Introduction to Statistics for Forensic Scientists

David Lucy
University of Edinburgh, UK

John Wiley & Sons, Ltd

Introduction to Statistics for Forensic Scientists

Introduction to Statistics for Forensic Scientists

David Lucy
University of Edinburgh, UK

John Wiley & Sons, Ltd

Telephone (+44) 1243 779777

Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Lucy, David.
Introduction to statistics for forensic scientists / David Lucy.
p. cm.
Includes bibliographical references and index.
ISBN-13 978-0-470-02200-0 (HB) ISBN-13 978-0-470-02201 9 (PB)
ISBN-10 0-470-02200-0 (HB) ISBN-10 0-470-02201 9 (PB)
1. Forensic sciences–Statistical methods. 2. Forensic statistics. 3. Evidence (Law)–Statistical methods.
I. Title.
HV8073.L83 2005
519.5024′36325 2005028184

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN-13 978-0-470-02200-0 (HB) ISBN-13 978-0-470-02201 9 (PB)
ISBN-10 0-470-02200-0 (HB) ISBN-10 0-470-02201 9 (PB)

Typeset in 10.5/13pt Times and Officina by TechBooks, New Delhi, India
Printed and bound in Great Britain by Antony Rowe Ltd., Chippenham, Wilts
This book is printed on acid-free paper responsibly manufactured from sustainable forestry
in which at least two trees are planted for each one used for paper production.

Contents

Preface

The detective story, whether it be in the form of a novel, a television programme, or a cinema film, has always exerted a fascination for people from all walks of life. Much of the appeal of the detective story lies in the way in which a series of seemingly disconnected observations fit a narrative structure where all pieces of information are eventually revealed to the reader, or viewer, make a whole and logical nexus. The story which emerges by the end of the plot as to how, and just as importantly why, the perpetrator committed the crime, is shown by some device, such as a confession by the "guilty" character, to be a true description of the circumstances surrounding the crime.

Detective stories have, at their core, some important and fundamental truths about how humans perceive what is true from what is false. The logical arguments used are woven together with elements of evidence taken from widely differing types of observation. Some observations will be hearsay, others may be more material observations such as blood staining. All these facts will be put together in some logical way to create a case against one of the characters in the story.

However, detective stories do have a tendency to neglect one of the more important elements of real investigation. That element is uncertainty. The interpretation of real observations is usually subject to uncertainty, for example, the bloodstain on the carpet may "match" the suspect in some biochemical way, but was the blood which made the bloodstain derived from the suspect, or one of the other possible individuals who could be described as a "match". Statistical science is the science of uncertainty, and it is only appropriate that statistics should provide at least part of the answer to some of the uncertain parts of evidence encountered in criminal investigations. That part of evidence upon which it is possible to throw some illumination is that evidence generated by forensic scientists. This tends to be numerical by nature, and is thus amenable to analysis by statisticians.

The are though, two roles for statistics in forensic science. The first is a need for forensic scientists to be able to take their laboratory data from experiments, and interpret that data in the same way that any observational scientist would do. This strand of statistical knowledge is commonly used by all sorts of scientists, and guides to it can be found in any handbook of applied statistical methods. The second role of statistical science is in the interpretation of observations from the case work with which forensic scientist may become involved. This strand of application of statistical methods in forensic science has been termed evidence evaluation. These days a number of books exist outlining statistical evidence evaluation techniques, all of them excellent, but unfortunately none of them aimed towards those who are relatively new to statistical science, and require a certain technical insight into the subject.

This volume attempts to bridge the gap in the literature by commencing with the use of statistics to analyse data generated during laboratory experiments, then progressing to address the issue of how observations made by, and reported to the forensic scientist may be considered as evidence.

Finally, I should like to acknowledge the assistance of Bruce Worton, Colin Aitken, Breedette Hayes, Gzregorz Zadora, James Curran, Nicola Martin, Nicola Clayson, Mandy Jay, Franco Taroni, John Kingston, Dave Barclay, Tom Nelson and Burkhard Schaffer. R (R development core team, 2004) was used to create all diagrams and calculations which appear in this volume. My gratitude is also due to all those forensic scientists who have allowed me the use of their data in this volume.

