Forensic Statistics

Multivariate trace comparison

J.L.F Göbbels

Forensic Statistics

Multivariate trace comparison

by

J.L.F Göbbels

To obtain the degree of Bachelor of Science
at Delft University of Technology,
To be defended publicly on 18 July, 2019 at 14:00.

Student ID:	4596498	
Project duration:	April, 2019 – July, 2019	
Thesis committee:	Dr. J. Söhl,	TU Delft, supervisor
	Dr. W. M. Ruszel,	TU Delft
	Drs. E. M. van Elderen,	TU Delft

Abstract

In today's complex, modern world, where more data and knowledge is available then ever before, consensus over approaches in statistics is far away. To understand what this means for forensic statistics, both the frequentist and Bayesian approach are considered in quantifying evidence. Classical frequentist approaches are elaborated according to an example of knife data, where both univariate and multivariate approaches are applied. As an alternative to the previous methods, the likelihood ratio is introduced. To arrive at a unified framework for calculating likelihood ratios, the European Union funded a project to calculate likelihood ratios using a software package with a user friendly interface called SAILR. Calculating the likelihood ratio is mostly done by using feature based models. A feature based model used by SAILR will be elaborated according to a glass comparison example. When the feature based model is illustrated, an extension on the model will be given using a non-parametric function. The discussed models will be elaborated and applied on test data from and international drugs comparison project using SAILR. As an alternative to feature based models, a score based model is introduced and compared to the previous drugs comparison results of the feature based model.

Preface

This thesis has been written in order to obtain the degree of Bachelor of Science. The research has been conducted under supervision of J. Söhl on behalf of the department of Statistics of the faculty EEMCS at the University of Technology Delft.

Initially, this thesis would only contain a short summary of already known methods in forensic statistics, but since the Netherlands Forensic Institute (NFI) provided me with the SAILR software packages and manuals, this thesis now also contains case work examples. Furthermore the NFI was very helpful in case of problems and were always willing to answer question even without there being a formal collaboration. Without their help, this thesis would never have contained an application of the theory described and therefore my sincere gratitude goes to them.

Secondly, I want to thank Jakob Söhl for all the freedom he gave me during this project. Even though we did not meet very often, you guided me in a very nice direction and advised me when my ideas became too ambitious for the limited time available.

Furthermore, I want to thank all my friends from EEMCS who supported me during this final period of my bachelor period and were always in for a cup of coffee or give me a boost whenever needed. Also, I would like to thank Jos Göbbels, who helped me to find the many grammatical errors I tend to make,

Finally, I would also like to thank E. van Elderen and W. Ruszel for taking a seat in the thesis committee.

J.L.F Göbbels
Delft, July 2019

Contents

1

Introduction

Only in 2017, the governmental crime lab for the state of Georgia in America confirmed 145 false positives for drugs comparison results in their state. One of these innocent suspects was a 19-year old Villa Rica teenager. After a $2 disposable drugs kit used by the police at that time mistook cleaning supplies in his car for XTC, it took over 18 months for a felony drug arrest to be dropped against him. Unfortunately, the damage at that point has already been done [33].

The comparison of forensic evidence can become mathematically very complex for forensic statisticians, who have a long history in applying complex statistical models to forensic evidence. Therefore it comes to no surprise that for court officials, who often have little to no mathematical background, interpreting the forensic findings can be hard or even misleading. Incorrect interpretation of statistical conclusions can even lead to legal aberration as in the case of the Villa Rica teenager, where court officials had little knowledge of the significance from the used drugs kit. In overcoming the misinterpretations, first consensus among forensic statisticians needs to be found. To harmonize the statistical models and software available, the European Union funded a project aimed at constructing a user friendly software package to calculate likelihood ratios. This unified framework is called SAILR, Software for Analysis and Implementation of Likelihood Ratios. To overcome misinterpretation in court, SAILR is provided with an option to convert the likelihood ratio to a verbal equivalent. Equipped with the SAILR software, different methods for calculating the likelihood ratio can be applied to drugs data made available by courtesy of The Netherlands Forensic Institute (NFI). By providing data under test for a known case, the relative value of the different methods can be compared. Using this test data, we can validate the different methods available.

Thesis outline

This research starts with formally outlining the hypotheses of the defense and the prosecution. In Chapter 2, both the common source and specific source problem are outlined. Chapter 3 evaluates the two commonly used schools of thoughts in forensic statistics: The frequentist point of view and the Bayesian approach. In Chapter 3, different frequentist approaches will be applied on an example of knife data. The properties of the available methods will be discussed and objections against these methods will be outlined. Whereas commonly only univariate data is considered, in this we will use multivariate data on eight different features. After these approaches and their properties are discussed, an introduction to likelihood ratios will be given using Bayes theorem. This ratio is then converted to a verbal likelihood ratio.

The likelihood ratio remains the basis for the final Chapters 6, 7 and 8. In Chapter 6, a basis for the feature based two-level model is made on the example of glass data. In forensic statistics, these two levels are known as the within source variation and the between source

variation. In Chapter 6, we assume a normal distribution for both levels. In Chapter 7, drugs tablet data from the NFI is applied. In this section, the two-level normal-normal model will be compared to the two-level normal-KDE model, where a kernel density estimate for the between source variation is applied instead of a normal density function. The results of both methods will be compared using data from a collaborative European project for evaluating amphetamines. Finally, in Chapter 8 an introduction to a relative new approach of calculating likelihood ratios is considered, using score based likelihood ratios.
While both feature based and score based models are elaborated frequently in forensic journals, they are almost never compared to each other. Therefore the same drugs data will be applied on both the score based and feature based models to compare both methods with each other by means of drugs comparison results.

In most forensic statistics report, only one method of evaluating evidence is elaborated extensively. This can be either one of the two feature based model or a specific score based model. This thesis however tries to capture both of the feature based models. In addition a frequency model, where only discrete data is available, is elaborated. Furthermore different score and distribution functions are elaborated using a score based model. Finally this thesis evaluates simple frequentist methods.
Because we will omit most of the mathematical derivations, this thesis can be seen as a large summary of all available methods in forensic statistics up to now, which can be used for instance by judges with undergraduate statistics knowledge who want to now more about the available methods in forensic statistics. To make all these methods more illustrative to the reader, real forensic data is applied to all methods. This provides the opportunity to compare the different models, which is done in Chapter 7 and 8.

2
Forensic comparison problem

After evidence is found at a crime scene, this evidence needs to be analyzed by a large variety of experts to reach a final verdict. Forensic evidence in general can be hard to interpret for court officials, luckily they are never directly asked to do so and therefore we can split experts into two different groups [27]: forensic experts, such as forensic statisticians and legal expert, such as judges, jurors and other court officials. Legal experts should evaluate all available evidence after it is quantified and combine this to reach a final verdict, whereas forensic experts are needed to actually quantify the evidence. In this section, the beginning of the forensic process is set up by quantifying the evidence found at the crime scene, where we formulate the hypotheses the forensic expert wants to evaluate.

2.1. Source problems

A forensic statistician takes part in the juristic process by taking into consideration two competing hypotheses, these two hypotheses are stated by:

- H_p : Hypothesis of the prosecution.
- H_d : Hypothesis of the defense.

In general, the hypothesis the prosecutor will claim to be correct will be something along the lines of "guilty" whereas the hypothesis of the defense will claim "innocent". In forensic statistics almost all problems can be split into two major comparison problems: the common source problem and the specific source problem.

2.1.1. Common source problem

Assume two fingerprints are found at a crime scene and we are interested in the question whether these two fingerprints originate from the same, unknown source. Answering this question could give answer to the question whether there are multiple suspects involved in a crime scene. The two unknown fingerprints are called e_{u_1} and e_{u_2}, these two sources of evidence can be linked to the following two competing hypotheses:

- H_p : The two sources e_{u_1} and e_{u_2} originate from the same unknown source.
- H_d : The two sources e_{u_1} and e_{u_2} originate from different unknown sources.

Because we do not know where the both sources originate from, we call the originating source unknown, so we are only interested in the question whether the two sources of evidence originate from the same unknown source.

2.1.2. Specific source problem

Suppose that a fingerprint is found at a crime scene. Furthermore a specific suspect is arrested and fingerprints of this suspect are taken. We call the fingerprint found at the crime scene e_u and the fingerprint of the suspect e_s, the two competing hypotheses are stated as follow:

- H_p : The unknown source e_u originates from the same specific source e_s.
- H_d : The unknown source e_u does not originate from the same specific source e_s.

In this type of source problems, the unknown source of evidence found at the crime scene (e_u) is referred to as the control source and the specific source taken from the suspect is referred to as the recovered source (e_s). This framework of hypothesis setups can be used multiple times in case multiple fingerprints are found at the crime scene, but also in case multiple suspects are accused of committing the crime, therefore the hypotheses can easily be broadened to a multivariable case. In this research we will only look at common source problems.

3
Frequentist and Bayesian thinking

Statistics is one of the mathematical disciplines that reaches beyond plain mathematics, it is both directly and indirectly used in a large variety of scientific disciplines such as physical science, biological science, health science and of course forensic science. Given that statistics is used in so many different areas of mathematics, it does not come to much surprise that there are many different opinions on how statistics should be done. In forensic science, uniform definitions and guidelines are more than ever needed to come to legitimate conclusion which can rely on approval from the rest of the forensic statistics community.

In this section, we will take a closer look at the two most applied schools of thought, namely Bayesian and frequentist methods, where a definition of how they define probability will be sketched. The outline will turn out to be of importance for the rest of this thesis outline, where a closer look will be taken at applications for both schools of thought.

3.1. Probability

The definition of probability is defined as [12]:

> "a quantity between 0 and 1 that represents the chance of an event occurring. Probability may sometimes be expressed as percentages, or as odds, without loss of information. A probability may also be used to express the belief that an event will occur. Such probabilities are often referred to as 'subjective'."

This definition gives insight to both the Bayesian and the frequentist approach of probability, whereas the definition "degree of belief" only applies on the Bayesian approach. The Bayesian method can therefore be considered subjective whereas the frequentist method can be seen as objective.

3.1.1. Frequentists

From a frequentist point of view towards probability, predictions should be made on the underlying truth of the experiments, only based on the data of the current event. As L. Pekelis wrote [3], frequentist arguments resemble the type of logic lawyers use in court. Tests where the differences in mean for the control variable and recovered variable are compared are typical examples of frequentist approaches. Take for example t-tests [3]. Because no prior experiment is taken into account, there cannot be any discussion over these prior data, therefore frequentist methods are commonly referred to as being more objective.

3.1.2. Bayesian

From a Bayesian point of view however, past knowledge of similar experiments is taken into account. The past knowledge, know as a prior, can be combined with the current experiment to draw a conclusion, known as the posterior. A typical Bayesian approach commonly follows the following guidelines [3];

1. Define the prior distribution that incorporates your subjective beliefs about a parameter.
2. Gather data
3. Update your prior distribution to obtain a posterior distribution. This posterior distribution represents an updated belief about the parameters after having seen the data.
4. Analyze the posterior distribution and draw conclusions from it.

Looking back at the definition of probability, the Bayesian approach of probability can be seen as a degree of belief that an even will occur, or more common in forensic statistics, the degree of belief that an event has occurred under certain circumstances. Because we take a prior into account, there can be discussion over this prior between different forensic experts, which makes the Bayesian approach subjected to discussion. the Bayesian decision frame is however still one of the most used methods in forensic statistics.

3.1.3. Parameter distinction

The best way to distinguish both methods is by looking at how both methods define the relevant parameters of a random variable [31]. The Bayesian approach treats parameters as being random variables whereas from a frequentist point of view, the parameters are not seen as random but as fixed variables, whose value is unknown. Because in the Bayesian approach, the parameters are treated as random variables, distributions can be assigned to them, where they can be updated by using the prior distributions. Therefore in the Bayesian view, a probability is assigned to a hypothesis. However from the frequentist point of view there does not exist any distribution regarding the parameters, therefore the hypothesis is tested without being assigned to a probability. The frequentist approach can be seen as a method where you want to know, given all possible sets of data that you could possibly observe, which parameter setting best represents the actual outcomes [3]. Because, as stated before, the parameters are unknown but fixed variables, the data will be the random variable which we can take expectations over. In the Bayesian approach, the parameters are random variables substituted to a distribution. This view of looking at probability is taught at most undergraduate statistics courses, where for example random variable X is distributed according to a normal distribution $X \sim \mathcal{N}(\mu,\sigma^2)$. The question of interest is, given the data, what the best parameter setting is, which can be seen as a weighted average based on the prior values.

3.2. Parameter setting

To make the parameter distinction in 3.1.3 more precise, consider a simple situation of observing a random variable X, which is constructed as having probability density function $f(x|\theta)$ with θ being the unknown parameter, carrying relevant information of the population of study. For example in the case of a normal distribution, this becomes the mean and standard deviation and the density function will be $f(x|\mu,\sigma)$ in this case. To approach inferential statistics about θ, which is the process of using data analysis to deduce properties of an underlying probability distribution [24], frequentist scientists consider parameter θ as a fixed but unknown number, where no probability statements can be made over. For a Bayesian scientist, θ is subject to many people's uncertainties. Fundamental in the Bayesian methods, is that all uncertainties must be described by probability distributions, where they provide a measure of personal degrees of belief in the occurrence of an event.

Frequentist think of probability in terms of a long-run relative frequency in which events occur under repeated observation, this way of thinking about probability makes it hard to think of probability as a single event, for example the probability that a certain suspect committed a crime at a certain time.

In contrast, the Bayesian method represents probability as a degree of belief in the assertion that the event will happen. Now it is simple to think of probability as a singular event. Bayes theorem specifies the way to do this.

