



**International Electronic Journal of
Mathematics Education**

Volume 4, Number 3, October 2009

www.iejme.com

Special issue on “Research and Developments in Probability Education”

Manfred Borovcnik & Ramesh Kapadia (Eds)

**ON CONDITIONAL PROBABILITY PROBLEM SOLVING RESEARCH –
STRUCTURES AND CONTEXTS**

M. Pedro Huerta

ABSTRACT. In this paper we summarize the research we have recently carried out on classifying problems of conditional probability. We investigate a particular world of school word problems we call ternary problems of conditional probability. With the help of a mathematical object, the trinomial graph, and the analysis and synthesis method, we propose a framework for a structural, didactical and phenomenological analysis of the ternary problems of conditional probability. Consequently, we have organized this world into several types of problems. With respect to students’ behaviour, we identify four types of thinking processes related to data format and the use of data. We also illustrate our approach by use of the diagnostic test situation, and in the particular context of health.

The main purpose of our work is to improve secondary school students’ understanding of conditional probability and their probability literacy by proposing a teaching approach based on problem solving within appropriate contexts. We believe that the framework we present in this paper could help teachers and researchers in this purpose.

KEYWORDS. Conditional probability, problem solving.

1. INTRODUCTION

In some research studies on conditional probability, students have to solve several problems, which are considered as different but in fact are structurally isomorphic and differ only in data format. The “disease problem”, as discussed below, is one example. Researchers draw conclusions about students' behaviour without taking into account the mathematical structure of the problem and the context in which it is put forward. They then generalise conclusions independently from the structure or the context, or both.

This is a methodological flaw, however, because students' behaviour depends also on,

- i. the mathematical structure of the problem (on what is known and what is unknown in a given situation),
- ii. the context in which the problem is formulated, and
- iii. the particular vocabulary with which data, (numerical or event based), and the relationships between the different data, are described.

To what extent students' behaviour depends on these factors (either considered as isolated or related) is an opportunity for research that has not yet been carried out. For this reason, we consider the aforementioned conclusions as being rather limited and restricted to the problem actually used in the experiment.

If we think about preparing students in conditional probability in secondary school, two questions may arise: why? and how? The answer to the first question takes the students' future into account, both at university and in the world outside. So, it is necessary to explore the contexts and phenomena in which conditional probability arises. Thinking in this way, we will determine not only what kind of skills secondary school students must have regarding conditional probability, but also what type of problems they will need to solve and in which contexts problems will have to be formulated. The answer to the second question lies in adopting a phenomenological and realistic approach to the teaching of conditional probability through the process of problem solving.

Of course, problem solving, and in particular conditional probability problem solving, are topics that are usually taught around the world with a greater or lesser degree of success. Shaughnessy (1992) pointed out the structure of teaching probability by affirming, “teaching probability and statistics *is* teaching problem solving”. Curricular standards in several countries would suggest that, in general, both mathematics and probability should be taught within a context. This includes a mathematical context, and connecting school mathematics with the

experiential reality of the students (NCTM 2000 in USA, BOE 2007 in Spain). In recent works, the relevance of context has also been corroborated in studying probability (see Watson 2005, Gal 2005, Carles & Huerta 2007). Probability in context and probability literacy are also related to a new notion of chance as a precursor to probability (Watson 2006).

Problem-solving may be analyzed from various perspectives. We use the term *scenario* according to the number of main inter-related subjects being studied. In scenario 1, the analysis is restricted exclusively to problems (specifically school problems). In scenario 2, research takes into account the problems and the students together in an inter-related way. In scenario 3, teachers appear on the scene; relationships between problems-students-teachers become the object of research. Our research project ^[1] will follow these three scenarios for analysis sequentially.

In this paper, we show what we already know about a particular world of conditional probability problems, mainly those related to scenario 1. This will include the framework used for their analysis and results of structural and contextual analysis. As a consequence of our analysis, we will advocate Freudenthal's phenomenology approach (Freudenthal 1983) to teaching. We will also present some results from a specific research project on a particular family of problems in scenario 2. Therefore, this work is relevant both for researchers and teachers.

2. FRAMEWORK FOR ANALYSIS OF THE STRUCTURE OF PROBLEMS

Research on conditional probability problems in scenario 1 lacks a conceptual framework. We start this section by focusing on the object of study, namely the ternary problems of conditional probability. But, before that it is necessary to clarify the meaning of some of the notions that will be used throughout this paper.

We say that a school problem is a *probability problem* if, in its formulation, at least one probability is mentioned, either as known or unknown data. Usually, but not always, these problems are word problems formulated in a context, which could be put forward in mathematical terms or in terms of a real situation. The following is an example of a word probability problem used in school textbooks:

Menu task. A delicatessen serves different menu items, of which 2 are different soups, 6 are different brands of sandwiches, and 4 are salads of different kinds. How likely is it for each event to happen if you choose one soup, one sandwich, and one salad at random from the menu?

We say that a probability problem is a *conditional probability problem* if in its formulation at least one conditional probability is mentioned, either as known or unknown data. Again, the expression of the known data referring to events, probabilities and conditionality depends on the context in which the problem is put forward. The following is an example of a word conditional probability problem used in schools:

Exit poll task. In a recent exit poll, 60% of respondents felt that the country was generally going in the right direction. 24% of the respondents were Republicans who felt that the country was generally going in the right direction. If a person in this poll felt that the country was generally going in the right direction, find the probability that the person was a Republican.

Ternary problems of conditional probability

The object of our research is to be a particular world of conditional probability problems. This world is generated by 2 basic events, A and B , and the 8 events derived from them by elementary operations. There are 16 probabilities and 18 relationships between them. Problems from this world are quite usual in conditional probability teaching environments in school and college. At the same time, some of these items are used in investigations on Bayesian reasoning. The following events, probabilities, and relationships are involved:

- The events: $A, B, \bar{A}, \bar{B}, A \cap B, A \cap \bar{B}, \bar{A} \cap B, \bar{A} \cap \bar{B}$;
- 16 probabilities:
 - four marginal or absolute probabilities: $p(A), p(B), p(\bar{A}), p(\bar{B})$;
 - four intersection probabilities: $p(A \cap B), p(A \cap \bar{B}), p(\bar{A} \cap B), p(\bar{A} \cap \bar{B})$;
 - and eight conditional probabilities: $p(A|B), p(\bar{A}|B), p(A|\bar{B}), p(\bar{A}|\bar{B}), p(B|A), p(\bar{B}|A), p(B|\bar{A}), p(\bar{B}|\bar{A})$; and
- 18 relationships between these probabilities; 8 are multiplicative and 10 are additive relationships.

Events		Probabilities				complementary	intersection	conditional
Basic	intersection	absolute	intersection	conditional				
A	$\frac{A \cap B}{A}$	$p(A)$	$p(A \cap B)$	$p(A B)$	$p(B A)$	$p(A B) + p(\bar{A} B) = 1$	$p(B) = p(A \cap B) + p(\bar{A} \cap B)$	$p(B) \times p(A B) = p(A \cap B)$ $p(B) \times p(\bar{A} B) = p(\bar{A} \cap B)$
	$\frac{A \cap \bar{B}}{A}$	$p(\bar{A})$	$p(A \cap \bar{B})$	$p(\bar{A} B)$	$p(\bar{B} A)$			
B	$\frac{A \cap B}{B}$	$p(B)$	$p(A \cap B)$	$p(A B)$	$p(B A)$	$p(B A) + p(\bar{B} A) = 1$	$p(A) = p(A \cap B) + p(\bar{A} \cap B)$	$p(A) \times p(B A) = p(A \cap B)$ $p(A) \times p(\bar{B} A) = p(\bar{A} \cap B)$
	$\frac{A \cap \bar{B}}{B}$	$p(\bar{B})$	$p(A \cap \bar{B})$	$p(\bar{A} B)$	$p(\bar{B} \bar{A})$			
						$p(B \bar{A}) + p(\bar{B} \bar{A}) = 1$	$p(\bar{A}) = p(\bar{A} \cap B) + p(\bar{A} \cap \bar{B})$	$p(\bar{A}) \times p(B \bar{A}) = p(\bar{A} \cap B)$ $p(\bar{A}) \times p(\bar{B} \bar{A}) = p(\bar{A} \cap \bar{B})$
						$p(A) + p(\bar{A}) = 1$		
						$p(B) + p(\bar{B}) = 1$		

Table 1. The world of ternary relations of conditional probability - [Click](#) to see more details.

All problems in this paper belong to this world of problems. The following is but one:

Symbolically formulated task. Given $p(A)$, $p(\bar{A})$, $p(\bar{B} | A)$, $p(B | \bar{A})$, calculate $p(A \cap \bar{B})$.

This formal task could be formulated within several contexts, though structurally it would still remain the same task. One example for a suitable context is the following:

Medical diagnosis task 1. A medical research lab proposes a screening test for a disease. To try out this test, it is given to 100 persons, 60 of whom are known to have the disease and 40 not to have the disease. A positive test indicates the disease and a negative test indicates no disease. Unfortunately, such a medical test can produce two kinds of errors:

1. A false negative test: For the 60 people who do have the disease, this screening test is negative for 2 persons indicating that they do not have it.
2. A false positive test: For the 40 people who do not have the disease, this screening test is positive for 10 people indicating that they have it.

