Chapter 1

Making sense out of figures

Figures do not lie, but liars figure.

Statistics is the methodology which scientists and mathematicians have developed for interpreting and drawing conclusions from data. Statistical methods are indispensable for investigations in many fields of knowledge, such as biological and social sciences, business, education, medicine, engineering, law, etc. Information on a research problem are collected in a form of measurements (numbers). An analysis of such data is necessary in order to obtain a better understanding of the phenomenon of interest. In this introductory chapter we give several interesting examples of the use of probability and statistics in real life. In many of them our intuition will not be of much help.

1.1 Examples

Simpson's Paradox. Graduate School Admission by Sex, Super-Duper University.

	No. Applied	No. admitted	% admitted
Department A			
Women	50	25	50%
Men	100	50	50%
Department B			
Women	100	30	30%
Men	50	15	30%
Total			
Women	150	55	37%
Men	150	65	43%

A fair gambling with a possible loaded coin. Pretend you have a coin for which you do not know probability of falling up heads. You suspect that the coin is loaded and that probability of heads is bigger than that of tails. Can you play a fair game by flipping the coin?

Yes. Flip the coin twice, ignore TT, HH outcomes and declare "heads" if you see HT and "tails" if you see TH. The probabilities of these two outcomes are identical.

Cryptographic Survey.[1] It is difficult to conduct sample surveys on sensitive issues because many people will not answer questions if the answers might embarrass them.

To ask a sample of tax payers, whether they have ever filed a fraudulent tax return, we ask each person to toss a fair coin once.

He/she is to answer **Yes** if either (i) the coin lands *heads* or (ii) the coin lands tails and he/she has ever filled a fraudulent return.

He/she is to answer **No** otherwise, i.e. if the coin lands tails and he/she has never filled a fraudulent report.

Expected proportion of **yes** *answers is* $\frac{1}{2}+\frac{1}{2}p$. *For instance, if we asked 100 tax payers, and record the proportion of yes answers 0.6, the estimate of* p *is 0.2, or 20 %.*

[1] I learn this from Peter Mueller

St. Petersburg Paradox. Pretend that you are invited to participate in the following game. First, you pay an entry fee of x dollars and then the game goes as follows. A fair coin is flipped until the first time "heads up." If this is at the kth flip, then you receive 2^k dollars. This ends the game. The question arose how much to pay for participation in this gamble. It has been observed that gamblers were not willing to pay more than 2 to 4 dollars to participate in such a gamble. However, the mathematical expectation of the gain of the gamble is infinite. In plain words, that means that one should be willing to pay as much as he is asked for (or he is able to). The only assumption is that the casino in which the game is played has an infinite capital.

Hence the paradox between the mathematical expectation of the gain and the observed willingness to pay. However, if the casino has \$ 10^{10} (10 billions) of capital, than the fair entry fee would be about 30 dollars.

Paradoxes de-Mere. In 1654 the Chevalier de Mere asked Blaise Pascal (1623-1662) the following two questions:

(i) Why it would be advantageous in a game of dice to bet on occurrence of a six in four trials but not advantageous in a game involving two dice to bet on the occurrence of a double six in twenty four trials?

The probabilities are 0.518 and 0.491, respectively.

(ii) In playing a game with three dice why the sum 11 is advantageous to sum 12 when both are results of six possible outcomes.
11: (1, 4, 6), (1, 5, 5), (2, 3, 6), (2, 4, 5), (3, 3, 5), (3, 4, 4).
12: (1, 5, 6), (2, 4, 6), (2, 5, 5), (3, 3, 6), (3, 4, 5), (4, 4, 4).

Taking into account all different permutations of the above triples the sum 11 has 27 favorable permutations while the sum 12 has 25 favorable permutation.

Deciding Authorship. The *Federalist Papers* played an important role in the history of the United States. Written in 1787 - 1788 by Alexander Hamilton, John Jay, and James Madison, the purpose of these 77 newspaper essays was to persuade the citizens of the State of New York to ratify the newly written Constitution of our emerging nation. These essays were signed with the pen name Publius, and published as a book which also contains eight essays by Hamilton. The question of whether Hamilton or Madison wrote twelve of the *Publius* essays has been a matter of dispute. Standard methods of historical research have not settled the problem.

Mosteller and Wallace (1964) applied statistical methods towards its solution. The problem is a difficult one, for Hamilton and Madison used the same style, standard phrases, and sentence structure, which were characteristic of most educated Americans of their time; for example, their average sentence lengths in the undisputed papers were 34.5 and 34.6 words, respectively. However, there were some subtle differences of style. Scholars had noticed, for instance, that Hamilton tended to use **while,** and Madison, **whilst.** This pair of "marker" words, however, was not decisive because in some papers neither word appeared. After painstaking analysis Mosteller and Wallace found 30 words differed substantially in the frequency of use by the two authors.

