Teaching Independence

KEYWORDS:
DEPENDENCE.

Henrik Dahl
Agder College, Kristiansand, Norway

Summary
Examples are given which illlustrate how independence enters into statisical problems and a demonstration of independence is presented. Coverage in statistical textbooks is examined

◆ INTRODUCTION◆

THE CONCEPT of independence has had a special role in probability and statistics throughout history. Kolmogorov (1950) in his famous book states that independence is what makes probability differ from measure theory:

> "Historically, the independence of experiments and random variables represents the very mathematical concept that has given the theory of probability its peculiar stamp".

Although Kolmogorov does not discuss how to implement independence in practice, he states the problem:

> "In consequence, one of the most important problems in the philosophy of the natural sciences is - in addition to the well known one regarding the essence of the concept of probability - to make precise the premises which would make it possible to regard any given real events as independent. This question, however, is beyond the scope of this book".

Von Mises (1957) tries to understand statistical independence by means of his "kollektives". He finds the well known definition of independence of events unsatisfactory, but is satisfied with his "kollektive" version of what is known as "product models".

Kac (1959) has the following comment about independence:

> "This notion originated in probability theory and for a long time was handled with vagueness which bred suspicion as to its being a bona fide mathematical notion."

Kac shows that independence exists as a mathematical object, at least in pure mathematics. However, the implementation of independence in practical situations is not treated by Kac.

◆ PROBLEMS IN TEACHING◆ INDEPENDENCE

Considering what has been stated above, it is not surprising that students have difficulty in understanding the meaning of independence. To teach this subject is a challenge. Students with some mathematical background, namely vector analysis, may benefit from connecting independence of random variables to orthogonality. If we view independent random variables as orthogonal vectors, their independence means that the vectors will project no randomness on each other. This connection between independence and orthogonality highlights independent experiments.

In elementary courses in probability and statistics students are confronted with the assumption that data are considered to be realisations of independent random variables. I will consider the problem of conveying to the students the meaning of this notion of independence of random variables in a way that, hopefully, makes them see the difference between situations where this assumption is reasonable and where it is not.

The formal definition of independence of the random variables $X_1, X_2,..., X$:

$$P(X_1 = x_1 \cap X_2 = x_2 \cap .... \cap X_n = x_n)$$
$$= P(X_1 = x_1).P(X_2 = x_2)...P(X_n = x_n)$$

for all possible $x_1, x_2, ..., x_n$

says little that can be useful to a newcomer to our subject in telling whether the data really have this structure or not.

Many students will not understand such formal definitions. However, these people may later in life get into positions where they have to consider statistical investigations in which the assumption of independence may be violated. To give such people some training in the difference between correct and incor-

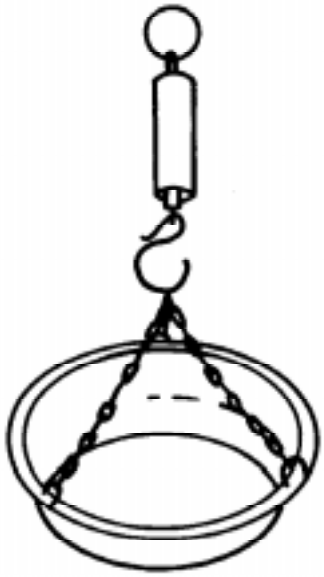
rect usage of the assumption of independence, I prefer to use examples.

◆EXAMPLES OF DEPENDENCE ◆ AND INDEPENDENCE

When using data to draw conclusions we need assurace that new data are really new "fresh information" not contaminated by data we have already collected. If this is not the case, new data may not represent new information.