Theorem 3.1 (Bayes Theorem)**.** *Let A and B be two events where $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. Then*

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A)}. \tag{3.1}$$

The uncertainty about a population parameter θ can be modeled by a probability distribution $\pi(\theta)$, called the prior distribution, this distribution captures the available data before the observations. Note that in most circumstances, the interpretation of the population parameter θ depends on the hypothesis and therefore can be seen as θ_p and θ_d.
The parameter θ is treated as random variable in order to describe personal uncertainty about its true value. In Bayesian analysis, parameters are treated as random, whereas observed data x are treated as fixed. Inference is based on the posterior distribution $\pi(\theta|x)$. Making use of Bayes theorem, the distribution on θ conditional on x becomes:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{f(x)}, \tag{3.2}$$

This way, Bayes theorem updates the initial information on the parameters θ by using the data in the observation x where $f(x|\theta)$ is the likelihood function, and $f(x)$ is the marginal distribution x.
Frequentists on the other side, treat data as random, even after observation. Data is seen as a repeatable random sample. Among frequentists, the parameter is considered as a fixed unknown constant to which no probability function can be assigned, since it is not random.

4 Hypothesis testing

Statistical inference can be seen as the process of using a "sample" of data to make a statement about a "population" [11]. The meaning of "sample" and "population" differ per context. In statistics, a sample is mostly referred to as a measurement taken from a larger population. In forensic statistics, in contrast to general statistics, the samples are the pieces of evidence left at the crime scene and therefore we are not free to choose them. In this section, different frequentist methods will be discussed including range tests, p-values and significant levels. Their application will be discussed in general statistics and their pro's and contra's in forensic statistics will be underlined.

We will make the theory described more precise by considering a data set consisting of multiple measurements on seven chemical elements of a certain knife. This dataset has been made available by the department of chemical and physical traces by courtesy of Peter Zoon of the NFI. The dataset has been adapted for experimental usage but originate from real data.

4.1. Knife data

The data comparison problem can be seen as a common source problem. We will compare ten different measurements of one knife to one measurement from another knife. Here seven different features indicated by $A-G$ will be compared;

- e_{u_1}: ten measurements on one knife.

- e_{u_2}: one measurement on one knife.

The first unknown source e_{u_1} will be denoted as the control data, which will be indicated X from now on. The second unknown source e_{u_2} will be denoted as the recovered data and indicated Y. Spectrum two to eleven are provided by the NFI, the data can be structured using tables;

Table 4.1: Measurements on the control data

measurement	item	measurement	A	B	C	D	E	F	G
Spectrum_2	1	1	0.59	0.00	13.77	0.65	84.65	0.34	0.00
Spectrum_3	1	2	0.57	0.00	13.28	0.62	85.17	0.35	0.00
Spectrum_4	1	3	0.59	0.00	14.37	0.68	84.07	0.29	0.00
Spectrum_5	1	4	0.57	0.00	13.37	0.64	85.17	0.26	0.00
Spectrum_6	1	5	0.52	0.00	14.29	0.68	84.26	0.25	0.00
Spectrum_7	1	6	0.55	0.00	13.93	0.64	84.61	0.27	0.00
Spectrum_8	1	7	0.57	0.00	13.90	0.55	84.67	0.30	0.00
Spectrum_9	1	8	0.60	0.00	13.41	0.60	85.06	0.33	0.00
Spectrum_10	1	9	0.58	0.00	13.34	0.73	85.06	0.28	0.00
Spectrum_11	1	10	0.57	0.00	14.03	0.71	84.38	0.31	0.00

Table 4.2: Measurements on the recovered data

measurement	item	measurement	A	B	C	D	E	F	G
Spectrum_100	3	1	0.55	0.00	14.98	0.60	83.44	0.43	0.00

4.2. Range tests

By observing the range of the control measurements, a comparison can be made with the results of the recovered measurement. This easy type of hypothesis testing is the basic step in forensic hypothesis testing.

4.2.1. Minimum maximum range tests

Range tests can be seen as a class of tests which compare statistical properties of every feature in the recovered data to the same statistical properties of the same measurements in the control data. The range set of an observation is defined as the interval of the lowest observed value to the highest observed value.

The simplest range test compares the recovered measurement to the control measurements range. If one of the recovered measurements falls outside the control range, it is assumed not to be originating from the control source. In case of the knife data, we see that feature C (14.98) in the recovered data falls above the maximum of the control data range (14.37). Furthermore, also features E and F fall outside the recovered range (83.44 below minimum 84.07 and 0.43 above maximum 0.35 respectively).

The simple method of range tests unfortunately does not give any insight into the strength of the evidence and is very susceptible for outlying measurements, therefore it can rather be seen as a simple warm up for more advanced testing methods, but should nevertheless not be ignored. The method however is understandable for people with a limited mathematical background and is easily extended to a multivariate test for multiple features.

4.2.2. Two sigma and three sigma tests

The minimum maximum range test compares the recovered measurement to an interval based on the minimum and maximum value of the control data. An outlying high or low value in the control measurements can disturb the interval. To overcome this flaw, an interval can be made based on the standard deviation. We take $\overline{x}$ to be the mean and σ_x to be the standard deviation of the control data. Both $(\overline{x}-2\sigma_x, \overline{x}+2\sigma_x)$ and $(\overline{x}-3\sigma_x, \overline{x}+3\sigma_x)$ can be taken for the confidence interval, which are called the 2σ and 3σ measurement. While 2σ intervals can detect small shifts from the mean, the false exclusion rate is far too high [12]. This false exclusion rate is measured in terms of the type I error:

Definition 4.1 (Type I error). *A type I error occurs when the null hypothesis* (H_0) *is true, but is rejected. This is referred to as false exclusion.*

Let p be the number of features tested, the probability of declaring at least one measurement from the recovered data to be from a different source than the control data while actually being from the same source is defined as [12]:

$$P = 1 - (1-\alpha)^p$$

Where α is 0.05 or 0.003 depending on whether respectively a 2σ or 3σ significance level is applied. This is because of the fact that if X is a normally distributed random variable with mean μ,

$$\mathbb{P}(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95,$$
$$\mathbb{P}(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997.$$

For seven different features, we would obtain seven different confidence intervals. In general if one of the features of the recovered measurement falls outside the range of measurements from the control data, we are not able to confirm a statement about the common source of both measurements. A possible alternative to multivariate confidence intervals are confidence ellipsoids. This method is in practice almost never used [12]. Note that in this approach, as well as other frequentist methods, the hypothesis with respect to the alternative hypothesis is not considered. Furthermore, for small evidence sizes, as common in forensic statistics, the statistical properties can be adversely affected, for instance because of outlying data points.

4.3. Hypothesis tests

To formally test a hypothesis, better tests than range tests need to be provided. It is of importance to introduce the power and significance of a test, which can be used for the validation of a hypothesis test. furthermore we will introduce two well-known quantities in statistics: p-values and confidence intervals.

4.3.1. P-values

One of the basic concepts in statistics is the notion of p-values [11]. The idea of p-values is based on first constructing an appropriate test statistic. The test statistic should be a good summary of what you are interested in and the distribution of the statistic when the null hypothesis is true should be known. The p-value is defined as:

Definition 4.2. *The probability under the assumption of no effect or no difference (null hypothesis), of obtaining a result equal to or more extreme than what was actually observed.*

The "p" stands for probability and measures how likely it is that any observed difference between groups is due to chance. The test statistic X is compared to the observed X_0 specified by the particular case. Put simple we might see a small p-value as a measure of evidence against the null hypothesis. A large p-value will however never be a measure confirming the null hypothesis. What can be seen as a measure against the null hypothesis depends on the situation. If we reject the null hypothesis when the test statistic is significantly larger than the observed statistic, this is called a *right-tailed probability.* If we take X_0 to be the value of the test statistic here, we can write the p-value formally as:

$$P = \mathbb{P}(X \geq X_0 \big| H_0 \text{ true}).$$

This is an example of a *one-tailed probability.* When we replace the "≥" sign by a "≤" sign, we would obtain a *left-tailed probability.* A third option is by looking at the *two-tailed probability,* where no direction in the alternative hypothesis is given. In this case we look for an extreme value:

$$P = \mathbb{P}(|X| \geq X_0 \big| H_0 \text{ true}).$$

4.3.2. Significance and power of a test

With respect to the previous mentioned type I error, a start can be made with defining what could be seen as large and small values. The significance of the test will be denoted as α, representing the probability of making a type I error that the forensic scientist is willing to accept, which is therefore subjective. In most situations in statistics, p-values less than 0.01 or 0.05 are considered to be small. In forensic statistics, this level is considered too high and we prefer a level of $\alpha \leq 0.01$. The level of $\alpha = 0.01$ can roughly be interpreted as less than a one in hundred chance of making a type I error. If a p-value is smaller than the significance level α, the test is said to be significant at the α level.
Decreasing the value of α indicates that we would make fewer false exclusions, however this is not without repercussion. Lowering the significance level α increases the probability of a type II error, which is defined as:

Definition 4.3 (Type II error). *The probability of accepting the null hypothesis when in fact the alternative hypothesis is true.*

The type II error will be denoted β and is known as the probability of false acceptance or false inclusion. The quantity $1-\beta$ is called the power of the test and the quantity α is called the significance level of the test. When we would decrease the significance level α, the number of false positives will drop. Altogether this means decreasing the significance level (α) of a test, decreases the power of the test $(1-\beta)$.

4.3.3. One- and two- sample t-test

An example of using p-values can be found in two sample t-tests. T-tests can be used in testing whether a control sample and a recovered sample originate from the same distribution with same parameters. Conclusions are drawn by taking into account that if two random variables originate from the same distribution, they are indistinguishable and may have a common source. The null hypothesis becomes H_p and the alternative hypothesis becomes H_d. Differences in the sample means are compared to the differences we would expect to have by random chance alone. The idea is based on the probability statement about the true, but unknown differences of the means for the both sources where the samples originate from. Let n_x be the number of measurements on the control variables given by $x_i, i = 1,...,n_x$ and let n_y be the number of measurement from the recovered sample given by $y_j, j = 1,...,n_y$. Both the control sample and the recovered sample are assumed to originate from a normal distribution with mean μ_x, μ_y and standard deviation σ_x, σ_y respectively. This is indicated as $x_i \sim N(\mu_x,\sigma_x)$ and $y_j \sim N(\mu_y,\sigma_y)$. The two sample t-test formally tests whether the distribution means are equal given $\sigma_x = \sigma_y = \sigma$. Therefore the null hypothesis becomes:

$$H_p : \mu_x = \mu_y \text{ or } H_p : \mu_x - \mu_y = 0. \tag{4.1}$$

The alternative hypothesis is given by

$$H_d : \mu_x \neq \mu_y \text{ or } H_d : \mu_c - \mu_r \neq 0. \tag{4.2}$$

the test statistic is now given by

$$T_0 = \frac{\overline{x} - \overline{y}}{\sqrt{(\frac{1}{n_y} + \frac{1}{n_x})\frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2}}}, \tag{4.3}$$

with sample means $\overline{x}$ and $\overline{y}$ and sample standard deviations s_c and s_r. The significance of the test is now found by comparing the test statistic T_0 to the distribution considering the null hypothesis is true. The null hypothesis is given by a t-distribution, which is characterized by the degrees of freedom. In the case of a two sample t-test, the degrees of freedom are given by $df = n_x + n_y - 2$. The p-value is now given by

$$P = \mathbb{P}(|X| \geq X_0 \big| H_0 \text{ true}). \tag{4.4}$$

This means for a prescribed significance level α, we reject the null hypothesis for feature k if

$$|\overline{x}_k - \overline{y}_k| > t_{n_x - n_y - 2}(\alpha)\sqrt{\frac{s_x}{n_x} + \frac{s_y}{n_y}}. \tag{4.5}$$

In the case of only one measurement on the recovered sample Y, the only measurement on the control sample becomes the mean and the standard deviation will be zero. If we still want to do a t-test, as is the case for the knives example, we need to switch to a one sample t-test. The test statistic becomes

$$t = \frac{\overline{x} - \mu}{\frac{s_x}{\sqrt{n_x}}}, \tag{4.6}$$

the total degrees of freedom become $df = n_x + n_y - 2 = n_x + 1 - 2 = n_x - 1$, note that the degrees of freedom represent the sample size and therefore in some sense the amount of evidence available [12]. We remove feature B and G because all values in both the recovered and control measurements are zero and therefore will return a p-value of zero. The following p-values are found now:

Table 4.3: P-values for knife data

feature	t	P-value
A	2.9091	0.01734
C	103.96	3.577e-15
D	5.9526	2.147e-4
E	673.72	2.2e-16
F	-23.264	2.384e-9

For feature D for example, we can state that on average we will encounter a result like this 215 times in one million and therefore probably has not occurred by random chance alone. Before officially rejecting the null hypothesis it must be below the prescribed significance level α. We can make a bridge to the sigma range tests by interpreting the t-statistic as the number of standard deviations that we will be off the mean when the null hypothesis is true.

4.3.4. Multivariate t-tests

The t-test is designed as a univariate test statistic and therefore we will find one p-value for every of the seven features. Different methods are available for taking into account multiple features, this can be either by still performing multiple tests but using correction, or using a multivariate test which takes possible correlation into account. We will take a look at both these methods where we will use Hotellings T^2 test for multivariate testing.

