1

Why probability and statistics?

Is everything on this planet determined by randomness? This question is open to philosophical debate. What is certain is that every day thousands and thousands of engineers, scientists, business persons, manufacturers, and others are using tools from probability and statistics.

The theory and practice of probability and statistics were developed during the last century and are still actively being refined and extended. In this book we will introduce the basic notions and ideas, and in this first chapter we present a diverse collection of examples where randomness plays a role.

1.1 Biometry: iris recognition

Biometry is the art of identifying a person on the basis of his or her personal biological characteristics, such as fingerprints or voice. From recent research it appears that with the human iris one can beat all existing automatic human identification systems. Iris recognition technology is based on the visible qualities of the iris. It converts these—via a video camera—into an "iris code" consisting of just 2048 bits. This is done in such a way that the code is hardly sensitive to the size of the iris or the size of the pupil. However, at different times and different places the iris code of the same person will not be exactly the same. Thus one has to allow for a certain percentage of mismatching bits when identifying a person. In fact, the system allows about 34% mismatches! How can this lead to a reliable identification system? The miracle is that different persons have very different irides. In particular, over a large collection of different irides the code bits take the values 0 and 1 about half of the time. But that is certainly not sufficient: if one bit would determine the other 2047, then we could only distinguish two persons. In other words, single bits may be random, but the correlation between bits is also crucial (we will discuss correlation at length in Chapter 10). John Daugman who has developed the iris recognition technology made comparisons between 222 743 pairs of iris codes and concluded that of the 2048 bits 266 may be considered as uncorrelated ([6]). He then argues that we may consider an iris code as the result of 266 coin tosses with a fair coin. This implies that if we compare two such codes from different persons, then there is an astronomically small probability that these two differ in less than 34% of the bits—almost all pairs will differ in about 50% of the bits. This is illustrated in Figure 1.1, which originates from [6], and was kindly provided by John Daugman. The iris code data consist of numbers between 0 and 1, each a Hamming distance (the fraction of mismatches) between two iris codes. The data have been summarized in two histograms, that is, two graphs that show the number of counts of Hamming distances falling in a certain interval. We will encounter histograms and other summaries of data in Chapter 15. One sees from the figure that for codes from the same iris (left side) the mismatch fraction is only about 0.09, while for different irides (right side) it is about 0.46.

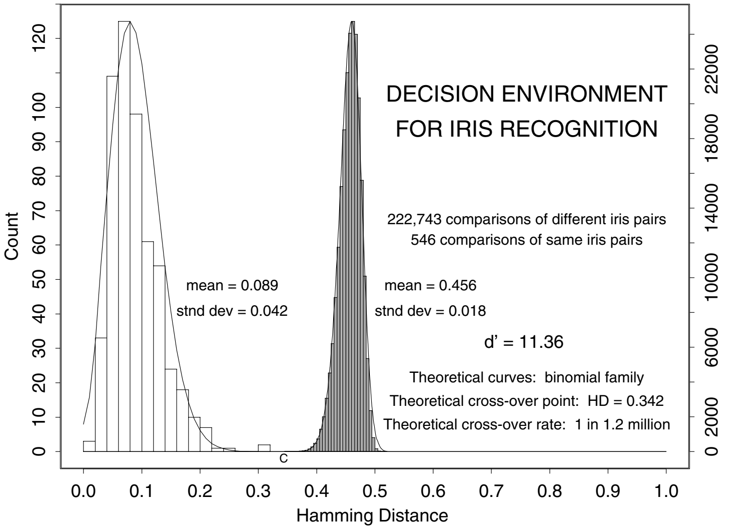


Fig. 1.1. Comparison of same and different iris pairs.

Source: J.Daugman. *Second IMA Conference on Image Processing: Mathematical Methods, Algorithms and Applications*, 2000. © Ellis Horwood Publishing Limited.