List of figures

List of tables

1 A short history of statistics in the law

The science of statistics refers to two distinct, but linked, areas of knowledge. The first is the enumeration of types of event and counts of entities for economic, social and scientific purposes, the second is the examination of uncertainty. It is in this second guise that statistics can be regarded as the science of uncertainty. It is therefore natural that statistics should be applied to evidence used for legal purposes as uncertainty is a feature of any legal process where decisions are made upon the basis of evidence. Typically, if a case is brought to a court it is the role of the court to discern, using evidence, what has happened, then decide what, if anything, has to be done in respect of the alleged events. Courts in the common law tradition are not in themselves bodies which can directly launch investigations into events, but are institutions into which evidence is brought for decisions to be made. Unless all the evidence points unambiguously towards an inevitable conclusion, different pieces of evidence will carry different implications with varying degrees of force. Modern statistical methods are available which are designed to measure this 'weight' of evidence.

1.1 History

Informal notions of probability have been a feature of decision making which date to at least as far in the past as the earliest writing. Many applications were, as related by Franklin (2001), to the process of law. Ancient Egypt seems to have two strands, one of which relates to the number of reliable witnesses willing to testify for or against a case, evidence which remains important today. The other is the use of oracles and is no longer in use. Even in the ancient world there seems to have been scepticism about the information divulged by oracles, sometimes two or three being consulted and the majority opinion followed. Subsequently the Jewish tradition made the assessment of uncertainty central to many religious and legal

Introduction to Statistics for Forensic Scientists David Lucy

practices. Jewish law is notable in that it does not admit confession, a wholly worthy feature which makes torture useless. It also required a very high standard of proof which differed according to the seriousness of any alleged offence. Roman law had the concept of onus of proof, but the wealthier sections of Roman society were considered more competent to testify then others. The Roman judiciary were allowed some latitude to judge in accordance with the evidence. In contrast to Jewish law, torture was widespread in Roman practice. In fact in some circles the evidence from a tortured witness was considered of a higher quality than had the same witness volunteered the evidence, particularly if they happened to be a member of the slave classes.

European Medieval law looked to the Roman codes, but started to take a more abstract view of law based on general principles. This included developments in the theory of evidence such as half, quarter and finer grades of proof, and multiple supporting strands forming what we today would call a case. There seem to have been variable attitudes to the use of torture. Ordeal was used in the earlier period to support the civil law in cases which were otherwise intractable. An important tool for evidence evaluation with its beginnings in the Western European Medieval was the development of a form of jury which has continued uninterrupted until the present day. It is obvious that the ancient thinkers had some idea that the evidence with which they were dealing was uncertain, and devised many ingenious methods of making some sort of *best* decision in the face of the uncertainties, usually revolving around some weighting scheme given to the various individual pieces of evidence, and some process of summation. Nevertheless, it is apparent that uncertainty was not thought about in the same way in which we would think about it today.

Informal enumeration types of analyses were applied as early as in the middle of the 17th century to observational data with John Gaunt's analysis of the London Mortality bills (Gaunt, 1662, cited in Stigler, 1986), and it is at this point in time that French mathematicians such as De Méré, Roberval, Pascal and Fermet started to work on a more recognizably modern notion of probability in their attempts to solve the problem of how best to divide up the stakes on interrupted dice games.

From there, ideas of mathematical probability were steadily developed into all areas of science using large run, or frequentist, type approaches. They were also applied to law, finding particular uses in civil litigation in the United States of America where the methods of statistics have been used, and continue to be used, to aid courts in their deliberations in such areas as employment discrimination and antitrust legislation (Fienberg, 1988).

First suggested in the latter part of the nineteenth century by Poincaré, Darboux and Appell (Aitken and Taroni, 2004, p. 153) was an intuitive and intellectually satisfying method for placing a simple value on evidence. This employed a measure called a likelihood ratio, and was the beginning of a more modern approach to evidence evaluation in forensic science. A likelihood ratio is a statistical method which can be used directly to assess the worth of observations, and is currently the predominant measure for numerically based evidence.

