Problems of Forensic Sciences 2006, LXV, 53–67

ISSN 1230-7483

STATISTICS IN FORENSIC SCIENCE. PART I. AN AID TO INVESTIGATION

Colin G. G. AITKEN

School of Mathematics, The University of Edinburgh, Edinburgh, United Kingdom

Abstract

This is the first of two papers describing roles for statistics in forensic science. Four applications for statistics in crime investigations are described. The outcomes of these applications are not of themselves evidence but can aid in investigations from which evidence may be derived and evaluated using the ideas described in the second paper of the pair.

Key words

Offender profiling; Pre-assessment; Relevant propositions; Familial relationships; DNA; Sample size estimation.

Received 16 April 2006; accepted 15 May 2006

1. Introduction

There are two main roles for statistics in forensic science. The first role is taken during the investigatory stage of a crime before a suspect has been identified and statistics can be used to assist in the investigation. The second role is taken during the trial stage. There is a defendant and statistics can be used to assist in the evaluation of the evidence. These two roles are discussed in two separate papers in this volume of *Problems of Forensic Sciences*. This is the first of the two papers and discusses the role of statistics as an aid to investigation. The second paper [2] discusses the role of statistics as an aid to the evaluation of evidence.

Four aspects of the use of statistics as an aid to investigation are considered here. These are:

– offender profiling;
– pre-assessment and relevant propositions;
– familial relationships;
– sample size estimation.

2. Offender profiling

Offender profiling, or specific case analysis, is the name given to the procedure by which characteristics of the crime are used to predict characteristics of the offender. The prediction is uncertain; it is probabilistic. Conditional probabilities are associated with certain characteristics of the offender, conditional on certain characteristics of the crime and its scene. It is hoped that these probabilistic predictions will assist the investigators in their prioritisation of resources in that the predictions will suggest directions in which they guide the investigation so as to optimise the probability of identifying the offender.

The characteristics of the crime which can be considered may include characteristics of the victim and of the crime and its scene and observations by the victim of the offender. Characteristics of the victim which may be considered include age, sex, and occupation. The victim may give an age estimate of the offender. Characteristics of the crime and its scene may include the cause of death, the location and the place at which the victim was last seen. The characteristics of the of-

fender may include age, marital status, the relationship to the victim, whether there are previous convictions or not. It may not be the case that sex is a characteristic that it is needed to predict as many of the cases in which profiling is used are sexual in nature and for the vast majority of these the offender is male.

2.1. Logistic regression

The techniques which may be used include epidemiological techniques. An example of the method is given in [3, 4] in the context of sexually motivated child murders. Consider the prediction of an offender's marital status, which, for the purposes of illustration, is considered as a binary variable. The two possible outcomes are "living with partner" or not. Let p be the probability that the offender is living with a partner. A score function is derived which relates the log odds that the offender is living with a partner, $\log_e \frac{p}{1-p}$ with the sum of the scores of those crime characteristics in the model that are present in the crime. The crime characteristics in the overall model are labelled, for a particular crime as 1 if present and 0 if absent. Scores are values associated with each of these crime characteristics. Note that the log odds are used rather than p itself. This is because p only takes values in the interval {0, 1} which provides a difficult constraint when a regression model is being fitted. The transformation to log odds $\log_e \frac{p}{1-p}$ loses no information as it is possible given a value for the log odds to determine the corresponding probability but the transformation provides a new variable which takes values from to + and so is not constrained.

Mathematically, consider k characteristics x_1 , …, x_k with associated scores , when all characteristics are absent, and $_1$,…, $_k$ for each of the k characteristics. Then,

$$\log_e \frac{p}{1-p} \qquad x_1 \ \ldots \ _k x_k. \qquad \{1\}$$

For example, if $k = 2$, if x_1 denoted the sex of the victim with value 0 for a boy and 1 for a girl and if x_2 denoted the age of the victim, under 11 years old (0) or 11 or more years old (1), then:

a) for a victim who was an 8-year-old boy,

$\log_e \frac{p}{1-p}$ and

b) for a victim who was a 12-year-old girl,

$\log_e \frac{p}{1-p}$.

From both of these equations it is possible to determine a value for p, the probability the offender is living with a partner.

During the course of research investigating the ideas using examples from sexually motivated child murders, a crime of a similar type to the cases under investigation was committed and for which a person was found guilty. The details were entered into the logistic regression model with the results given in Table I[1].

2.2. Bayesian networks

A Bayesian network is a graphical representation of the relationships amongst the various characteristics of offender, victim and crime. A graph, in this context, is a set of nodes and directed arcs. Each node represents a particular characteristic. Two nodes are linked by a directed arc whose direction represents an influential relationship. The absence of an arc between two nodes implies that the two characteristics associated with these nodes are conditionally independent of each other, that is, they are independent conditional on knowledge of the values of the other characteristics. There is also a restriction that the directed arcs cannot form a closed loop so that it cannot be possible to start from a particular node and follow arcs to return to that node.

Consider Figure 1, taken from [4]. The arrows linking victim sex to offender marital status and victim age to offender marital status are a reflection of the belief that the sex and age of a victim in a sexually motivated child murder are dependent on the marital status of the offender. There is no direct arc joining offender marital status to offender preconvictions. This is indicative of the belief that the marital status of an offender is independent of whether he has previous convictions or not, once the other characteristics (age and sex of victim, the method of killing, the place where the victim was last seen and the relationship between the victim and offender) have been accounted for. The choice of the number of characteristics and the characteristics themselves can involve some subjectivity. The same example of sexually motivated child murders with a ten-node network is discussed in [5].

[1] Reasons of confidentiality prohibit publication of the scores and but further details are available from the Communication Development Unit, Research Development and Statistics Department, Home Office, Room 264, 50, Queen Anne's Gate, London SW1H 9AT

TABLE I. PREDICTED AND ACTUAL OUTCOMES OF FIVE OFFENDER CHARACTERISTICS

Predicted outcome	Probability	Actual outcome	Correctness
Age 0 20	0.65	21	Reasonable
Single	0.91	Single	Correct
Known to victim	0.96	Acquaintance	Correct
Point of contact of offender and victim within 5 miles of offender's residence	0.79	Yes	Correct
Previous convictions	0.92	Yes	Correct

Reproduced by permission of the Forensic Science Society.

TABLE II. PROBABILITIES FOR OFFENDER CHARACTERISTICS FOR A FEMALE VICTIM, AGED 0 7 YEARS, FOUND STRANGLED IN HER OWN HOME AND OUTWITH HER OWN HOME

Characteristic	Outcome	Probability		
		Initial	Revised	
			In own home	Outwith own home
Living with partner	Yes	0.24	0.33	0.36
	No	0.76	0.67	0.64
Relationship	Known to victim	0.57	0.66	0.11
	Unknown to victim	0.43	0.34	0.89
Previous convictions	Yes	0.73	0.73	0.70
	No	0.27	0.27	0.30

Reproduced by permission of the Forensic Science Society.