Example 1

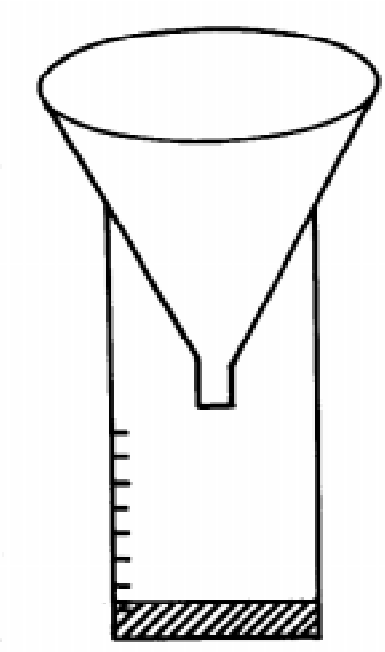
Pupils are given an assignment to weigh a stone several times using a spring balance. During the process they are supposed to find out how reliable the spring balance is for weighing. The pupils only make one weighing of the stone. Let this weight be X (grams). To make the teacher believe that they have completed the assignment they also claim that they obtained the weights:



X+l, X +2, ...X+20,
X—1, X—2, ... X—20.
To trick the teacher into believing that everything is alright they write the 41 "weights" in haphazard order. Whatever statistical analysis is used to investigate the reliability of the weighing process, with this data set it will be useless. The data set has nothing to say about the reliability of the weighing process. This example shows that we have to make some sort of assumption like independence. Without limitations on the contamination between observations, they cannot be used for inference. One of the basic ideas of statistics is that we can get more reliable statistical conclusions by using more data. The square root law says that to increase the precision of a statistical investigation by a certain factor, we have to increase the number of observations by the square of the factor. The square root law presupposes independence of the observations.

Example 2

A questionnaire is used to collect data from four departments of a firm. If the people who fill in the form confer before they complete the questionnaire, dependence will creep in. The data set will then not contain four pieces of information but, perhaps, only one piece of information (like Example 1).



Example 3

Rainfall is measured in Kristiansand on 12, 13, 14, 15, 16, 17, 18, 19, and 20 February 1992. These data can not be considered independent, because rainy or dry weather has a tendency to persist for some days.

Example 4

Rainfall is measured in Kristiansand on 12.2.1992, 12.2.1993, 12.2.1994 and12.2.1995. These data can reasonably be considered to be independent because the dependence of the weather will not be great for days one year apart.
To make it reasonable to consider data which we have collected as independent, we have to initialise the measurement process after each observation. We also have to make sure that no information leaks from one experiment to other experiments.

Example *5*

A story is told of a police officer who each year made the draw for a club lottery. Each year 1000 tickets were sold numbered from 0 to 999 and only one ticket was awarded a prize. A person with some statistical background saw a display of the numbers of the winning tickets for some years and noticed a remarkably even representation of the intervals 0-99, 100-199, 200-299, 300-399, 400-499, 500-599, 600-699, 700-799, 800-899, 900-999. When the police officer was confronted with this, he explained that he had arranged to have the different intervals equally represented to make the drawing random! The well-meaning police officer had introduced dependence in the hope of getting randomness.

Example 6

Users of statistics often assume independence when this is not realistic. The following example is from New Scientist (September 1991).
The Airbus A-320 aircraft has five flight computers working in parallel. This means that if one of the

flight computers breaks down, the next one takes over, and so on. It is claimed that the breakdown probability of each of the flight computers is 0.00001 per hour flight. From this it is concluded that the probability of computer breakdown is O.0000i~ per hour's flight. Apparently someone has thought it appropriate to assume independence between breakdowns of the five flight computers. The fact that two A-320 aeroplanes have crashed, probably because of computer failure, does raise serious doubts concerning the independence of the breakdown of different flight computers.

Example 7

Throughout my career I have experienced several misuses of statistical methods. The following is a serious example of misuse of statistical methods involving independence which I encountered in a psychological investigation back in 1971.

Subjects go through a psychological test consisting of several questions which they have to answer. This test was replicated 20 times. The data was collected as a series of 0's and l's, 1 indicating a correct answer and 0 an incorrect answer. All observations were assumed independent. I wondered if the subjects learned as they were given the same questions time and again and thus improved. Inspection of the data showed that the frequency of "correct" increased steadily through the series. When I concluded that this clearly showed that the assumption of "independent repetitions of the same experiment" was violated, I was met by the argument that this meant that "independent repetitions of the same experiment" was impossible to meet in psychological research! Rayner (1993) discusses two-sample t-tests and the issue of "paired" or "independent" samples.