Since the inception of DNA evidence in forensic science in the courts of the mid 1980s, lawyers, and indeed forensic scientists themselves, have looked towards statistical science to provide precise evaluation of the worth of evidence which follows the explicitly probabilistic approach to the evidential value of DNA matches.

1.2 Some recent uses of statistics in forensic science

A brief sample of the *Journal of Forensic Sciences* between the years 1999 and 2002 shows that about half of the papers have some sort of statistical content. These can be classified into: regression and calibration, percentages, classical hypothesis tests, means, standard deviations, classification and other methods. This makes knowledge of numerical techniques at some level essential, either for publication in the literature, or knowledgeable and informed reading.

The statistical methods used in the surveyed papers were:

1. Regression and calibration – regression is finding the relationship between one thing and another. For example, Thompson *et al.* (1999) wished to compare amounts of explosive residue detected by GC-MS with that detected by LC-UV. To do this they undertook a regression analysis which told them that the relationship was almost 1:1, that is, they would have more or less the same measurement from either method. Calibration is in some senses the complement of regression in that what you are trying to do is make an estimate of one quantity from another. Migeot and De Kinder (2002) used calibration to make estimates of how many shots an assault rifle had fired since its piston was last cleaned by the number of carbon particles on the piston.

2. Percentages and enumeration statistics – counts and proportions of objects, employed universally as summary statistics.

3. Means, standard deviations and t-tests – a mean is a measure of location[†]. For example, Solari and Abramovitch (2002) used stages in the development of teeth to estimate ages for Hispanic detainees in Texas. They assigned known age individuals to 10 stages of third molar development and calculated the mean age for the individuals falling in that age category. What they were then able to do was to assign any unknown individual to a developmental category, thus suggesting an expected age for that individual.

   Standard deviations are measures of dispersion about a mean. In the example above, Solari and Abramovitch (2002) also calculated the standard deviation for age for each of their developmental categories. They were then able to gain some idea of how wrong they would be in assigning any age to an unknown individual.

[†] Location in this context is a measure of any central tendency, for instance, male stature in the United Kingdom tends towards 5′8″.

t-tests tell you how different are two samples based on the means and standard deviations of those samples. For example, Koons and Buscaglia (2002) used *t*-tests on elemental compositions from glass found at a crime scene to that found on a suspect to tell whether the two samples of glass possibly came from the same source.

4. Classification – this allows the researcher to assign categories on the basis of some measurement. Stojanowski and Siedemann (1999) used neck bone measurements from known sex skeletons and a discriminant function analysis to calculate a feature rule which would allow them to categorize skeletal remains as male, or female.

5. Other methods – these include χ^2[‡] tests and Bayesian methods.

1.3 What is probability?

When we speak of probability what is it we mean? Everybody uses the expression 'probably' to express belief favouring one possible outcome, or world state, over other possible outcomes, but does the term probability confer other meanings?

Examining the sorts of things which constitute mathematical ideas of probability there seem to be two different sorts. The first are the aleatory[§] probabilities, such events as the outcomes from dice throwing and coin tossing. Here the system is known, and the probabilities deduced from knowledge of the system. For instance, with a fair coin I know that in any single toss it will land with probability 0.5 heads, and probability 0.5 tails. I also know that in a long run of tosses roughly half will be heads, and roughly half tails.

A second type of probability is epistemic. This is where we have no innate knowledge of the system from which to deduce probabilities for outcomes, but can by observation induce knowledge of the system. Suppose one were to examine a representative number of people and found that 60% of them were mobile telephone users. Then we would have some knowledge of the structure of mobile telephone ownership amongst the population, but because we had not examined every member of the population to see whether or not they were a mobile telephone user, our estimate based on those we had looked at would be subject to a quantifiable uncertainty.