Hotellings two sample T^2 statistic is defined as

$$T^2 = \frac{n_x n_y}{n_x + n_y}(\overline{\mathbf{x}} - \overline{\mathbf{y}})^T S_k^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}}), \tag{4.7}$$

where S_k is an estimate for the pooled covariance matrix,

$$S_k = \frac{\sum_{j=1}^{n_x}(\mathbf{x_j} - \overline{\mathbf{x}})(\mathbf{x_j} - \overline{\mathbf{x}})^{\mathbf{T}} + \sum_{j=1}^{n_y}(\mathbf{y_j} - \overline{\mathbf{y}})(\mathbf{y_j} - \overline{\mathbf{y}})^{\mathbf{T}}}{n_x + n_y - 2} \tag{4.8}$$

and

$$T^2 \sim F_{(p, n_x + n_y - 2)}. \tag{4.9}$$

We reject the null hypothesis if

$$T^2 \geq T^2_{(1-\alpha, p, n_x + n_y - 2)}. \tag{4.10}$$

The Hotellings T^2 statistic can also be used for a one sample T^2 statistic, which we will do to calculate the p-value of the knife data. To calculate the multivariate p-value, we need to remove features B and G because they make the pooled covariance matrix non-invertible. We find the following:

Table 4.4: Hotellings T^2 for knife data

features	T^2	P-value
ACDEF	57.477	8.132e-4

Statistical hypothesis testing is based on rejecting the null hypothesis if the p-value of the observed data under the null hypothesis is low. If multiple hypotheses are tested, the chance of a rare event increases, and therefore, the likelihood of incorrectly rejecting a null hypothesis (i.e., making a type I error) increases [31]. If we use the convention of a p-value of 0.05, the case of p comparisons as in the example,

$$\alpha_{FW} = 1 - (1 - \alpha_{pc})^p,$$

where p is equal to the number of comparisons performed and α_{pc} is equal to the specified per contrast error rate, which is taken to be 0.05. Without any correction, $\alpha_{FW} = 1-(1-\alpha_{pc})^p = 1-(1-0.05)^3 = 0.143$, which makes the probability of erroneously rejecting the null hypothesis (type I error) at least once amongst the family of analyses equal to 14.3%. To avoid this phenomenon, a Bonferroni correction can be used:

Definition 4.4. *Let $H_1,\dots,H_m$ be a family of hypotheses and $p_1,\dots,p_m$ their corresponding p-values. Let m be the total number of null hypotheses and m_0 the number of true null hypotheses. The familywise error rate (FWER) is the probability of rejecting at least one true H_i, that is, of making at least one type I error. The Bonferroni correction rejects the null hypothesis for each $p_i \leq \frac{\alpha}{m}$, thereby controlling the FWER at $\leq \alpha$.*

Assume the within-source variability is constant and the degrees of freedom are equal to $N-m-2$, the means $\mathbf{\bar{x}_k}$ and $\mathbf{\bar{y}_k}$ are significantly different at the $100\alpha/p\%$ level in a two-sided t-test if

$$|\overline{x}_k - \overline{y}_k| > t_{(n_x+n_y-2)}(\frac{\alpha}{2p})s_k\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}. \tag{4.11}$$

Here s_k is the pooled within-group standard deviation within group k. Pooled standard deviations need to be taken because the sample can (and in general will) consist of different numbers of measurements.
Although p-values are commonly used in statistics, they are also largely criticized [31], mostly over their incorrect interpretation, therefore it is important to emphasize what the p-value is **not** [35]:

- The p-value is not the probability that the null hypothesis is true, or the probability that the alternative hypothesis is false.
- The p-value is not the probability that the observed effects were produced by random chance alone.

The p-value gives the probability of observing test statistic by random chance alone and therefore if this probability is very low, either that the null hypothesis is true and a highly improbable event has occurred *or* that the null hypothesis is false.

Another point of criticism on p-values is the "fall-of-the-cliff effect" pointed out by Ken Smalldon. The points here is the fact that at a rejection level of 0.05, when a p-value of 0.049 is obtained, the null hypothesis will be rejected whereas it will not be rejected when observing a p-value of 0.051. This reasoning is hard to explain to court officials and an alternative is therefore to provide confidence intervals.

4.4. Confidence intervals

A confidence interval is equivalent with a significance level for a hypothesis test. A confidence level is mostly stated as a percentage, such as a 95% confidence interval. Statements considering a confidence interval are stated as: "We are 100(1-α)% confident that the interval contains the true value or parameter of interest". Note that this idea relies on repeated sampling of an infinite population, which is the core thought of the frequentist approach. A confidence interval is not a statement of probability and this is also one of the points of criticism from the Bayesian scientists. Incorrect statements of the confidence interval are for example:

- "There is a α% chance the true value is in my interval"
- "α% of all data measurements in the population fall within the interval"

for a parameter estimation of θ, denoted $\hat{\theta}$, the confidence interval will look like

$$\hat{\theta} \pm z_\alpha se(\hat{\theta})$$

Where $se(\hat{\theta})$ is the standard error of the estimate and z_α is a multiplier depending on the chosen interval.
A α% confidence interval for the difference in mean will look like

$$\overline{x} - \overline{y} = t^*_{df}\left(1 - \frac{\alpha}{2}\right) se(\overline{x} - \overline{y}),$$

under the assumption that the samples originate from population with the same variance, the standard error is defined by

$$se(\overline{x} - \overline{y}) = \sqrt{\left(\frac{1}{n_y} + \frac{1}{n_x}\right)\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}}$$

with the degrees of freedom equal to

$$df = n_x + n_y - 2.$$

For the case when the assumption of same variance is dropped, the standard error becomes

$$se(\overline{x} - \overline{y}) = \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}$$

with

$$df = \frac{(\frac{s_x}{n_x} + \frac{s_y}{n_y})^2}{\left(\frac{(\frac{s_x^2}{n_x})^2}{n_x - 1} + \frac{(\frac{s_y^2}{n_y})^2}{n_y - 1}\right)}.$$

in general for a two sided tail test, if a 100(1-α)% confidence interval contains the hypothesized value of interest, the associated p-value from a hypothesis test will be greater than α. Although confidence intervals are mostly used as an alternative to p-values because they should be easier to understand, research under both undergraduates and master students showed no significant difference in the understanding of the both sources [21]. Therefore a new method in forensic statistics using Bayes theorem (see def. 3.1) tries to capture both hypotheses using a likelihood ratio.

5 Likelihood ratios

Evidence found at a crime scene can be summarized using multiple statistical approaches. A major objection against the frequentist methods in Chapter 3 is that only one hypothesis is taken into account. To make sure both the hypothesis of the prosecution and the defense are taken into account, a generally accepted method in forensic statistics is making use of likelihood ratios. By making use of likelihood ratio $\frac{\mathbb{P}(H_p|e,I)}{\mathbb{P}(H_d|e,I)}$, the hypothesis of both the prosecutor and the defender are considered. Here e is the available evidence and I is the background material.

In this chapter, first of all Bayes theorem, as stated in Section 3.1 will be applied to this likelihood ratio in section 5.1 to construct a Bayesian decision framework. The likelihood ratio will be made suitable for both a Bayesian and a frequentist way of doing calculations, which will turn out to be of great importance when doing calculations as is detailed in Sections 6 and 7. Finally, the likelihood ratios will be converted to verbal language in the same form the ENFSI translates the likelihood ratios.

5.1. Applying Bayes theorem

The likelihood ratio to consider is noted as $\frac{\mathbb{P}(H_p|e,I)}{\mathbb{P}(H_d|e,I)}$. From now on we will refer to this as the posterior odds. This odd is posterior in the sense that the evidence e is already given. Using Bayes theorem, the posterior odd will be converted to a prior odd. From a frequentist point of view, no priors should be considered, in this setting only the likelihood ratio is considered.

$$\frac{\mathbb{P}(H_p|e,I)}{\mathbb{P}(H_d|e,I)} = \frac{\mathbb{P}(e,I|H_p)\mathbb{P}(H_p)}{\mathbb{P}(e,I)} \frac{\mathbb{P}(e,I)}{\mathbb{P}(e,I|H_d)\mathbb{P}(H_d)} = \frac{\mathbb{P}(e|H_p,I)}{\mathbb{P}(e,I|H_d,I)} \frac{\mathbb{P}(I|H_p)}{\mathbb{P}(I|H_d)} \frac{\mathbb{P}(H_p)}{\mathbb{P}(H_d)}. \tag{5.1}$$

In general, the background information I is omitted for ease of notation [2]. The likelihood ratio can now be split into two parts, the value of evidence and the prior probability, the following equation summarizes this:

$$\underbrace{\frac{\mathbb{P}(H_p|e)}{\mathbb{P}(H_d|e)}}_{\text{posterior odds}} = \underbrace{\frac{\mathbb{P}(e|H_p)}{\mathbb{P}(e|H_d)}}_{\substack{\text{value of}\\ \text{evidence}}} \underbrace{\frac{\mathbb{P}(H_p)}{\mathbb{P}(H_d)}}_{\text{prior odds}}, \tag{5.2}$$

Given this equation, we are able to say more about the roles of both the forensic expert and the judge. The posterior ratio is split into two parts, the prior probability and the value of evidence, which is a likelihood ratio. The forensic expert is concerned with determining the value of evidence. The legal expert is concerned with the prior odds, which are subjective and should not be analyzed by the forensic expert.

5.2. Verbal likelihood ratio

The likelihood ratio indicates how many times more probable the evidence is given that the prosecution hypothesis is true compared to the defense hypothesis. This ratio turns out to be hard to interpret for court officials. An example of false interpretation is the $boomerangeffect$, where weak evidence supporting the hypothesis of the prosecutor is wrongly interpreted as evidence supporting the hypothesis of the defence. To overcome misinterpretations among court officials, sometimes only verbal likelihood ratios are given. In 2014, the NFI suggested a verbal likelihood framework to translate numerical values to verbal likelihood ratio scales.

Likelihood ratio (LR)	**Verbal LR scale**
1-2	The forensic findings do not support H_d over H_p
2-10	The forensic findings provide **weak** support for H_p rather than for H_d
10-100	The forensic findings provide **moderate** support for H_p rather than for H_d
100-1,000	The forensic findings provide **moderate strong** support for H_p rather than for H_d
1,000-10,000	The forensic findings provide **strong** support for H_p rather than for H_d
10,000-1,000,000	The forensic findings provide **very strong** support for H_p rather than for H_d
>1,000,000	The forensic findings provide **extremely strong** support for H_p rather than for H_d

Table 5.1: ENFSI verbal likelihood

The inverse of the scale is used as support for H_d over H_p. Formulating LRs verbally tries to make forensic statistics more objective and easier to interpret for court officials. It was first introduced by Evett [17] and worked out in more detail by Nordgaard [27].

5.3. SAILR software

To calculate LRs in a uniform way, the ENFSI has set a unified software framework to calculate LRs, called SAILR; Software for Analysis and Implementation of Likelihood Ratios, this program calculates LRs in a frequentist way, which can be done both feature based (Chapter 7 and 6) or score based (Chapter 8.2). The SAILR software has a user-friendly graphical interface that calculates the likelihood ratios given the chosen model and specifications. The model provides validation histograms and is equipped with a verbal LR scale in both English (option called ENFSI), Swedish and Dutch. At the time of writing SAILR 1.3.0 is available upon request, further versions are still in development.

6 Hierarchical model on forensic glass

Non-biological forensic evidence is generally referred to as trace evidence, such as paint, fibers or glass but also impressions of fingerprints or shoe marks fall into this broad category. Glass evidence forms one of the most largely used trace evidences in forensic statistics and besides the complex field of DNA, is the part of evidence interference that makes most use of statistics. Glass evidence interpretation can be used on a large spread of glass traces. Of course we can think of a classic example of a glass window being broken during a burglary, but also broken windscreen glass or broken bottles that arise during the flight from a crime scene can be used for this type of trace evidence. In this chapter a start will be made with describing the Bayesian hierarchical random effect model used in forensic statistics. This model has been applied on forensic sources for the first time by Lindsey in 1977 on the refractive index of glass and therefore we will start the outline of the hierarchical model using this foundation in forensic statistics. In this chapter, only the variance estimation models will be described without applying the model on real glass evidence samples. In Chapter 7 however, the model will be applied for comparison of MDMA tablets.

6.1. Techniques in glass comparison

To compare both our control and recovered sample, we need to quantify our evidence. Evidence can be either discrete such as color or continuous properties such as thickness and density of chemical components. The two most used methods for comparison are refractive index (RI) and elemental analysis. The elemental analysis suffers from two major shortcomings. First of all it is very slow, second of all, it is destructive, which in the case of a very small recovered sample becomes a major drawback for this type of evidence analysis. Because of these two large drawbacks, RI will be used more often in practice. The RI of a material is a dimensionless number that describes how fast light propagates through the material, which depends on the density of the material [30].

6.1.1. Test outline

Suppose that at the crime scene, the forensic experts have taken a sample of size n_x control fragments on remaining glass, while the forensic laboratory has recovered n_y fragments from the suspect's clothing. For each of the fragments, a RI measurement has been performed, which gives $x_1, x_2, \ldots, x_{n_x}$ measurements on the control sample and $y_1, y_2, \ldots, y_{n_y}$ measurements on the recovered sample. For now we have taken one measurement on each source, but we could have taken multiple measurements on one source. Performing multiple measurements on the recovered source can be difficult because the glass fragments recovered from the suspect's clothing will be very small in general. In reality, also the number of measurements on the control sources will be limited because of technical or financial limitations.