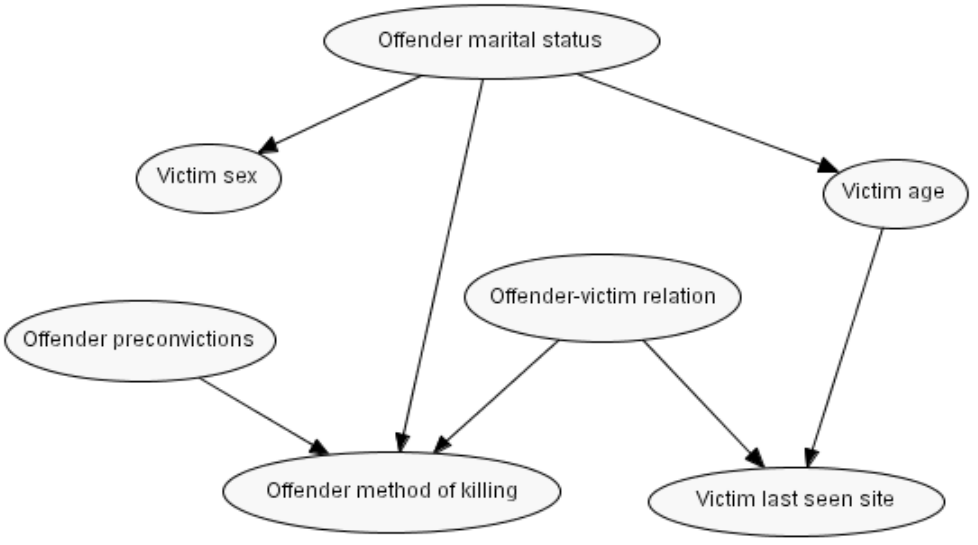

Fig. 1. A seven-node network showing the relationships amongst three offender characteristics (previous convictions, relationship to victim and marital status) and four victim and crime scene characteristics (age, sex, method of killing and site last seen). Six of the nodes are binary nodes with the following response: offender martial status (living with partner or not), victim sex (male or female), victim last seen site (home or elsewhere), offender preconvictions (yes or no), offender method of killing (strangulation or other), offender-victim relation (known or stranger). One node has three categories: victim age (0 7, 8 12, 13+) (reproduced by permission of the Forensic Science Society).

There are three stages to the construction of a network. The first involves the specification of the characteristics (nodes) to be included. This can be done by the statistician in consultation with detectives experienced in the type of crime for which the network is needed. The second stage considers the relationships amongst the characteristics, namely which nodes are to be joined by arcs. Finally, in the third stage, values for the conditional probabilities as determined by the locations of the arcs have to be determined. This may be done with reference to a dataset of characteristics of the crimes or, subjectively, in consultation with experienced detectives.

Once the network has been constructed it is possible to use it to provide probabilities for offender characteristics given victim and crime scene characteristics. Two examples are given in Table II. Software for the provision of networks which can analyse nodes of categorical or normally distributed continuous data is available from http://www.hugin.com.

The main change in the revised probabilities from the initial probabilities arises with the relationship between the offender and the victim. If the victim is murdered in her own home the probability that the offender is known to the victim is 0.66, i.e., the odds are 0.66/0.34 or approximately 2 to 1 on that the offender is to be found amongst those people known to the victim. However, if the victim is murdered outwith her own home, the probability that the offender is known to the victim is 0.11, i.e., the odds are 0.11/0.89, or approximately 8 to 1 against that the offender is to be found amongst those people known to the victim.

2.3. Summary

Logistic regression and Bayesian networks are two ways in which statistics can aid an investigation. It is stressed that these methods are aids to an investigation and are not evidence in themselves. They can provide assistance to investigators to help them identify suspects. However, they are not evidence and the output of their use should not be produced as such in court.

3. Pre-assessment and relevant propositions

Evidence is evaluated through the use of a factor known as the likelihood ratio which converts the odds in favour of one proposition (often known as the prosecution proposition) relative to another proposition (often known as the defence proposition), prior to presentation of the evidence (prior odds) into odds in favour of the proposition posterior to presentation of the evidence (posterior odds). Some details are given in [2]. Before the evidence is evaluated it is necessary to consider the nature of the two propositions, bearing in mind their relevance to the case under investigation. This consideration takes the form of what is known as a pre-assessment of the case, a procedure for which the development of the ideas may be consulted in [11, 12, 13, 15, 16].

A hierarchy of propositions has been developed. This hierarchy has three levels. Level 3 is the offence or crime level, level 2 the activity level and level 1 the source level. Another level has been added below the source level. This has been termed sublevel 1. The need for such a sublevel is illustrated with a DNA example [8]. DNA is found at a crime scene but it is not always certain from what body fluid the DNA may have come. A level 1 (source level) proposition would be that "the semen came from the suspect" (i.e., the suspect is the source of the semen). The sublevel I proposition would be that "the DNA came from the suspect". A level 2 (activity level) proposition for this DNA example would be that the suspect was present at the crime scene. A level 3 (crime level) proposition would be that the suspect committed the crime.

The case pre-assessment process can be summarised by the following steps:

1. collection of information that the scientist may need about the case;
2. consideration of the questions that the scientist can reasonably address, and, consequently, the level of propositions for which the scientist can reasonably choose to assess evidence;
3. identification of the relevant parameters which will appear in the likelihood ratio;
4. assessment of the strength of the likelihood ratio expected given the background information;
5. determination of the examination strategy;
6. conduct of tests and observation of outcomes;
7. evaluation of the likelihood ratio and report of the value.

An illustration of the use of the first four of these steps in practice for an example of fibres analysis is given in [6] using an example from [9]. Many of the details from the example as given in [6] are omitted here to emphasise the points regarding the pre-assessment process:

1. Collection of information that the scientist may need about the case. There was a post office robbery. Witnesses said that one of the men was wearing a dark green balaclava mask Some distance from the office a dark green balaclava was found. Mr *U* was arrested the following day. He denied all

knowledge of the incident. Reference samples of fibres from the balaclava were taken as well as combings from his head hair. Mr *U* has not yet been charged with the robbery because there is very little evidence against him.

2. Consideration of the questions that the scientist can reasonably address, and, consequently, the level of propositions for which the scientist can reasonably choose to assess evidence. A pair of source level propositions would be (a) that fibres in the hair combings from Mr *U* came from the balaclava and (b) that the fibres in the hair combings from Mr *U* came from some other garment or fabric. A pair of activity level propositions would be (c) that Mr *U* wore the mask at the time of the robbery and (d) that Mr *U* has never worn the mask. Consideration of the activity level propositions requires the scientist to have access to good background information on the offence. For example, this could include the time of the offence, the time of arrest, and the time at which the samples were taken.
3. Identification of the relevant parameters which will appear in the likelihood ratio. Consideration of the activity level requires information on the transfer, persistence and recovery of fibres. Data on fibre transfer to hair and on persistence are available (for example, [7, 10, 17]). With few data on these parameters, such as may be the situation when considering evidence on hairs rather than fibres, then it is necessary to restrict consideration to the source level proposition.
4. Assessment of the strength of the likelihood ratio expected given the background information. Various possibilities are available for the outcome of the analysis of the hair combings in comparison with the fibres on the green balaclava. It is assumed that either no group of fibres is found or only one group of fibres is found. If one group of fibres were found it could either match (in some sense) or not match the fibres in the green balaclava. If the group did match then it could either be a small group or a large group. Formulae for the probabilities of these possible outcomes, under each of two propositions at the activity level, (a) that Mr *U* wore the mask at the time of the robbery and (b) Mr *U* has never worn the mask are given in [6]. Likelihood ratios using probabilities proposed by [9] are also given in [6]. If no group of fibres is found in the hair combings, these probabilities give a value of the likelihood ratio of 100 in support of the proposition that Mr *U* has never worn the mask. If one group of fibres is found and it does not match the fibres in the green balaclava that also has a likelihood ratio of 100 in support of the proposition that Mr *U* has never worn the mask. If a small group of matching fibres is found then there is a likelihood ratio of about 3 in support of the proposition that Mr *U* wore the mask at the time of the robbery. If a large group of matching fibres is found then there is a likelihood ratio of about 840 in support of the proposition that Mr *U* wore the mask at the time of the robbery.

These likelihood ratios then provide information for consideration of the question "is it useful to proceed with the analysis of the hair combings?"

The first four steps of the procedure described above are part of the investigatory stage and are another example of the role of statistics in such a stage of forensic science.

