IN SEARCH OF THE SPECIFICITY AND THE IDENTIFIABILITY OF STOCHASTIC THINKING AND REASONING

Herman Callaert
Center for Statistics, Limburgs Universitair Centrum, Belgium

Research on stochastic thinking and reasoning, with its implications on the teaching of statistics, should have professional input both from educational psychology as well as from the field of probability and statistics. Studying research papers on stochastic thinking and reasoning, one wonders which thought processes can be identified as being specific for the stochastic aspect of the experiments that are carried out. Such identification could be helpful, possibly leading to guidelines for optimizing statistical teaching strategies. Through a couple of examples, this paper tries to understand the complexity of apparently simple experiments. The chosen examples are very classical (about the "equiprobability bias" and about the "representativeness heuristic"), and many research papers have already appeared on these topics. It is shown that the mathematical complexity as well as the wording can easily play a confounding role when setting up experiments. An indication is given of how this could lead to new research questions.

1 Introduction

Research on stochastic thinking and reasoning is often based on experiments, which are carefully planned so that the mechanism can be discovered of how people (often children or students) learn and construct meaning. The choice of the experiments and the way in which they are carried out is rooted in the vast experience of research in psychology and education, and they greatly benefit from this professional framework. At the same time, probability and statistics is a research area in its own right, with its own professional rules and its specific academic terminology. Input from both sides could be beneficial for everyone involved in the fascinating research field of stochastic thinking, as well as in the teaching of statistics.

The field of probability and statistics is enormous. To fix the idea, one could start with some basic concepts in probability. Without going into a mathematical formulation, one could say that "probability is a way of describing formally with numbers the long-run regularity of random behavior". How then should the instructor proceed when teaching probability? What type of experiments can clarify the thinking and reasoning of students when randomness is at play? Can we separate the stochastic part from the deterministic part in simple experiments? What about complex experiments, consisting of intermediate steps? What happens when these intermediate steps are more or less disguised by focusing mainly on the behavior of the final outcome only?

2 Compound experiments and confounding

Is there a relation between the final outcome of experiments and the identifiability of stochastic reasoning? Let's look at some examples, from very simple to more complex.

When buying a ticket of the national lottery, most people do not believe that they have a 50% chance of winning the jackpot. However, there are only two final outcomes for any person: one either wins or one looses.

We remark here that the final outcome is related, in a simple and direct way, to an underlying experiment that can easily be conceptualized (drawing at random a ticket from an urn containing a huge number of different tickets). Apparently nobody has difficulties with the equiprobability bias in this type of experiments.

When working with a "one die" gadget in a microworld setting, Pratt (1998) reports that children have a reasonable view on equiprobability in this stochastic experiment. They use a choose-from [1 2 3 4 5 6] workings box for obtaining equally likely results. Conversely, when being asked to favor 1's and 2's the children themselves edit the workings box into a choose-from [1 1 2 2 3 4 5 6]. Why is there no equiprobability bias here? Remark again that the final outcome is related, in a simple and direct way, to an underlying experiment that can easily be conceptualized.

When it comes to the sum of two dice, Pratt (2000) reports that, in pre-interviews, children (Anne and Rebecca) declared that no total for two dice was harder or easier to obtain than any other. Apparently, Lecoutre's (1992) equiprobability bias is at play here, and it is interesting to try and understand the children's reasoning. Anne's articulation of the equiprobability bias stems from the fairness local resource (the totals must be equally likely because the dice individually are fair). Rebecca applied the unsteerability resource of the total of two dice ("Cos it's random; you can't control which number it lands on").

Remark here that the final outcome is not at all related, in a simple and direct way, to an underlying experiment that can easily be conceptualized. Interplay between easy stochastics and complicated mathematics might obscure the identifiability of the root for the equiprobability bias. Let's have a closer look at the components that might influence the experiment.

When working with just one die, children edited the workings box into [1 1 2 2 3 4 5] and expected that this would favor 1's and 2's. Although unsteerability was not removed in this situation, the children nevertheless did not expect equiprobability. Could this be related to the "closeness" of a simple random experiment and the final outcome? Could it be that Rebecca had mathematical difficulties for relating the sum of two dices to a simple urn model (or workings box) from which it would be easy to predict which numbers will be "favored"?