Even if the control and recovered samples have the same origin, there will be a difference in the set of measurements for the two samples because of variation between the fragments, due to internal variation and variability in the measurement techniques. It is crucial to determine

whether the variation between the sources is due to internal variation and variability of the measurement techniques or that the variation is caused because the samples originate from different sources.
To test whether the mean μ_x of the samples of the control variable and the mean μ_y of the recovered sample are equal, we construct hypothesis $H_0 : \mu_c - \mu_r = 0$. Now we can start with defining the random effect model, which is a feature based model.

6.2. Random effect model

To test the stated null hypotheses, the two competing hypotheses are determined by parametric models for the data up to the point of a finite dimensional vector space for the indexing parameter θ. The SAILR program offers the possibility to make use of a non-parametric kernel density estimations (KDE), which will be discussed in chapter 7. If a distinction between two different glass components must be made, which is a common source problem, it is important to decide in advance which discriminatory elements in glass we are going to use to detect similarities between the two fragments of glass. For a univariate comparison, the previous mentioned RI can be used [30]. For a multivariate comparison, p different relevant features can be taken for each of the elements. These features will consist of the proportion of certain chemicals in the glass fragments.

6.2.1. Between and within source variation

In the random effect model, we take into account background data. By using background data, an estimate for the between source variation can be made. Within each of the m different sources, n_i different measurements are made, therefore the background data (of course also consisting of the same p features) becomes;

$$\mathbf{Z_{ij}} = \begin{bmatrix} z_{ij1} \\ z_{ij2} \\ \vdots \\ z_{ijp} \end{bmatrix} \qquad i = 1,\ldots,m; \quad j = 1,\ldots,n_i. \tag{6.1}$$

$$\text{with} \quad \mathbf{\bar{z}_{i\bullet}} = \sum_{j=1}^{n_i} \frac{1}{n_i}\mathbf{z_{ij}} \tag{6.2}$$

$$\text{and} \quad \mathbf{\bar{z}_{\bullet\bullet}} = \frac{1}{m}\sum_{i=1}^{m} \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{z_{ij}}. \tag{6.3}$$

Note that $\mathbf{\bar{z}_{i\bullet}}$ is a matrix consisting of the means from all i sources for all p feature and $\mathbf{\bar{z}_{\bullet\bullet}}$ is a vector consisting of the overall means for all p features The same matrix and vector can be constructed for $\mathbf{X}$ and $\mathbf{Y}$. With respect to the background data $\mathbf{Z}$, the random effect model assumes two sources of variation; first of all, the *within-source* variation, this is the variance within source i over measurements n_i. Secondly, the *between-source* variation, this is the variation over all measurements. For the within-source variation, it is reasonable to assume the variation is only caused because of noise and therefore taken to be normally distributed. The between-source variation can be either normally distributed or estimated with a kernel density estimate.

6.3. Glass variation

Back to the glass example, consider m different sources with n_i measurements per source i. For example m windows with n_i measurements per window. The total number of measurements is now equal to $N = \sum_{i=1}^{m} n_i$. For each of the fragments, a certain number of ratios of chemical elements is tested, denoted by p. Consider for example Calcium (Ca), Potassium (K), Silicon (Si) and Iron (Fe). We can now look at the log ratios of Ca/K, Ca/Si and Ca/Fe. The choice for these three ratios is based on expertise of forensic experts, who argue that these three ratios are the most discriminatory. The choice for the natural logarithm ensures that the analysis is not relying on which of the two elements is chosen to be the denominator and which to be the nominator. The natural logarithm furthermore reduces positive skewness and makes the data more likely to be normally distributed, which is a requirement for the within-source variation.

6.3.1. Within-source

The mean vector within source i is denoted by $\theta_\mathbf{i}$ and the matrix of within-source covariance is denoted by Σ. We now have that $\mathbf{Z_{ij}}$ given $\theta_\mathbf{i}$ and Σ, is normally distributed, so:

$$(\mathbf{Z_{ij}}|\theta_i,\Sigma) \sim \mathcal{N}(\theta_i,\Sigma), \quad i=1,\dots,m; \quad j=1,\dots,n_i.$$

6.3.2. Between-source

The general mean vector between all sources is denoted by μ and the between-source covariance matrix is denoted by T. The distribution of mean vector within source i, $\theta_\mathbf{i}$, can be expressed in terms of the between source variation. For now this variation is taken to be normal, but with SAILR we could also use a kernel density estimation (KDE). Assuming normality for the between-source variation gives:

$$(\theta_\mathbf{i}|\mu,T) \sim N(\mu,T), \quad i=1,\dots,m. \tag{6.4}$$

The distribution of the measurements on the control and recovered data, $\mathbf{X}$ and $\mathbf{Y}$, are also taken to be normal, conditional on the source. This gives rise to the distribution of $\mathbf{X}$ and $\mathbf{Y}$ being normal with means θ_x and θ_y and covariance matrix $D_x = n_x^{-1}\Sigma$ and $D_y = n_y^{-1}\Sigma$, following:

$$(\bar{\mathbf{X}}_{\bullet\bullet}|\theta_x,D_x) \sim \mathcal{N}(\theta_x,D_x) \tag{6.5}$$

$$(\bar{\mathbf{Y}}_{\bullet\bullet}|\theta_y,D_y) \sim \mathcal{N}(\theta_y,D_y). \tag{6.6}$$

Because of the previous assumption of between-source normality,

$$(\bar{\mathbf{X}}_{\bullet\bullet}|\mu,T,D_x) \sim \mathcal{N}(\mu,T+D_x) \tag{6.7}$$

$$(\bar{\mathbf{Y}}_{\bullet\bullet}|\mu,T,D_y) \sim \mathcal{N}(\mu,T+D_y). \tag{6.8}$$

6.4. Variance estimation

When modelling hierarchical simple random effects, the analysis of variance technique is commonly used to estimate the within source covariance matrix Σ and between source covariance matrix T. The estimations are based on the background information using the following identity [22]

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mathbf{Z_{ij}}-\bar{\mathbf{Z}}_{\bullet\bullet})(\mathbf{Z_{ij}}-\bar{\mathbf{Z}}_{\bullet\bullet})^T = \sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mathbf{Z_{ij}}-\bar{\mathbf{Z}}_{\mathbf{i}\bullet})(\mathbf{Z_{ij}}-\bar{\mathbf{Z}}_{\mathbf{i}\bullet})^T + \sum_{i=1}^{m} n_i(\bar{\mathbf{Z}}_{\mathbf{i}\bullet}-\bar{\mathbf{Z}}_{\bullet\bullet})(\bar{\mathbf{Z}}_{\mathbf{i}\bullet}-\bar{\mathbf{Z}}_{\bullet\bullet})^T \tag{6.9}$$

The left-hand side here is called the total sum of squares (TSS). The two terms on the right hand side are called the within group sums of squares (SS_W) and between group sums of squares (SS_B). The expectation of SS_W can found to be [22]:

$$E(SS_W) = \Sigma(N-m) \tag{6.10}$$

with N the total number of observations and m the number of sources. An estimate for the within source covariance matrix becomes

$$\hat{\Sigma} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\mathbf{Z_{ij}}-\bar{\mathbf{Z}}_{\mathbf{i}\bullet})(\mathbf{Z_{ij}}-\bar{\mathbf{Z}}_{\mathbf{i}\bullet})^T}{N-m} \tag{6.11}$$

For the estimate of the between source covariance matrix, we take $\mu = \bar{\mathbf{z}}_{\bullet\bullet} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{Z}_{ij} = \frac{1}{m}\sum_{i=1}^{m}\bar{\mathbf{Z}}_{\mathbf{i}\bullet}$. This is an unweighted mean because every sources is given equal weight. For ease of notation, $\bar{\mathbf{Z}}_{\bullet\bullet}$ will be denoted $\bar{\mathbf{Z}}$ and $\bar{\mathbf{Z}}_{i\bullet}$ will be denoted $\bar{\mathbf{Z}}_i$. The estimation of T now becomes;

$$\mathbb{E}(SS_B) = \sum_{i=1}^{m} n_i \{\mathbb{E}(\bar{\mathbf{Z}}_\mathbf{i}\bar{\mathbf{Z}}_\mathbf{i}{}^T) - \mathbb{E}(\bar{\mathbf{Z}}_\mathbf{i}\bar{\mathbf{Z}}^T) - \mathbb{E}(\bar{\mathbf{Z}}\bar{\mathbf{Z}}_\mathbf{i}{}^T) + \mathbb{E}(\bar{\mathbf{Z}}\bar{\mathbf{Z}}^T)\}$$

$$\text{where}\quad \mathbb{E}(\bar{\mathbf{Z}}_\mathbf{i}\bar{\mathbf{Z}}_\mathbf{i}{}^T) = Cov(\bar{\mathbf{Z}}_\mathbf{i}, \bar{\mathbf{Z}}_\mathbf{i}) + \mathbb{E}(\bar{\mathbf{Z}}_\mathbf{i})\mathbb{E}(\bar{\mathbf{Z}}_\mathbf{i}) = n_i^{-1}\hat{\Sigma} + \hat{T} + \mu\mu^T$$

$$\mathbb{E}(\bar{\mathbf{Z}}_\mathbf{i}\bar{\mathbf{Z}}^T) = Cov(\bar{\mathbf{Z}}_\mathbf{i}, \bar{\mathbf{Z}}) + \mathbb{E}(\bar{\mathbf{Z}})\mathbb{E}(\bar{\mathbf{Z}}_\mathbf{i}) = Cov(\bar{\mathbf{Z}}_\mathbf{i}, \frac{1}{m}\sum_{i=1}^{m} \bar{\mathbf{Z}}_{\mathbf{i}\bullet}) + \mu\mu^T$$

$$\text{with}\quad Cov(\bar{\mathbf{Z}}_\mathbf{i}, \frac{1}{m}\sum_{i=1}^{m} \bar{\mathbf{Z}}_{\mathbf{i}\bullet}) = Cov(\bar{\mathbf{Z}}_\mathbf{i}, \frac{n_i}{m}\bar{\mathbf{Z}}_i) = \frac{n_i}{m} Cov(\bar{\mathbf{Z}}_\mathbf{i}, \bar{\mathbf{Z}}_\mathbf{i})$$

$$\text{by independence of}\quad \bar{\mathbf{Z}}_\mathbf{i} \quad\text{and}\quad \bar{\mathbf{Z}}_\mathbf{i}{}^T$$

$$= \frac{n_i}{N}(T + n_i^{-1}\Sigma) + \mu\mu^T$$

$$\mathbb{E}(\bar{\mathbf{Z}}\bar{\mathbf{Z}}^T) = Cov(\bar{\mathbf{Z}}, \bar{\mathbf{Z}}) + \mathbb{E}(\bar{\mathbf{Z}})\mathbb{E}(\bar{\mathbf{Z}}) = Cov(\frac{1}{m}\sum_{j=1}^{m} \bar{\mathbf{Z}}_{\mathbf{j}\bullet}, \frac{1}{m}\sum_{j=1}^{m} \bar{\mathbf{Z}}_{\mathbf{j}\bullet}) + \mu\mu^T$$

$$= \sum_{j=1}^{m} \Big(\frac{n_j}{N})^2 (T + n_j^{-1}\Sigma\Big) + \mu\mu^T$$

which comes down to

$$E(SS_W) = \sum_{i=1}^{m} n_i \Big\{ n_i^{-1}\Sigma + T + \mu\mu^T - 2(\frac{n_i}{N}(T + n_i^{-1}\Sigma + \mu\mu^T) + \sum_{j=1}^{m} \frac{n_j}{N}(T + n_j^{-1}\Sigma) + \mu\mu^T \Big\} \tag{6.12}$$

$$= \sum_{i=1}^{m} n_i \Big\{ \Sigma(n_i^{-1} - 2\frac{n_i}{N} n_i^{-1} + \sum_{j=1}^{m} (\frac{n_j}{N})^2 n_j^{-1}) \Big\} + \sum_{i=1}^{m} n_i \{ T(1 - 2\frac{n_i}{N} + \sum_{j=1}^{m} (\frac{n_j}{N})^2)\}$$

$$= \Sigma\Big(m - 2\sum_{i=1}^{m} \frac{n_i}{N} + N\sum_{i=1}^{m} \frac{n_i}{N^2} \Big) + T\Big(N - 2\sum_{i=1}^{m} \frac{n_i^2}{N} + N\sum_{i=1}^{m} \frac{n_i^2}{N^2} \Big),$$

such that

$$\hat{T} = \frac{\sum_{i=1}^{m} n_i (\mathbf{Z}_\mathbf{i} - \mathbf{Z})(\mathbf{Z}_\mathbf{i} - \mathbf{Z})^T - \hat{\Sigma}(m - 2\sum_{i=1}^{m} \frac{n_i}{N} + \sum_{i=1}^{m} \frac{n_i}{N})}{N - 2\sum_{i=1}^{m} \frac{n_i^2}{N} + N\sum_{i=1}^{m} \frac{n_i^2}{N^2}} \tag{6.13}$$

$$= \frac{\sum_{i=1}^{m} n_i (\mathbf{Z}_\mathbf{i} - \mathbf{Z})(\mathbf{Z}_\mathbf{i} - \mathbf{Z})^T - \hat{\Sigma}(m-1)}{N - 2\sum_{i=1}^{m} \frac{n_i^2}{N} + N\sum_{i=1}^{m} \frac{n_i^2}{N^2}}.$$

Now denote

$$MS_w = \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k_i} \frac{1}{n_i - 1}(\mathbf{Z}_\mathbf{ij} - \bar{\mathbf{Z}}_{\mathbf{i}\bullet})(\mathbf{Z}_\mathbf{ij} - \bar{\mathbf{Z}}_{\mathbf{i}\bullet})^\mathbf{T} = \hat{\Sigma}, \tag{6.14}$$

$$MS_b = \frac{1}{m-1}\sum_{i=1}^{m} \frac{1}{n_i}(\bar{\mathbf{Z}}_{\mathbf{i}\bullet} - \bar{\mathbf{Z}}_{\bullet\bullet})(\bar{\mathbf{Z}}_{\mathbf{i}\bullet} - \bar{\mathbf{Z}}_{\bullet\bullet})^\mathbf{T}. \tag{6.15}$$