Scientists often use this sort type of generalization to suggest possible mechanisms which underly the observations. This type of empiricism employs, by necessity, some form of the uniformitarian assumption. The uniformitarian assumption implies that processes observed in the present will have been in operation in the past, and will be in operation in the future. A form of the uniformitarian assumption is, to some extent, an inevitable feature of all sciences based upon observation, but it is the absolute

[‡] Pronounced 'chi-squared'.
[§] Aleatory just means by chance and is not a word specific to statistics.

cornerstone of statistics. Without accepting the assumption that the processes which cause some members of a population to take on certain characteristics are at work in the wider population, any form of statistical inference, or estimation, is impossible.

To what extent probabilities from induced and deduced systems are different is open to some debate. The deduced probability cannot ever be applied to anything other than a notional system. A die may be specified as fair, but any real die will always have minor inconsistencies and flaws which will make it not quite fair. To some extent the aleatory position is artificial and tautological. When a fair die is stipulated then we know the properties in some absolute sense of the die. It is not possible to have this absolute knowledge about any actual observable system. We simply use the notion as a convenient framework from which to develop a calculus of probability, which, whenever it is used, must be applied to probability systems which are fundamentally epistemic. Likewise, because all inferences made about populations are based on the observation of a few members of that population, some degree of deduced aleatory uncertainty is inevitable as part of that inference.

As all real probabilities are induced by observation, and are essentially frequencies, does this mean that a probability can only ever be a statement about the relative proportions of observations in a population? And, if so, is it nonsense to speak of the probability for a single event of special interest?

An idea of a frequency being attached to an outcome for a single event is ridiculous as the outcome of interest either happens or does not happen, From a single throw of a six-sided die we cannot have an outcome in which the die lands 1/6 with its six face uppermost, it either lands with the six face uppermost, or it does not. There is no possible physical state of affairs which correspond to a probability of 1/6 for a single event. Were one to throw the six-sided die 12 times then the physical state corresponding to a probability of 1/6 would be the observation of two sixes. But there can be no single physical event which corresponds to a probability of 1/6.

The only way in which a single event can be quantified by a probability is to conceive of that probability as a product of mind, in short to hold an idealist interpretation of probability (Hacking, 1966). This is what statisticians call subjective probability (O'Hagen, 2004) and is an interpretation of probability which stipulates that probability is a function of, and only exists in, the mind of those interested in the event in question. This is why they are subjective, not because they are somehow unfounded or made up, but because they rely upon idealist interpretations of probability.

A realist interpretation of probability would be one which is concerned with frequencies and numbers of outcomes in long runs of events, and making inferences about the proportions of outcomes in wider populations. A realist interpretation of probability would not be able to make statements about the outcome of a single event as any such statement must necessarily be a belief as it cannot exist in the observable world, and therefore requires some ideal notion of probability. Realist positions imply that there is something in the observed world which is causing uncertainty, uncertainty being a property external to the mind of the observer. Some might argue that these external probabilities are propensities of the system in question to behave

in a specific way. Unfortunately the propensity theory of probability generates the same problem for a realist conception when applied to a single event because a propensity cannot be observed directly, and would have to be a product of mind. In many respects realist interpretations can be more productive for the scientist because of the demands that some underlying explanatory factor be hypothesized or found. This is in contrast to idealist positions where a cause for uncertainty is desirable, but not absolutely necessary, as the uncertainty resides in the mind.

This distinction between realist and idealist is not one which is seen in statistical sciences, and indeed the terms are not used. There are no purely realist statisticians; all statisticians are willing to make probabilistic statements about single events, so all statisticians are to some degree idealistic about their conception of probability. However, a debate in statistics which mirrors the realist/idealist positions is that of the frequentist/Bayesian approaches. There is a mathematical theorem of probability called Bayes' theorem, which we will encounter in Section 9.2, and Bayesians are a school of statisticians named after the theorem. The differences between Bayesians and frequentists are not mathematical, Bayes' theorem is a mathematical theorem and, given the tenets of probability theory, Bayes' theorem is correct. The differences are in this interpretation of the nature of probability. Frequentists tend to argue against subjective probabilities, and for long-run frequency based interpretations of probability. Bayesians are in favour of subjective notions of probability, and think that all quantities which are uncertain can be expressed in probabilistic terms.