Suppose the test is given to a person whose disease status is unknown. What is the probability that he/she has the disease and the test is negative?

With respect to the number of involved relationships, it is possible to distinguish two kinds of conditional probability problems. These can be termed *one-step* and *more-than-one-step problems*. A *one-step-problem* of conditional probability can be solved using only *one* of the 18 relationships mentioned above, although it is usually one of the multiplicative ones. Medical diagnosis task 1 is a one-step problem because only one relationship is required: $p(A \cap \bar{B}) = p(\bar{B} | A) \times p(A)$, where A means to have the disease and B to have a positive test result, and $p(\bar{B} | A)$ and $p(A)$ are both known.

But many conditional probability problems are *more-than-one-step problems*. In order to solve them more than one of the 18 relationships are required. In general, almost all problems shown in this work are examples of this kind of problem, as in the following:

Symbolically formulated task. Given $p(A)$, $p(\bar{A})$, $p(\bar{B} | A)$, $p(B | \bar{A})$ —calculate $p(\bar{A} | \bar{B})$

The same problem formulated within a context could read as follows:

Medical diagnosis task 2. Context as in medical diagnosis 1 but now with the following question: If the test result is negative, what is the probability that the person does not have the disease?

We say that a relationship between probabilities is a *ternary relationship* if three

probabilities are related to each other by a linear equation such as:

$$p(A \cap \bar{B}) = p(\bar{B} | A) \times p(A).$$

This equation relates the probabilities $p(A)$, $p(\bar{B} | A)$ and $p(A \cap \bar{B})$ together. The relationship

$$p(A \cap B) + p(A \cap \bar{B}) = p(A)$$

is, again, a ternary relationship; knowing two of the probabilities implies knowing the third. There are two basic types of ternary relations, namely additive and multiplicative. The relationship

$$p(A) + p(\bar{A}) = 1,$$

is slightly different from the two above as to know one probability implies to know the other.

Ternary problems of conditional probability

Ternary problems of conditional probability are defined as problems that fulfil the following conditions:

- One conditional probability is involved, either as known data or as unknown data or both.
- Three probabilities are known.
- All probabilities, both known and unknown are connected by ternary relationships.

To solve ternary problems of conditional probability, one may relate — stepwise — two known and one unknown probability by a ternary relation. The process is repeated until the requested probability may be calculated. So, ternary problems of conditional probability are more-than-one step problems that, in order to be formulated, require four probabilities; three of them known, one unknown, and at least one of all four is a conditional probability. Problems in this section are examples of ternary problems of conditional probability.

If three of the known pieces of data are connected by a ternary relationship, the problem is said to be *redundant*. By a subsequent reformulation of it, at least one item of the related data can be removed. The following example (Kahneman & Tversky 1982, p. 156) is redundant:

Taxi problem. A cab was involved in a hit and run accident at night. There are two cab companies that operate in the city, a Blue cab company, and a Green Cab company. You are given the following data:

- a) 85% of the cabs in the city are Green and 15% are Blue,
- b) a witness identified the cab as Blue.

The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colours 80% of the time and failed 20% of the time. What is the probability that the cab involved in the accident was Blue rather than Green?

Shaughnessy proposes to use 2 by 2 contingency tables for presenting such problems:

	Witness accuracy (80%)		Total
	Witness says Blue	Witness says Green	
Blue Cabs	120	30	150
Green Cabs	170	680	850
Total	290	710	1000

Table 2a. Data for the taxi problem in a two-way table.

Of course, a problem formulated in a format like this is a redundant problem. There is more data than is necessary to find an answer to the proposed question, since some of the data is related by (additive) ternary relationships. Therefore, the following presentation is also possible, amongst others, in order to answer the same question. By means of a simple analysis of known and unknown data, we can conclude that there are 6 possible ternary problems that can be put forward in this situation. The following is one example (see Table 2b).

	Witness accuracy		Total
	Witness says Blue	Witness says Green	
Blue Cabs	120		
Green Cabs			850
Total	290		1000

Table 2b. Data for the taxi problem reformulated – presented in terms of ternary problem.

Data structures in ternary problems of conditional probability

By simple analysis, it can be demonstrated that 3 is the minimum number of known items of data required to formulate and solve ternary problems of conditional probability. That is, given three probabilities, appropriately chosen from the set of 16 possible probabilities, any of the other remaining probabilities can be calculated.

Given any ternary problem of conditional probability, such as the examples in this paper, it may be shown that the problem itself can be described by means of three known probabilities, irrespective of the context in which it is put forward. Both problems above (medical diagnosis tasks 1 and 2) can be described by the three items of data that are known: $p(A)$, $p(\bar{B} | A)$, $p(B | \bar{A})$. Although the probability $p(\bar{A})$ is explicitly mentioned in the text of both problems, it is not necessary to solve it because it is related to $p(A)$. So, in fact, both problems can be formulated with three probabilities: one absolute probability and two conditional probabilities.

Consequently, for each ternary problem, the data structure can be described by a vector (x, y, z) for the known data: x represents the number of absolute probabilities, y the number of intersection probabilities, and z the number of conditional probabilities, so it holds that $x+y+z = 3$. By choosing x, y and z in a suitable way from the set of 16 probabilities, we are theoretically able to identify 9 types of data structure in ternary problems of conditional probability. Here, the following data structures are displayed with emphasis on the third component:

$(2, 1, 0), (1, 2, 0), (0, 3, 0), (0, 2, 1), (1, 1, 1), (2, 0, 1), (0, 1, 2), (1, 0, 2),$ and $(0, 0, 3)$. Medical diagnosis tasks 1 and 2 are of type $(1, 0, 2)$; the taxi problem in our reformulation is of type $(2, 1, 0)$.

Classification and structure of ternary problems of conditional probability

However, though it is useful to describe the data structure by the triples (x, y, z) , they do not fully capture the structure of a problem. It is also necessary to consider also the question put forward in the problem. We will convert this question into data, but this time as unknown data.

So, initially, each ternary problem of conditional probability has a structure that is basically characterized by three known items of data and one related unknown item of data. Depending on the structure, ternary problems of conditional probability can be classified into families and subfamilies, as Lonjedo (2007) does, in fact do, according to the following criteria:

- Level (L) of a problem is the number of known conditional probabilities in the formulation of a problem. There are four possible levels of L_i , corresponding to this number, $i = 0, 1, 2, 3$. A problem is L_3 -level if in its formulation 3 known conditional probabilities are put forward.
- Related to each level, Category (C) of a problem is the number of known absolute probabilities in the formulation of the problem. There are three possible categories of C_j , corresponding to the following possibilities; $j = 0, 1,$ and 2 .
- Finally, Type (T) of a problem represents the unknown data in the problem. There are three possible Types of (T_k) : $T_1 =$ conditional probability, $T_2 =$ absolute probability, $T_3 =$ intersection probability.

So, every ternary problem of conditional probability belongs to an L-family of problems described by means of an (L_i, C_j, T_k) -triple. If a problem, for example, belongs to an (L_1, C_1, T_3) -triple, this indicates that in the formulation of the problem there are three known quantities that

can be read in a probabilistic sense. There should be one conditional probability (L_1), one absolute probability (C_1), and consequently one intersection probability, with one quantity being unknown, namely an intersection probability (T_3). See Table 3 for the result of the classification of all ternary problems of conditional probability into four families and twenty subfamilies (Lonjedo 2007).

	L_0			L_1			L_2			L_3		
C_0	C_0T_1	\emptyset	\emptyset	C_0T_1	C_0T_2	C_0T_3	C_0T_1	C_0T_2	C_0T_3	C_0T_1	C_0T_2	C_0T_3
C_1	C_1T_1	\emptyset	\emptyset	C_1T_1	C_1T_2	C_1T_3	C_1T_1	C_1T_2	C_1T_3	\emptyset	\emptyset	\emptyset
C_2	C_2T_1	\emptyset	\emptyset	C_2T_1	\emptyset	C_2T_3	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset	\emptyset

Table 3. Classification of ternary problems of conditional probability into families and subfamilies.

The symbol \emptyset means that there is no ternary problem belonging to that subfamily. In effect, there can be no problem in the (L_0, C_1, T_2) -triple, for example, as it would not be a conditional probability problem. No conditional probability is given (L_0); no conditional probability is asked for (T_2). However, there are different reasons that allow us to affirm that, for instance, there are no ternary problems of conditional probability that could be classified into the (L_2, C_2, T_1) -triple: if one of these problems belonged to that sub-family, then two conditional probabilities and two absolute probabilities would be known. A problem such as this is redundant, and in order to classify it, a previous analysis would be required. We will come back to this to see the result of such a classification in the next section.

Ternary problems and trinomial graphs

Cerdán & Huerta (2007) use *trinomial graphs* in order to study ternary problems of conditional probability by means of the analysis and synthesis method. These problems have been modelled using a trinomial graph as shown in Figure 2 (click [here](#)), whereby dashed lines represent ternary additive relationships, and solid lines represent multiplicative relationships. Nodes are labelled by the probabilities of the related “events”. In the graph, dark nodes represent known data and white nodes signify unknown data. Solving a problem in the graph consists of transforming white nodes into dark nodes by means of an algorithm. This algorithm, which we call the *destruction* algorithm of the graph, allows users familiar with it to find the solution to any posed problem. Thus, because the graph in Figure 2 shows all the 18 ternary relationships as a connection between 3 nodes each, when we have two dark nodes at an edge we also have a third as in any ternary relationship.

