Métodos Estatísticos em Ciência Forense

Eric Pavarim Lima
Alex Rodrigo dos Santos Souza
IMECC
vigência de setembro de 2023 a agosto de 2024

22 de Agosto de 2024

1 Introdução e enunciado do problema

O presente projeto tem como objetivo o estudo e revisão de métodos estatísticos em problemas da área forense. Para isso fora de extrema importância a criação de uma base sólida e gradual em Inferência Bayesiana, visto que, o conceito de evidência no cunho forense faz uso primordial de tal teoria, citada mesmo nos primeiros artigos da história da chamada estatística forense[3].

Além disso, há de mencionar outras atividades contempladas como uma análise breve em um conjunto de dados sobre fragmentos de vidros utilizando o software RStudio[7], a análise descritiva foi essencial para que essa área de fragmentos de vidro, tão datada em artigos seja proximamente explorada com êxito; uma simulação para devida escolha de hipóteses por razões de verossimilhança em contextos diferentes, evidenciando a importância da percepção do problema forense encontrado, assim como o correto teste de hipóteses; um estudo adicional e breve do problema de espectroscopia na área forense, em especial na forma de interpretação, separação e estimação das curvas de absorbância em substâncias.

2 Resumo das atividades

Com ênfase no estudo em Inferência Bayesiana, revisão de teorias probabilísticas e estudo de problemas forenses tal qual suas intersecções com a estatística, foi utilizado como principal referência Aitken e Taroni (1995)[1], com isso, o enfoque primordial fora a leitura e aprendizado de tais temas. Os estudos originaram-se no problema de fragmentos de vidro, isto é, identificação de alguma origem mútua nos fragmentos de vidro distintos (ou seja, verificar se ambos têm ou não a mesma origem). Complementar a isso, uma análise descritiva de um banco de dados forense fora realizada a fim da percepção de fatores determinantes entre amostras distintas para que, posteriormente, haja a aplicação de métodos forense-estátisticos devidos.
Seguindo os estudos, a revisão dos perfis de DNA e impressões digitais foram realizadas, dada a importância do tema para o mundo forense, visto que, ambas evidências são extremamente comuns. Em conjunto com o aprendizado, a aplicação da razão de verossimilhança ocorreu no contexto de simulações de razões de verossimilhança gaussianas multivariadas (no panorama forense, em impressões digitais).

Ao fim dos estudos, a relevância da estatística e seus métodos na ciência forense fora valorizada, assim, o projeto seguiu , de forma cronológica, uma familiaridade com diversos problemas reais e adicionalmente, uso computacional para melhor entendimento dos mesmos.

3 Metodologia

Para que a leitura de materiais com enfoque forense seja realizada, criou-se a necessidade de revisar e aprender conceitos de probabilidade e estatística. A inferência Bayesiana no contexto forense tem a necessidade provinda da constante atualização de evidências assim como a consideração de poucos dados, isto é, estritamente oposto a inferência frequentista que necessita de grande volume de dados para aplicação dos mais diversos teoremas.

3.1 Análise de fragmentos

- Definição 1: Se um evento R tem probabilidade $Pr(R)$ de ocorrer, então:
  - *Odds/betting quocients* é a medida de incerteza que mede quão maior é a probabilidade de um evento ocorrer em relação ao seu complementar.

    $$\frac{P(R)}{1-P(R)}$$

    Este conceito tem extrema importância no quantificação de evidências, que têm efeito direto na probabilidade de alguma certa suposição sobre um suspeito (antes do ínicio do julgamento) ou do acusado (enquanton o julgamento está em processo). Um dos casos mais intuitivos é o dos eventos H_p (hipótese de culpa) ou H_d (hipótese de inocência), assim, dado certas informações prévias I, temos:
    $\frac{Pr(H_p|I)}{Pr(H_d|I)}$, isto é, as chances a favor da culpa, dado informações prévias, contra a inocência.
    Como abordado no capítulo 3.3 da referência Aitken e Taroni (1995)[1] , a maior importância dessa definição dá-se pela fácil interpretação visto que *odds* representam quão maior são as chances de um evento occorer em comparação com seu complementar.
    Um dos maiores problemas da junção entre a área legal e a estatística, por exemplo no famoso caso de Sally Clark[2] em que, fora interpretado de forma errônea conceitos de probabilidade como independência de eventos e probabibilidade condicional, ocasionando uma prisão injusta de uma mulher inocente. Este caso ilustra bem o cuidado que deve ocorrer na junção da área jurídica com a matemática, principalmente estatística, portanto, conceitos que têm interpretação clara mesmo para profissionais de outras áreas ganham valor nesse contexto.