4. Familial relationships from DNA

Suppose a crime has been committed. A DNA profile has been obtained which can be deduced to be that of the criminal. No match is obtained from a DNA database. The problem is to evaluate the evidence of a match of a relative of that person on the database.

Suppose the relative is a brother. Denote the member of the database as X and the brother of X as B. Consider a four allele locus. Alleles are denoted with lower case Roman letter a, b, c and d. The corresponding profile frequencies are denoted p_a, p_b, p_c and p_d. It is assumed these frequencies are known for each loci and allele under consideration. It is possible to determine expressions for $Pr((X \text{ is } (y, z) \mid B \text{ is } (w, x))$, assuming first that X is a brother of B and second that X and B are unrelated.

Table III contains the probabilities for one locus with four alleles. As an illustration as to how the analysis works, assign numerical values to the probabilities for the four alleles, say $p_a = 1/4$, $p_b = 1/12$, $p_c = 1/2$, $p_d = 1/6$. Assume the criminal (B) has a profile (a, b). Look at X on the database. Possible values for the likelihood ratio are given in Table IV. The likelihood ratio is the ratio of $Pr((X = (y, z) \mid B = (a, b))$; X and B brothers) to $Pr((X = (y, z) \mid B = (a, b))$; X and B unrelated).

For this example, the greatest support of the proposition that X and B are brothers is given to people X on the database whose alleles are (a, b), those of B. This is as expected.

When more than one locus is being considered:

– determine the likelihood ratio (LR) for each pair of alleles on each locus for each person X under con-

TABLE III. PROBABILITIES FOR DNA PROFILES FOR BROTHERS *B* AND *X* FOR A SINGLE LOCUS WITH FOUR ALLELES *a*, *b*, *c* AND *d* WITH PROFILE FREQUENCIES DENOTED p_a, p_b, p_c AND p_d. PROFILE FOR *B* IS DENOTED (*w*, *x*) AND FOR *X* IS DENOTED (*y*, *z*)

Alleles		$Pr(X \text{ is } (y, z) \mid B \text{ is } (w, x)$ and X and B are brothers)	Comments
B (w, x)	X (y, z)		
a, *b*	*c*, *d*	$\frac{p_c p_d}{2}$	No alleles in common between *B* and *X*
a, *b*	*b*, *c*	$\frac{p_c(3p_b \quad p_a)}{2}$	One allele in common
a, *b*	*a*, *b*	$\frac{1 \quad p_a \quad p_b \quad 2p_a p_b}{4}$	Both alleles in common
a, *a*	*a*, *a*	$\frac{(1 \quad p_a)^2}{4}$	Both alleles in common and homozygous
a, *a*	*c*, *d*	$\frac{p_c p_d}{2}$	No alleles in common and *B* homozygous
a, *a*	*a*, *b*	$\frac{p_b(1 \quad p_a)}{2}$	One allele in common and *B* homozygous
c, *d*	*a*, *a*	$\frac{p_a^{\ 2}}{4}$	No alleles in common and *X* homozygous
a, *b*	*a*, *a*	$\frac{p_a(1 \quad p_a)}{4}$	One allele in common and *X* homozygous

TABLE IV. DETERMINATION OF THE LIKELIHOOD RATIOS (*LR*) THAT THE EVIDENCE OF THE PROFILES OF *X* AND *B* IS SO MANY TIMES MORE LIKELY IF THEY ARE BROTHERS THAN IF THEY ARE NOT. SINGLE LOCUS WITH FOUR ALLELES *a*, *b*, *c* AND *d* WITH $p_a = 1/4$, $p_b = 1/12$, $p_c = 1/2$, $p_d = 1/6$. CONDITIONAL PROBABILITIES ASSUMING *X* AND *B* ARE BROTHERS ARE DERIVED FROM TABLE III. CONDITIONAL PROBABILITIES ASSUMING *X* AND *B* ARE UNRELATED ARE $2p_yp_z$ FOR HETEROZYGOTE (y, z) *X* AND p_y^2 FOR HOMOZYGOTE (y, y) *X*

(y, z)	$Pr(X = (y, z) \mid B = (a, b)$; *X* and *B* brothers)	$Pr(X = (y, z) \mid B = (a, b)$; *X* and *B* unrelated)	*LR*
(*a*, *b*)	0.3438	0.0417	8.24
(*a*, *c*)	0.2083	0.2500	0.83
(*c*, *c*)	0.1250	0.2500	0.50
(*a*, *a*)	0.0781	0.0625	1.25
(*a*, *d*)	0.0694	0.0833	0.83
(*b*, *c*)	0.0625	0.0833	0.75
(*c*, *d*)	0.0417	0.1667	0.25
(*b*, *d*)	0.0417	0.0278	1.50
(*b*, *b*)	0.0226	0.0069	3.28
(*d*, *d*)	0.0069	0.0278	0.25
Sum	1.0000	1.0000	

sideration, (X_1, …, X_n, say) in comparison with the known alleles of *B* for that locus;
- multiply these *LR*s together for the pairs of alleles for each person X_i ($i = 1, \ldots, n$) over all loci, assuming independence of loci (linkage equilibrium);
- the product is then the strength of the evidence linking each person X_i ($i = 1, \ldots, n$) on the database with *B*, the criminal.

Investigations may then be made of brothers, if any, of those people X_i which had the highest values for the product.

The use of familial testing has already been used with success in the UK[2].

5. Sample size estimation

Often, in the course of an investigation, it is necessary to examine a large consignment of identical looking items to determine the proportion which is illegal. For example, the consignments may be of:
- white tablets: how many to examine in order to make an inference about the proportion which are illicit?
- CDs: how many to examine in order to make an inference about the proportion which are pirated?
- computer images: how many to examine in order to make an inference about the proportion which are pornographic?

Statistically, these three situations are the same and the same approach may be used in each. There is an obvious saving in resources if it is not necessary to examine all the members of the consignment.

Assumptions are needed in order to construct a model from which inferences about the consignment may be drawn from the sampling procedure. The simplest procedure is obtained from the most restrictive assumptions and it is this which will be described here. It is assumed that:
- each member of the consignment may be classified into one of two and only two categories (a so-called binary classification);
- the consignment is homogenous in all respects, other than the one of interest;
- the probability that a member of the consignment belongs to a particular category is constant over all members of the consignment;
- the probability that a member of the consignment belongs to a particular category is independent of the assignment to a category of any other member of the consignment.

For white tablets, these assumptions mean that the tablets are categorised as either licit or illicit and that they all have the same colour, logo, texture, weight and shape. For CDs, these assumptions mean that the CDs are categorised as either legal or pirated and that they all look the same; they could have different labels on them for different recordings but the prior probability of a particular recording being pirated or not has to be constant for the different recordings. For computer images, these assumptions mean that the images are categorised as either pornographic or not and that, before opening, the files can be believed to have the same prior probabilities of being pornographic or not.

With these assumptions, a Bayesian procedure may be developed to determine the sample size necessary to satisfy a certain criterion. The criterion relates to a pre-specified proportion of the consignment which is illegal (a generic term used here to cover possibilities which are dependent on the context, such as illicit drugs, pirated CDs and pornographic computer images) and the probability with which it is desired to determine that the true proportion is greater than this pre-specified proportion. The pre-specified proportion is so-called because it is specified prior to inspection of the consignment. The sampling procedure has to be random. All members of the consignment have to have an equal probability of being selected for the sample, independent of all other members of the consignment.