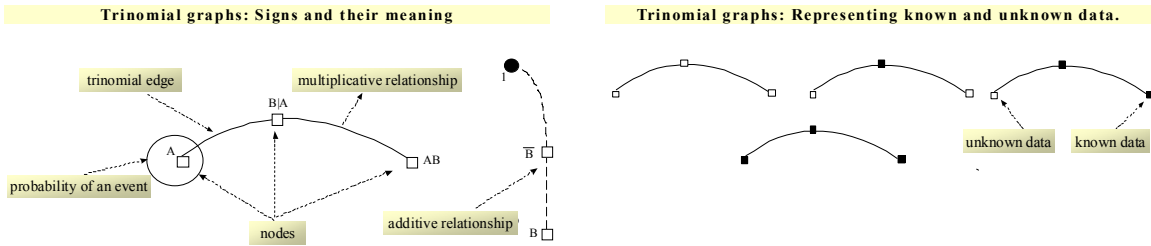


Figure 1. Signs in trinomial graphs - [Click to enlarge](#).

This graph shows all 18 ternary relations as edges between 3 nodes each. Once the probabilities of 2 nodes are known, one can calculate the third. Thus, one may proceed from already known nodes to calculate further nodes until one can calculate the requested probability.

A problem has a solution if and only if its related trinomial graph can be “destroyed”.

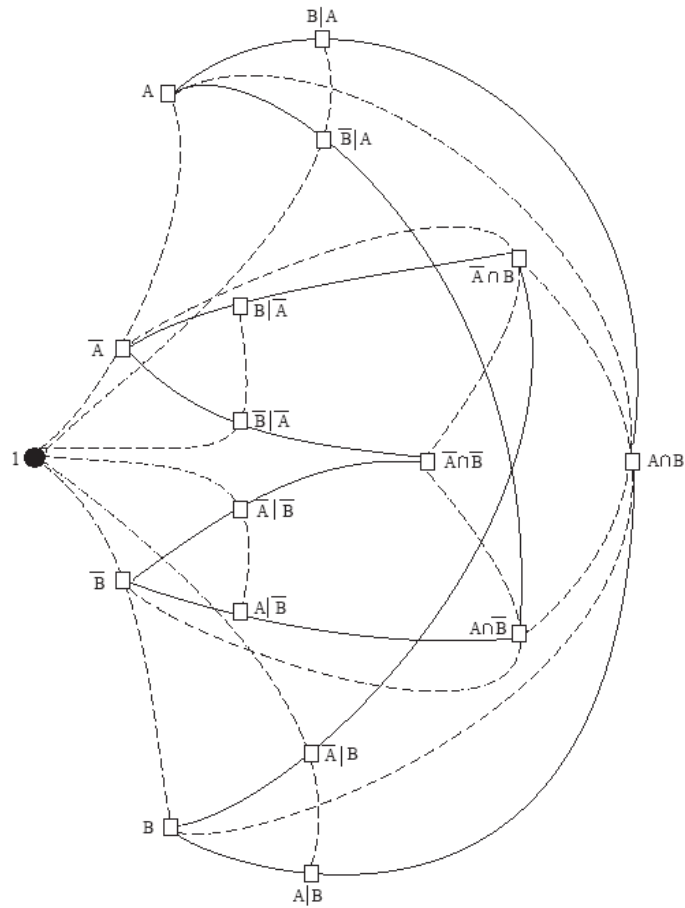


Figure 2. Trinomial Graph of ternary problems of conditional probability - [click to enlarge](#).

Ternary Problems of Conditional Probability are problems, which satisfy the following conditions:

1. There is at least one conditional probability, either as an item of data, a question, or both.
2. There are three known quantities.
3. All quantities, both known and unknown, are connected by ternary relationships as:

$p(A) + p(\bar{A}) = 1$, $p(A \cap B) + p(A \cap \bar{B}) = p(A)$	additive relationships; and
$p(A B) \times p(B) = p(A \cap B)$	multiplicative relationships
4. The question in the problem is an unknown quantity that is related to other quantities by more than one of the relationships above.

There are several processes —representing paths in the graph— to find the solution of a problem by means of the destruction algorithm. That is, there will be more than one path to find a solution in the graph and hence the graph associated to the resolution of each problem is not unique. One of them is minimal in the sense that it implies the minimum number of relationships in finding the requested unknown data. The resulting graph we call the *trinomial graph of the problem*. We use this concept to define isomorphism and complexity of problems:

- Given two problems P and Q, we will say that P is *structurally isomorphic* to Q if both share the same (minimal) trinomial graph.
- Given any problem P, the number of additive and multiplicative relationships in its minimal trinomial graph defines the degree of complexity of P. We also use this measurement in order to compare problems belonging to the same family or subfamily.

These relationships allow us to study isomorphism between problems belonging to each family and sub-family in Table 3 above and evaluate the complexity of these problems. As a consequence, each family and subfamily of problems can be divided into equivalence classes.

Analysis and synthesis method

The *analysis and synthesis method* proceeds in an opposite way to the destruction algorithm. While the latter proceeds from known to unknown data, the analysis method proceeds from unknown towards known data, subsequently identifying which data and relationships are involved in the process of solving. The synthesis of a problem consists of reconstructing the path shown by the analysis, this time using known data and usually by means of an expression that relates all known to the unknown data. We demonstrate the method by an example.

Structural version of medical diagnosis task 2. Given $p(A)$, $p(\bar{A})$, $p(\bar{B}|A)$, $p(B|\bar{A})$

—calculate $p(\bar{A}|\bar{B})$

In the trinomial graph in Figure 2, we represent the known data with dark nodes and emphasise the question by colouring the corresponding node in red (Figure 3). The process of the analysis may be followed as a movie (click to see the [movie](#)). The analysis starts at the question $p(\bar{A}|\bar{B})$ (red node in the graph). In order to calculate this node, $p(\bar{B})$ and $p(\bar{A} \cap \bar{B})$ are required (line number 1 for the required relationship in a different colour to represent its status). We should proceed in this way throughout the graph up to the point where all unknown probabilities can be calculated using the known probabilities in the problem, if this is the case.

In the movie, the succession of lines, from 1 to 5, constitutes the analysis of data and relationships needed in order to solve the question in the problem. The dashed line numbered by 0 informs us that the problem is redundant because either $p(A)$ or $p(\bar{A})$ could be omitted in the text of the problem, although both must be used in solving it.

The synthesis consists of an expression containing the answer to the question as a function of the known data and the relationships involved:

$$p(\bar{A} | \bar{B}) = \frac{p(\bar{A}) - p(\bar{A}) \times p(B | \bar{A})}{p(A) \times p(\bar{B} | A) + p(\bar{A}) - p(\bar{A}) \times p(B | \bar{A})}$$

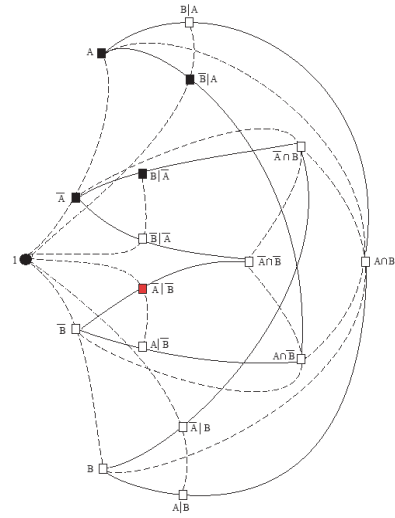


Figure 3. Trinomial graph for medical diagnosis task 2 - [Click to enlarge](#).

Sometimes, the analysis method is useful in order to classify problems into families and subfamilies, mainly when the problem is redundant. Look at the following example. Let P be a problem, which may be described by the triple (L_2, C_2, T_1) . It has already been stated that this family has no ternary problems because they are redundant, and in order to classify them, a previous analysis is required.

Redundant problem P. $p(A), p(B), p(A | B), p(B | A)$ are all known —calculate $p(A | \bar{B})$.

Again, we represent known data by dark nodes and emphasise the question by colouring the corresponding node in red (see Figure 4).

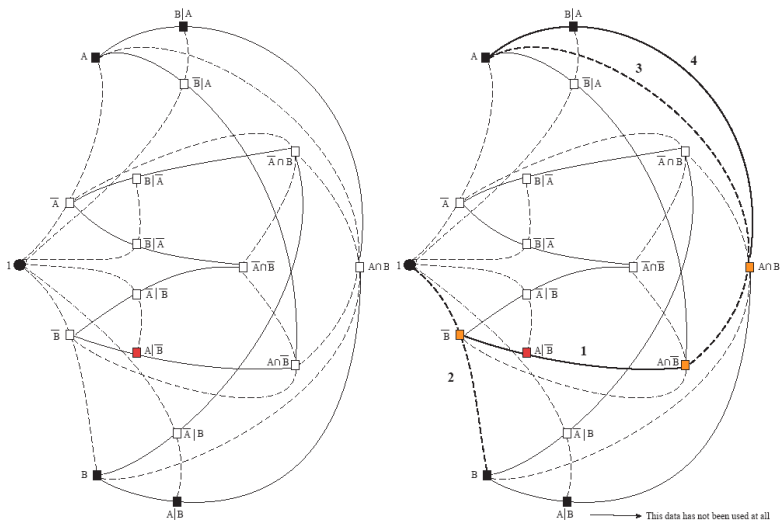


Figure 4a. Trinomial graph for redundant problem.