Table below gives the rate of use of nine words, by the two authors: the rate 3.24 is obtained by taking the number of times **upon** was used by Hamilton divided by the number of 1000's of words in the Hamilton essays. Note that Madison used **upon** infrequently but used **on** more frequently than Hamilton.

The frequency of occurrence of the 30 words were combined into an index in such a way that the index was large for papers known to have been written by Hamilton and small for Madison papers. In fact, the score ranged from 0.3120 to 1.3856 for Hamilton papers and from -0.8627 to 0.1462 for Madison papers. Except for one paper, the scores of the disputed papers went from -0.7557 to -0.0145. Thus, these disputed papers were assigned to Madison. The paper with a score of 0.3161 was also assigned to Madison on the basis of further investigation and with less assurance.

	Frequency Per 1000 Words	
Word	Hamilton	Madison
Upon	3.24	0.23
Also	0.32	0.67
An	5.95	4.58
By	7.32	11.43
Of	64.51	57.89
On	3.38	7.75
There	3.20	1.33
This	7.77	6.00
To	40.79	35.21

Place your bet if you enjoy gambling. Do not expect to win. The data supplied by *Quarterly Report, Dec 1987. South Dakota Lottery* give among several income-statement entries the following two:

Ticket Sales	\$ 11,812,905
Prize Payments	\$ 5,322,975

For a player, that means that only 45 cents is expected in return on each dollar invested in the lottery tickets. Why then one should play such an unfavorable game?

The Gambler, the Mathematician and the Statistician.[2]

After flipping a fair coin 10 times an unusual outcome was recorded: ten heads in a row! Three players, the gambler, the mathematician and the statistician were asked to place their bets on the 11th flip. Here are their rationales:

Gambler: I'll put my bet on tails. After 10 heads in a row tails have larger probability to show up.

Mathematician: The coin is a piece of metal without memory. If it is fair, probabilities of heads and tails in 11th flip are the same. I do not have any preferences.

Statistician: Did you say 10 heads in a row? I am placing my bet on heads; the coin may be loaded.

A need for randomness In 1936 the Literary Digest took the largest poll ever taken: they had 2.4M respondents to their opinion poll about the Franklin Roosevelt/Alf Landon election that year, of which about 57% indicated a preference for Landon. With such a large sample size the sampling error becomes very small, 0.0006. That means, the preference for Landon in the population of US voters is predicted to be $57 \pm 0.06\%$. In fact 38% voted for Landon. What went wrong? The Literary Digest failed to choose a random sample. Insetead, they used a "sample of convenience" consistiong of those who responded to their questionare from among the telephone and magazine subscribers whose addresses they could find. Randomizing is not just something that statisticians like to complain about - it is something that, if ignored, can lead to wrong answers to scientific questions.

1.2 Exercises

1. Two competing hospitals, A and B, have released the data given in the two tables below. The symbols +/– correspond to the condition of incoming patient (Good/Bad). Symbols S and U correspond to the condition of the patient after the treatment (Satisfactory/Unsatisfactory).

Hospital A.

	+	–	Total
S	41	39	80
U	5	10	15
Total	46	49	95
%	89.13%	79.59%	84.21%

Hospital B.

	+	–	Total
S	32	11	43
U	4	3	7
Total	36	14	50
%	88.89%	78.57%	86%

(a) If you were a patient in good condition, what hospital would you prefer?
(b) If you are the manager of the hospital B, what percentages you would put in a local newspaper add?

2. Devise your own strategy for asking your class if they ever cheated on any exam while at Duke. Use a fair die instead of a coin.

[2] Do not relate to *The Good, the Bad, and the Ugly.*

3. (a) A ticket for the important game of your favorite Durham Bulls team (vs. Salem Buccaneers) costs \$ 10 and you have only \$ 5. You really want to see the game. Would you play the game in which you may double your capital with the probability 1/3 and loose it with the probability of 2/3.
(b) What if you are offered to play the same game and the bet is \$ 100.
(c) What is your choice between two options: getting \$ 1,000,000 with the probability 1 (sure event), or getting \$ 2,100,000 with the probability 1/2.

4. Devise how would you simulate (a) a fair coin toss with a possibly loaded die, (b) a fair die roll with a possibly loaded coin.

5. Let's Make a Deal. In the popular television game show *Let's Make a Deal,* Monty Hall is the master of ceremonies. At certain times during the show, a contestant is allowed to choose one of three identical doors A, B, C, behind only one of which is a valuable prize (a new car). After the contestant picks a door (say, door A), Monty Hall opens another door and shows the contestant that there is no prize behind that door. (Monty Hall knows where the prize is and always chooses a door where there is no prize.) He then asks the contestant whether he or she wants to stick with their choice of door or switch to the remaining unopened door. Should the contestant switch doors? Does it matter?