- Teorema de Bayes: Sejam S e R dois eventos e $Pr(R) \neq 0$ a probabilidade do evento R ocorrer, então:

  $$P(S|R) = \frac{P(R|S)\times P(S)}{P(R)}]$$

  Pode-se reescrever esse teorema no formato de *odds*. Denote por $\overline{S}$ o eventos complementar de S. Então:

  $$\frac{Pr(S|R)}{Pr(\overline{S}|R)} = \frac{Pr(R|S)}{Pr(R|\overline{S})}.\frac{Pr(S)}{Pr(\overline{S})}$$

O lado esquerdo denota as *odds* a favor de S dado R $\frac{Pr(S|R)}{Pr(\overline{S}|R)}$, chamadas de *posterior odds*, enquanto o lado direito possui dois termos, as odds a favor de S$\frac{Pr(S)}{Pr(\overline{S})}$ ou *prior odds* e o termo $\frac{Pr(R|S)}{Pr(R|\overline{S})}$ denotado por fator de Bayes ou *likelihood ratio*.
O fator de Bayes é sempre maior ou igual a zero (não possuindo limite superior) e é interpretado as *odds* de um evento (R nesse caso) dado a suposição ou não de dois eventos exclusivos entre si (S e $\overline{S}$).
Exemplo:[6]
Sejam três eventos:
- X_A referente ao fenótipo Kell, sendo a ocorrência A indicador do indivíduo Kell+ e $\overline{A}$,Kell-
- X_B referente ao fenótipo Duffy, sendo a ocorrência B indicador do indivíduo Duffy+ e $\overline{B}$,Duffy-
- X_C indicando a cor, sendo C a pessoa rosa e $\overline{C}$, azul.

Após alguns cálculos chega-se no valor 4 para o fator de Bayes que relaciona esse evento A com as *odds* a favor do individuo ser rosa. Em termos práticos, considera-se esse exemplo um problema de identificação, e imagine que as chances de um individuo ser rosa ou azul sejam equivalentes e não haja nenhuma informação prévia sobre o antígeno Kell, assim as odds a favor de qualquer uma das cores são neutras. Porém, a descoberta do individuo ser Kell+ (evento A) ocasiona em um peso para as odds a favor da cor rosa, tal evidência auxilia na provável identificação (como uma atualização das odds iniciais). O individuo tem 4 vezes mais chances de ser rosa se o mesmo for Kell+ dado esse fator de Bayes.

- Definição 2: Sejam S e R dois eventos e $Pr(R) \neq 0$ a probabilidade do evento R ocorrer, então:

$$P(S|R) = \frac{P(R|S) \times P(S)}{P(R)}]$$

- Definição 3:
  - Sejam S e R dois eventos com probabilidades $Pr(S)$ e $Pr(R)$ de ocorrer. A probabilidade condicional (um evento ocorrer dado a ocorrência de outro) é definida por:

$$Pr(S|R) = \frac{Pr(S \cap R)}{Pr(R)}$$

  A inferência bayesiana tem papel fundamental na aplicação forense, principalmente, por dois fatores principais: O primeiro dá-se pela necessidade de considerar pequenas e quase nulas evidências, diferentemente da inferência frequentista que exige grande número de dados, o que nem sempre é possível resgatar e obter em situações forenses. Outro fator importantíssimo para tal abordagem é a constante necessidade de atualização a medida que novos dados, evidências são considerados ou descobertos, o que faz uso quase que primordial dessa atualização de probabilidades proposta nesse segmento da estatística.

- Definição 4:
  - Sejam $X_1, X_2, \dots X_n$ variáveis aleatórias com distribuição $f(x|\theta)$, tal que $\theta \in \Theta$ é um parâmetro desconhecido. Dizemos que θ segue uma distribuição a priori $\pi(\theta)$ definida.

- Definição 5:
  - Sejam $X_1, X_2, \dots X_n$ variáveis aleatórias com distribuição $f(x|\theta)$, tal que $\theta \in \Theta$ é um parâmetro desconhecido. Assuma que θ tem distribuição a priori $\pi(\theta)$. Então a distribuição a posteriori de θ é dada por:
    $\pi(\theta|x) = \frac{\pi(\theta)f(x|\theta)}{\int_\Theta \pi(\theta)f(x|\theta)d\theta} = \frac{\pi(\theta)f(x|\theta)}{m(x)} \propto \pi(\theta)f(x|\theta)$
    Em termos práticos, basta utilizar-se da proporção visto que $m(x)$ é uma constante, obtida dos dados observados.