Figure 4b. Final state of algorithm - [Click to enlarge](#).

The steps, we apply in our algorithm may be seen from Figure 4; they are

- In order to know $P(A | \bar{B})$, then $p(\bar{B})$ and $p(A \cap \bar{B})$ are required (1).
- $p(\bar{B})$ is easily known because of data $p(B)$ (2).
- By means of relationship (3), to know $p(A \cap \bar{B})$ requires knowing $p(A)$ and $p(A \cap B)$.
- Other options are of course possible, but to consider them is irrelevant for the time being. $p(A)$ is a known item of data and $p(A \cap B)$ is known by means of relationship (4).
- The analysis of the problem is finished because the unknown probability $p(A | \bar{B})$ has been calculated using the known probabilities.

So, in the synthesis of the problem, we see that it has one solution that does not take $p(A | B)$ into account at all:

$$p(A | \bar{B}) = \frac{p(A) \times [1 - p(B | A)]}{1 - p(B)}$$

As $p(A | B)$ has not been used, problem P can be classified into the (L_1, C_2, T_1) -triple.

To end this section, in which we have put forward the first part of our framework to analyse the world of conditional probability problems, let us again consider a well-known problem (see Tversky & Kahneman 1982, p. 154)

The disease problem. If a test to detect a disease whose prevalence is 1/1000 has a false positive rate of 5%, what is the chance that a person found to have a positive result actually has the disease, assuming you know nothing about the person's symptoms or signs.

Tversky & Kahneman have put forward this problem to 60 persons (students and staff at Harvard Medical School); only 11 participants gave the answer 2% (which was considered as the correct answer by the authors) while 30 participants gave 95% as the answer. In fact, under the hypothesis that the test diagnoses every person who has the disease correctly as positive, the correct answer would be 2%. However, this is a hypothesis for which no information or corroboration is given, neither from the wording of the problem nor from the context.

If we read the problem in a probabilistic sense, it could be symbolically formulated as follows: Given $p(A)$ and $p(+ | \bar{A})$ —calculate $p(A | +)$, where A means to have the disease and $+$ to give positive result in the test. That is, only two items of data are explicitly mentioned in the problem, and hence known.

But, it can be demonstrated by means of the trinomial graph method (analysis in Figure 5) that with only these two pieces of data the problem is indeterminate. It does not matter which question we ask, and in particular which question is asked above. So, this question could have as many answers as hypotheses that might be formulated about the accuracy of the test, and among them is that which presupposes that the test correctly diagnoses people who have the disease, that is to say, $p(+|A) = 1$.

If we then take an arbitrary hypothesis about the value of $p(+|A)$ into account, the problem now turns to the following:

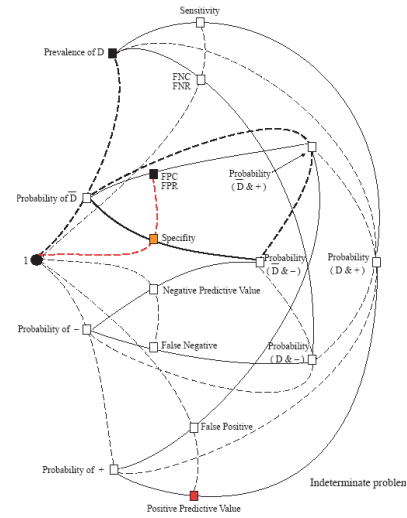


Figure 5. Indeterminate problem
- Click to enlarge.

Modified disease problem. Given $p(A)$, $p(+|\bar{A})$ and $p(+|A)$ —calculate $p(A|+)$.

This problem belongs to the (L_2, C_1, T_1) -triple. Again, the analysis and synthesis may be applied (see Figure 6). Moreover, we can also observe how the answer to this question depends on the third known item of data — $p(+|A)$ — necessarily imported by the solver of the problem.

Consequently, in our opinion, the answer that some participants gave could also be caused by not having enough information, misinterpreting the false positive rate and confusing it with the false positive, or both (see Figure 6 to the right).

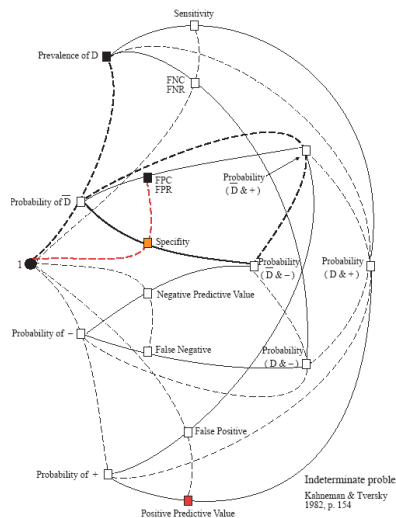


Figure 6a. Modified disease problem –
Click to enlarge.

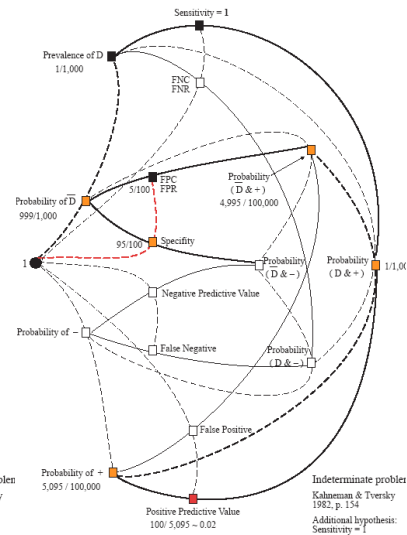


Figure 6b. Misinterpretation of false
positive rate.

Situations, contexts, and problems

A school-word problem can be seen as an instance of something more general. We use the term *situation* to mean something more general and *context* to refer a particular situation that is responsible for a restriction of the semantic field of a notion or concept (Puig & Cerdán 1988). The semantic field related to a concept contains a set of both mathematical and non-mathematical meanings. It allows us to make sense both of the events and event operations, as well as probabilities and probability relationships.

Given a general or particular situation or context, we have a *problem* each time a number of known probabilities and one unknown probability is sought. The graph in Figure 2 represents the most general situation we can consider from the 8 events, 16 probabilities and 18 relationships described beforehand. This general situation is a mathematical one, the context being mathematical-symbolic. If in this graph we have three dark nodes, then we have a problem, in a mathematical-symbolic context, each time we consider whichever of the white nodes is left.

A diagnostic test is such a situation. This situation can also be described by means of the aforementioned 8 events, 16 probabilities and 18 relationships. What is relevant now is how these events, probabilities and relationships come up in the situation. Testing something or someone by means of an unreliable test provokes uncertainty in its results; hence, all pre-test diagnostics must be revised and updated after the test is applied and its result is known.

A test, of course, can be applied to many things and in many particular situations; from manufactured products to determine their quality (defective or not); to witnesses in a trial to determine the reliability (or accuracy) of their testimony; to people in order to detect a disease; to women to detect pregnancy, and so on. So, a diagnostic test in health is a particular situation or context, as are diagnostic tests in law, in the manufacturing industry, and the other contexts above. All of them share the same meanings within the general situation, but differ from the meanings tied to the events, probabilities and their relationships in the specific context.

Moreover, depending on the context in which the problems are put forward, there will be problems that are more meaningful than others. Only in the general situation does it not matter which problem could make sense in being put forward.

In any situation, either general or particular, almost everything could be known: events, and the relationships between events and probabilities and the relationships between probabilities. When something is not known in a situation or context, we then say that we have a problem. The disease problem is a problem formulated in the context of diagnostic tests in health, and the

taxicab problem is a problem formulated in the context of diagnostic tests in law.

Exploring situation and context in which conditional probability is used could have, at least in our opinion, a twofold aim, depending on its uses. In research, full knowledge of the situation could allow researchers to formulate suitable problems in order to investigate the students' behaviour depending on any given situation. In teaching conditional probability, this knowledge could help teachers teach the topic in context by means of solving problems, by posing problems that help students to learn conditional probability in a more meaningful way, and by exploring phenomena in contexts where conditional probability is used.

We, as researchers, propose a method to analyse situations and contexts in which problems are examples within the situations and contexts themselves. This method completes our framework. The problems we are going to consider are, of course, ternary problems of conditional probability formulated in some specific context. This method also takes into account the notion of phenomena (Freudenthal 1983, p. 28) as well as the relationship between phenomena and organisational means that Freudenthal uses to describe a wider meaning of mathematical concepts and to help mathematics learning.

3. ANALYSIS OF STRUCTURES AND CONTEXTS

In this section we will report on some results from our recent studies. In the first part, the research took place in scenario 1. In the second part, in scenario 2. This research is continuing.