A numerical example of the criterion, applied to a consignment of white tablets, would be to determine the sample size to be inspected so as to have a 95% probability that at least 50% of the consignment was illicit if all the tablets inspected were found to be illicit. In this example, the pre-specified proportion is 50% and the probability with which it is desired to determine that the true proportion is greater than this pre-specified proportion is 0.95. The answer to this criterion is 4. Thus, if four tablets are selected at random from the consignment, examined and all found to be illicit, then there is a 95% probability that at least 50% of the consignment is illicit. This sample size, 4, is independent of the size of the consignment; it is the same whether the consignment contains a 1000 members or a million members. It may be slightly different if the consignment is small, say less than 50, but the difference will be to make the sample size smaller, not larger.

The sample size 4 may appear to be too small to make the claim about the consignment that is inferred

[2] It was reported in the British press (*The Times, The Guardian)* on April 20th 2004, that a person had been jailed for manslaughter using evidence based on the link between DNA retrieved at the scene of the crime and the DNA profile of a relative of the accused.

from it. However, two points may be made. First, it is not a very big claim. For a consignment of one million tablets, it claims that there is 95% probability that at least half-a-million tablets are illicit. This still leaves some scope for there to be a large number of licit tablets. Secondly, consider the tossing of a coin. For each toss, there are two possible outcomes, what are called in the UK heads and tails, depending on which side lands uppermost. If the coin is fair, the two outcomes are equally likely and have a probability of 1/2 of occurring each time the coin is tossed, independently of the outcomes of previous tosses. The coin is tossed four times. If the coin is fair, the probability of achieving four heads on four tosses is $(1/2)^4$ which is 1/16 or 0.0625. This is a small probability, close to 0.05, the complement of 0.95, the 95% associated with the example of the white tablets. Thus, four tosses of a coin which all resulted in four heads may cast doubt on the assumption that the coin is fair, in a manner analogous to the result obtained from sampling four white tablets at random and discovering they were all illicit.

Further details of these ideas of sampling are given in [1] and in guidelines of the European Network of Forensic Science Institutes (ENFSI) Drugs Working Group, published in 2004 [14].

6. Conclusion

Four applications of statistics in the investigation of possible crimes have been described. It has to be remembered that these applications are to investigations. It is rare that their results would be used as evidence. For example, the offender profile of a criminal is not evidence to be presented in court to support the proposition that a defendant who fits the profile is the criminal. The pre-assessment of transfer evidence is not itself evidence to support a particular proposition. The result of the inspection of the evidence is an assessment of the strength of the evidence but the process by which it was decided to do the inspection is not evidence in itself. The use of familial testing in a DNA database to help identify suspects is not in itself evidence. The determination of a sample size is not in itself evidence.

The role of statistics in the evaluation of evidence is the subject of the second papers in this pair of paper on the role of statistics in forensic science.

References

1. Aitken C. G. G., Sampling how big a sample?, *Journal of Forensic Sciences* 1999, 44, 750 760.
2. Aitken C. G. G., Statistics in forensic science II: The evaluation of evidence, *Problems of Forensic Sciences* 2006, 65, 68 81.
3. Aitken C. G. G., Connolly T., Gammerman A. [et al.], Predicting an offender's characteristics: an evaluation of statistical modeling, Police Research Group, Special Interest Series Group 4; London 1995.
4. Aitken C. G. G., Connolly T., Gammerman A. [et al.], Statistical modeling in specific case analysis, *Science & Justice* 1996, 36, 245 255.
5. Aitken C.G.G., Gammerman A., Zhang G. [et al.], Bayesian belief networks with an application in specific case analysis, [in:] Computational learning and probabilistic reasoning, Gammerman A. [ed.], John Wiley & Sons, London 1996.
6. Aitken C. G. G., Taroni F., Statistics and the evaluation of evidence for forensic scientists, John Wiley and Sons, Chichester 2004.
7. Ashcroft C. M., Evans S., Tebbett I. R., The persistence of fibers in head hair, *Journal of the Forensic Science Society* 1988, 28, 289 293.
8. Buckleton J., Triggs C. M., Walsh S. J. [ed.], Forensic DNA evidence interpretation, CRC Press, Boca Raton 2005.
9. Champod C., Jackson G., European Fibres Group Workshop: Case assessment and Bayesian interpretation of fibres evidence, Proceedings of the 8th Meeting of European Fibres Group, Krakow 2000, 33 45.
10. Cook R., Webb-Salter M. T., Marshall L., The significance of fibres found in head hair, *Forensic Science International* 1997, 87, 155 160.
11. Cook R., Evett I. W., Jackson,G. [et al.], A model for case assessment and interpretation, *Science & Justice* 1998, 38, 151 156.
12. Cook R., Evett I. W., Jackson G. [et al.], A hierarchy of propositions: deciding which level to address in casework, *Science & Justice* 1998, 38, 231 239.
13. Cook R., Evett I. W., Jackson G. [et al.], Case pre-assessment and review of a two-way transfer case, *Science & Justice* 1999, 39, 103 122.
14. European Network of Forensic Science Institutes Drugs Working Group (2004), Guidelines on representative drug sampling, http://www.enfsi.org.
15. Evett I. W., Jackson G., Lambert J. A. [et al.], The impact of the principles of evidence interpretation and the structure and content of statements, *Science & Justice* 2000, 40, 233 239.
16. Evett I. W., Jackson G., Lambert J. A., More on the hierarchy of propositions: exploring the distinction between explanations and propositions, *Science & Justice* 2000, 40, 3 10.

17. Salter M. T., Cook R., Transfer of fibres to head hair, their persistence and retrieval, *Forensic Science International* 1996, 81, 211 221.

Corresponding author
Colin Aitken
The University of Edinburgh
School of Mathematics
Mayfield Road
Edinburgh EH9 3JZ, United Kingdom
e-mail: c.g.g.aitken@ed.ac.uk

Problems of Forensic Sciences 2006, LXV, 53–67

STATYSTYKA W NAUKACH SĄDOWYCH. CZĘŚĆ I. STATYSTYKA JAKO UŻYTECZNE NARZĘDZIE NA ETAPIE DOCHODZENIOWYM ŚLEDZTWA

1. Wprowadzenie

Metody statystyczne w naukach sądowych stosowane są zarówno w fazie dochodzeniowej śledztwa, czyli zanim podejrzany zostanie zidentyfikowany, jak też w fazie procesowej, kiedy podejrzany jest już znany. W pierwszym przypadku metody statystyczne znajdują zastosowanie jako narzędzie pomocnicze mające na celu dostarczenie informacji pozwalającej ustalić sprawcę przestępstwa. Natomiast w fazie procesowej metody statystyczne mogą być użyte w celu oceny wartości poszczególnych dowodów. Te dwa rodzaje wykorzystania metod statystycznych w naukach sądowych omówione zostały w oddzielnych artykułach opublikowanych w niniejszym numerze czasopisma *Z zagadnień nauk sądowych*. W pierwszym z nich omówiono rolę metod statystycznych jako narzędzia stosowanego na etapie dochodzeniowym śledztwa. Drugi artykuł [2] porusza z kolei problem użycia metod statystycznych w oszacowaniu wartości dowodowej różnego rodzaju śladów kryminalistycznych.

Można wyróżnić następujące zastosowania metod statystycznych na etapie dochodzeniowym śledztwa:

- profilowanie sprawcy;
- wstępne oszacowanie wartości dowodowej materiału dowodowego;
- ustalanie więzów pokrewieństwa na podstawie analizy profili DNA;
- oszacowanie wielkości próbki koniecznej do przeprowadzenia badań.

2. Profilowanie sprawcy

Profilowanie sprawcy to czynność, w trakcie której na podstawie przebiegu zdarzenia (przestępstwa) określana jest charakterystyka sprawcy. Należy zaznaczyć, że ustalone w trakcie procesu profilowania cechy sprawcy przestępstwa nie mogą zostać uznane za pewniki, lecz powinny być traktowane w kategoriach prawdopodobieństwa, a konkretnie prawdopodobieństwa warunkowego. Cechy te są bowiem uzależnione od okoliczności i miejsca popełnienia przestępstwa. Ponadto przewidywanie to może pomóc osobie prowadzącej śledztwo w określeniu kierunków, w których dochodzenie powinno zostać poprowadzone, by uzyskać największą możliwość zidentyfikowania sprawcy.