You may still wonder how it is possible that irides distinguish people so well. What about twins, for instance? The surprising thing is that although the color of eyes is hereditary, many features of iris patterns seem to be produced by so-called epigenetic events. This means that during embryo development the iris structure develops randomly. In particular, the iris patterns of (monozygotic) twins are as discrepant as those of two arbitrary individuals.

For this reason, as early as in the 1930s, eye specialists proposed that iris patterns might be used for identification purposes.

1.2 Killer football

A couple of years ago the prestigious *British Medical Journal* published a paper with the title “Cardiovascular mortality in Dutch men during 1996 European football championship: longitudinal population study” ([41]). The authors claim to have shown that the effect of a single football match is detectable in national mortality data. They consider the mortality from infarctions (heart attacks) and strokes, and the “explanation” of the increase is a combination of heavy alcohol consumption and stress caused by watching the football match on June 22 between the Netherlands and France (lost by the Dutch team!). The authors mainly support their claim with a figure like Figure 1.2, which shows the number of deaths from the causes mentioned (for men over 45), during the period June 17 to June 27, 1996. The middle horizontal line marks the average number of deaths on these days, and the upper and lower horizontal lines mark what the authors call the 95% confidence interval. The construction of such an interval is usually performed with standard statistical techniques, which you will learn in Chapter 23. The interpretation of such an interval is rather tricky. That the bar on June 22 sticks out off the confidence interval should support the “killer claim.”

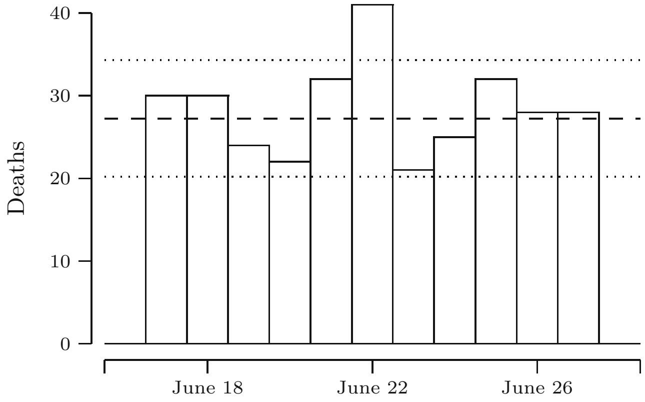


Fig. 1.2. Number of deaths from infarction or stroke in (part of) June 1996.

It is rather surprising that such a conclusion is based on a *single* football match, and one could wonder why no probability model is proposed in the paper. In fact, as we shall see in Chapter 12, it would not be a bad idea to model the time points at which deaths occur as a so-called Poisson process.

Once we have done this, we can compute how often a pattern like the one in the figure might occur—without paying attention to football matches and other high-risk national events. To do this we need the mean number of deaths per day. This number can be obtained from the data by an estimation procedure (the subject of Chapters 19 to 23). We use the sample mean, which is equal to $(10 \cdot 27.2 + 41)/11 = 313/11 = 28.45$. (Here we have to make a computation like this because we only use the data in the paper: 27.2 is the average over the 5 days preceding and following the match, and 41 is the number of deaths on the day of the match.) Now let p_{high} be the probability that there are 41 or more deaths on a day, and let p_{usual} be the probability that there are between 21 and 34 deaths on a day—here 21 and 34 are the lowest and the highest number that fall in the interval in Figure 1.2. From the formula of the Poisson distribution given in Chapter 12 one can compute that $p_{\text{high}} = 0.008$ and $p_{\text{usual}} = 0.820$. Since events on different days are independent according to the Poisson process model, the probability p of a pattern as in the figure is

$$p = p_{\text{usual}}^5 \cdot p_{\text{high}} \cdot p_{\text{usual}}^5 = 0.0011.$$

From this it can be shown by (a generalization of) the law of large numbers (which we will study in Chapter 13) that such a pattern would appear about once every $1/0.0011 = 899$ days. So it is not overwhelmingly exceptional to find such a pattern, and the fact that there was an important football match on the day in the middle of the pattern might just have been a coincidence.