Situations, contexts, phenomena, and their influence

Among the few studies on conditional probability focusing on Freudenthal's phenomenological approach, that by Eddy (1982) may be referred to. However, it cannot be considered as a didactical-phenomenological approach, because it does not refer to the teaching of probability.

A more recent project is that of Carles & Huerta (2007). Here, the focus is on phenomenological analysis of problems, with reference to the framework discussed in earlier sections of this paper. The main question they tried to answer may be formulated as follows: As a problem is seen as an instance of something more general, legitimate issues for its research are:

- How are these more general situations and contexts related to the problem put forward?
- What phenomena in situations and contexts give rise to the problem?
- How do these phenomena manifest in the posed problem?
- How can they be organized by events, probabilities and relationships between them?

Due to the fact that we investigate problems in diverse teaching-learning environments, our information originates from several sources, including College textbooks or websites covering keywords like: Probability, Conditional Probability, Conditioned Probability, Bayes' Theorem. In all cases, the aim of our search was *conditional probability problems* used in the 2005–2006 school-year. The majority of conditional probability problems we found were ternary problems; these became our object of research. By analyzing the documents we classified the problems according to the following criteria, which will be clarified subsequently:

- Context (in which the problem is formulated)
- Phenomena referring to events (that is to say, organized by means of events)
- Phenomena referring to probabilities (that is to say, organized by means of probabilities)
- Specific terminology, classification (attending to the data and problem structure of the data in the text of the problem and the presentation format of the data)
- Specific teaching environment or reference.

Context. This is a particular situation in which problems are put forward. In a context, a particular concept such as conditional probability has a specific meaning or is used in a specific sense. For example, a diagnostic test is a situation. In our work, this situation is not only thought about for problems like those in this paper, but also for any situation where something has to be tested in order to determine if it fits a certain standard, which may be related to quality or health control parameters etc. Tests are not completely reliable. Therefore test results are always associated with certain risks, which are expressed by probabilities.

Phenomena (referring to events). In a particular context, phenomena relate to those statements that can be recognized as having an uncertain outcome. Such statements can be organized by means of reference sets (Freudenthal 1983, p.41), i.e., events in a probabilistic sense and operations between these events. Examples are “being ill”, “being ill and have a negative diagnostic test”, etc. Neither of these phenomena will be recognized as a “conditional event” even though it is usual to refer to them as if they were events. E. g., the phrase “being ill knowing that he /she has a negative diagnostic result” cannot be organized by means of a reference set.

Phenomena (referring to probabilities). In a particular context, apart from quantities, we refer to signs, words, phrases, and statements that express a measurement or the need for a measurement regarding the uncertainty of a phenomenon.

For example, *sensitivity* is a term that refers to probabilities. By means of the sensitivity of a test, usually in percentages or probabilities, we express the probability that the result of a test will be positive if the patient actually has the disease. *Prevalence of a disease* is another example. Its dictionary meaning is not related to probabilities. However, in contexts as considered in this paper, the word acquires a particular meaning in a probabilistic sense. So, the phenomenon of *prevalence of a disease* can be organized by means of probabilities usually expressed in percentages or numbers between 0 and 1. In a problem, the prevalence may refer to known or unknown data.

Exemplary phenomenological analysis

To illustrate matters, we analyze the following problems; FPR/FPC means false positive rate/coefficient – a person is wrongly diagnosed as positive; FNR/FNC means false negative rate (see the [glossary](#) for details).

P₁. It is known that in a certain city one out of every 100 citizens is a tubercular person. A test was administered to a citizen. When a person is tubercular, the test gives a positive result in 97% of cases. When he/she is not tubercular, only 0.01% of the cases give positive results. If the test is positive for this person, what is the probability that he/she actually is tubercular?

P₂. A diagnostic test for diabetes has an FPC (or FPR) of 4% and an FNC (or FNR) of 5%. Assume that the prevalence of diabetes in a specific town is 7%.

- a) What is the probability that a person is diabetic if his/her test was positive?
- b) What is the probability that a person is not diabetic if his/her test was negative?

P₃. A diagnostic test for uterine cancer has a false positive coefficient (FPC, or rate FPR) of 0.05 and false negative coefficient (FNC, or rate FNR) of 0.01. A woman who belongs to a group with pre-test probability of 0.15 for the disease has a negative result in her test. Calculate the probability that she is not ill.

P₄. A specific tuberculin test is applied to diagnose whether a person is infected by tuberculosis or not. The sensitivity and specificity of the test are very high, with values of 0.97 and 0.98 respectively. In a certain town there is a very high proportion of false positives; to be more specific, 0.9 of those who tested positive are in fact not tubercular. Calculate the prevalence of the disease.

Problem Situation Context	Phenomena referring to		Specific terms	Classification and data format
	events	conditional probabilities		
P ₁	<ul style="list-style-type: none"> • Be tubercular • Not be tubercular • Test gives a positive result 	<ul style="list-style-type: none"> • If a person is tubercular, the test yields a positive result with a high probability (in % format) • If a person is not tubercular, the test yields a positive result with a small probability (in % format) • If the test is positive there is a probability (<1) that the person actually is tubercular 	<ul style="list-style-type: none"> • Positive result in tests 	<ul style="list-style-type: none"> L₂C₁T₁ family (1, 0, 2) form • Rate • Percentages • $p(D +)$
	P ₂	<ul style="list-style-type: none"> • Be diabetic • A person does not suffer from diabetes but is positive on test • A person suffers from diabetes but is negative on test 	<ul style="list-style-type: none"> • FPC or FPR (False Positive Coefficient or Rate) • FPC or FNR (False negative Coefficient or Rate) 	<ul style="list-style-type: none"> • FPC or FPR • FNC or FNR • Prevalence of diabetes • Test is positive • Test is negative • Diagnostic Test
P ₃		<ul style="list-style-type: none"> • Not suffer from uterine cancer • Suffer from uterine cancer • Test positive in diagnostics without uterine cancer • Test negative in diagnostics with uterine cancer • A person tested negative who does not suffer from uterine cancer 	<ul style="list-style-type: none"> • False positive coefficient or rate (FPC or FPR) • False negative coefficient or rate (FNC or FNR) 	<ul style="list-style-type: none"> • False positive coefficient or rate • False negative coefficient or rate • Pre-test probability • Negative result in test • Diagnostic Test
	P ₄	<ul style="list-style-type: none"> • Be infected by tuberculosis. • Not be infected by tuberculosis. 	<ul style="list-style-type: none"> • Sensitivity • Specificity • False positive 	<ul style="list-style-type: none"> • Sensitivity • Specificity • False positive • Prevalence of disease • Tuberculin Test

Table 4. Aspects of the phenomenological analysis of the problems 1 to 4.

The result of the phenomenological analysis for the problems listed above is summarized in Table 4. The problems we mentioned above could be analyzed with the help of tables such as those numbered 5 and 6, and the graph in Figure 7. These tables organize the phenomenological analysis made in the diagnostic test in a health context and the graph allows us to apply the analysis-synthesis method and thereby make analytical readings of the problems.

In Table 5 we show the reference sets in this context, which are a means of organization for the phenomena we describe, particularly in a health context.

In Table 6, phenomena have been organized by means of probabilities of the aforementioned reference sets, now events in a mathematical sense, both by means of absolute probabilities and conditional probabilities. For that it is necessary to read data using a probabilistic sense.

Differences between Table 6 and a table for another context in the same situation, such as quality control for example, must be placed in phenomena and specific terms, but not in organizational means and format of data expression. Introducing organizational means as a column in Table 6 is based on the relationships between phenomena and organizational means, as Freudenthal (1983, p. 32) identified some years ago, such as a play of pairs phenomena-organizational means. In our case, this supposes reference sets for “to suffer an illness or disease”, “not to suffer this illness or disease”, “to test positive in the diagnosing procedure”, “to test negative”, designated by capital letters or specific signs: respectively “ D ”, “ \bar{D} ”, “+”, “-“. These sets refer to a set of people for whom some of the phenomena are present.

These sets, by themselves, are probabilistically unimportant, unless we consider them as Borel sets [2], and therefore we can consequently operate with them, and can take into consideration the complements, unions and intersections (Freudenthal 1983, p. 43). On the other hand, this supposes expressing judgements about these reference sets in terms of probabilities, such as this: “the probability of D is...” which is usually expressed by means of signs such as $p(D)$.

Pre-test				Post-test			
D	\bar{D}	+	-	$D \cap +$	$\bar{D} \cap +$	$D \cap -$	$\bar{D} \cap -$
Be ill	Not be ill	Test positive	Test negative	Suffer	Not suffer	Suffer	Not suffer
Be infected	Not be infected			from a specific disease			
Suffer	Not suffer	in the diagnostics, regardless of the health status		and test positive	and test positive	and test negative	and test negative
from a specific disease				(or similar phrases referring to the conjunction)			

Table 5. Reference sets in the Diagnostic Test in Health Context.