Which gives

$$= \frac{(m-1)MS_b - (m-1)MS_w}{N - 2\sum_{i=1}^{m} \frac{n_i^2}{N} + N\sum_{i=1}^{m} \frac{n_i^2}{N^2}} \tag{6.16}$$

$$= \frac{(m-1)MS_b - (m-1)MS_w}{N - \frac{\sum_{i=1}^{m} n_i^2}{N}}$$

$$= \frac{MS_b - MS_w}{\kappa} \quad\text{with}\quad \kappa = \frac{1}{m-1}\Big(N - \frac{\sum_{i=1}^{m} n_i^2}{N} \Big). \tag{6.17}$$

The maximum likelihood estimate for the population variance T is now estimated. This matrix is equal to the between-batch variance of the background data and is corrected for random error i.e. noise, which is called the within-batch variance. Summarizing for the between source variance matrix T we find,

$$\begin{aligned} \hat{T} &= \frac{MS_b - MS_w}{\kappa} \\ \text{with} \quad MS_w &= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{k_i}\frac{1}{n_i-1}(\mathbf{Z_{ij}} - \bar{\mathbf{Z}}_{\mathbf{i}\bullet})(\mathbf{Z_{ij}} - \bar{\mathbf{Z}}_{\mathbf{i}\bullet})^{\mathbf{T}} \\ MS_b &= \frac{1}{m-1}\sum_{i=1}^{m}\frac{1}{n_i}(\bar{\mathbf{Z}}_{\mathbf{i}\bullet} - \bar{\mathbf{Z}}_{\bullet\bullet})(\bar{\mathbf{Z}}_{\mathbf{i}\bullet} - \bar{\mathbf{Z}}_{\bullet\bullet})^{\mathbf{T}} \\ \kappa &= \frac{1}{m-1}\Big(\sum_{i=1}^{m} - \frac{\sum_{i=1}^{m} n_i^2}{\sum_{i=1}^{m} n_i}\Big) \end{aligned}$$

Now that we have defined both variances, we are able to construct the *hierarchical random effects model*, in forensic statistics commonly referred to as *two-level normal-normal* model. Let z_{ij} denote a p-dimensional vector of measurements on the j^{th} component from the i^{th} source for $i = 1,\ldots,m; \quad j = 1,2,\ldots,n_i$. The simple random effect model is given by

$$Z_{ij} = \mu_a + a_i + w_{ij}, \tag{6.18}$$

where $a_i \overset{iid}{\sim} \mathcal{N}_k(\mathbf{0},\hat{\Sigma})$ and $w_{ij} =\overset{iid}{\sim} \mathcal{N}_k(\mathbf{0},\hat{T})$ are independent of each other. This model can be seen as a combination of multivariate normal random vectors. Therefore Z_{ij} follows a simple random effect model. The vector $\overline{\mathbf{Z}}_{\mathbf{i}\bullet}$ now has distribution:

$$\overline{\mathbf{Z}}_{\mathbf{i}\bullet} \overset{iid}{\sim} \mathcal{N}_k(\mu_a, \Sigma_C) \tag{6.19}$$

$$\Sigma_c = \begin{bmatrix} \hat{\Sigma}+\hat{T} & \hat{T} & \ldots & \hat{T} \\ \hat{T} & \hat{\Sigma}+\hat{T} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \hat{T} \\ \hat{T} & \ldots & \hat{T} & \hat{\Sigma}+\hat{T} \end{bmatrix} \tag{6.20}$$

Returning to the control and recovered data we have:

- X: Control data measurements with
  - $\mathbf{X} = \begin{bmatrix} x_{11} & \ldots & x_{n_x1} \\ x_{12} & \ddots & x_{n_x2} \\ \vdots & \ddots & \vdots \\ x_{1p} & \ldots & x_{n_xp} \end{bmatrix}$
  - $\mathbf{X_j} = [x_{j1}, x_{j2}, \ldots, x_{jp}]$ with x_j the p measurements on measurement j
  - summary $\hat{\theta}_x = \bar{x} = \frac{1}{n_x}\sum_{j=1}^{n_x} x_j$
- Y: Recovered data measurements with
  - $\mathbf{Y} = \begin{bmatrix} y_{11} & \cdots & y_{n_y1} \\ y_{12} & \ddots & y_{n_y2} \\ \vdots & \ddots & \vdots \\ y_{1p} & \cdots & y_{n_yp} \end{bmatrix}$
  - $\mathbf{Y_j} = [y_{j1}, y_{j2}, \ldots, y_{jp}]$ with y_j the p measurements on measurement j
  - summary $\hat{\theta} = \bar{y} = \frac{1}{n_y}\sum_{j=1}^{n_y} y_j$

6.4.1. Rewriting the likelihood ratio

Making use of continuous data provides more difficulty in the model. Recall the likelihood ratio in (5.2):

$$LR = \frac{\mathbb{P}(e|H_p)}{\mathbb{P}(e|H_d)}.$$

A classical model in forensic comparison often uses distance measures between the characteristics of the compared items in the evidence e, these two characteristics are the control item and the recovered item, this way we can split the evidence e into the characteristics X and the characteristics of Y. The score based models define the distance between the different features and then implement the distance scores into a function, which will be done in chapter 8.2. For the feature based models, the summary statistics will be compared, which will be done by rewriting the likelihood ratio

$$LR = \frac{\mathbb{P}(X,Y|H_p)}{\mathbb{P}(X,Y|H_d)} = \frac{\mathbb{P}(X|H_p)}{\mathbb{P}(X|H_d)} * \frac{\mathbb{P}(Y|X,H_p)}{\mathbb{P}(Y|X,H_d)}. \tag{6.21}$$

This likelihood ratio can be rewritten by making use of the fact that the probability of observing the characteristics of X does not depend on the hypothesis and therefore $\frac{\mathbb{P}(X|H_p)}{\mathbb{P}(X|H_d)} = 1$. Under the hypothesis of the defense, X and Y originate from different sources and therefore the two samples will be independent of each other. Conditioning on one of these gives no further information, therefore $\mathbb{P}(Y|X,H_d) = \mathbb{P}(Y|H_d)$. Applying this we find;

$$LR = \frac{\mathbb{P}(X,Y|H_p)}{\mathbb{P}(X,Y|H_d)} = \frac{\mathbb{P}(X|H_p)}{\mathbb{P}(X|H_d)} * \frac{\mathbb{P}(Y|X,H_p)}{\mathbb{P}(Y|X,H_d)} = \frac{\mathbb{P}(Y|X,H_p)}{\mathbb{P}(Y|H_d)}. \tag{6.22}$$

In the case of continuous characteristics, which is the case when we perform measurements instead of discrete counts, the probabilities in both the denominator and nominator are replaced by probability density functions. Here $f(x,y)$ is the joint density function and $f(x)$ and $f(y)$ are marginal distribution functions,

$$LR = \frac{f(y|x,H_p)}{fy|H_d}. \tag{6.23}$$

To make this applicable for the available data, the averages $\overline{x}$ and $\overline{y}$ are taken for every batch. Moreover to make the likelihood ratio multivariate, averaged vectors $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are used. A *two level random effect model* is now used to account for both the within and between source variation. This model is often referred to as a *two-level normal-normal* model because it models two levels of variation; the variation within each batch and the variation between different batches, assuming normality for both of the variations.
The general model now becomes

$$LR = \frac{f(\bar{\mathbf{y}}|\bar{\mathbf{x}},H_p)}{f(\bar{\mathbf{y}}|H_d)} = \frac{\int f(\bar{\mathbf{y}}|\bar{\mathbf{x}},H_p)d\boldsymbol{\theta}}{\int f(\bar{\mathbf{y}}|H_d)d\boldsymbol{\theta}} \tag{6.24}$$

In this equation $\boldsymbol{\theta}$ represents the parameters of each batch, which in the case of a normal distribution are the batch mean and the batch standard deviation. Because we use continuous data we can integrate over the probability density functions.

6.5. Two-level normal-normal model

To apply the two-level normal-normal model, we first look at the univariate case before considering the multivariate equivalent.

6.5.1. Two-level normal-normal univariate

Remember that the characteristics of items in X and Y are normally distributed within their source. We start with the univariate case where we only look at one feature, so $p = 1$:

$$X_i \sim N(\theta_X, \sigma_X^2) \quad Y_i \sim N(\theta_Y, \sigma_Y^2)$$

The measurements will now only vary due to random effect. The variances of each batch are assumed to be known, whereas the means are assumed to follow a normal distribution, which is to say the batch means follow a normal prior;

$$p(\theta_i) \sim N(\mu_0, \tau_0^2)$$

For the posterior distribution, the prior is updated by making use of the distribution of the mean $\bar{x}$, a Bayesian construction framework is applied [19]:

$$p(\theta|\bar{x}) \sim N(\mu_n, \tau_n^2)$$

$$\text{with} \quad \mu_n = \frac{\frac{\mu_0}{\tau_0} + \frac{n_x}{\sigma^2}\bar{x}}{\frac{1}{\tau_0^2} + \frac{n_x}{\sigma^2}}$$

$$\text{and} \quad \tau_n^2 = \frac{\tau_0^2\sigma^2}{\sigma^2 + n_x\tau_0^2}$$

When applying the posterior and prior distribution to the likelihood ratio of (6.24) with mean $\bar{y}$ of the recovered measurements the following likelihood ratio is obtained,

$$LR = \frac{\int f(\bar{y}|\bar{x}, H_p)d\theta}{\int f(\bar{y}|H_d)d\theta} = \frac{\int f(\bar{y}|\theta, H_p)p(\theta|\bar{x}, H_p)d\theta}{\int f(\bar{y}|\theta, H_d)p(\theta|H_d)d\theta} \tag{6.25}$$

$$= \frac{\frac{1}{\sqrt{2\pi}u_n}\exp\left\{-\frac{(\bar{y}-\mu_n)^2}{2u_n^2}\right\}}{\frac{1}{\sqrt{2\pi}u_0}exp\left\{-\frac{(\bar{y}-\mu_0)^2}{2u_0^2}\right\}} = \frac{u_0}{u_n}\exp\left\{\frac{1}{2}\left(\frac{(\bar{y}-\mu_0)^2}{u_0^2} - \frac{(\bar{y}-\mu_n)^2}{u_n^2}\right)\right\}$$

$$\text{with} \quad u_0^2 = \tau_0^2 + \frac{\sigma_y^2}{n_y}$$

$$\text{and} \quad u_n^2 = \tau_n^2 + \frac{\sigma_y^2}{n_y}.$$

The first term in the exponent measures rarity and therefore the likelihood ratio should become large when the rarity is large. The second term discounts for random error effect and therefore the likelihood ratio becomes small when there is a large random error.

6.5.2. two-level normal-normal multivariate

To analyze all p features, a multivariate model is considered, the likelihood ratio is constructed the same way as before:

$$LR = \frac{f(\mathbf{Y}|\mathbf{X}, H_p)}{f(\mathbf{Y}|H_p)} = \frac{f(\{y_d, y_{th}, \ldots, y_w\}|\{x_d, x_{th}, \ldots, x_w\}, H_p)}{f(\{y_d, y_{th}, \ldots, y_w\}|H_d)} \tag{6.26}$$

The likelihood ratio is now calculated using multivariate distributions for the measurements within the batch hence $X \sim \mathcal{N}(\theta, \Sigma)$, $X \sim (\theta, \Sigma)$, and multivariate normal distribution for the batch means $(\theta_i|\Sigma) \sim \mathcal{N}(\mu_0, T_0)$. Without going into detail, the likelihood ratio is now given by;

$$LR = \frac{|\mathbf{U_0}|^{\mathbf{1/2}}}{|\mathbf{U_n}|^{\mathbf{1/2}}}exp[\frac{1}{2}((\bar{\mathbf{y}} - \boldsymbol{\mu}_\mathbf{0})^T\mathbf{U_0}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_\mathbf{0}) - (\bar{\mathbf{y}} - \boldsymbol{\mu}_\mathbf{n})^T\mathbf{U_n}^{-1}(\bar{\mathbf{y}} - \boldsymbol{\mu}_\mathbf{n}))] \tag{6.27}$$

$$\text{with} \quad \mathbf{U_0} = \mathbf{T_0} + \mathbf{n_y^{-1}}\Sigma_\mathbf{y} \tag{6.28}$$

$$\mathbf{U_n} = \mathbf{T_n} + \mathbf{n_y^{-1}}\Sigma_\mathbf{y} \tag{6.29}$$

$$\boldsymbol{\mu}_\mathbf{n} = \mathbf{T_0}(\mathbf{T_0} + \mathbf{n_x^{-1}}\Sigma_\mathbf{x})^{-\mathbf{1}}\bar{\mathbf{x}} + \mathbf{n_x^{-1}}(\mathbf{T_0} + \mathbf{n_x^{-1}}\Sigma_\mathbf{x})^{-\mathbf{1}}\boldsymbol{\mu}_\mathbf{0} \tag{6.30}$$

$$\mathbf{T_n} = \mathbf{T_0} - \mathbf{T_0}(\mathbf{T_0} - \mathbf{n_x^{-1}}\Sigma_\mathbf{x})^{-\mathbf{1}}\mathbf{T_0}. \tag{6.31}$$

7 CHAMP drugs project

The illegal production of MDMA tablets (3,4-methyleendioxymethamfetamine) [16] remains a major problem to the Dutch government and its citizens. A report from the European Monitoring Center for Drugs and Drug Addiction (EWD) from 2017 stated that the drug production of MDMA tablets, commonly referred to as *Ectasy (XTC)*, is centered around the Netherlands and Belgium. The Dutch general prosecutor revealed that in 2017, seventeen MDMA-labs and twelve tableting locations have been coiled [25]. To overcome the problem of international confidential information exchange, The European Network of Forensic Science Institutes (ENFSI) was founded in 1995 with the purpose of improving the mutual exchange of information in the field of forensic science.
The CHAMP project (Collaborative Harmonization of Methods for Profiling of Amphetamine Type Stimulants) is a collaborative ENFSI project between multiple forensic institutes aimed at finding a uniform method of testing amphetamines. In this chapter, measurements on XTC tablets will be considered using the feature based measurement methods in SAILR. In Section 7.1, an outline of the available data and an outline of the SAILR interface will be provided. The data will be applied in Section 7.2 on the hierarchical model described in Chapter 6 on three different models; one for discrete data only and two for continuous data.
My profound gratitude goes to the NFI for providing me with data samples from the CHAMP project and access to the SAILR packages.