This leads to a rather interesting position for forensic scientists. On the one hand they do experimental work in the laboratory where long runs of repeated results are possible; on the other hand they have to interpret data as evidence which relates to singular events. The latter aspect of the work of the forensic scientist is explicitly idealistic because events in a criminal or civil case happened once and only once, and require a subjective interpretation of probability to interpret probabilities as degrees of belief. The experimental facet of forensic science can easily accommodate a more realist view of probability.

The subjective view of probability is the one which most easily fits commonsense notions of probability, and the only one which can be used to quantify uncertainty about single events. There are some fears amongst scientists that a subjective probability is an undemonstrated probability without foundation or empirical support, and indeed a subjective probability can be that. But most subjective probabilities are based on frequencies observed empirically, and are not, as the term subjective might imply, somehow snatched out of the air, or made up.

There is a view of the nature of probability which can side-step many of the problems and debates about the deeper meaning of just what probability is. This is an instrumentalist position (Hacking, 1966) where one simply does not care about the exact interpretation of probability, but rather one simply views it as a convenient intellectual devise to enable calculations to be made about uncertainty. The instrumentalist's position implies a loosely idealist background, where probability is a product of mind, and not a fundamental component of the material world.

2 Data types, location and dispersion

All numeric data can be classified into one or more types. For most types of data the most basic descriptive statistics are a measure of central tendency, called location, and some measure of dispersion, which to some extent is a measure of how good is a description the measure of central tendency. The concepts of location and dispersion do not apply to all data types.

2.1 Types of data

There are three fundamental types of data:

1. *Nominal* data are simply classified into discrete categories, the ordering having no significance. Biological sex usually comes in male/female, whereas gender can be male/female/other. Things such as drugs can be classified by geographical area such as South American, Afghan, Northern Indian or Oriental. Further descriptions by some measure of location, and dispersion, are not really relevant to data of this type.

2. *Ordinal* data are again classified into discrete categories; this time the ordering does have significance. The development of the third molar (Solari and Abramovitch, 2002) was classified into 10 stages. Each class related to age, and therefore the order in which the classes appear is important.

3. *Continuous* data types can take on any value in an allowed range. The concentration of magnesium in glass is a continuous data type which can range from 0% to about 5% before the glass becomes a substance which is not glass. Within that range magnesium can adopt any value such as 1.225% or 0.856%.

Table 2.1 Table of year and Δ^9-THC (%) for marijuana seizures: these data are simulated (with permission) from ElSohly *et al.* (2001) and are more fully listed in Table 2.2

Seizure	Year	Δ^9-THC (%)
1	1986	9.26
2	1987	7.58
3	1987	7.65
4	1986	10.29
5	1986	8.29
6	1987	7.85
7	1986	8.40
⋮	⋮	⋮

Table 2.2 Table of year and Δ^9-THC (%) for marijuana seizures: these data are simulated (with permission) from ElSohly *et al.* (2001)

Year	Δ^9-THC (%)	
1986	9.26	10.30
	8.29	8.40
	8.32	8.84
	8.82	8.41
	9.74	9.02
	10.70	7.05
	7.91	7.72
	8.41	8.93
	7.21	9.95
	6.29	8.16
1987	7.59	7.66
	7.85	7.91
	6.61	7.42
	6.91	8.46
	8.34	8.12
	7.97	7.15
	9.09	7.93
	7.93	

The type of data sometimes restricts the approaches which can be used to examine and make inferences about those data. For example, the idea of central tendency, and a dispersion about the central tendency, *is* not really relevant to nominal data, whereas both can be used to summarize ordinal and continuous data types.

There are a few of points of terminology with which it is necessary to be familiar:

- Nominal and ordinal data types are known collectively as *discrete*, because they place entities into discrete exclusive categories.

- All the above data types are called *variables*.