Phenomena (referring to probabilities)	Specific Terms	Organizational means	Data Format
Mistakes produced by test	FPR	$p(+ \bar{D})$	Necessarily in percentages and probabilities
	FNR	$p(- D)$	
Success produced by test "VALIDITY"	Sensitivity	$p(+ D)$	
	Specificity	$p(- \bar{D})$	
Mistakes produced in the diagnostic procedure "DIAGNOSTIC ERRORS"	negative diagnosis is false (often called "false negative")	$p(D -)$	
	positive diagnosis is false (often called "false positive")	$p(\bar{D} +)$	
Success produced in the diagnostic procedure "PREDICTIVE VALUES"	PPV or Positive Predictive Value Positive diagnosis is correct	$p(D +)$	
	NPV or Negative Predictive Value Negative diagnosis is correct	$p(\bar{D} -)$	
to have the disease (pre-test)	Prevalence of the disease	$p(D)$	Reasonably in natural numbers or absolute or natural frequencies
not to have the disease (pre-test)	Prevalence of no disease	$p(\bar{D})$	
Results from the diagnostic test	To test positive	$p(+)$	
	To test negative	$p(-)$	
to have the disease <i>and</i> to test positive in the diagnostic procedure	We didn't find out or it does not exist	$p(D \cap +)$	
not to have the disease <i>and</i> to test positive in the diagnostic procedure		$p(\bar{D} \cap +)$	
to have the disease <i>and</i> to test negative in the diagnostic procedure		$p(D \cap -)$	
not to have the disease <i>and</i> to test negative in the diagnostic procedure		$p(\bar{D} \cap -)$	

Table 6. Results of the phenomenological analysis in Diagnostic Test in Health Context.

The world of ternary problems of conditional probability in context

Figure 7 represents the world of ternary problems of conditional probability in the context of health. The graph shows all relationships. Lines representing ternary relationships in the graph are displayed in order to allow for global and local analysis. The conditional probabilities, in context, are:

sensitivity, specificity, false positive and false negative (rate).

They act on the prevalence of D or \bar{D} in order to update them according to the result of the diagnostic test.

We talk about *updating* an absolute probability $p(D)$ into new *conditional* probabilities of the disease.

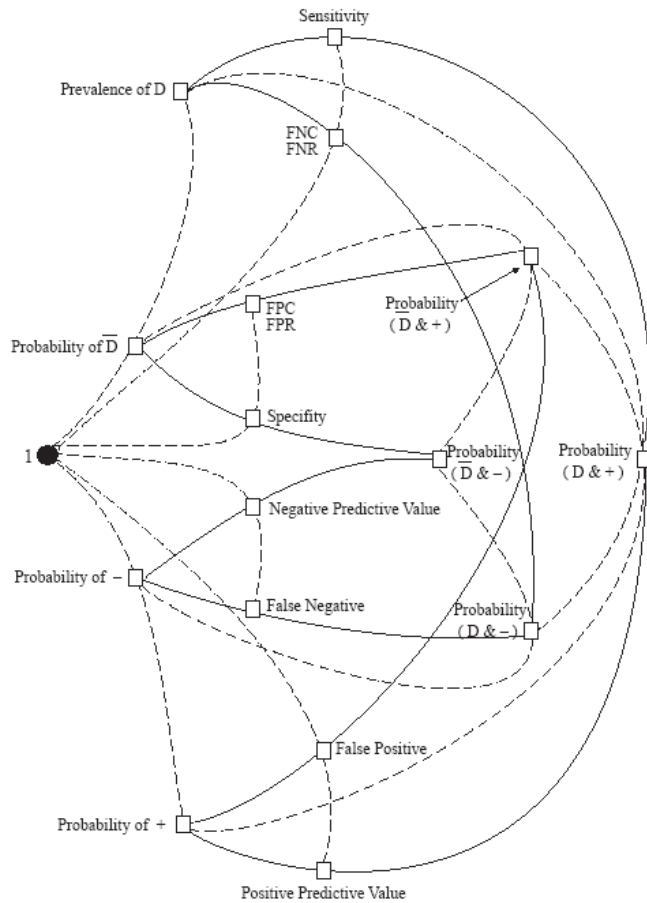


Figure 7. Graph of the world of ternary problems in the diagnostic test in health context – [click to enlarge](#).

Updating prevalence of D consists of solving two problems

Given prevalence of D , the probability of $+$ and the sensitivity of the test, calculate: a) the PPV b) the False negative.

Prevalence of D	Refers to	Context of health	Context of terms mathematics	Context of involved relationships health	Context of mathematics
Prior to test	All people	Prevalence of D	Absolute probability: $p(D)$	—	—
Probability of D conditioned by test result:					
Updated	A person who has been tested	a) Positive predictive value (PPV)	$p(D +)$	$= PPV = \frac{\text{prevalence } D \times \text{sensitivity}}{\text{probability } (+)}$	$= \frac{p(D) \times p(+ D)}{p(+)}$
		b) False negative	$p(D -)$	$= \text{False negative} = \frac{\text{prevalence } D \times (1 - \text{sensitivity})}{1 - \text{probability } (+)}$	$= \frac{p(D) \times [1 - p(+ D)]}{1 - p(+)}$

The analytical reading of the various mistakes and correct “statements” are summarized in Table 6 using probabilities. We have 8 conditional probabilities and 4 absolute or marginal probabilities, being complementary in pairs. Some of these complementary relationships are in themselves important in this context:

$$\text{FPR} + \text{specificity} = 1$$

$$\text{FNR} + \text{sensitivity} = 1$$

$$\text{probability } (-) + \text{probability } (+) = 1$$

The process of updating probabilities

The process of updating is expressed next to Figure 7. In order to talk about updating the probability of a disease by means of a diagnostic test, it is necessary to take into account the pre-test prevalence of the disease (usually taken as statistical information about a wider group to which the person to be tested belongs to) as well as specificity and sensitivity of the diagnostic test. Usually, the terms false negative and positive predictive value denote those updated probabilities; these terms refer to the post-test percentages or probabilities of suffering from the disease, reflecting the diagnostic result of the test and the pre-test prevalence of the disease.

Of course, all problems can be read in a similar way as will be done with the problem P_1 . In this case, the expression “one out of every 100 citizens is tubercular” could be read as follows: “in a town the prevalence of tuberculosis is one out of 100 citizens”. A mathematical reading of this supposes choosing a reference set T for the set of citizens suffering from tuberculosis. With a probabilistic sense one may affirm that $p(T) = 0.01$. The sensitivity of the test is expressed by means of the expression “When a person is tubercular the test gives a positive result in 97% of cases”, and the false positive rate by the expression “When he/she is not tubercular, only 0.01% of the cases give positive results”. That is, the sensitivity is 97% and the FPR is 0.01%.

The corresponding mathematical readings of the data, events and probabilities, are possible only if these readings are done with a probabilistic sense, because nothing is known about any sampling out of the population in which T is included. The probabilities to be positive have to be updated from new information about T and \bar{T} , resulting in that now the sensitivity is read $p(+|T) = 0.97$ and the FPR, $p(+|\bar{T}) = 0.0001$. The question in problem P_1 asks about the positive predictive value (PPV) of the test by means of the expression “If the test is positive for these people, what is the probability that he/she is tubercular?”, that is, about an updating of the prevalence of T from the additional information of the results of the test. In a probabilistic sense,

it asks about $p(T | +)$. Then, if we introduce the three known probabilities to the graph in Figure 7 by applying the analysis-synthesis method, we obtain:

$$PPV = \frac{\text{prevalence } T \times \text{sensitivity}}{\text{prevalence } T \times \text{sensitivity} + (1 - \text{prevalence } T) \times FPC}$$

we determine that PPV is a ratio between the quantities: the rate of the prevalence of tubercular people and positives related to the total of positives, whose mathematical reading is the well known Bayes' Formula:

$$p(T | +) = \frac{p(T) \times p(+ | T)}{p(T) \times p(+ | T) + [(1 - p(T)) \times p(+ | \bar{T})]}$$

The analysis of P_4 is different from P_1 – P_3 . While P_1 is structurally isomorphic to P_2 a); P_2 b) is isomorphic to P_3 ; however, P_4 belongs to the L_3 - family and is therefore not isomorphic to any of the others. P_1 , P_2 and P_3 are not globally isomorphic in pairs; there are some contextual differences that do not make them completely isomorphic.

Problem P_4 can be read in a probabilistic sense. The 3 conditional probabilities $p(B | A)$, $p(\bar{B} | \bar{A})$ and $p(\bar{A} | B)$ are known —calculate $p(A)$.

Introducing these data into the basic trinomial (see Figure 8) we observe that the graph has no entry point. That is, we have no line containing two dark nodes. So, we can not *destroy* the graph starting with known data. If we apply the analysis and synthesis method on it, the synthesis might explain why it is not possible to solve the problem with only the known data:

$$p(A) = \frac{[p(\bar{B} | \bar{A}) - p(A)] \times [1 - p(\bar{A} | B)]}{p(\bar{A} | B) \times p(B | A)}$$

This expression shows an answer to the question in the problem, taking the known data into account. However, it depends on unknown data. So, problem P_4 has an algebraic reading and, because of this, it necessarily obliges solvers to work with an unknown probability as if it were known. This probability could even be the probability that is asked for, or any other. The resulting algebraic equation that solves the problem will be dependent on the specific choices made.