- Definição 6:
  - Uma distribuição a priori é conjugada de outra se a posteriori entre as duas é uma distribuição da mesma família da distribuição a priori.

    Exemplo :
    Seja $X_1, \dots X_n$ variáveis aleatórias com função de probabilidade $f(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, ou seja, $X_i \sim Binomial(n, \theta)$.
    Além disso, $\theta \in \Theta$ um parâmetro desconhecido com distribuição a priori $\pi(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$, i.e, $\theta \sim Beta(\alpha, \beta)$. A distribuição a posteriori de θ dá-se por $f(\theta|x) = \frac{1}{B(\alpha,n+\beta)}\theta^{\alpha-1}(1-\theta)^{n+\beta-1}$, portanto, $\theta|X \sim Beta(\alpha, n+\beta)$
    Uma das aplicações forenses[1] é o caso do parâmetro θ de interesse, simbolizando a frequência do alelo M no sistema de grupo sanguíneo MN. A distribuição a priori é Beta e os dados indicando a presença ou não do alelo têm como função de verossimilhança a Binomial. Dessa forma, a posteriori é previsível sendo uma distribuição Beta também.

- Definição 7:
  - Sejam O_1 e O_2 duas *odds* distintas, logo:

    $$\theta = \frac{O_1}{O_2} \geq 0$$

    θ é definido como *odds ratio* e tem papel fundamental na interpretação forense e mensura quantas vezes as odds O_1 é maior em relação as odds O_2.

    Ilustra-se no capítulo 5.7[1] um caso de dois grupos de cédulas contaminadas com cocaína, o primeiro grupo de cédulas encontradas em determinadas investigações de tráfico e o segundo grupo de circulação geral. Além disso, p_x e p_y são as probabilidades de uma cédula aleatória estar contaminada no respectivo grupo e seus estimadores $\hat{p_x}$ e $\hat{p_y}$ são calculados. A expressão $\theta = \frac{p_x(1-p_y)}{p_y(1-p_x)}$ é a *odds ratio* e tem significado de as *odds* a favor das cédulas do primeiro grupo estarem contamidas com cocaína são θ maiores do que as *odds* a favor das cédulas do segundo grupo estarem contaminadas com cocaína.

- Definição 8:
  - Sejam H_p e H_d duas hipóteses conflitantes, I as informações prévias e E a evidência em questão.

    $$V = \frac{Pr(E|H_p,I)}{Pr(E|H_d,I)}$$

    V é definido como o valor da evidência e é a *likelihood ratio* que atua na atribuição de um peso sobre as *prior odds* $\frac{Pr(H_p|I)}{Pr(H_d|I)}$ a partir da atualização de evidências E para que, consequentemente, as *posterior odds* sejam atualizadas.
    As evidências na área forense têm uma forma matemática diferente da forma que a incerteza é tratada, isto é, a visão sobre verossimilhança é mais importante. Em outras palavras, a força ou não de uma evidência é quantificada não por probabilidades, mas sim por esse fator de atualização.

3.2 Análise de DNA

No contexto de perfis de DNA, o principal problema encontra-se na identificação de suspeitos a partir de uma fita, evidência de DNA encontrada na cena do crime (vale salientar que o uso de DNA como evidência para juri não é permitido, aconselhado nos Estados Unidos [8]). Entretanto, a junção entre Estatística e o esse contexto têm algumas propriedades interessantes para análise do mesmo.