Anne's reaction starts with a correct observation of the (initial) components of the random experiment. Each die individually generates 6 different numbers with the same probability, and hence is fair. And then of course, if you only use "fair" components, the end result should be "fair" too!

At this point, we leave stochastics and enter deterministic mathematics. The gap between the sum of two dice, and the numbers on one single die, is enormous, and it

might be related to the gap between mathematical language, and language used in everyday life.

To start with, mathematics has (in the multidimensional setting) distinct words and distinct notations for the collection of elements. For the outcomes of rolling two dice, there are exactly 21 different outcome sets {x,y}, and exactly 36 different ordered pairs (x,y). There is nothing inconsistent or conflicting about this, but non-mathematical literature is often not so explicit in its wording. Take for example the following statement (Pratt (2000)): "It seems, for example, noncontentious that the proper grasp of the probability of obtaining a 7 with two dice involves knowing how many configurations of two dice result in a total of 7; the probability of a 7 is directly proportional to this number". In mathematical terminology, the word "configurations" has to be specified as meaning either "sets of two elements" or "ordered pairs". Although distinct, both concepts can be used for arriving at the probability of obtaining a 7, but direct proportionality only holds in the latter case (one could wonder why direct proportionality is needed anyway).

Next comes the mathematical operation of adding numbers. Since this operation is commutative, it invites children not to pay attention to order. In Pratt (2000) it is described that Anne says "Well, 1 and 2 and 2 and 1 are the same ...; they come to the same number". At that point, the Researcher intervenes: "They come to the same total, but are they the same as far as the spinners are concerned?". This might have been a crucial intervention, and it could have been avoided by choosing a different mathematical operation. Instead of adding numbers, one could subtract numbers. And since subtraction is not commutative, one has to specify which number has to be subtracted from which (e.g. the number on the first die minus the number on the second die). This would lead to all numbers between –5 and +5 as possible final outcomes, and with probabilities just like those for the sum.

Switching from summation to subtraction might force the student to consider the 36 ordered pairs (x,y)=(number on first die, number on second die). But different pairs may give rise to the same outcome, as is the case e.g. for (5,1) and (6,2) since $5-1=6-2=4$. This property may be an additional confounding factor, and it is present as well in the addition as in the subtraction operation.

Let's try to capture the bare essence of the experiment of throwing two dice. One could think of two cubical objects, physically different but indistinguishable to the human eye. Each object has six faces, all of them physically different, but also here the human eye can't tell the difference. Throwing the first die then means that (ideally) any of the 6 faces has the same probability of being on the upper side. The same property holds for the second die. Hence, combining each possible face of the first die with each possible face of the second die leads to 36 different outcome possibilities for this experiment. For fair dice and independent throws, all those outcomes have the same probability 1/36.

Next comes the fact that one has to mark the faces, so that the human eye is able to see the differences (the faces are physically different anyway, also without markings). A first problem might arise already at this stage. How should one mark? Should one use colors, numbers, or what? Are some colors or numbers associated with "luck", "easier to obtain", or "rare"? When using numbers, it might be wise to look for integers which are "as neutral as possible", so that they only serve as a "label" for a particular face of a particular die. Avoiding both the classical numbers {1 2 3 4 5 6} and also the numbers which arise as the sum of two dice, could be a wise strategy. It might help the students to look at the experiment "from scratch", without interference of possible prejudices stemming from games of chance they have played before with "regular" dice.

As a first attempt, write {13 14 15 16 17 18} on the faces of the first die, and label the faces of the second die with {19 25 31 37 43 49}. The basic stochastic process of each die is analogous as before, generating six numbers with equal probability. As far as the outcome space is concerned for throwing these particular dice, there are 36 different ordered pairs (x,y)=(first,second), and there are also 36 different (unordered) sets of outcomes {x,y}. On top of this, the sum of the two dice yields 36 different outcome possibilities (every number between 32 and 67), and each outcome has the same probability 1/36. Hence, if Anne and Rebecca would have said in the pre-interview that "no total is harder or easier to obtain than any other", they would have been right now. The problem with their motivation of why this is true will still need further study. Also, explaining the children's answers in terms of the equiprobability bias doesn't seem obvious at this point.