1.3 Cars and goats: the Monty Hall dilemma

On Sunday September 9, 1990, the following question appeared in the "Ask Marilyn" column in *Parade*, a Sunday supplement to many newspapers across the United States:

> Suppose you're on a game show, and you're given the choice of three doors; behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?—Craig F. Whitaker, Columbia, Md.

Marilyn's answer—one should switch—caused an avalanche of reactions, in total an estimated 10 000. Some of these reactions were not so flattering ("You are the goat"), quite a lot were by professional mathematicians ("You blew it, and blew it big," "You are utterly incorrect .... How many irate mathematicians are needed to change your mind?"). Perhaps some of the reactions were so strong, because Marilyn vos Savant, the author of the column, is in the *Guinness Book of Records* for having one of the highest IQs in the world.

The switching question was inspired by Monty Hall's "Let's Make a Deal" game show, which ran with small interruptions for 23 years on various U.S. television networks.

Although it is not explicitly stated in the question, the game show host will *always* open a door with a goat after you make your initial choice. Many people would argue that in this situation it does not matter whether one would change or not: one door has a car behind it, the other a goat, so the odds to get the car are fifty-fifty. To see why they are wrong, consider the following argument. In the original situation two of the three doors have a goat behind them, so with probability 2/3 your initial choice was wrong, and with probability 1/3 it was right. Now the host opens a door with a goat (note that he can always do this). In case your initial choice was *wrong* the host has only one option to show a door with a goat, and switching leads you to the door with the car. In case your initial choice was *right* the host has two goats to choose from, so switching will lead you to a goat. We see that switching is the best strategy, doubling our chances to win. To stress this argument, consider the following generalization of the problem: suppose there are 10 000 doors, behind one is a car and behind the rest, goats. After you make your choice, the host will open 9998 doors with goats, and offers you the option to switch. To change or not to change, that's the question! Still not convinced? Use your Internet browser to find one of the zillion sites where one can run a simulation of the Monty Hall problem (more about simulation in Chapter 6).

In fact, there are quite a lot of variations on the problem. For example, the situation that there are four doors: you select a door, the host always opens a door with a goat, and offers you to select another door. After you have made up your mind he opens a door with a goat, and again offers you to switch. After you have decided, he opens the door you selected. What is now the best strategy? In this situation switching only at the last possible moment yields a probability of 3/4 to bring the car home. Using the law of total probability from Section 3.3 you will find that this is indeed the best possible strategy.

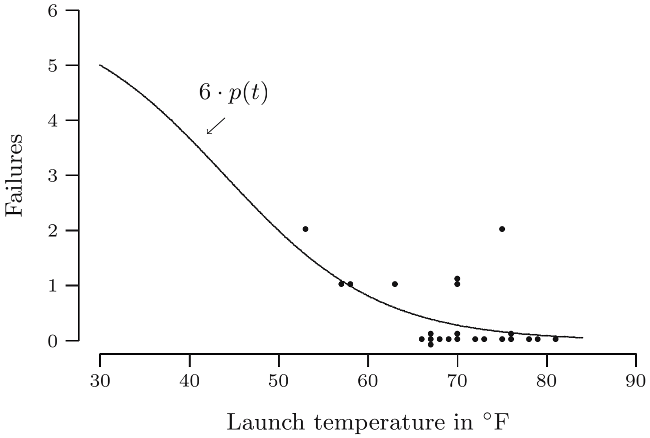
1.4 The space shuttle *Challenger*

On January 28, 1986, the space shuttle *Challenger* exploded about one minute after it had taken off from the launch pad at Kennedy Space Center in Florida. The seven astronauts on board were killed and the spacecraft was destroyed. The cause of the disaster was explosion of the main fuel tank, caused by flames of hot gas erupting from one of the so-called solid rocket boosters.

These solid rocket boosters had been cause for concern since the early years of the shuttle. They are manufactured in segments, which are joined at a later stage, resulting in a number of joints that are sealed to protect against leakage. This is done with so-called O-rings, which in turn are protected by a layer of putty. When the rocket motor ignites, high pressure and high temperature build up within. In time these may burn away the putty and subsequently erode the O-rings, eventually causing hot flames to erupt on the outside. In a nutshell, this is what actually happened to the *Challenger*.