Independence has the function of excluding contamination of information between observations. This does not mean that observations as they are recorded do not gradually give a clearer picture of the phenomenon that we are studying!

◆RECOGNISING INDEPENDENCE ◆

We seldom have enough information to perform a formal test of independence. Without a specific alternative it is not apparent what test should be used to test independence. The chi-squared test for independence has small power with small data sets and may have too much power for big data sets. The Durbin-Watson test for independence only reveals dependence coupled to the order of the observations. The real meaning of independence is not as easily understood as other terms in probability and statistics.

To demonstrate the practical meaning of independence I have made demonstration material consisting of three strings of beads of two colours, yellow and blue. The number of beads in each string is nearly 200. In one of these the beads have been threaded independently and form a series of Bernoulli trials. In the other two the colours of consecutive beads are dependent.

I ask the students to identify the string where the colours of consecutive beads are independent. The students soon recognise that the three strings of beads differ in the mixing of the colours. In one the colours are very well mixed, making few long runs of the same colour. In another the colours are not well mixed making a lot of long runs of the same colour. The third string seems to have an intermediate level of mixing.

Few students think the string with long runs of same colour corresponds to independence. In this they are right. In fact the generator used for this string has a serial correlation of +0.7. However, it is not so easy to agree on which of the other two has independence. This is interesting since the one where the colours are very well mixed has been generated by a mechanism having a serial correlation as low as —0.7.

It appears that some people think independence implies some force to avoid long runs. Put a different way, people expect randomness to take pity on person who has lost several games.

◆INDEPENDENCE IN TEXTBOOKS◆

I am rather worried about the signals statisticians give concerning the importance of independence. To verify or falsify such an impression I conducted a survey of the coverage given to independence in some textbooks of probability and statistics in our library at Agder College. Of course, our library does not have all the relevant textbooks, so the survey should not be taken too seriously. However I believe that even such a primitive investigation can give some indications. If a non-statistician wonders whether independence plays an important part in statistics and examines the books I have used to find out, I think he will get the same impression as I did, namely that independence seems to play a minor part in probability and statistics. The result of the investigation was:

Table 1. Frequencies of number of pages discussing independence.

No.of	0	1	2	3	4	5	6	7	8	9	10	...	26
Freq	3	11	4	3	2	4	3	2	I	0	2	...	1

This is a case where the outlier is very interesting. It is the book of Hodges and Lehmann (1970) which has a whole chapter reserved to the discussion of when models using independence are realistic.

◆THOROUGH DISCUSSIONS ◆ OF INDEPENDENCE

Whilst searching for independence in textbooks of probability and statistics, I also came across other books discussing this topic. In particular, the two volumes of De Finetti (1974 and 1975) have a lot to say about the topic. De Finetti discusses at least five different concepts relevant to dependence at independence:
Logical independence/dependence
Linear independence/dependence
Nonlinear independence/dependence
Stochastic independence/independence
Conditional stochastic independence/dependence

In addition, Dc Finetti has his own weaker condition of "independent identically distributed" called exchangeability. However, I feel that this notion is too technically demanding for introductory courses.

I will end by discussing the question raised earlier concerning the apparent paradox that independent identically distributed observations cannot gradually give a clearer picture of the phenomenon that we are studying. In a Bayesian framework this is resolved by noting that independence here really means *independence given the parameter.* In a non-Bayesian framework the best I can say is that independence in this situation means independence for a person who knows the value of the parameter.

References

Kolmogorov, A.N. (1950). Foundations of the Theory of Probability. Chelsea.
Von Mises R. (1957). Probability, Statistics and Truth. Dover.
Kac, M. (1959). Statistical independence in probability, analysis and number theory. Wiley.
Rayner, J.C.W. (1993). Assumptions are important: the paired and pooled t tests. *Teaching Statistics* **15** (1), 15-17.
Hodges, Lehmann (1970). Basic Concepts of Probability and Statistics. Holden Day.
De Finetti B. (1974). Theory of Probability Volume 1, Wiley.
De Finetti B. (1975). Theory of Probability Volume 2, Wiley.