Another easy and unfamiliar experiment would consist in finding the product of two regular dice. This yields numbers between 1 and 36 with a very irregular pattern for their chances.

Starting with an identical random process in the underlying components (the two dice are "fair") can lead to many different outcomes and chances, and this just depends on deterministic decisions like: which numbers does one write on the faces of the die, and which mathematical operation (addition, subtraction, multiplication) does one use? It might be helpful to consider many variations on "the total of two dice" to find out what is typical in the reasoning about random processes. Trying to reduce the number of confounding elements might be important when setting up research experiments in this field.

3 Language, similarity, and confounding

Statistics and probability are rooted in experiences of everyday life, and hence seem familiar. This familiarity is both of great value and of great danger. Many people are convinced that, in the information age that we live in, a basic knowledge of probability and statistics is as essential as the ability to read and write. This basic knowledge however can't be gained just by everyday experiences and intuition. That many concepts are not at all intuitive seems more natural for mathematics than for

statistics. In daily live, people are surrounded by patterns in flowers, shells and pineapples, and by beautiful works of art, but nobody assumes that these experiences lead intuitively to an understanding of the golden section or to familiarity with Fibonacci numbers. The same attitude should hold for statistics too. Statistics is not a branch of mathematics, but it still is a mathematical science (Moore, 1998), and hence, familiarity with mathematical rigor and with mathematical language is crucial in the discovery of difficulties associated with stochastic thinking and reasoning. It is not easy to avoid the confounding effect of everyday language in experiments that are set up for studying stochastic thinking. Examples are plenty, and we briefly refer to a classical one, often conducted in the framework of statistical misconceptions.

A standard experiment, studied over and over again in numerous research papers, deals with sequences of binary outcomes, such as heads and tails. Callaert (2002) reports about such an experiment, trying to gain insight into what is called "the representativeness heuristic" by Kahneman and Tversky (henceforth abbreviated K&T).

Consider the following items:

Item 1: A fair coin is tossed six times. At each toss, the coin lands either H (=heads) or T (=tails). The results are recorded in the order they appear.

The outcome H T H T T H

_ is less likely than

_ is as likely as

_ is more likely than

the outcome H H H H H H

Item 2: A fair coin is tossed six times. At each toss, the coin lands either H (=heads) or T (=tails). An outcome that contains three heads and three tails

_ is less likely than

_ is as likely as

_ is more likely than

an outcome that contains six heads

Although the wording is very precise, the poor responses of the students are unbelievably shocking. The small-scale study by Callaert among high school graduates entering university, found that less than 60% answered item 1 correctly. Students with a strong mathematical background however apparently had learned about this "tricky" question, and almost all of them gave a correct answer to item 1. However, more than 50% of those same students failed to produce the right answer to item 2! So, after all, what did they really understand?

Why are the above test items so difficult? Could one identify (possibly new) aspects that might deserve further research? Is it true that mistakes on item 1 can for a

substantial part be explained by the representativeness heuristic? K&T claim that, according to this heuristic, the subjective probability of an event, or a sample, is determined by the degree to which it: (i) is similar in essential characteristics to its parent population; and (ii) reflects the salient features of the process by which it is generated. This statement gives itself rise to fundamental problems. In addition, one might suspect some confounding, related to the wording of the problem. The statement in test item 1 is meant as a rigorous mathematical expression (it is about ordered six-tuples), but it can easily be interpreted by the student as a statement about (i) the behavior of a fair coin, (ii) the behavior of six tosses of a coin, (iii) the behavior of heads and tails in six tosses of a coin.