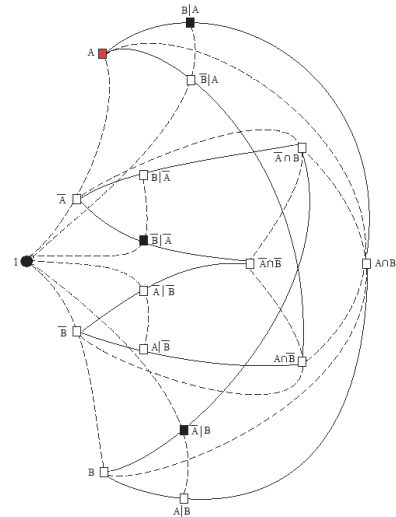


Figure 8. Trinomial graph for P_4
 - [Click to enlarge](#). Here is the [solution movie](#).

If we work on this problem in context, the graph with the analysis reveals information not only about relations between probabilities but about events, too. So, in order to calculate the prevalence of D , one needs to know the sensitivity of the test and to have information about people that both have the disease *and* test positive. But, at the same time, it is also required to have information about those people that both do not have the disease *and* test positive as it is already known that the specificity of the test is not 100%. Thus, to have information about the two intersection events will give us information about people testing positive.

However, we do not have this information about either of these events. If we did, the question in the problem could be solved. Therefore, we need to formally incorporate more information. Suppose then, that x out of every 100 citizens give positive results in a test. Hence, in the corresponding graph, an orange node turns black; we get an entry point and can destroy it (Click for a [movie](#) to follow the process of the analysis). As a consequence, we find an answer to the question. Of course, it depends on x :

$$p(D) = \frac{0.1x}{0.97}.$$

But, at the same time we also have

$$p(\bar{D}) = \frac{0.9x}{0.02}.$$

And because $p(D) + p(\bar{D}) = 1$, we can then calculate the value of x , in the resulting equation:

$$\frac{0.1x}{0.97} + \frac{0.9x}{0.02} = 1, \text{ being } x = 0,0222;$$

and for the required probability of D : we get: $p(D) = 0,0023$.

Some results about students' behaviour solving problems from the L_1 family

In scenario 2, very little research, if any, has been conducted taking into account the structures and contexts with which problems can be formulated as independent variables. If problem solving is seen as a process that starts when a problem is given to students and finishes when students answer the proposed question, then the variables describing this process, called process and product variables, are clearly dependent on those influential variables. These two dependent variables, namely process and product, are task variables.

In general, research on conditional probability pays attention to the participants' cognitive processes in solving tasks of assigning conditional probabilities to events. More recently, data format as an influential factor has become an object of interest. Evans et al (2000), Girotto &

González (2001) all agree that the presentation format of the data in the text of the problems influences students' success. They draw conclusions using *one* problem belonging to the (L_2, C_1, T_1) -triple, formulated in the diagnostic test in health context. They come to the conclusion that if the data is expressed in natural frequencies, as defined in Hoffrage et al (2002), students have substantially higher success rates.

But, unfortunately this research does not investigate the nature of the task and it does not restrict its conclusion to those problems that are open to a formulation with natural frequencies. Furthermore, Estrada & Díaz (2006) indicate difficulties in reading contingency tables with natural frequencies, and relate these difficulties to the understanding of everyday language used in formulating a conditional probability problem. The problem they put forward to participants is a redundant problem of the L_0 -family. All four questions involved are one-step problems, and only two of them are conditional probability problems but not one of them is a ternary problem.

Our approach is documented in Lonjedo (2007). We study students' behaviour solving a particular family of ternary problems of conditional probability, analyzing amongst other things whether it is dependent on problem structure, on the complexity of problems, or the context in which problems are formulated. Another influential factor under scrutiny is the data format in which probabilities are formulated. Students perform better on problems in which data is expressed in frequencies rather than in probabilities format. Related to this result, and as a consequence, we observed that students find the assignment strategy ^[3] more efficient than the calculation strategy when the data in the formulation of problems is expressed in a frequencies format. This is so even if students belong to higher levels (Huerta & Lonjedo 2006). We also conclude that students using an assignment strategy do not use given data in a probabilistic sense.

In ongoing research (see Huerta & Lonjedo 2007) we identify students' modes of reasoning depending on both the structure of data and the data format. Problems of the L_1 -family were used to design a problem solving test. We formulated 8 pairs of problems, corresponding to the 8 (L_1, C_i, T_k) sub-families from L_1 . Two problems from each one of the sub-families are structurally isomorphic. All of them were formulated in context. Eight of the problems were formulated in contexts in a situation we could call a statistical situation. The other eight problems were formulated in a mathematical situation but not in a symbolic context.

The statistical situation is quite familiar to the students because it describes situations in which a sample is divided in two, each one of which has two different characteristics. In a situation like this, it is reasonable to assume that the data format should be either absolute frequencies (referring to absolute and intersection probabilities), or natural frequencies (for the

conditional probability), or percentages (whatever the type of probability involved). Problems formulated in the mathematical situation, were reasonably formulated in terms of probabilities without the use of symbolical notation. As a consequence, the context in which the problems were put forward requires the use of expressions different from those used in their isomorphic problem.

Six tests of four problems each were administered to secondary school students (age 14-18) and mathematical college students with different levels of competence in mathematics. Each of the 16 problems was solved by more than 30 students, distributed among these school levels. Only students from upper secondary school and college (age 16+) are taught about conditional probability, whereas students from lower secondary school have received no such instruction.

In general, the success rates are very low (for details see Huerta & Lonjedo 2007). Problems with probabilities are significantly more difficult than those formulated in percentages. Formulated in percentages, problems from the (L_1, C_i, T_1) -triple seem to be more difficult than those from (L_1, C_i, T_3) . However, this result may be justified because the sample of students includes those with a very low competence in probability. Data in context could be misinterpreted and, as a consequence, misused. In the problems, a conditional probability has always been involved, either as known or unknown data, or both, and this might produce confusion among students.

We were not only interested in the students' success (a research variable related to the product variable), but also in the detailed resolution processes, whether successful or not. Therefore we have designed a set of descriptors to analyze students' solving behaviour (Huerta & Lonjedo 2007). For successful students we distinguish four types of thinking (see Fig. 9).

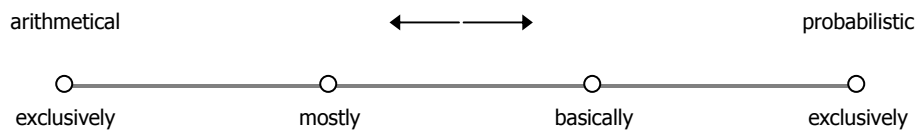


Figure 9. Four categories of thinking along the opposite pair of 'arithmetical' and 'probabilistic'.

This classification depends on whether students use data in the text of problem without a probabilistic sense, in a transition towards a probabilistic sense, or with a probabilistic sense. For unsuccessful students we can relate their difficulties,

- to semantic variables, mainly in expressions about conditionality,
- and mistakes related to these semantic difficulties that made the students misinterpret data referring to conditional probability.

Our research indicates that thinking processes are strongly related to data format and this is related to the context in which the problem is put forward. In order to confirm some of these hypotheses, we designed a new test with 6 reformulated problems from the initial test of 16. Reformulating implies a new problem in the same context, structurally isomorphic to the old one, but this time using a new data format, both for events and probabilities. This can be more meaningful to the students and helpful in using data to find an answer to the problem. In order to do this, a more natural language in the context has to be used.

Students' success rates with these new problems with data in frequencies were significantly greater than in the first test. They used mainly arithmetical reasoning. Those who were successful with data given in percentages or probabilities used mainly probabilistic reasoning. Generally, in solving these problems students use data explicitly mentioned without translating them from one format into another. That is, they solve the problems in the context in which the problems are formulated. Only in rare cases did students translate frequencies into probabilities and use probabilistic reasoning, changing from a non-mathematical to a mathematical situation. These were mainly mathematics students. On the other hand, success rates for isomorphic problems increase if data in the formulation of the problem is expressed in frequencies and percentages and words referring to conditionality are avoided (e.g. *and* and *also* are words that refer to an intersection event).

4. CONCLUSION

As mentioned in the introduction, research on problem solving can be carried out in more than one scenario depending on the different aspects that are taken into account. Here, the aspects are only problems, problems-students or problems-students-teachers, therefore giving us three possible scenarios.

It is reasonable that conditional probability problem solving is not only dependent on the nature of data (data format) but on what the data is and its relation to itself and to the question which is put forward. It is just this, the data, events, probabilities, and relationships that we call the structure of problems. So, from a research point of view, the structure of a conditional probability problem might be considered as an independent variable in every investigation on this topic, in which participants have to solve tasks related to conditional probability. It seems to us that this is not done in the research reported in literature. As a consequence, how is the students'

behaviour dependent on the structure of the posed problem? This is a research question that has not yet been dealt with and answered.