Do cech związanych z popełnieniem przestępstwa, które mogą zostać uwzględnione w modelu statystycznym stosowanym w procesie profilowania sprawcy, należą: charakterystyka pokrzywdzonego, charakterystyka miejsca zdarzenia oraz obserwacje dokonane przez ofiarę, które dotyczą cech napastnika. Charakterystyka pokrzywdzonego powinna obejmować informację o jego wieku, płci i zawodzie. Zarówno charakterystyka przebiegu przestępstwa, jak też i miejsca zdarzenia, winna zawierać np. informacje o przyczynie zgonu lub o miejscu, w którym pokrzywdzony był widziany po raz ostatni. Charakterystyka napastnika może zawierać jego wiek, status cywilny, relacje z pokrzywdzonym oraz informacje o jego karalności.

2.1. Regresja logistyczna

Metody statystyczne, które mogą zostać użyte w profilowaniu sprawcy, to między innymi te, które wcześniej znalazły zastosowanie w epidemiologii. Przykłady ich wykorzystania opisane zostały w innych pracach [3, 4] w kontekście morderstw dzieci dokonywanych na tle seksualnym. Dla przykładu rozważmy określenie stanu cywilnego napastnika, który rozpatrywany jest tutaj jako zmienna binarna, tzn. posiadająca tylko dwa możliwe stany: napastnik „ma partnera" lub napastnik „nie ma partnera". Niech p będzie prawdopodobieństwem, że napastnik „ma partnera". Uzyskane w takim przypadku wyrażenie, oparte na regresji logistycznej, łączy logarytm szans podany wyrażeniem $\log_e \frac{p}{1\ p}$ z sumą rezultatów przypisanych każdej z cech charakteryzujących przestępstwo, które uwzględnione zostały w stosowanym modelu statystycznym. Poszczególne cechy wchodzące w skład charakterystyki przestępstwa oznaczone są w ten sposób, że 1 oznacza sytuację, gdy dana cecha wystąpiła w analizowanym przypadku, a 0, gdy nie wystąpiła. Należy zaznaczyć, że w obliczeniach użyty jest logarytm szans, a nie bezwzględne wartości p, ponieważ p przyjmuje wartości tylko w przedziale $\{0, 1\}$, co w znacznym stopniu ogranicza stosowanie modelu opartego na analizie regresji. Transformacja wartości p do logarytmu szans $\log_e \frac{p}{1\ p}$ nie powoduje jednak utraty informacji o prawdopodobieństwie p, ponieważ możliwe jest podanie wartości tego prawdopodobieństwa, gdy znamy wartość logarytmu szans. Wykonywana transformacja p do logarytmu szans wprowadza nową zmienną, która przyjmuje

wartości w przedziale od 8 do +8 i tym samym nie jest niczym ograniczona.

Rozważmy k cech $x_1, \ldots, x_k$ i związany z nimi rezultat uzyskany wówczas, gdy wszystkie z rozpatrywanych cech nie wystąpiły w analizowanym przypadku oraz rezultaty $_1, \ldots, _k$, które uzyskano w przypadku zaobserwowania każdej z k cech. Uwzględniając powyższe informacje, uzyskamy:

$$\log_e \frac{p}{1 \quad p} \qquad x_1 \quad \ldots \quad {}_k x_k \,. \qquad \{1\}$$

Dla przykładu, gdy $k = 2$, a x_1 oznacza płeć pokrzywdzonego, przy czym wartość 0 oznacza chłopca, zaś 1 oznacza dziewczynkę i gdy x_2 oznacza wiek pokrzywdzonego (poniżej 11 lat przypisujemy 0, powyżej 11 lat przypisujemy 1), to wówczas:

a) dla pokrzywdzonego, który jest 8-letnim chłopcem, uzyskamy $\log_e \frac{p}{1 \quad p}$, natomiast

b) dla pokrzywdzonej, która jest 12-letnią dziewczyną, uzyskamy $\log_e \frac{p}{1 \quad p}$.

Podobne równania można uzyskać w przypadku, gdy ustalamy stan cywilny napastnika i na podstawie tych równań określamy wartość prawdopodobieństwa p, że podejrzany „ma partnera" lub go „nie ma".

Należy dodać, że podczas prowadzenia badań nad możliwością zastosowania powyższej metody w przypadku morderstwa dziecka o podłożu seksualnym wydarzył się w rzeczywistości podobny przypadek, zaś podejrzany został uznany za winnego[1].

2.2. Sieci bayesowskie

Analizę w oparciu o sieć bayesowską można zastosować również w procesie profilowania sprawcy przestępstwa. W takim przypadku sieć bayesowską traktować należy jako graficzną ilustrację relacji pomiędzy różnymi cechami opisującymi napastnika oraz ofiarę przestępstwa. Wykres tworzy zbiór węzłów i strzałek. Każdy z węzłów reprezentuje konkretną cechę. Dwa węzły mogą być połączone za pomocą strzałki, której kierunek obrazuje występującą pomiędzy nimi zależność. Brak strzałki łączącej dwa węzły wskazuje, że cechy reprezentowane przez te dwa węzły nie są od siebie zależne. Istnieje jedno ograniczenie w budowaniu sieci bayesowskich: kierunki strzałek nie mogą być ułożone w taki sposób, aby tworzyły zamknięte pętle. Innymi słowy, rozpoczynając przy jednym węźle (tzw. węźle początkowym) i podążając za kierunkiem strzałek, nie możemy dotrzeć do węzła początkowego.

Rozważmy sytuację zaprezentowaną na rycinie 1, którą zaczerpnięto z innej publikacji [4]. Strzałki wiążące płeć oraz wiek poszkodowanego ze stanem cywilnym napastnika odzwierciedlają tezę, że płeć i wiek poszkodowanego w przypadku morderstw dzieci na tle seksualnym są zależne od stanu cywilnego napastnika. Nie istnieje strzałka łącząca bezpośrednio stan cywilny sprawcy z jego karalnością w przeszłości. Oznacza to, że zakładamy, iż stan cywilny jest niezależny od tego, czy sprawca był wcześniej skazany przez sąd, czy też nie; jest także niezależny od innych cech, takich jak wiek i płeć poszkodowanego, sposób popełnienia morderstwa, miejsce, w którym ofiara była widziana po raz ostatni czy relacji panujących pomiędzy ofiarą i sprawcą. Wybór liczby oraz jakości rozpatrywanych cech może opierać się na subiektywnej ocenie problemu dokonywanej przez osobę analizującą go. Ten sam przypadek morderstwa na tle seksualnym, lecz opisany za pomocą sieci złożonej z dziesięciu węzłów, omówiono w literaturze [5].

Podczas tworzenia sieci bayesowskiej wyróżnić można trzy etapy. W pierwszym następuje wyszczególnienie cech (węzłów), które winny być uwzględnione. To zadanie statystyk powinien wykonać po konsultacjach z biegłym sądowym mającym doświadczenie i wiedzę w analizowanym rodzaju przestępstw. Drugi etap wymaga określenia relacji pomiędzy poszczególnymi cechami, tzn. określenia, które węzły mają być połączone strzałkami i jaki powinien być kierunek tych strzałek. W ostatnim, trzecim etapie, określane są wartości prawdopodobieństw zdarzeń opisanych w poszczególnych węzłach. Dotyczy to również prawdopodobieństw warunkowych wskazujących na możliwość wystąpienia jakiegoś zdarzenia pod warunkiem, że zaszło inne zdarzenie. Ustalenie tych prawdopodobieństw powinno być oparte na wiedzy uzyskanej w trakcie analizy odpowiednich baz danych lub też zostać ustalone w sposób subiektywny, po konsultacjach z biegłymi sądowymi posiadającymi wiedzę na temat analizowanego problemu.