Let's try to expand on the above interpretations. At the same time, and for instructional purposes, let's try to relate those interpretations to tools in a microworld setting (as done e.g. by Pratt in his experiments).

i) The behavior of a fair coin is described theoretically by the Bernoulli distribution with success parameter equal to 1/2. A computer simulation consists of any process resulting in outcomes that can be dichotomized so that both parts have equal probability (like randomly generating integers, and associating heads with even numbers and tails with odd numbers). The outcome to be denoted is either H or T. Repeat this process many times. In the long run, the proportion of heads approaches 1/2 (law of large numbers). This statement is about the global behavior of the sequence of outcomes. It is not about local characteristics of segments of such a sequence. It is not about "exactly six tosses". Hence, the parent population of "six tosses" is not at all "the behavior of a fair coin". This observation makes the interpretation of "representativeness" as described by K&T difficult to understand. Should instructors tell the students that outcomes do not resemble the parent population, or should they tell them that they first have to search for the correct parent population? For the question in test item 1, the parent population is far from the "tossing a coin" parent population, described in many research papers.

ii) The behavior of six tosses of a coin is described theoretically by the uniform distribution over the $2^6 = 64$ different outcome possibilities. A computer simulation starts by repeating the process in (i) exactly six times. The outcome to be denoted at that moment is the observed ordered six-tuple (# # # # # #) with # equal to either H or T. Then, repeat the above procedure many times. In the long run, the relative frequency of any of the six-tuples approaches 1/64. Carrying out the above experiment with tools in a microworld setting might help the student to arrive at the correct answer on test item 1. However, there is an additional confounding element here, and it has to do with the complexity of the notation of the outcomes. Experiments with a uniform distribution over the numbers 1, 2, 3, ... ,64 are easier to grasp than experiments with a uniform distribution over 64 elements of ordered six-tuples. When being confronted with a large amount of outcomes in a simulation experiment, it is much easier to get a correct

feeling about the relative frequency of a 9 or a 27, than about the relative frequency of a (H T H T T H) or a (H T H T H H). A possible problem, which might be called "similarity bias", is the fact that it is more difficult to recall six elements in their exact order than to recall a single number. Hence, a student might say that in the simulation process, she has seen many more situations "looking like" (H T H T T H) than situations "looking like" (H H H H H H). And she is right of course.

iii) The behavior of the number of heads and tails in six tosses of a coin is described theoretically by the binomial distribution. A computer simulation is exactly as in (ii) with one big difference. The outcome to be noted now is not a six-tuple, but a number between zero and six, indicating the number of heads in those six tosses. When this procedure is repeated many times, the relative frequency of any of the numbers between 0 and 6 approaches the corresponding binomial probability. The relative frequency of the number 6 (six heads) approaches 0.015625, while the relative frequency of the number 3 (three heads and three tails) approaches 0.3125, which is twenty times larger. This knowledge is helpful for arriving at the correct answer to test item 2.

4 Discussion

In his reflection on teaching probability and statistics, Shaughnessy (1992) emphasizes the need to (i) know more about how students think about probability, (ii) identify effective methods of instruction, and (iii) develop reliable methods of assessment that more accurately reflect students' conceptual understanding. That this is no easy task is reflected through the above examples. The use of well-designed experiments in the framework of microworlds might create a rich environment, both for learning stochastic concepts as for researching this learning process.

REFERENCES

Callaert, H. (2002). Understanding Statistical Misconceptions. . *Proceedings of the Sixth International Conference on Teaching Statistics,* ed. B. Phillips, Voorburg (NL). The International Statistical Institute.

Kahneman, D. and Tversky, A. (1972). Subjective Probability: A Judgment of Representativeness. *Cognitive Psychology,* 3, 430-454.

Lecoutre, M.P. (1992). Cognitive models and problem spaces in 'purely random' situations. *Educational Studies in Mathematics,* 23, 557-568.

Moore, D. (1998). Statistics Among the Liberal Arts. *Journal of the American Statistical Association,* 1253-1259.

Pratt, D. (1998). Expressions of Control in Stochastic Processes. *Proceedings of the Fifth International Conference on Teaching of Statistics. Vol 2*. Voorburg (NL). The International Statistical Institute.

Pratt, D. (2000). Making Sense of the Total of Two Dice. *Journal for Research in Mathematics Education,* 31, 602-625.

Shaughnessy, J. M. (1992). Research in Probability and Statistics: Reflections and Directions. *Handbook of Research for Mathematics,* ed. D. Grouws, New York: Macmillan, pp.115-147.