The framework we present in this paper has been designed for empirical research on probability problem solving and basically consists of:

- identifying a *particular world* of conditional probability school-problems in which data, both known and unknown, are related by means of ternary relationships;
- studying these problems in a symbolic and mathematical context, determining different *forms of known data* in which problems may be formulated, and *families of problems* depending on types of known data and questions in problems considered as unknown data;
- introducing the *trinomial graph* as a methodological tool, which serves for various purposes; it is the key concept for determining isomorphism between problems; it may help defining the complexity of problems, depending on the minimum number of relationships required to solve a problem; it may also help to resolve problems;
- introducing the *analysis and synthesis*, an innovative method for the analysis of conditional probability problems, allowing us to analyse and identifying problems that have an *arithmetic reading*, as problems P_1 – P_3 do; or have an *algebraic reading*, as problem P_4 does; or are *under-determined* (the disease problem) or *redundant* (the taxicab problem).

Moreover, it is reasonable that students' behaviour in solving problems of conditional probability is not only dependent on the *structure* of the problem but also on the *context* in which problems are put forward. Some studies report on such a dependence, but not as an independent variable. The consideration of contexts as an independent variable in research, as well as the notion of context itself, is normally used in a different sense from that presented here. So, our framework is completed with the following issues on context based on Freudenthal's didactical phenomenology:

- To distinguish between situations, contexts and problems: diagnostic test situation, diagnostic test in health context, problems P_1 – P_4 in this context.
- To consider didactical phenomenology as a means to analyse problems both in situations and contexts,
- To elaborate schemes for the phenomenological analysis.

The research in scenario 2, on conditional probability problem solving that we are currently carrying out, considers product and process variables (related to students' productions

when they are solving problems) as dependent variables of these two influential factors: structure and context. These factors can be divided into more specific variables that make empirical research more coherent. For instance, we could investigate whether students' behaviour in solving conditional probability problems from the L_0 -family depends on the context in which the problem is formulated. However, we do not abandon other independent variables, such as data format, about which we already know some aspects of students' behaviour (e. g. Lonjedo 2007).

On the other hand, the problems in our studies may be considered as problems of applications for conditional probability, by which we mean problems that are solved *after* formal teaching of the concepts. But, teaching of this topic based on an exploration of the phenomena involving its use in different contexts is precisely the opposite from formal teaching. Hence, as Freudenthal (1983) suggests, teaching the topic of conditional probability in secondary school might begin with students exploring phenomena involved with the topic. This could be followed by teaching the formal topic as a means of organizing the phenomena previously studied. The problems in this paper could be an example of this. While in secondary school the focus should remain on contexts, students at college level should learn about the formal concepts, not as concepts per se but as a means of modelling already familiar problems within a context. Our ideas about a phenomenological-based teaching are reported in Carles & Huerta (2007).

One of the most commonly used contexts has been analyzed in this paper. We termed it the diagnostic test. It can be recognized in several different settings: in textbooks, in research on students' behaviour in solving conditional probability problems, and so on. Generally, data and relationships between data are not previously analyzed in relation to the context. But, if, when thinking about teaching conditional probability, we previously analyze problems as suggested in this paper, we can determine what type of problems can reasonably be put forward to our students at every level of education. We can also understand in which context these problems must be stated in order to improve the students' understanding of conditional probability.

The world of conditional probability problems in a particular situation called diagnostic test combines both professional and educational use. The question is "Which teaching model do we choose in order to improve the students' literacy on conditional probability?" Is it based on a formal approach, for instance, first Bayes' formula and then applications via solving problems? Or is it a phenomenological-based model that focuses first on exploring phenomena via solving problems and only after that, searches for means of organizing all these phenomena? In this paper we defend the second position. In this sense, the world of problems we present and the way in which we approach them can be considered as suggestions for designing a teaching approach.

REFERENCES

- B.O.E. (2006). Real Decreto 1513/2006, por el que se establecen las enseñanzas mínimas correspondientes a la Educación Primaria, *Boletín Oficial del Estado*, nº 293. Madrid, Ministerio de Educación y Ciencia.
- B.O.E. (2007). Real Decreto 1631/2006, por el que se establecen las enseñanzas mínimas correspondientes a la Educación Secundaria Obligatoria, *Boletín Oficial del Estado*, nº 5. Madrid, Ministerio de Educación y Ciencia.
- Carles, M. & Huerta, M. P. (2007). Conditional probability problems and contexts. The diagnostic test context. In D. Pitta-Pantazi & G. Philippou (Eds.), *European Research in Mathematics Education V. Proc. CERME 5* (pp. 702–710). Nicosia: The University of Cyprus.
- Carles, M. (2007). *Estudios preliminares de los problemas de probabilidad condicional en contexto. El caso del test de diagnóstico*, Memoria para el examen de DEA. Universitat de València. (Preliminary studies on conditional probability problems in context. The case of diagnostic tests, Memory of the pre-doctoral DEA exam. University of Valencia).
- Cerdán, F. & Huerta, M. P. (2007). Problemas ternarios de probabilidad condicional y grafos trinomiales. *Educación Matemática*, 19 (1), 27–62.
- Eddy, D.M. (1983). Probabilistic reasoning in clinical medicine: Problems and opportunities. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 249–267). Cambridge: Cambridge Academic Press.
- Estrada, A.; Díaz, C. (2006). Computing probabilities from two-way tables: an exploratory study with future teachers. *Proc. ICOTS-7*. Retrieved June 7, 2009, from <http://www.stat.auckland.ac.nz/~iase/publications/17/C413.pdf>.
- Evans, J., Handley, S. J., Perham, N., Over, D. E., & Thompson, V. A. (2000). Frequency versus probability formats in statistical word problems. *Cognition*, 77, 197–213.
- Freudenthal, H. (1983). *Didactical phenomenology of mathematical structures*. Dordrecht: Kluwer Academic Press.
- Gal, I. (2005). Towards “probability literacy” for all citizens: Building blocks and instructional dilemmas. In Graham A. Jones (Ed.), *Exploring probability in School: Challenge for teaching and learning* (pp. 39–63). Berlin: Springer.
- Giroto, V. & González, M. (2001). Solving probabilistic and statistical problems: a matter of information structure and question form. *Cognition*, 78, 247–276.
- Hoffrage, U.; Gigerenzer, G.; Graus, S. & Martignon, L. (2002). Representation facilitates reasoning: what natural frequencies are and what they are not. *Cognition*, 84, 343–352.
- Huerta, M. P. & Lonjedo, M^a A. (2006). The nature of the quantities in a conditional probability problem. Its influence on the problem solving behaviour. In M. Bosch (Ed.), *European Research in Mathematics Education IV. Proc. CERME 4* (pp. 528–538). Barcelona: Fundemi IQS – Universitat Ramon Llull.
- Huerta, M.P. & Lonjedo, M.A. (2007). The same problem in three presentation formats: Different percentages of success and thinking processes. In D. Pitta-Pantazi & G. Philippou (Eds.), *European Research in Mathematics Education V. Proc. CERME 5* (pp. 732–741). Nicosia: The University of Cyprus.

- Lonjedo, M. A. (2007). *Análisis de los problemas ternarios de probabilidad condicional de enunciado verbal y de sus procesos de resolución* (Analysis of the word ternary problems of conditional probability and their resolution processes). Doctoral Dissertation. Universitat de València.
- NCTM, (2000). *Principles and Standard for school mathematics*. Reston, VA: N.C.T.M. Retrieved June 7, 2009 from <http://standards.nctm.org/>.
- Puig, L. & Cerdán, F. (1988). *Problemas aritméticos escolares*. Madrid: Síntesis.
- Shaughnessy, J. M. (1992). Research in probability and statistics: Reflections and directions. In D. Grouws (Ed.), *Handbook of Research on Mathematics Teaching and Learning* (pp. 465–494). New York: MacMillan.
- Tversky, A. & Kahneman, D. (1983). Evidential impact of base rates. In D. Kahneman, P. Slovic & A. Tversky (Ed.), *Judgment under uncertainty: Heuristics and biases* (pp. 153–160). Cambridge: Cambridge Academic Press.
- Watson, J. (2005). The probabilistic reasoning of middle school students. In Graham A. Jones (Ed.), *Exploring probability in School: Challenge for teaching and learning* (pp. 145–169). New York: Springer.
- Watson, J. (2006). Chance – Precursor to Probability. In J. Watson (2006), *Statistical Literacy at School* (pp. 127–185). Mahwah, NJ: Lawrence Erlbaum Associates.

NOTES

- [1] EDU2008-03140/EDU Project, funded by Ministerio de Ciencia e Innovación, España.
- [2] Consequently, it is time to use signs from set theory in order to denote events. For example, changing the expressions from “to have no disease”, or “no D ” to the complementary set \bar{D} .
- [3] Any problem belonging to the L_0 -family may be re-formulated in such a way that all information can be displayed in a 2 by 2 contingency table format with frequencies if the context in which the problem is put forward allows us to do it. These problems will be conditional probability problems if at least one conditional probability is asked for. To answer a question like this may be an assignment matter, that is to say assigning to it a ratio between two numbers in the table. In this case, calculations with probabilities may not be required at all as in the taxicab problem in Shaughnessy (1992, p. 471).

ANNEX – ALL APPENDICES

[Glossary of terms](#)

All other appendices to this paper may be found from [this link](#).

Author : M. Pedro Huerta
E-mail : manuel.p.huerta@uv.es
Address : Departament de Didàctica de la Matemàtica, Universitat de València. Spain