Sporządzenie sieci umożliwia uzyskanie wartości liczbowych, które ułatwiają wytypowanie najbardziej prawdopodobnych cech sprawcy oraz cech charakteryzujących zarówno ofiarę, jak też miejsce zdarzenia, a ustalonych na podstawie informacji, które uwzględniono w sieci. Dwa przykłady takiego postępowania zamieszczono w tabeli II. Program komputerowy, który dokonuje analizy sieci bayesowskich opartych na węzłach opisanych przez dane dyskretne lub dane charakteryzowane rozkładem normalnym, jest dostępny na stronie http://www.hugin.com.

Główną zmianą zaobserwowaną w wartościach prawdopodobieństw uzyskanych po analizie przeprowadzonej w oparciu o sieć bayesowską w stosunku do pierwotnych

[1] Istnieje zakaz publikacji danych poufnych o wartościach i , ale pozostałe szczegóły uzyskać można w Communication Development Unit, Research Development and Statistics Department, Home Office, Room 264, 50, Queen Anne's Gate, London SW1H 9AT.

prawdopodobieństw jest wzrost prawdopodobieństwa zdarzenia, że ofiara znała sprawcę. Jeżeli ofiara została zamordowana w jej własnym domu, to wówczas prawdopodobieństwo, że ofiara znała sprawcę, wynosi 0,66, a tym samym szansa, że sprawca należy do kręgu osób, które ofiara znała, równa się od 0,66 do 0,34, lub w przybliżeniu 2 do 1. Natomiast gdy ofiara została zamordowana poza jej domem, to wówczas prawdopodobieństwo, że sprawca znał ofiarę, wynosi 0,11, a tym samym szansa, że sprawca nie jest osobą, która znała ofiarę, jest równa 0,11/0,89, lub w przybliżeniu 8 do 1.

2.3. Podsumowanie

Regresja logistyczna i sieci bayesowskie są metodami, za pomocą których statystyka może wspomóc osoby prowadzące śledztwo na etapie dochodzenia. Należy zaznaczyć, że metody te mają pomóc w dochodzeniu, tj. dać wskazówki, które pozwolą na zidentyfikowanie sprawcy, choć nie stanowią one dowodu, zaś rezultat ich zastosowania nie może być wykorzystany w sądzie.

3. Wstępne oszacowanie materiału dowodowego – faza dochodzeniowa śledztwa

Wartość dowodowa różnego rodzaju materiału dowodowego jest oszacowywana za pomocą ilorazu wiarygodności, w którym informacja o dowodzie rozważana jest w kontekście hipotezy prokuratora i hipotezy alternatywnej, zwanej hipotezą obrony, co również omówiono w odrębnej publikacji [2]. Zanim dowód będzie oszacowany, konieczne jest określenie obu hipotez z uwzględnieniem informacji o przebiegu przestępstwa. Proces ten zwany jest wstępnym oszacowaniem materiału dowodowego na etapie dochodzeniowym śledztwa. Więcej informacji o tej metodzie można znaleźć w licznych publikacjach [11, 12, 13, 15, 16].

Hierarchia hipotez posiada trzy poziomy. Poziom 3 związany jest ze sprawcą i (lub) przestępstwem, poziom 2 związany jest z aktywnością sprawcy, a poziom 1 z rezultatami badań materiału dowodowego (tzw. poziom źródła). W ramach poziomu 1 stworzono dodatkowy podpoziom 1. Konieczność jego uwzględnienia można zilustrować na przykładzie badania DNA [8]. Na miejscu zdarzenia znaleziono ślady nasienia, które posłużyły do oznaczenia profilu DNA. Hipoteza prokuratora w ramach poziomu 1 (poziom źródła) może być sformułowana następująco: „sperma pochodzi od podejrzanego", tj. podejrzany jest źródłem spermy. Natomiast w ramach podpoziomu 1 hipoteza prokuratora może być ujęta następująco: „DNA pochodzi od podejrzanego". Hipoteza prokuratora w ramach poziomu 2 (tzw. poziom aktywności) brzmi, że podejrzany był na miejscu przestępstwa. W ramach poziomu 3 (poziom przestępstwa) hipoteza prokuratora będzie taka, że to podejrzany dokonał zarzucanego mu czynu.

Proces wstępnego oszacowania materiału dowodowego można przedstawić sumarycznie w następujących punktach:

1. zebranie informacji o sprawie, które są istotne dla biegłego;
2. sformułowanie pytań, na które biegły może udzielić odpowiedzi oraz zaproponowanie hipotez, które może on rozpatrywać w celu oszacowania wartości dowodowej analizowanego materiału;
3. określenie cech, które powinny być uwzględnione w modelu służącym do wyznaczenia ilorazu wiarygodności;
4. oszacowanie spodziewanej wartości ilorazu wiarygodności, który można uzyskać na podstawie informacji uzyskanych o przestępstwie;
5. określenie strategii badań;
6. przeprowadzenie testów i analiza wyników;
7. oszacowanie ilorazu wiarygodności na podstawie uzyskanych rezultatów przeprowadzonych badań i jego interpretacja dla potrzeb wymiaru sprawiedliwości.

Realizację pierwszych czterech etapów przedstawiono poniżej na przykładzie analizy włókien, którą dokładniej omówiono w odrębnej publikacji [6] z uwzględnieniem danych zawartych w publikacji Champoda i Jacksona [9]. Wiele szczegółów związanych z zapożyczonym przykładem [6] zostało tutaj pominiętych w celu uproszczenia i klarownego przedstawienia istoty procesu oszacowania materiału dowodowego w fazie dochodzeniowej śledztwa.

1. Zebranie informacji o sprawie, które są istotne dla biegłego. Dokonano rabunku na poczcie. Naoczny świadek stwierdził, że jeden z mężczyzn nosił ciemnozieloną kominiarkę. W pewnej odległości od poczty została znaleziona ciemnozielona kominiarka. Następnego dnia *U* został aresztowany. Zaprzeczył, że ma cokolwiek wspólnego ze zdarzeniem. Pobrano próbkę porównawczą włókien z kominiarki, jak również wyczesano włókna z włosów z głowy podejrzanego. *U* nie postawiono na tym etapie śledztwa zarzutu udziału w napadzie na pocztę, ponieważ dowody przeciwko niemu były słabe.
2. Sformułowanie pytań, na które biegły może udzielić odpowiedzi i zaproponowanie hipotez, które może on rozpatrywać w celu oszacowania wartości dowodowej analizowanego materiału dowodowego. Na poziomie źródła można rozważyć w tym przypadku następujące hipotezy: (a) włókna wyczesane z włosów *U* pochodzą z kominiarki i (b) włókna wyczesane z włosów *U* pochodzą z innej części garderoby lub tkaniny. Na poziomie aktywności można rozważać następujące dwie hipotezy: (c) *U* miał na sobie kominiarkę w czasie rabunku i (d) *U* nigdy nie nosił

tej kominiarki. Należy dodać, że rozważanie hipotez na poziomie aktywności wymaga od biegłego dostępu do informacji o okolicznościach zdarzenia, na przykład informacji o czasie, w którym dokonano przestępstwa lub też informacji o czasie, jaki upłynął pomiędzy aresztowaniem podejrzanego i zabezpieczeniem próbki do badań.