After the explosion, an investigative commission determined the causes of the disaster, and a report was issued with many findings and recommendations ([24]). On the evening of January 27, a decision to launch the next day had been made, notwithstanding the fact that an extremely low temperature of 31°F had been predicted, well below the operating limit of 40°F set by Morton Thiokol, the manufacturer of the solid rocket boosters. Apparently, a "management decision" was made to overrule the engineers' recommendation not to launch. The inquiry faulted both NASA and Morton Thiokol management for giving in to the pressure to launch, ignoring warnings about problems with the seals.

The *Challenger* launch was the 24th of the space shuttle program, and we shall look at the data on the number of failed O-rings, available from previous launches (see [5] for more details). Each rocket has three O-rings, and two rocket boosters are used per launch, so in total six O-rings are used each time. Because low temperatures are known to adversely affect the O-rings, we also look at the corresponding launch temperature. In Figure 1.3 the dots show the number of failed O-rings per mission (there are 23 dots—one time the boosters could not be recovered from the ocean; temperatures are rounded to the nearest degree Fahrenheit; in case of two or more equal data points these are shifted slightly.). If you ignore the dots representing zero failures, which all occurred at high temperatures, a temperature effect is not apparent.



Source: based on data from Volume VI of the Report of the Presidential Commission on the space shuttle Challenger accident, Washington, DC, 1986.

Fig. 1.3. Space shuttle failure data of pre-*Challenger* missions and fitted model of expected number of failures per mission function.

In a model to describe these data, the probability $p(t)$ that an individual O-ring fails should depend on the launch temperature t. Per mission, the number of failed O-rings follows a so-called binomial distribution: six O-rings, and each may fail with probability $p(t)$; more about this distribution and the circumstances under which it arises can be found in Chapter 4. A *logistic* model was used in [5] to describe the dependence on t:

$$p(t) = \frac{\mathrm{e}^{a+b\cdot t}}{1 + \mathrm{e}^{a+b\cdot t}}.$$

A high value of $a + b \cdot t$ corresponds to a high value of $p(t)$, a low value to low $p(t)$. Values of a and b were determined from the data, according to the following principle: choose a and b so that the probability that we get data as in Figure 1.3 is as high as possible. This is an example of the use of the method of maximum likelihood, which we shall discuss in Chapter 21. This results in $a = 5.085$ and $b = -0.1156$, which indeed leads to lower probabilities at higher temperatures, and to $p(31) = 0.8178$. We can also compute the (estimated) expected number of failures, $6 \cdot p(t)$, as a function of the launch temperature t; this is the plotted line in the figure.

Combining the estimates with estimated probabilities of other events that should happen for a *complete* failure of the field-joint, the estimated probability of such a failure is 0.023. With six field-joints, the probability of at least one complete failure is then $1 - (1 - 0.023)^6 = 0.13$!

1.5 Statistics versus intelligence agencies

During World War II, information about Germany's war potential was essential to the Allied forces in order to schedule the time of invasions and to carry out the allied strategic bombing program. Methods for estimating German production used during the early phases of the war proved to be inadequate. In order to obtain more reliable estimates of German war production, experts from the Economic Warfare Division of the American Embassy and the British Ministry of Economic Warfare started to analyze markings and serial numbers obtained from captured German equipment.

Each piece of enemy equipment was labeled with markings, which included all or some portion of the following information: (a) the name and location of the marker; (b) the date of manufacture; (c) a serial number; and (d) miscellaneous markings such as trademarks, mold numbers, casting numbers, etc. The purpose of these markings was to maintain an effective check on production standards and to perform spare parts control. However, these same markings offered Allied intelligence a wealth of information about German industry.