- There are nominal and ordinal (occasionally continuous) variables which are used to classify other variables, these are called *factors*. An example would be Δ^9-THC concentrations in marijuana seizures from various years in the 1980s given in Table 2.1. Here '% Δ^9-THC' is a continuous *variable* and 'year' is an *ordinal variable* which is being used as a *factor* to classify Δ^9-THC.

2.2 Populations and samples

Generally in chemistry, biology and other natural sciences a sample is something taken for the purposes of examination, for example a fibre and a piece of glass may be found at the scene of a crime; these would be termed samples. In statistics a sample has a different meaning. It is a sub-set of a larger set, known as a population. In the table of dates and % Δ^9-THC in Table 2.1, the % Δ^9-THC column gives measurements of the % Δ^9-THC in a sample of marijuana seizures at the corresponding date. In this case the population is marijuana seizures.

Populations and samples must be hierarchically arranged. For instance one could examine the 1986 entries and this would be a sample of % Δ^9-THC in a 1986 population of marijuana seizures. It could also be said that the sample was a sample of the population of all marijuana seizures, albeit a small one. Were all marijuana observed for 1986 this would be the population of marijuana for 1986, which could for some purposes be regarded as a sample of all marijuana from the population of marijuana from the 1980s. The population of marijuana from the 1980s could be seen as a sample of marijuana from the 20th century.

It is important to realize that the notions of population and sample are not fixed in nature, but are defined by the entities under examination, and the purposes to which observation of those entities is to be put. However, populations and samples are always hierarchically arranged in that a sample is always a sub-set of a population.

2.3 Distributions

Most generally a distribution is an arrangement of frequencies of some observation in a meaningful order. If all 20 values for THC content of 1986 marijuana seizures are grouped into broad categories, that is the continuous variable % THC is made into an ordinal variable with many values, then the frequencies of THC content

in each category can be tabulated. This table can be represented graphically as a histogram[†].

A histogram of simulated Δ^9-THC frequencies from 1986 taken from Table 2.2, is represented in Figure 2.1. In Figure 2.1 the horizontal axis is divided into 14 categories of 0.5% each, the vertical axis is labelled 0 to 10, and indicates the counts, or frequency, of occurrences in that particular category. So for the first two categories ($5 \rightarrow 6\%$) there are no values, the second category ($6.0 \rightarrow 6.5\%$) occurs with a frequency 1, and so on.

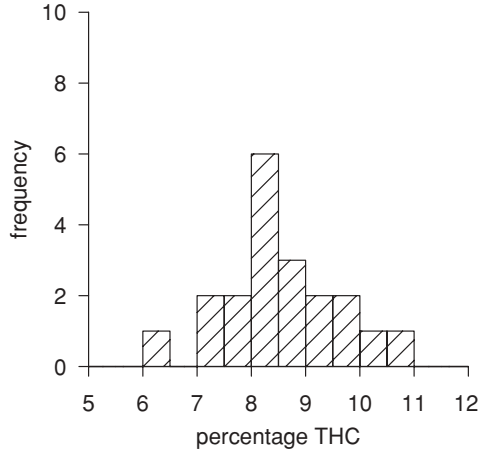

Figure 2.1 Histogram of simulated Δ^9-THC (%) values for a sample of marijuana seizures dating from 1986

The histogram in Figure 2.1, which gives the sample frequency *distribution* for Δ^9-THC in marijuana from 1986, has three important properties:

1. It has a single highest point at about 8.25% THC, the two ends of the distribution (tails) having progressively lower frequencies as they get further from the highest point. This property is called *unimodal* and indicates that there is some tendency amongst 1986 marijuana consignments towards a THC content of about 8.25%.

2. The distribution is more or less symmetric about the 8.25% value.

3. The distribution is dispersed about the 8.25% point in some measurable way.

The histogram in Figure 2.2 is the sample distribution for THC in marijuana from 1987. Here the % THC tends towards a value of about 7.75%; the same properties of dispersion about this value and a sort of symmetry can be seen as in Figure 2.1.

[†] A histogram is not to be confused with a bar chart, which looks similar, but in a bar chart height represents frequency rather than the area of the rectangles. Usually in a histogram the categories are of equal 'width', but this is not always the case.