3. Określenie cech, które powinny być uwzględnione w modelu służącym do wyznaczenia ilorazu wiarygodności. Rozważanie hipotez na poziomie aktywności wymaga informacji o tym, że doszło do przeniesienia, pozostania i ostatecznie odzyskania włókien. Dane na temat przenoszenia się włókien na włosy i pozostawania na nich są dostępne w literaturze [7, 10, 17]. W przypadku, gdy istnieje niewiele danych, na przykład w sytuacji, gdy rozpatrywanym dowodem są włosy zamiast włókien, to wówczas możliwe jest jedynie rozważanie hipotez związanych wyłącznie z materiałem dowodowym, czyli analiza hipotez na poziomie źródła.
4. Oszacowanie spodziewanej wartości ilorazu wiarygodności, który będzie można uzyskać na podstawie informacji uzyskanych o przestępstwie. Wyniki analizy porównawczej włókien wyczesanych z włosów i włókien z ciemnozielonej kominiarki mogą być różne. Załóżmy, że nie ujawniono żadnych włókien lub znaleziona została tylko jedna grupa włókien. Jeżeli tak się stało, to może ona pasować (pod względem rozważanych cech, np. morfologicznych) lub też nie pasować do włókien z zielonej kominiarki. Jeżeli grupa włókien pasuje pod względem rozważanych cech do włókien z ciemnozielonej kominiarki, to wówczas liczebność grupy włókien może być określona jako mała lub duża. We wcześniejszej publikacji [6] zamieszczono wartości prawdopodobieństw, że opisane powyżej sytuacje mogły zajść w zależności od tego, która z analizowanych na poziomie aktywności hipotez zaszła: (c) U miał nałożoną kominiarkę w czasie rabunku, (d) U nigdy nie nosił kominiarki. Wartości ilorazu wiarygodności uzyskane na podstawie prawdopodobieństw zaproponowanych przez Champoda i Jacksona [9] są również zamieszczone we wspomnianej wyżej pracy [6]. W przypadku, gdy żadne włókno nie zostało znalezione w próbce uzyskanej w trakcie wyczesania włosów, to wówczas prawdopodobieństwa te posiadają wartość ilorazu wiarygodności równą 100, która wspiera hipotezę, że U nigdy nie nosił tej kominiarki. Jeśli znaleziona została pojedyncza grupa włókien i nie były one podobne do włókien z ciemnozielonej kominiarki, to wówczas również uzyskuje się iloraz wiarygodności równy 100, który wspiera hipotezę, że podejrzany nigdy nie nosił tej kominiarki. W razie, gdy została znaleziona nieliczna grupa włókien zgodnych z włóknami z kominiarki, to wówczas iloraz wiarygodności jest równy 3 i wspiera hipotezę, że U nosił tę kominiarkę podczas rabunku. Natomiast jeżeli została znaleziona duża grupa włókien zgodnych z włókami z kominiarki, to wówczas iloraz wiarygodności jest równy 840 i wspiera hipotezę, że U nosił tę kominiarkę podczas rabunku. Uzyskane wartości ilorazu wiarygodności dostarczają pewnych informacji, które z kolei umożliwiają udzielenie odpowiedzi na pytanie: „czy użyteczne będzie wykonanie analizy próbki uzyskanej w trakcie wyczesania włosów osoby podejrzanej?"

Pierwsze cztery kroki opisanej wcześniej procedury są częścią dochodzenia prowadzonego przez wymiar sprawiedliwości. Poniżej przedstawiono inne przykłady metod statystycznych, które mogą okazać się pomocne w tym procesie.

4. Ustalanie stopnia pokrewieństwa na podstawie rezultatów analizy DNA

Załóżmy, że popełniono przestępstwo. Uzyskano profil DNA, który może być uznany za pochodzący od przestępcy. Nie ma on swojego odpowiednika w bazie danych. Problemem do rozważenia jest oszacowanie wartości dowodu z profilu DNA w przypadku znalezienia w bazie danych profilu DNA niemal zgodnego z profilem DNA podejrzanego, ale pochodzącego od jego krewnego.

Załóżmy, że krewnym podejrzanego jest jego brat. Oznaczamy osobę z bazy danych jako X, a brata X jako B. Rozważmy cztery allele. Allele są opisane za pomocą liter a, b, c oraz d. Odpowiadające tym profilom częstości występowania w odpowiedniej populacji odniesienia można opisać jako p_a, p_b, p_c i p_d. Zakładamy, że znane są częstotliwości występowania alleli w każdym z rozważanych loci. Wówczas możliwe jest określenie wyrażeń dla $Pr(X \text{ jest } (y, z) \mid B \text{ jest } (a, b))$, przy założeniu, że po pierwsze, X jest bratem B, i po drugie, że X i B nie są spokrewnieni.

W tabeli III zamieszczono prawdopodobieństwa dla jednego lokus z czterema allelami. Aby zilustrować, jak wygląda taka analiza, przypiszmy konkretne wartości prawdopodobieństw uzyskane dla każdego z alleli, np. $p_a = 1/4$, $p_b = 1/12$, $p_c = 1/2$, $p_d = 1/6$. Załóżmy, że przestępca (B) ma profil (a, b). Sprawdzamy X w bazie danych. Możliwe do uzyskania wartości ilorazu wiarygodności przedstawiono w tabeli IV. Iloraz wiarygodności jest tutaj stosunkiem $Pr(X = (y, z) \mid B = (a, b))$; X i B są braćmi) do $Pr((X = (y, z) \mid B = (a, b))$, tj. X i B nie są spokrewnieni.

W rozważanym przypadku największe wsparcie uzyskujemy dla hipotezy, że X i B są braćmi, jeżeli dla osoby X, ujętej w bazie danych, uzyskane allele to (a, b), czyli takie same jak dla B, czego można się było spodziewać.

W przypadku, gdy rozważanych jest więcej niż jeden lokus, to wówczas:
- wyznaczamy iloraz wiarygodności (*LR*) dla każdej pary alleli poszczególnych *loci* i dla każdej z rozważanych X osób, (X_1, …, X_n,) w porównaniu ze znanymi allelami dla B dla tego lokus;
- mnożymy wartości *LR* uzyskane dla każdej z par alleli poszczególnych osób X_i (i = 1, …, n) poprzez wszystkie analizowane *loci*. Zakładamy niezależność *loci* genetycznych;
- rezultat mnożenia jest siłą dowodu łączącą każdą z osób X_i (i = 1, …, n) zawartych w bazie danych z osobą B, tzn. z przestępcą.

Na podstawie uzyskanych wartości dochodzenie może być ukierunkowane na te osoby X_i, w przypadku których uzyskano najwyższe wartości *LR*.

Zastosowanie powyższej metodyki do ustalenia powiązań rodzinnych na podstawie profilu DNA było z powodzeniem stosowane w Wielkiej Brytanii[2].

5. Określenia wielkości próbki koniecznej do pobrania w celu wykonania badań

W praktyce laboratoryjnej często niezbędne jest dokonanie analizy dużej partii identycznych przedmiotów w celu określenia, jaka część z nich jest posiadana niezgodnie z prawem. Na przykład duży zbiór identycznych przedmiotów stanowić mogą:
- białe tabletki: ile należy pobrać do badań w celu stwierdzenia, jaka ich część zawiera nielegalną substancję (np. narkotyk)?
- płyty kompaktowe: ile należy przebadać w celu określenia, jaka część płyt CD z zabezpieczonej kolekcji zawiera nielegalne nagrania?
- zdjęcia cyfrowe w komputerze: ile z nich należy przebadać w celu określenia, jaka część fotografii zawiera treści pornograficzne?

Z punktu widzenia statystyki powyższe problemy są identyczne i w celu ich rozwiązania może być zastosowana ta sama metoda statystyczna. Pozwala to zaoszczędzić czas i pieniądze, ponieważ analizie nie muszą zostać poddane wszystkie przedmioty wchodzące w skład partii danego rodzaju materiału dowodowego.