7.1. Comparison

In a comparison model, control data **X** and recovered data **Y** are compared with each other, if recovered data **Y** is assumed not to originate from control data **X**, it is assumed to originate from background data **Z**. In this setting, hypotheses are formulated the following way:

- H_p : **X** and **Y** originate from the same source.
- H_d : **X** and **Y** originate from different sources, hence **Y** originates from the background data **Z**.

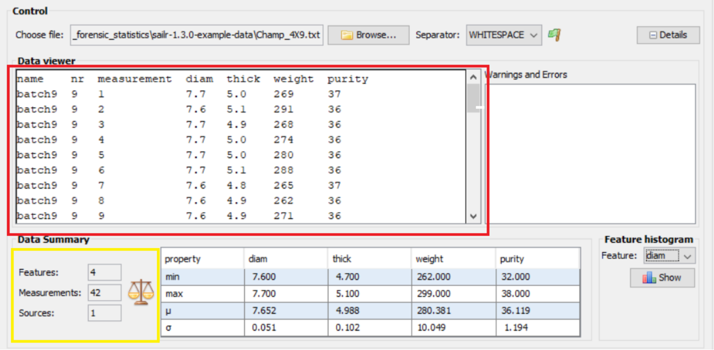
Using the SAILR packages, every source in the recovered data **Y** is independently compared to every control source in **X** using the same background dataset **Z**. If **X** consist of m_x different sources and **Y** consist on m_y different sources, SAILR will return $m_x \times m_y$ different likelihood ratios comparing all different sources. All measurements in **X**, **Y** and **Z** consist of p different features with $k = 1,\dots,p$. In the case of the CHAMP project, four different features will be measured: diameter, thickness, weight and purity. Note that the first three feature are physical features whereas the last one is a chemical feature.

In the case of the considered XTC comparison example, the control data consists of just one source, the number of measurements on this source is a matter of choice from the investigator. The number of measurements on the recovered data is determined by what is available and the investigator therefore has little choice.

Just as in the glass example (Chapter 6), background data will exist of multiple measurements on multiple sources. The number of different sources is denoted by m, where each source has n_i repeated measurements.

7.1.1. SAILR visualization

For the illustrative purpose of this theory, the CHAMP data is analyzed. First of all the control data is analyzed. Measurements from CHAMP **batch 9** are taken, found in `Champ_4X9.txt` provided by the NFI. The visualization tool implemented in SAILR returns;



Choose file: _forensic_statistics\sailr-1.3.0-example-data\Champ_4X9.txt

Separator: WHITESPACE

```
name     nr  measurement  diam  thick  weight  purity
batch9   9   1            7.7   5.0    269     37
batch9   9   2            7.6   5.1    291     36
batch9   9   3            7.7   4.9    268     36
batch9   9   4            7.7   5.0    274     36
batch9   9   5            7.7   5.0    280     36
batch9   9   6            7.7   5.1    288     36
batch9   9   7            7.6   4.8    265     37
batch9   9   8            7.6   4.9    262     36
batch9   9   9            7.6   4.9    271     36
```

Features: 4

Measurements: 42

Sources: 1

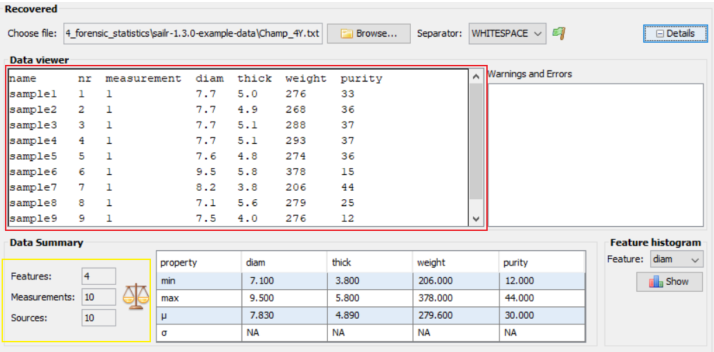
property	diam	thick	weight	purity
min	7.600	4.700	262.000	32.000
max	7.700	5.100	299.000	38.000
μ	7.652	4.988	280.381	36.119
σ	0.051	0.102	10.049	1.194

Feature: diam

Figure 7.1: SAILR visualization of control data

In red, the raw data is given, in yellow the number of features, measurements and sources is given. SAILR furthermore summarizes properties as mean, standard error, minimum and maximum in a clear table for each of the p features. In total 42 measurements on the same batch are taken.

For the recovered samples, `Champ_4Y.txt` is analyzed.



Choose file: 4_forensic_statistics\sailr-1.3.0-example-data\Champ_4Y.txt

Separator: WHITESPACE

```
name      nr  measurement  diam  thick  weight  purity
sample1   1   1            7.7   5.0    276     33
sample2   2   1            7.7   4.9    268     36
sample3   3   1            7.7   5.1    288     37
sample4   4   1            7.7   5.1    293     37
sample5   5   1            7.6   4.8    274     36
sample6   6   1            9.5   5.8    378     15
sample7   7   1            8.2   3.8    206     44
sample8   8   1            7.1   5.6    279     25
sample9   9   1            7.5   4.0    276     12
```

Features: 4

Measurements: 10

Sources: 10

property	diam	thick	weight	purity
min	7.100	3.800	206.000	12.000
max	9.500	5.800	378.000	44.000
μ	7.830	4.890	279.600	30.000
σ	NA	NA	NA	NA

Feature: diam

Figure 7.2: SAILR visualization of recovered data

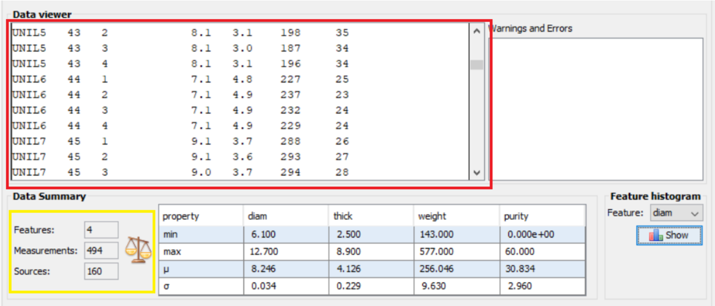
Here the same four features are considered on ten different measurements. The recovered data is constructed as;

- Five samples from **batch 9**
- One sample from **batch 1**
- One sample from **batch 7**
- One sample from **batch 12**
- Two different random samples from the street samples batch **Z**

Constructing the recovered measurements this way provides insight in the strength of the test results. For the five samples from batch 9 it is likely that their likelihood ratios will be high.

Samples six, seven and eight originate from different batches and therefore their likelihood ratio is expected to be low. Finally the last two samples originate from an unknown background batch and therefore no prior assumptions can be made. Because all measurements in the recovered measurements are from a different source, a standard error is meaningless and not applicable.

Finally `Champ_4Z_street.txt` is taken for the background information. This large background information set combines data from different research institutes. In total this set consists of 494 measurements on 160 different sources.



```
UNIL5    43    2              8.1    3.1    198      35
UNIL5    43    3              8.1    3.0    187      34
UNIL5    43    4              8.1    3.1    196      34
UNIL6    44    1              7.1    4.8    227      25
UNIL6    44    2              7.1    4.9    237      23
UNIL6    44    3              7.1    4.9    232      24
UNIL6    44    4              7.1    4.9    229      24
UNIL7    45    1              9.1    3.7    288      26
UNIL7    45    2              9.1    3.6    293      27
UNIL7    45    3              9.0    3.7    294      28
```

Features: 4
Measurements: 494
Sources: 160

property	diam	thick	weight	purity
min	6.100	2.500	143.000	0.000e+00
max	12.700	8.900	577.000	60.000
μ	8.246	4.126	256.046	30.834
σ	0.034	0.229	9.630	2.960

Figure 7.3: SAILR visualization of background data

In Figure 7.3, part of the control data is shown. We see that in this part $i = 44$ and $n_{44} = 4$, because there are 4 measurements on this source.

7.2. Hierarchical random effect model

We will now apply three feature based models. First, a relatively simple example where only discrete data is available will be considered. Secondly we apply the constructed model from Chapter 6 and finally we will adapt the two-level normal-normal model to a two-level normal-KDE model, where we drop the assumption of a normal distribution for the between source variation and fit a non-parametric kernel density estimator.

7.3. Frequencies model for discrete data

In the event when only discrete data is available, a frequency model can be used for constructing a likelihood ratio. Likelihood ratios only exist when both the control variable **X** and recovered variable **Y** are single measurements, so a single XTC tablet cannot be measured to be both red and blue for example. Of course both **X** and **Y** can consist of multiple sources and therefore we again obtain $m_x \times m_y$. For the denominator, $f(y|x, H_p) = 1$ because under H_p it is assumed that x and y have exactly the same measurements because there is no within source variation. For the nominator, assuming that x and y do not originate from the same source, we take the relative frequency of observing equivalence given all measurements in the background data. Using background sample **Z** and equation (6.22) we find:

$$LR = \frac{f(y|x, I, H_p)}{f(y|I, H_d)} = \frac{1}{f(y|I, H_d)} = \frac{1}{\frac{P_Z}{n_Z P_X}} = \frac{n_Z P_X}{P_Z},$$

$$P_Z = \sum_{i=1}^{m} \sum_{j=1}^{k_i} \mathbb{1}_{\{z_{i,j}=y\}} \quad \text{number of measurements background } \mathbf{Z} \text{ equal to } \mathbf{Y},$$

$$P_X = 1_{\{X=Y\}} = \begin{cases} 1 \text{ if } X \text{ and } Y \text{ are equal,} \\ 0 \text{ otherwise.} \end{cases}$$

Note that his method will never turn into a negative likelihood. For an example of this method we are not able to use the CHAMP data because this consists of continuous features only, therefore we use NFI datasets `X_NFI7.txt`, `Y_NFI7.txt` and `Z_NFI7.txt` Selecting the frequencies method returns 4 likelihood ratios.

Control item	Recovered item	LR comparison result
1	1	93.0
2	1	0
1	2	0
2	2	15.5

Using ENFSI scale, we find that there is moderate strong support for control sources 1 and 2 to match recovered source 1 and 2 respectively, whereas for control source 1 and 2 there is extremely strong support not to have come from the same source.
Making use of continuous data provides more difficulty in the model.

7.4. Two-level normal-normal model

In the case of XTC tablets comparison, we first need to specify what we will take to be the type, items and characteristics. We will use the data described in Section 7.1.1.

- Type; different batches
- Item; different tablets
- Characteristics; different measurements

Furthermore we will take $p = 4$ different measurements. We can specify the three types of data into;

- *Control data* **X**: For one control batch X, n_x measurements are taken. In this illustrative example, 42 different measurement were made on Batch 1 of the CHAMP data, therefore, with $n_x = 42$ we have

$$\mathbf{x_j} = (x_{j1}, \ldots, x_{jp})^T; \quad j = 1, \ldots, 42; \quad p = 1, \ldots, 4.$$

- *Recovered (or questioned) data* **Y**: For the recovered data Y, ten different samples are constructed from different batches as constructed above, of course in practice you do not know where the samples originate from.

$$\mathbf{y_i} = (y_{i1}, \ldots, y_{ip})^T; \quad i = 1, \ldots, 10; \quad p = 1, \ldots, 4.$$

- *Background (or reference) data* **Z**: n_i different measurements are taken on m different batches on the background data, or general population. $m = 160$ different batches are used with n_i measurements per batch, running from 2 to 6 different measurements in the batch. We denote the total number of measurements by N.

$$\mathbf{z_{ij}} = (z_{ij1}, \ldots, z_{ijp})^T; \quad i = 1, \ldots, 160; \quad k_i = 1, \ldots, n_i \quad p = 1, \ldots, 4.$$

Once again we denote the average of the background data within Batch i by

$$\bar{\mathbf{z}}_{\mathbf{i}\bullet} = \sum_{j=1}^{n_i} \frac{1}{n_i} \mathbf{z_{ij}} \tag{7.1}$$

and we take for the overall mean between all batch

$$\bar{\mathbf{z}}_{\bullet\bullet} = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \mathbf{z_{ij}}. \tag{7.2}$$

Using the constructed method in Chapter 6 now returns ten different likelihood ratios; one for every source in the recovered data.

