ldots, \hat{y}_r$, the classifiers constructed so far.

3) The new $r+1$ step inclusion probabilities are defined by

$$p^{(r+1)}(i) = (1 + k(i)^4) / \sum_{j=1}^{m} (1 + k(j)^4)$$

4) Finally, after R steps the aggregated classifier is determined

$$\hat{y}_A = I(\hat{f}_A \geq 1/2), \quad \text{with} \quad \hat{f}_A = 1/R \sum \hat{f}_r$$

For bagging we use the same manner of aggregating classifiers, but the inclusion probabilities are simply $1/m$ at each step.

The simulation study may now be summarized as follows. Repeat the following steps 100 times

1. Generate a training set T containing 150 observations from class 0 and 150 observations from class 1.
2. Estimate the unaggregated function $\hat{f}$ and classifier $\hat{y}$ using T.
3. Apply arcx4 to T to obtain an aggregated function $\hat{f}_a$ and classifier $\hat{y}_a$, and apply bagging to T to construct an aggregated function $\hat{f}_b$ and classifier $\hat{y}_b$ (the subscript "a" stands for arcing and "b" for bagging). We chose $R = 25$ so the aggregated classifiers are based on averages of 25 functions.

	Breiman				Squared error		
	$\bar{e}$	e_B	bias	var	$\mathrm{E}(f-\hat{f})^2$	$(f-\mathrm{E}\hat{f})^2$	$\mathrm{E}(\hat{f}-\mathrm{E}\hat{f})^2$
unaggregated	21.6%	3.7%	4.4%	13.6%	.138	.068	.070
arced	10.4%	3.7%	1.8%	4.9%	.095	.089	.006
bagged	14.1%	3.7%	6.1%	4.3%	.082	.072	.010

Table 14. Prediction error of arcing and bagging and its bias-variance decomposition

4. Use the classifiers so obtained to predict the class of 1000 observations in an independent test.

After these computations have been performed we have 100 predictions for each observation in the test set for the arced classifier as well as the bagged classifier and the unaggregated classifier. Furthermore we know the bayes error for each point in the test set since we specified the distribution from which the class 0 points were drawn, the distribution from which the class 1 points were drawn, and the prior probabilities of both classes. This allows us the compute the probability of class 1 at point $\mathbf{x}$ using Bayes rule as follows

$$P(y=1\,|\,\mathbf{x}) = \frac{p(\mathbf{x}\,|\,y=1)P(y=1)}{p(\mathbf{x}\,|\,y=1)P(y=1)+p(\mathbf{x}\,|\,y=0)P(y=0)}$$

where p denotes a probability density.

The class 0 observations were drawn from a 10-dimensional normal distribution with mean 0 and covariance matrix 5 times the identity. The class 1 observations were drawn from a 10-dimensional normal distribution with mean 0.5 and identity covariance matrix.

$$\mathbf{x}\,|\,y=0 \sim \mathcal{N}^{10}(\mu_{\mathbf{0}}, \Sigma_0), \quad \mathbf{x}\,|\,y=1 \sim \mathcal{N}^{10}(\mu_{\mathbf{1}}, \Sigma_1)$$

Since the class-conditional distributions are normal with distinct covariance matrices, the optimal decision boundary is a quadratic function [13]. Again we used a simple neural network as the classifier in the simulation study. We don't discuss the details of neural networks here but simply state that it may be used to estimate a classification function from the training set. Since neural networks tend to have substantial variance, one would expect that averaging through either bagging or arcing will reduce overall prediction error through variance reduction.

The results of the simulation study are summarized in table 14. Column 1 contains the average error rate (denoted by $\bar{e}$) of the three methods on the test set. We conclude from this column that both arcing and bagging result in a substantial reduction of the error rate compared to the "conventional" unaggregated classifier. Furthermore arcing seems to have a slight edge over bagging. These conclusions are in line with more elaborate studies to be found in [3]. Column 2 contains the Bayes error rate on the test set which we were only able to calculate because we specified the data generating mechanism ourselves. In any practical data analysis setting the Bayes error rate is of course unknown, although methods have been suggested for its estimation (see [16]). The third and fourth column contain the decomposition of reducible error into its bias and variance components, using the additive decomposition of Breiman (see section 5.2). We see that both aggregation methods lead to a substantial variance reduction. Furthermore, arcing leads to a reduction in bias as well, but bagging leads to some increase in bias in this study. More elaborate studies suggest that both arcing and bagging tend to reduce bias somewhat, but the major improvement in overall performance is caused by a substantial reduction in variance.