W celu skonstruowania modelu statystycznego, który umożliwi opracowanie procedury pobierania reprezentatywnej próbki do badań, należy dokonać kilku założeń. Najprostszą procedurę oszacowania wielkości próbki reprezentatywnej utworzymy w przypadku, gdy dokonamy kilku restrykcyjnych założeń. Procedura taka zakłada, że:
- każdy element analizowanej partii materiału dowodowego może zostać zakwalifikowany tylko do jednej z dwóch kategorii (tj. próbka legalna – próbka nielegalna);
- partia materiału jest jednorodna we wszystkich aspektach innych niż przedmiot analizy, czyli stwierdzenie, czy próbka zawiera substancje (treści) legalne (nielegalne) bądź nie;
- wartość prawdopodobieństwa, że element partii materiału należy do danej kategorii, jest stała dla każdego z elementów tego zbioru;
- wartość prawdopodobieństwa, że dany element partii materiału należy do danej kategorii, jest niezależna od zdarzenia, że do tej kategorii może być przypisany inny element analizowanego zbioru próbek.

W przypadku białych tabletek złożenia te oznaczają, że tabletki są klasyfikowane do jednej z dwóch grup: tabletka zawiera środek legalny – tabletka zawiera środek nielegalny. Ponadto wszystkie tabletki z analizowanego zbioru mają ten sam kolor, logo, teksturę, wagę i kształt. W przypadku partii płyt kompaktowych płyty są klasyfikowane jako zawierające nagrania legalne lub „piracki". Niemniej jednak mogą się one różnić naklejkami i (lub) mieć nagrane różne treści, ale zakłada się takie samo stałe prawdopodobieństwo, że dana płyta CD zawiera nagranie, które jest „pirackie", bez względu na to, jaka nielegalna treść nagrana jest na płycie. W przypadku zdjęć zgromadzonych w komputerze są one klasyfikowane przed ich otwarciem jako zawierające lub nie treści pornograficzne.

Stosując powyższe założenia, podejście bayesowskie może być zastosowane w celu określenia liczby elementów partii analizowanego materiału, które powinny zostać poddane analizie. Wykonanie analizy statystycznej wymaga ustalenia przez biegłego sądowego, *a priori*, prawdopodobieństwa ( ), z jakim chce on ustalić, jaka część ( ) badanego zbioru materiału dowodowego jest nielegalna. Zakłada się ponadto, że proces pobierania reprezentatywnej próbki do badań powinien być procesem losowym, a prawdopodobieństwo, że każdy element części kwestionowanego materiału może być wybrany do badań, musi być jednakowe dla każdego z elementów analizowanej partii.

Posłużmy się przykładem liczbowym uwzględniającym powyższe kryteria do analizy partii białych tabletek. Należy np. określić, jaka powinna być liczebność zbioru tabletek, które trzeba poddać analizie w przypadku założenia, że z 95% prawdopodobieństwem chcemy ustalić (= 0.95), iż co najmniej 50% elementów partii białych tabletek zawiera nielegalną substancję (= 0,5). Odpowiedź brzmi: 4 tabletki. Tym samym, jeżeli 4 tabletki wybrane na drodze losowej z dużej partii tabletek po analizie zostaną zaklasyfikowane jako zawierające narkotyki, to

[2] Brytjskie gazety (*The Times, The Guardian)* doniosły w dniu 20 kwietnia 2004 r., że osoba oskarżona o nieumyślne spowodowanie śmierci została skazana na karę więzienia na podstawie dowodu opartego na powiązaniu DNA ujawnionego na miejscu zdarzenia z profilem DNA krewnego osoby oskarżonej,znajdującym się w bazie profili DNA.

wówczas istnieje 95% prawdopodobieństwo, że co najmniej 50% tabletek w analizowanej partii zawiera także nielegalną substancję. Wielkość próbki, $n = 4$, jest niezależna od rozmiaru partii analizowanego materiału i prawdopodobieństwo jest takie samo, jak w przypadku, gdy zbiór tabletek będzie złożony z tysiąca lub miliona elementów. Natomiast wielkość próbki może być trochę inna, gdy partia tabletek jest niewielka, np. liczy mniej niż 50 elementów. Wówczas liczba tabletek, które należy pobrać do analizy, będzie odpowiednio mniejsza.

Wielkość próbki pobranej do badań (4 tabletki) może wydawać się zbyt mała w porównaniu do liczebności zbioru tabletek dostarczonych do badań. Rozważmy jednak przykład rzutu monetą. W każdym rzucie monetą możliwe są dwa wyniki: orzeł i reszka. Jeżeli moneta jest symetryczna, to oba rezultaty są jednakowo prawdopodobne i z prawdopodobieństwem równym 1/2 mogą się one zdarzyć w każdym z rzutów monetą, niezależnie od rezultatów wcześniejszych rzutów. Przyjmijmy, że rzucamy monetą 4 razy. Jeżeli moneta jest symetryczna, to prawdopodobieństwo uzyskania czterech reszek wynosi $(1/2)^4$, czyli 1/16 lub 0,0625. Taka wartość prawdopodobieństwa jest bardzo mała, bliska 0,05, i porównać ją można do wartości 0,95, czyli 95% prawdopodobieństwa w przytoczonym przykładzie o białych tabletkach. Tak więc w przypadku, gdy wykonamy cztery rzuty monetą, z których każdy zakończy się wyrzuceniem reszki, to można poddać w wątpliwość założenie, że moneta jest symetryczna. Podobnie należy traktować rezultaty uzyskane w przytoczonym przykładzie z białymi tabletkami.

Dalsze szczegóły dotyczące pobierania reprezentatywnej próbki do badań w drodze losowej zamieszczone są w innej publikacji [1] oraz w przewodniku opracowanym przez Grupę Roboczą Ekspertów ds. Badania Narkotyków działającą w ramach Europejskiej Sieci Instytutów Nauk Sądowych, który opublikowano w roku 2004 [14].

6. Wnioski

W niniejszej publikacji omówione zostały cztery przykłady zastosowania metod statystycznych, które mogą być użyteczne w fazie dochodzeniowej śledztwa. Należy pamiętać, że przykłady te odnoszą się wyłącznie do wspomnianego etapu i istnieje niewielka szansa, by uzyskane rezultaty zostały użyte jako dowód w procesie sądowym. Na przykład uzyskany profil psychologiczny sprawcy przestępstwa nie jest dowodem, który może być przedstawiony w sądzie w celu wsparcia hipotezy, że oskarżony, który pasuje do tego profilu, popełnił dane przestępstwo. Wyniki oszacowania, czy możliwe jest znalezienie na dostarczonej do badań odzieży podejrzanego śladów (np. włókien), które mogły ulec przeniesieniu na tę odzież w trakcie dokonywania przestępstwa, nie mogą być traktowane jako poparcie bądź dla rozpatrywanych w sądzie hipotez prokuratora, bądź hipotez obrony. Spowodowane jest to faktem, że rezultat wstępnego badania dowodu jest co prawda częścią procesu oszacowania wartości dowodowej np. włókien, ale proces, który doprowadził do tego badania, nie jest dowodem samym w sobie. Wskazanie przestępcy na podstawie ustalonego podobieństwa profilu DNA do profilu DNA jego krewnego, który znajduje się w bazie danych, może pomóc zidentyfikować podejrzanego, ale rezultat takiego testu nie jest dowodem. Oszacowanie wielkości próbki reprezentatywnej, koniecznej do pobrania w celu wykonania badań, również nie jest dowodem w procesie sądowym.

W drugim artykule opublikowanym w niniejszym numerze czasopisma *Z zagadnień nauk sądowych* zostanie opisana rola statystyki w oszacowaniu wartości dowodowej materiału dowodowego.