Table 7.1: Likelihood ratios of ten recovered items from Batch 9 of the CHAMP project, according to a two-level normal-normal model

comparison result	LR	Verbal LR
Recovered: nr=1	529.758	**Moderately strong** support for Hp rather than for Hd
Recovered: nr=2	470.881	**Moderately strong** support for Hp rather than for Hd
Recovered: nr=3	1.066e+03	**Strong** support for Hp rather than for Hd
Recovered: nr=4	673.769	**Moderately strong** support for Hp rather than for Hd
Recovered: nr=5	538.985	**Moderately strong** support for Hp rather than for Hd
Recovered: nr=6	0.000e+00	**Extremely strong** support for Hd rather than for Hp
Recovered: nr=7	1.482e-66	**Extremely strong** support for Hd rather than for Hp
Recovered: nr=8	4.037e-57	**Extremely strong** support for Hd rather than for Hp
Recovered: nr=9	6.439e-18	**Extremely strong** support for Hd rather than for Hp
Recovered: nr=10	0.048	**Moderately strong** support for Hd rather than for Hp

The first five likelihood ratios are high. This is exactly what we expected because the recovered measurements are chosen to be from the control measurements. The following three likelihood ratios are very low, which is also in line with what we expected. Finally we observe that the last two likelihood ratios are very low, nothing is yet to conclude from this data. The two-level normal-normal model seems to perform quite well.

7.5. Two-level normal-KDE

Assuming a normal distribution for the within source variation (within the batches), is proven to be a valid assumption. A normal distribution for the between source variation (between the batch means) in practice often is not a valid assumption. The introduction of "super tablets", which are tablets with an extraordinary high level of MDMA disrupts the normality assumption between different sources. To make measurements more accurate, non-parametric statistics can be used. We will try to fit a kernel density estimation for the means.

7.5.1. Two-level normal-KDE univariate

The probability distribution is estimated by making use of the background means $Z_1,\dots,Z_m$. a Gaussian kernel will be fit here for the m different sources in Z. The principle of non-parametric fitting is based on smoothing a histogram that would be obtained when the data would be shown with an empirical distribution which assigns mass size $\frac{1}{m}$ to each Z_i. The mass is now smoothed using a kernel function K and bandwidth h. By definition the kernel function is stated as;

$$f(\theta) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{h}K\Big(\frac{\theta - Z_i}{h}\Big). \tag{7.3}$$

The choice of a normal (Gaussian) kernel means;

$$K(w) = \frac{1}{\sqrt{2\pi}}exp\Big\{-\frac{1}{2}w^2\Big\}. \tag{7.4}$$

The bandwidth parameter h scales the kernel K, it can therefore be seen as a smoothing parameter. When we choose a large bandwidth, the mass will be spread around the data point more extensively. A large bandwidth gives a over smoothed estimate whereas a small bandwidth gives an under smoothed estimate. Estimating the optimal bandwidth is done by minimizing the mean integrated squared error (MISE) elaborated by Silverman in 1986 [29]. The optimal bandwidth h is then given by

$$h_{opt} = \Big\{\frac{4}{(p+2)m}\Big\}^{\frac{1}{p+4}}. \tag{7.5}$$

We will refer to h_{opt} as h from now on. Al together this gives final KDE;

$$\hat{f}(\theta) = \frac{1}{mh}\sum_{i=1}^{m}\frac{1}{\sqrt{(2\pi)}}exp\Big\{-\frac{1}{2h^2}(\theta - \bar{z}_i)^2\Big\}. \tag{7.6}$$

We will not go into detail how the likelihood is obtained. Note that the idea is based on fact that in the numerator, under H_d, independence between X and Y is assumed and therefore can be split into two parts.

$$LR = m\frac{(\sigma_y^2/n_y + h^2)^{1/2}}{(\sigma_y^2/n_y + \tau_h^2)^{1/2}} * \frac{\sum_{i=1}^m \exp(-a_i/2)\exp(-1/2[(\bar{y}-\mu_{hi})^2/(\sigma_y^2/n_y+\tau_h^2)])}{\sum_{i=1}^m \exp(-a_i/2)\sum_{i=1}^m \exp(-b_i/2)} \tag{7.7}$$

$$\text{with} \quad \mu_{hi} = \frac{\frac{\bar{z}_i}{h^2} + \frac{n_x\bar{x}}{\sigma_x^2}}{\frac{1}{h^2} + \frac{n_x}{\sigma_x^2}}, \quad \tau_h^2 = \frac{h^2\sigma_x^2}{\sigma_x^2 + n_x h^2}$$

$$a_i = \frac{(\bar{x}-\bar{z}_i)^2}{\sigma_x^2/n_x + h^2} \quad \text{and} \quad b_i = \frac{(\bar{y}-\bar{z}_i)^2}{\sigma_y^2/n_y + h^2}$$

7.5.2. Two-level normal-KDE multivariate

When we consider the multivariate case with $p = 4$ different features, the likelihood ratios becomes

$$LR = \frac{f(\mathbf{y}|\mathbf{x}, H_p)}{f(\mathbf{y}|H_d)}. \tag{7.8}$$

The likelihood ratio uses a Gaussian kernel. The same optimal bandwidth h is applied now. According to Bolck this gives [16]:

$$m\frac{|\mathbf{U}_{hn}|^{-\frac{1}{2}}\sum_{i=1}^m\left(\exp\left\{-\frac{1}{2}\left(\bar{\mathbf{x}}-\bar{\mathbf{z}}_i\right)^t(\mathbf{U}_{hx})^{-1}\left(\bar{\mathbf{x}}-\bar{\mathbf{z}}_i\right)\right\}\exp\left\{-\frac{1}{2}\left(\bar{\mathbf{y}}-\mu_{hi}\right)^t(\mathbf{U}_{hn})^{-1}\left(\bar{\mathbf{y}}-\mu_{hi}\right)\right\}\right)}{|\mathbf{U}_{h0}|^{-\frac{1}{2}}\left(\sum_{i=1}^m\exp\left\{-\frac{1}{2}\left(\bar{\mathbf{x}}-\bar{\mathbf{z}}_i\right)^t(\mathbf{U}_{hx})^{-1}\left(\bar{\mathbf{x}}-\bar{\mathbf{z}}_i\right)\right\}\right)\left(\sum_{i=1}^m\exp\left\{-\frac{1}{2}\left(\bar{\mathbf{y}}-\bar{\mathbf{z}}_i\right)^t(\mathbf{U}_{h0})^{-1}\left(\bar{\mathbf{y}}-\bar{\mathbf{z}}_i\right)\right\}\right).} \tag{7.9}$$

$$\begin{aligned} \text{with} \quad \boldsymbol{\mu}_n &= \mathbf{T}_0\left(\mathbf{T}_0 + n_x^{-1}\Sigma_x\right)^{-1}\bar{\mathbf{x}} + n_x^{-1}\Sigma_x\left(\mathbf{T}_0 + n_x^{-1}\Sigma_x\right)^{-1}\boldsymbol{\mu}_0 \\ \mathbf{T}_n &= \mathbf{T}_0 - \mathbf{T}_0\left(\mathbf{T}_0 + n_x^{-1}\Sigma_x\right)^{-1}\mathbf{T}_0 \\ \mathbf{U}_0 &= \mathbf{T}_0 + n_y^{-1}\Sigma_y \text{ and } \mathbf{U}_n = \mathbf{T}_n + n_y^{-1}\Sigma_y \end{aligned} \tag{7.10}$$

Applying the same dataset as in Section 7.4 gives

Table 7.2: Likelihood ratios of ten recovered items from Batch 9 of the CHAMP project, according to a two-level normal-normal model

comparison result	LR	Verbal LR
Recovered: nr=1	546.152	**Moderately strong** support for Hp rather than for Hd
Recovered: nr=2	689.456	**Moderately strong** support for Hp rather than for Hd
Recovered: nr=3	1.050e+03	**Strong** support for Hp rather than for Hd
Recovered: nr=4	617.528	**Moderately strong** support for Hp rather than for Hd
Recovered: nr=5	748.179	**Moderately strong** support for Hp rather than for Hd
Recovered: nr=6	0.000e+00	**Extremely strong** support for Hd rather than for Hp
Recovered: nr=7	1.466e-66	**Extremely strong** support for Hd rather than for Hp
Recovered: nr=8	1.385e-57	**Extremely strong** support for Hd rather than for Hp
Recovered: nr=9	1.146e-17	**Extremely strong** support for Hd rather than for Hp
Recovered: nr=10	0.032	**Moderate** support for Hd rather than for Hp

The results of Table 7.2 are almost identically to the results of Table 7.1. We observe that the first five likelihood ratios are slightly higher whereas the following three likelihood ratios are slightly lower. This indicates that the two-level normal-KDE model performs slightly better than the two-level normal-normal model. Note that the verbal LR are still equivalent.

8 Score based likelihood ratios

Besides the feature based models, a new approach is in development, which is called the score based model. This method overcomes some of the flaws of feature based models such as the large computational effort required for this type of models. The model is however not yet very accurate and there are many score and distribution functions available, which makes it hard to arrive at a uniform model.

8.1. Score function

To calculate a likelihood ratio by using a score based model, first of all a score function needs to be chosen. This score function represents a distance or similarity between the recovered data **Y** and the control data **X**. The score function is denoted by $\delta(\mathbf{x},\mathbf{y})$ and we will take a look at some of the 11 different score functions that SAILR provides. After the score function has been chosen and determined, the density function needs to be chosen to compare the distances. In the likelihood ratio, the scores using the within source distribution are compared with the scores using the between source distribution. Note that both distributions use the same function but adjusted to different settings, which can be both parametric or non-parametric. For the within source distribution, every measurement in the background population Z is compared to every measurement within its source, whereas the between source distribution compares background population Z with measurements in other sources.

For multiple measurements on one item, the average will be taken. For multiple sources, different likelihood ratios will be calculated per source, resulting in m_x x m_y likelihood ratios. The likelihood is then calculated by;

$$LR = \frac{f_{within}(d(\bar{\mathbf{x}},\bar{\mathbf{y}}))}{f_{between}(d(\bar{\mathbf{x}},\bar{\mathbf{y}}))} \tag{8.1}$$

$$\text{with} \quad \bar{\mathbf{y}} = \sum_{j=1}^{n_y} \mathbf{y_j} \tag{8.2}$$

$$\text{and} \quad \bar{\mathbf{x}} = \sum_{j=1}^{n_x} \mathbf{x_j}. \tag{8.3}$$

8.2. Distances and scores

In total SAILR provides 11 different distance functions, we will take a look at three of them. The number of features is given by $p = 4$.

- **Euclidean:** The simplest distance measure is the Euclidean measure, this measure is defined by

$$\delta(X,Y) = \sqrt{\sum_{i=1}^{p} (X_i - Y_i)^2}. \tag{8.4}$$

Because of its simplicity, the score performs problematic as soon as the data introduces difficulties. For example, if the multivariate data is not equally scaled and different units are used, the distance is no longer trustworthy.

- **Bray-Curtis Distance:** This score is often used with count data, it might perform bad for non-count data which can contain negative values. However for the CHAMP data, there are no negative values so this score is worth trying. The score is given by

$$\delta(X,Y) = \frac{\sum_{i=1}^{p} |x_i - y_i|}{\sum_{i=1}^{p} |x_i + y_i|}. \tag{8.5}$$

- **Canberra distance:** Finally we will take a look at the Canberra distance, the distance is proven to perform well in certain forensic continuous comparison problems and therefore applied to the CHAMP data [34]. The distance is defined by

$$\delta(X,Y) = \sum_{i=1}^{p} \frac{|x_i - y_i|}{|x_i| + |y_i|}. \tag{8.6}$$

For the distribution function, three options are available, first the already encountered kernel density estimate and furthermore two parametric functions: Gamma and Weibull.

8.3. Comparison

For every score function with measurements $\mathbf{X}, \mathbf{Y}$ and $\mathbf{Z}$, we consider the two hypothesis H_p and H_d. For hypothesis H_p, the evidence e is the score between X and Y assuming the measurements originate from the same source. Because the measurements originate from the same source, we can assume the evidence e originates from a within source variation score e_w. For hypothesis H_d, the score between X and Y must come from the between source variation because X and Y are assumed to originate from different sources.

The following procedure will be executed to arrive at a final score:
First, all within scores e_w under hypothesis H_p are compared. Compared in this case means that we calculate the distance functions that were chosen. For every source $i = 1,\dots,m$, n_i different measurements are available, $k_i = 1,\dots,n_i$. For every source i, we first compare measurement $k_i = 1$ to measurements $k_i = 2,\dots,n_i$. Next we compare measurement $j = 2$ to measurements $j = 3,\dots,n$ until finally we only compare measurement $j = n_{i-1}$ to $j = n_i$. This results in $(n_i - 1) + (n_i - 2) + \cdots + 2 + 1 = \frac{1}{2} n_i (n_i - 1)$ different measurements per source i and because this is constructed for every source m we finally have $\frac{1}{2}\sum_{i=1}^{m} n_i(n_i-1)$ different within source measurements under H_p.

Under H_d, the scores of X and Y are compared assuming between source variation. It is assumed the samples originate from different sources and therefore there is no within source variation between them. For every source i, the first measurement $k_i = 1$ is compared to all measurements from other sources. For measurement one in source one this results in $\sum_{j=2}^{m} n_j$ measurements. We will do this for every measurement in source one, which are in total $n_i \sum_{j=2}^{m} n_j$ measurements. For another source i with n_i measurements, this means we will conduct $n_i \sum_{j=1, j\neq i}^{m} n_j$ measurements for source i. In total this will lead to $\frac{1}{2}\sum_{i=1}^{m} n_i \sum_{j=1, j\neq i}^{m} n_j$ different scores under the between source variation, taking account for double counts.