The first products to be analyzed were tires taken from German aircraft shot over Britain and from supply dumps of aircraft and motor vehicle tires captured in North Africa. The marking on each tire contained the maker's name, a serial number, and a two-letter code for the date of manufacture. The first step in analyzing the tire markings involved breaking the two-letter date code. It was conjectured that one letter represented the month and the other the year of manufacture, and that there should be 12 letter variations for the month code and 3 to 6 for the year code. This, indeed, turned out to be true. The following table presents examples of the 12 letter variations used by four different manufacturers.

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Dunlop	T	I	E	B	R	A	P	O	L	N	U	D
Fulda	F	U	L	D	A	M	U	N	S	T	E	R
Phoenix	F	O	N	I	X	H	A	M	B	U	R	G
Sempirit	A	B	C	D	E	F	G	H	I	J	K	L

Reprinted with permission from "An empirical approach to economic intelligence" by R.Ruggles and H.Brodie, pp.72-91, Vol. 42, No. 237. © 1947 by the American Statistical Association. All rights reserved.

For instance, the Dunlop code was Dunlop Arbeit spelled backwards. Next, the year code was broken and the numbering system was solved so that for each manufacturer individually the serial numbers could be dated. Moreover, for each month, the serial numbers could be recoded to numbers running from 1 to some unknown largest number N, and the observed (recoded) serial numbers could be seen as a subset of this. The objective was to estimate N for each month and each manufacturer separately by means of the observed (recoded) serial numbers. In Chapter 20 we discuss two different methods of estimation, and we show that the method based on only the maximum observed (recoded) serial number is much better than the method based on the average observed (recoded) serial numbers.

With a sample of about 1400 tires from five producers, individual monthly output figures were obtained for almost all months over a period from 1939 to mid-1943. The following table compares the accuracy of estimates of the average monthly production of all manufacturers of the first quarter of 1943 with the statistics of the Speer Ministry that became available after the war. The accuracy of the estimates can be appreciated even more if we compare them with the figures obtained by Allied intelligence agencies. They estimated, using other methods, the production between 900 000 and 1 200 000 per month!

Type of tire	Estimated production	Actual production
Truck and passenger car	147 000	159 000
Aircraft	28 500	26 400
Total	175 500	186 100

Reprinted with permission from "An empirical approach to economic intelligence" by R.Ruggles and H.Brodie, pp.72-91, Vol. 42, No. 237. © 1947 by the American Statistical Association. All rights reserved.

1.6 The speed of light

In 1983 the definition of the meter (the SI unit of one meter) was changed to: *The meter is the length of the path traveled by light in vacuum during a time interval of* 1/299 792 458 *of a second.* This implicitly defines the speed of light as 299 792 458 meters per second. It was done because one thought that the speed of light was so accurately known that it made more sense to define the meter in terms of the speed of light rather than vice versa, a remarkable end to a long story of scientific discovery. For a long time most scientists believed that the speed of light was infinite. Early experiments devised to demonstrate the finiteness of the speed of light failed because the speed is so extraordinarily high. In the 18th century this debate was settled, and work started on determination of the speed, using astronomical observations, but a century later scientists turned to earth-based experiments. Albert Michelson refined experimental arrangements from two previous experiments and conducted a series of measurements in June and early July of 1879, at the U.S. Naval Academy in Annapolis. In this section we give a very short summary of his work. It is extracted from an article in *Statistical Science* ([18]).

The principle of speed measurement is easy, of course: measure a distance and the time it takes to travel that distance, the speed equals distance divided by time. For an accurate determination, both the distance and the time need to be measured accurately, and with the speed of light this is a problem: either we should use a very large distance and the accuracy of the distance measurement is a problem, or we have a very short time interval, which is also very difficult to measure accurately.