The last three columns of table 14 contain the decomposition of mean square estimation error into its bias and variance components. This decompostition provides some additional insight. Again both aggregation methods lead to an improvement of overall mean square estimation error. Furthermore we see that this improvement is entirely due to a substantial decrease in the variance component of error. Note that arcing leads to some increase in squared bias but to a decrease in bias according to Breiman's definition. This suggest that the bias induced by arcing is of the "harmless" kind, i.e. negative boundary bias in Friedman's terminology. Since boundary bias is predominantly negative, the strong reduction in variance (in the mean squared error sense) leads to a substantial reduction of the error rate.

7 Summary

Probability theory is the primary tool of statistical inference. In the probability calculus we reason deductively from known population to possible samples and their respective probabilities under a particular sampling model (we only discussed simple random sampling). In statistical inference we reason inductively from the observed sample to unknown (usually: up to a fixed number of parameters) population. The three main paradigms of statistical inference are frequentist inference, likelihood inference and Bayesian inference. Simplifying matters somewhat, one may say that frequentist inference is based on the sampling distribution of a sample statistic, likelihood inference is based on the likelihood function, and Bayesian inference on the posterior distribution of the parameter(s) of interest.

One of the goals of statistical data analysis is to estimate functions $y = f(\mathbf{x})$ from sample data, in order to predict realisations of random variable Y when $\mathbf{x}$ is known. We may evaluate methods to estimate such functions by comparing their expected prediction error under repeated sampling from some fixed population (distinct samples gives rise to a different estimates $\hat{f} \mid T$). Decomposition of reducible prediction error into its bias and variance components explains to a large extent the relative performance of different methods (e.g. linear regression, neural networks and regression trees) on different problems. Well-known phenomena such as overfitting and variance reduction techniques such as model averaging (e.g. bagging and arcing) follow naturally from the decomposition into bias and variance.

With the increased computer power, computationally expensive inference methods such as bootstrapping are becoming widely used in practical data analysis. From a frequentist perspective, the bootstrap constructs an empirical sampling distribution of the statistic of interest by substituting "sample" for "population" and "resample" for "sample", and mimicing the repeated sampling process on the computer. In many cases this frees the analyst from "heroic" assumptions required to obtain analytical results.

We have attempted to provide the reader with an overview of statistical concepts particularly relevant to intelligent data analysis. In doing so, we emphasized and illustrated the basic ideas, rather than to provide a mathematically rigorous in-depth treatment. Furthermore, this chapter serves as a basis for the more advanced chapters of this volume that give comprehensive overviews of topics only touched upon here.

References

1. D.A. Berry. *Statistics: a Bayesian perspective.* Wadsworth, Belmont (CA), 1996.
2. L. Breiman. Bagging predictors. *Machine Learning*, 26(2):123–140, 1996.
3. L Breiman. Bias, variance, and arcing classifiers. Technical Report 460, University of California, 1996.
4. A.C. Davison and D.V. Hinkley. *Bootstrap methods and their application.* Cambridge University Press, Cambridge, 1997.
5. A.W.F Edwards. *Likelihood.* The John Hopkins University Press, Baltimore, 1992.
6. B. Efron and R.J Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, New York, 1993.
7. J.H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19:1–141, 1991.
8. J.H. Friedman. On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77, 1997.
9. A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis.* Chapman & Hall, London, 1995.
10. W.H. Greene. *Econometric Analysis (second edition).* Macmillan, New York, 1993.
11. D.J. Hand. *Construction and assessment of classification rules.* John Wiley, Chichester, 1997.
12. R.C. Holte. Very simple classification rules perform well on most commonly used data sets. *Machine Learning*, 11:63–90, 1993.
13. G.J. McLachlan. *Discriminant Analysis and Statistical Pattern Recognition.* John Wiley, New York, 1992.
14. T.M. Mitchell. *Machine Learning.* McGraw-Hill, Singapore, 1997.
15. D.S. Moore. Bayes for beginners: some reasons to hesitate. *The American Statistician*, 51(3):254–261, 1997.
16. B.D. Ripley. *Pattern Recognition and Neural Networks.* Cambridge University Press, Cambridge, 1996.
17. R.M. Royall. *Statistical evidence: a likelihood paradigm.* Chapman & Hall, London, 1997.