Now that we have found n_w different e_w scores and n_b different e_b scores, we are able to compute n_w likelihood ratios under H_p and n_b different likelihood ratios under H_d. To get from all these scores to a likelihood ratio, we calculate the numerator $\mathbb{P}(e_w|H_p)$, which is the probability of finding a within score e_w in an e_w distribution. For the denominator $\mathbb{P}(e_w|H_d)$, we calculate the probability of finding the same e_w score in an e_b distribution. The probability is calculated using the distribution function, which in our case will be chosen to be the kernel density estimate. This returns likelihood ratio

$$LR = \frac{\mathbb{P}(e_w|H_p)}{\mathbb{P}(e_w|H_d)}. \tag{8.7}$$

8.4. CHAMP data comparison

Using the three different distribution functions and three different score functions in Section 8.2, we can calculate nine arrays of ten different likelihood ratios. Just as in Chapter 7, using SAILR we find;

Table 8.1: Likelihood ratios for different score and distribution functions

		Euclidean	Bray-Curtis	Canberra
KDE	1	12.208	13.847	20.165
	2	2.688	5.495	32.280
	3	5.060	12.016	30.175
	4	2.601	5.008	25.076
	5	6.165	15.624	32.547
	6	0.002	0.020	7.312e-25
	7	0.010	0.014	0.012
	8	3.106	4.534	0.067
	9	0.706	0.500	5.210e-15
	10	0.674	0.394	0.150
Gamma	1	8.882	12.873	27.251
	2	2.923	5.350	92.798
	3	5.863	10.840	73.705
	4	2.820	4.975	43.845
	5	7.328	15.277	95.631
	6	5.118e-4	6.937e-4	9.926e-6
	7	0.005	0.002	0.003
	8	3.421	4.602	0.048
	9	0.687	0.619	7.012e-5
	10	0.653	0.392	0.100
Weibull	1	7.287	10.572	14.210
	2	2.688	4.952	32.468
	3	5.060	9.147	27.945
	4	2.601	4.642	19.747
	5	6.165	12.187	33.105
	6	0.002	0.002	6.732e-5
	7	0.010	0.004	0.007
	8	3.106	4.331	0.081
	9	0.706	0.687	3.099e-4
	10	0.674	0.449	0.153

Alarming in the data is that all three distributions using Euclidean and Bray-Curtis distance give a likelihood ratio > 1 for Batch 8 (which originates from Batch 12). We know likelihood ratio eight compares samples from different batches (Section 7.1.1) and therefore the likelihood ratio should be below one. Note furthermore that the Canberra distance performs significantly better compared to the other measures; likelihood ratios one until five are higher and ratios six until eight are lower for all distribution functions compared to the Euclidean and Bray-Curtis distance.

When we compare the score based functions to the feature based functions, we observe that the feature based functions perform significantly better. Feature based models compare the probability of observing the evidence given that the control and recovered samples come from the same source or come from different sources. In contrast, score based models compare the probability of observing pairwise similarity between the control and recovered samples given that they originate from the same source with the probability of pairwise similarity given that the samples come from different sources. [1]

A big advantage for score based models is that they reduce multivariate information to a univariate distance which in the case of a lot of different features can be a major improvement in computational time. Furthermore, covariance estimation between sources is possible with only few samples available. Shortcomings of the method are that the values of the likelihood ratio are based on pairwise scores rather than the similarity and rarity of the features as is the case for feature based models. Given that the method has many advantages definitely makes it worth doing more research on, for example on different score functions.

9 Conclusion

To qualify evidence, we have seen that hypotheses can be tested using classical statistical inference methods such as P-values, confidence intervals and both univariate and multivariate t-tests. Since these tests do not consider background information or qualify the value of evidence, the result of these test is not suited to be presented in court. The inference methods can however still be used for pre-laboratory research meaning that when we find very strong evidence against the prosecution hypothesis, it might not be worthwhile to complete further statistical analysis which can save both money and time. A generally accepted method for qualifying the value of evidence can be found in the form of likelihood ratios. We have seen, using test data from a known case of illicit drugs, that both the two-level normal-normal and two-level normal-KDE model perform very well in quantifying the evidence, where the two-level normal-KDE model performs a little better. As a final comparison measure, score-based methods have been evaluated. We have compared multiple score functions with all available distribution functions to conclude that score-based likelihood ratios do not yet meet the standards for forensic statistics.

Recommendations and future development

- Although the tests described in Chapter 3 might be too simplistic to present in court, possibilities to broaden the test to a non-parametric estimation function can still be discovered. This could possibly lead to new methods for forensic evidence comparison.
- In Chapter 6, both the RI and comparison on multiple chemical features are considered. Because the RI method is both technically and financially more attractive, this method is preferred over the multiple feature comparison. Using test data we could compare both methods and determine if the RI provides representative likelihood ratios.
- In this research only the comparison method in SAILR was elaborated. SAILR does however provide a possibility for discrimination between two background samples and a validation method for one background sample. These two operations can be evaluated with the same data as used in this report.
- The score-based comparison models can reduce computational time enormously. Using the current distance functions available, the score-base models do not yet meet the standards for forensic statistics. In future development, score functions can be further explored and implemented in SAILR packages.
- The methods used for validating the comparison methods were only based on looking for a likelihood ratio as high as possible for equivalent evidence sources and as low as possible for different evidence sources. There are however significantly better methods to compare the forensic evidence. Two of these methods are Tippett plots and the Empirical cross-entropy. Both methods are relatively new, but their results yet are positive and definitely worth doing more research on.

Bibliography

[1] Aitken, C. G. G. . Bayesian Hierarchical Random Effects Models in Forensic Science. In *Front. Genet.*, 2018.

[2] Aitken, C.G.G and Lucy, D. Evaluation of Trace Evidence in the Form of Multivariate Data. *Journal of the Royal Statistical Society Series C*, 53:109–122, 01 2004. doi: 10.1046/j.0035-9254.2003.05271.x.

[3] Birkett, A. Bayesian vs Frequentist A/B Testing – What's the Difference?, 2019. URL `https://conversionxl.com/blog/bayesian-frequentist-ab-testing/`.

[4] Bolck, A. and Alberink, I. Variation in likelihood ratios for forensic evidence evaluation of XTC tablets comparison. *Journal of Chemometrics*, 25:41 – 49, 01 2011. doi: 10.1002/cem.1361.

[5] Bolck, A. and Ni, H. and Lopatka, M. Evaluating score-and feature-based likelihood ratio models for multivariate continuous data: Applied to forensic MDMA comparison. *Law, Probability and Risk*, 14, 09 2015. doi: 10.1093/lpr/mgv009.

[6] Bolck, A. and Weyermann, C. and Dujourdy, L. and Esseiva, P. and Berg, J. Different likelihood ratio approaches to evaluate the strength of evidence of MDMA tablet comparisons. *Forensic science international*, 191:42–51, 08 2009. doi: 10.1016/j.forsciint.2009.06.006.

[7] Campbell, G and Curran, J.M. The Interpretation of Elemental Composition Measurements from Forensic Glass Evidence II. *Science & justice : Journal of the Forensic Science Society*, 37:245–249, 01 2009. doi: 10.1016/S1355-0306(97)72197-X.

[8] Campbell, G and Curran, J.M. The Interpretation of Elemental Composition Measurements from Forensic Glass Evidence III. *Science & justice : Journal of the Forensic Science Society*, 49:2–7, 04 2009. doi: 10.1016/j.scijus.2008.09.001.

[9] Chan K.P.S. and Aitken C.G.G. Estimation of the Bayes' factor in a forensic science problem. *Journal of Statistical Computation and Simulation*, 33(4):249–264, 1989. doi: 10.1080/00949658908811201. URL `https://doi.org/10.1080/00949658908811201`.

[10] Curran, J.M. The statistical interpretation of forensic glass evidence. *International Statistical Review*, 71:497–520, 12 2003.

[11] Curran, J.M. *Introduction to Data Analysis with R for Forensic Scientists*. CRC press, Boca Raton, 2011.

[12] Curran, J.M. The Frequentist Approach to Forensic Evidence Interpretation. *Encyclopedia of Forensic Sciences*, pages 286–291, 12 2013. doi: 10.1016/B978-0-12-382165-2.00194-X.

[13] Curran, J.M. and Triggs, C and Almirall, J and Buckleton, S. J and Walsh, K. The Interpretation of Elemental Composition Measurements from Forensic Glass Evidence II. *Science & Justice - SCI JUSTICE*, 37:245–249, 10 1997. doi: 10.1016/S1355-0306(97)72198-1.

[14] Dahiru, Tukur. P-value, A true test of statistical significance? A cautionary note. *Annals of Ibadan postgraduate medicine*, 6:21–26, 06 2008. doi: 10.4314/aipm.v6i1.64038.

[15] Dorp, I.N. van. Statistical modelling of forensic evidence. Masters thesis, Delft University of Technology, 2018. URL `https://repository.tudelft.nl/islandora/object/uuid%3A26b62fb7-97ed-438f-88e1-f7995ab4c73c?collection=education`.

[16] European Monitoring Centrum for Drugs and Drug Addiction. Methylenedioxymethamphetamine (MDMA or 'Ecstasy') drug profile, 2019. URL `http://www.emcdda.europa.eu/publications/drug-profiles/mdma`.

[17] Evett, I. Towards a uniform framework for reporting opinions in forensic science casework. *Science & Justice - SCI JUSTICE*, 38:198–202, 07 1998. doi: 10.1016/S1355-0306(98)72105-7.

[18] Galbraith, C. and Smyth, P. Analyzing user-event data using score-based likelihood ratios with marked point processes. *Digital Investigation*, 22:S106–S114, 08 2017. doi: 10.1016/j.diin.2017.06.009.

[19] Gelman, A. and Carlin, John B. and Stern, H.S. and Rubin, D.B. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, 2nd ed. edition, 2004.

[20] Goodman, S. Introduction to Bayesian methods I: Measuring the strength of evidence. *Clinical trials (London, England)*, 2:282–90; discussion 301, 02 2005. doi: 10.1191/1740774505cn098oa.

[21] Hoekstra, R. and Morey, R.D. and Rouder, J.N. and Wagenmakers, E. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5):1157–1164, Oct 2014. ISSN 1531-5320. doi: 10.3758/s13423-013-0572-3.

[22] Kool, F.S. Feature-based models for forensic likelihood ratio calculation. Masters thesis, Delft University of Technology, 2016. URL `https://repository.tudelft.nl/islandora/object/uuid%3A5c088097-b0f0-4342-9737-202c81e7212d?collection=education`.

[23] Kool, F.S and Steenhuis, R and Neijmeijer, R. and Bolck, A. User Manual SAILR - Version 1.3.0, 03 2017.

[24] Laake, P. and Fagerland, M.W. *Research in Medical and Biological Sciences*. Academic Press, Amsterdam, Second Edition edition, 2015. ISBN 978-0-12-799943-2. doi: https://doi.org/10.1016/B978-0-12-799943-2.00011-2.

[25] Man, T de. NRC checkt: 'Wereldwijd is Nederland de grootste producent van xtc, 2017. URL `https://www.nrc.nl/nieuws/2018/06/04/nrc-checkt-wereldwijd-is-nederland-de-grootste-producent-van-xtc-a1605281`.

[26] Neijmeijer, R. Assessing Performance of Score-Based Likelihood Ratio Methods for Forensic Data. Masters thesis, Universiteit Leiden, 2016.

[27] Nordgaard, A. and Rasmusson, B. The likelihood ratio as value of evidence–more than a question of numbers. *Law Probability and Risk*, 11:303–315, 12 2012. doi: 10.1093/lpr/mgs016.

[28] Ommen, D.M. and Saunders, C.P. and Neumann, C. A Note on the Specific Source Identification Problem in Forensic Science in the Presence of Uncertainty about the Background Population. *arXiv e-prints*, art. arXiv:1503.08234, 03 2015.

[29] Silverman, B. W. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.

[30] SQA Forensic Science. Report On: Glass Analysis Information: Refractive Index Measurement - Laser, 2019. URL `https://www.sqaacademy.org.uk/pluginfile.php/34022/mod_resource/content/2/Glass/laser.html`.

[31] Taroni, F. and Biedermann, A. and Bozza, S. Statistical hypothesis testing and common misinterpretations: Should we abandon p-value in forensic science applications? *Forensic Science International*, 259:e32–e36, 02 2016. doi: 10.1016/j.forsciint.2015.11.013.

[32] Thompson, W. and Vuille, J. and Biedermann, A. and Taroni, F. The Role of Prior Probability in Forensic Assessments. *Frontiers in genetics*, 4:220, 10 2013. doi: 10.3389/fgene.2013.00220.

[33] R. Travis. Look how often field drug tests send innocent Georgians to jail, 2017. URL `http://www.fox5atlanta.com/news/i-team/look-how-many-times-field-drug-tests-send-innocent-georgians-to-jail`.

[34] Uitdehaag, S. and Wiarda, W. and Donders, T. and Kuiper, I. Forensic Comparison of Soil Samples Using Nondestructive Elemental Analysis. *Journal of Forensic Sciences*, 62(4): 861–868, 2017. doi: 10.1111/1556-4029.13313. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.13313`.

[35] Wasserstein, R.L. and Lazar, N.A. The ASA's statement on p-values: context, process, and purpose [Editorial]. *Amer. Statist.*, 70(2):129–133, 2016. ISSN 0003-1305. doi: 10.1080/00031305.2016.1154108. URL `https://doi.org/10.1080/00031305.2016.1154108`.