In Michelson's time it was known that the speed of light was about 300 000 km/s, and he embarked on his study with the goal of an improved value of the speed of light. His experimental setup is depicted schematically in Figure 1.4. Light emitted from a light source is aimed, through a slit in a fixed plate, at a rotating mirror; we call its distance from the plate the radius. At one particular angle, this rotating mirror reflects the beam in the direction of a distant (fixed) flat mirror. On its way the light first passes through a focusing lens. This second mirror is positioned in such a way that it reflects the beam back in the direction of the rotating mirror. In the time it takes the light to travel back and forth between the two mirrors, the rotating mirror has moved by an angle α, resulting in a reflection on the plate that is displaced with respect to the source beam that passed through the slit. The radius and the displacement determine the angle α because

$$\tan 2\alpha = \frac{\text{displacement}}{\text{radius}}$$

and combined with the number of revolutions per seconds (rps) of the mirror, this determines the elapsed time:

$$\text{time} = \frac{\alpha/2\pi}{\text{rps}}.$$

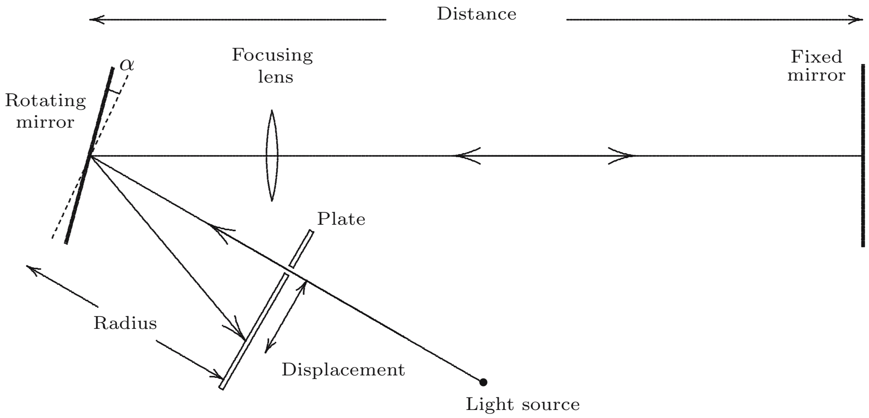


Fig. 1.4. Michelson's experiment.

During this time the light traveled twice the distance between the mirrors, so the speed of light in air now follows:

$$c_{\text{air}} = \frac{2 \cdot \text{distance}}{\text{time}}.$$

All in all, it looks simple: just measure the four quantities—distance, radius, displacement and the revolutions per second—and do the calculations. This is much harder than it looks, and problems in the form of inaccuracies are lurking everywhere. An error in any of these quantities translates directly into some error in the final result.

Michelson did the utmost to reduce errors. For example, the distance between the mirrors was about 2000 feet, and to measure it he used a steel measuring tape. Its nominal length was 100 feet, but he carefully checked this using a copy of the official "standard yard." He found that the tape was in fact 100.006 feet. This way he eliminated a (small) systematic error.

Now imagine using the tape to measure a distance of 2000 feet: you have to use the tape 20 times, each time marking the next 100 feet. Do it again, and you probably find a slightly different answer, no matter how hard you try to be very precise in every step of the measuring procedure. This kind of variation is inevitable: sometimes we end up with a value that is a bit too high, other times it is too low, but on average we're doing okay—assuming that we have eliminated sources of systematic error, as in the measuring tape. Michelson measured the distance five times, which resulted in values between 1984.93 and 1985.17 feet (after correcting for the temperature-dependent stretch), and he used the average as the "true distance."

In many phases of the measuring process Michelson attempted to identify and determine systematic errors and subsequently applied corrections. He also systematically repeated measuring steps and averaged the results to reduce variability. His final dataset consists of 100 separate measurements (see Table 17.1), but each is in fact summarized and averaged from repeated measurements on several variables. The final result he reported was that the speed of light in vacuum (this involved a conversion) was 299 944 ± 51 km/s, where the 51 is an indication of the uncertainty in the answer. In retrospect, we must conclude that, in spite of Michelson's admirable meticulousness, some source of error must have slipped his attention, as his result is off by about 150 km/s. With current methods we would derive from his data a so-called 95% confidence interval: 299 944 ± 15.5 km/s, suggesting that Michelson's uncertainty analysis was a little conservative. The methods used to construct confidence intervals are the topic of Chapters 23 and 24.