A estrutura do DNA consiste em duas fitas (longos polímeros de desoxirribose e fosfato) que são conectadas em um espiral duplo (hélice) por ligações de hidrogênio entre bases complementares de nitrogênio. A sequência dessas bases que irá ditar a informação genética específica que a molécula cobre. Uma partição do DNA que codifica para a síntese da célula de uma proteína específica é chamado de *gene*. Apenas quatro bases formam o DNA (Adenina (A), Guanina (G), Citosina (C) e Timina (T)). Além disso, as bases unem-se a suas complementares, de tal forma que, cada fita é inerentemente complementar à outra, Adenina liga-se apenas a Timina e Guanina a Citosina e vice-versa.
Diferentes formas do gene são chamados *alelos*. Se um mesmo alelo está presente em ambos cromossomos de um par, a pessoa é homozigota. Se dois alelos são diferentes, heterozigota. A composição genética de uma pessoa é chamada de genótipo.
Na análise forense, o genótipo de um grupo de lócus gênicos analisados é chamado de perfil do DNA.
Para adereçar a probabilidade do DNA de uma pessoa selecionada aleatoriamente ter o mesmo perfil que o DNA encontrado na cena do crime, por exemplo, deve-se saber a frequência daquele perfil na população (o teste póstumo ou a busca em um banco de dados pré-estabelecido não oferece diferenças para a mensuração da evidência em si [8])
Há uma convenção na área de genética que consiste em adotar cada gene (ou marcador do lócus) como uma letra e cada alelo com um número (índice). Por exemplo, o décimo alélo no lócus A seria denotado A_{10}.
Além disso, a frequência do alelo A_i é denotada p_i tal que $\sum p_i = 1$ sob todos os alelos presentes no lócus.

- Definição 9 [1]:
  - Frequências alélicas podem ser calculadas de observações de uma amostra. Para calcular as frequências esperadas, parte-se de um suposição, o chamado equilíbrio de Handy-Weinberg que diz que:
    Em um lócus de uma população com k alelos, indexados por i e j, e $m = k(k+1)/2$, isto é, a quantidade de genótipos. Uma amostra de tamanho n é retirada e o número de membros da amostra para cada um dos genótipos é contado, com x_{ij} membros do genótipo i, j (denotando alelos componentes i, j para $1 \leq i, j \leq k$), onde o indivíduo é heterozigoto se $i \neq j$ e homozigoto se $i = j$.
    As frequências relativas x_{ij}/n (proporção de cada genótipo da amostra) é denotada P_{ij}. A proporção p_i de alelos do tipo i é dada por:

    $$p_i = P_{ii} + \frac{1}{2} \sum_{j \neq i}^{j=k} P_{ii}$$

    em que, $P_i i$ referente à paridade idêntica e $\frac{1}{2} P_{ij}$ pois em heterozigotos, apenas metade dos alelos são i (e claro $i \neq j$). A suposição de que os alelos em um lócuos, um do pai e outro da mãe, são independentes leva à distribuição de equilíbrio para as frequências relativas de uma população (também conhecida como Equilíbrio de Handy-Weinberg ou Panmixia/Cruzamento aletório).
    O a frequência genotípica de AA é p^2, de Aa é $2pq$ e de aa é q^2. Além disso, $p^2 + 2pq + q^2 = (p+q)^2 = 1$.
    De modo geral, seja p_i e p_j as proporções populacionais dos alelos A_i e A_j, para $i, j = 1, \ldots, k$ onde k é o número de alelos em um lócus em questão. As frequências genotípicas esperadas P_{ij} são obtidas das seguintes equações, assumindo equilíbrio de Handy-Weinberg:

$$P_{ij} = \begin{cases} 2\mathrm{p}_i p_j, & \mathrm{i} \neq j \\ \mathrm{p}_i^2, & \mathrm{i}{=}\mathrm{j} \end{cases}$$

– Para além disso temos que a probabilidade Q, isto é, a probabilidade de que duas pessoas escolhidas aleatoriamente tenham o mesmo genótipo (desconhecido, não específicado) é a soma dos quadrados das frequências de todos os genótipos. Para um caso geral com n alelos:

$$Q = \sum_{i=1}^{n} Pii^2 + \sum_{i=1}^{n} \sum_{j<i} P_{ij}^2$$

a partir dessa definição, pode-se calcular o poder de discriminação, ou porder de exclusão, de um genótipo como sendo $1 - Q$. Quanto maior o poder de discriminação, melhor o genótipo é em distinguir entre pessoas (ou seja, mais 'raro').

Para m sistemas independentes com respectivos valores $Q_1, Q_2, \ldots, Q_m$. A probabilidade DP_m de ser apto a distiguir dois indivíduos usando os m testes é, como antes:

$$DP_m = 1 - \prod_{l=1}^{m} Q_l$$

- Definição 10:

– Em um contexto de DNA, seja E_c a amostra do crime, E_s do suspeito e I a informação prévia, e sejam as suposições, em um nível de fonte:
- H_p : o suspeito é a fonte da evidência;
- H_d : outra pessoa, não relacionada ao suspeito, é a fonte (isto é, o suspeito não é a fonte da evidência).
Ambos os perfis de DNA são de tipo A, por exemplo. A razão de verossimilhanças pode, então, ser expressa como:

$$\frac{P(E_c{=}A|E_s{=}A,H_p,I)}{P(E_c{=}A|E_s{=}A,H_d,I)}$$

um adendo é, assumindo que o sistema de tipagem de DNA é suficiente para identificação do dono do DNA quando há duas amostras distintas, isto é, haverá, sim, correspondência de duas amostras quando as mesmas são do mesmo dono (H_p) e que não haverá falsos negativos. A amostra recuperada é do tipo A se é conhecida que o suspeito é do tipo A, se H_p é assumida como verdadeira, nesse cenário, $P(E_c = A|E_s = A, H_p, I) = 1$.
Além disso, é amplamente assumido que perfis de DNA de duas pessoas distintas (o suspeito e o dono do rastro quando H_p é verdadeira) são independentes. Então $P(E_c = A|E_s = A, H_d, I) = P(E_c = A|I)$. Nesse caso, apenas a probabilidade de perfil ($2p_i p_j$ quando heterozigoto suspeito ou p_i^2 para homozigoto) de que uma pessoa desconhecida tenha o perfil A é necessário (Essa é uma simplificação amplamente aceita [1]).
Embora o uso de perfis de DNA para estatística forense fora estudado, o mesmo não é tão encorajado devido ao *NRC - US National Research Council* [14,15]). Sendo assim, cabe como alternativa o uso de impressões digitais, visto que, ambas são evidências de força e eficácia similar [12].

3.3 Impressões digitais

No contexto de impressões digitais, inicialmente o objetivo era identificação de indivíduos para propósitos administrativos ou legais, sendo assim, todas impressões eram adquiridas sob situações de controle e geralmente de altíssima

qualidade e precisão (as chamadas impressões de controle). O *NGI - U.S Next Generation Identification System* é um sistema que recebe por volta de 5 milhões de requisições mensais [12], sendo mais da metade para uso civil. Entretanto, no contexto forense, o cenário é contrastante , visto que, as impressões em cenas de crime muitas vezes estão desgastadas, parciais, ou distorcidas (são chamadas de *latent prints* ou *fingermarks*) e carregam imprecisão e variância consigo, daí a atuação fundamental da estatística.
Ao decorrer dos anos, diversos autores tiveram tentativas de quantificação de impressões digitais, tanto por funções de Kernel, quanto por mapeamentos geométricos de características ou por razões de verossimilhança baseadas em *scores* [9].
Seguindo para a parte inferencial, fora desenvolvido dois cenários formais [12] para origem dos chamados *latent prints*.

- Definição 11:

  – O primeiro cenário chama-se *Common Source Scenario* e considera que duas impressões têm a mesma origem sem especificar formalmente qual origem seja ou não. Este cenário compara duas evidências e_{u1} e e_{u2} (por exemplo, duas impressões coletadas em dois cenários de crime distintos para ponderar a possibilidade de algum criminoso em série) com objetivo de determinar se foram feitas pela mesma pessoa desconhecida ou não (verificar se há possível relação ou a quantidade de culpados distintos). As hipóteses consideradas são:
  - H_{0CS} : e_{u1} e e_{u2} foram deixados pela mesma pessoa, embora desconhecida;
  - H_{1CS} : e_{u1} e e_{u2} foram deixados por duas diferentes pessoas (desconhecidas).
  Além disso, para estudo de convergências e configurações, foram considerados os seguintes modelos generativos (representados por dois modelos de efeito hierárquicos):

$$e_{u1} = \mu + d + u_1, \text{ onde } d \sim N(0, \sigma_d^2) \text{ e } u_1 \sim N(0, \sigma_{u_1}^2)$$
$$e_{u2} = \mu + d + u_2, \text{ onde } d \sim N(0, \sigma_d^2) \text{ e } u_2 \sim N(0, \sigma_{u_2}^2)$$

  em que μ é a média da distribuição de características das impressões de toda população, d é a distância entre as características em diferentes indivíduos e a média da população e u_1 e u_2 são efeitos aleatórios que afetam a aparência final (depois do desenvolvimento, transferência, fotografia etc.) das impressões. Note que u_1 e u_2 podem ser diferentes como duas impressões diferentes podem ser afetadas por diferentes conjuntos de fatores.
  Daí, os modelos bivariados e_{u1} e e_{u2} ganham as seguintes formas:

$$\begin{pmatrix} e_{u_1} \\ e_{u_2} \end{pmatrix} | H_{0_{CS}} \sim NMV\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma_d^2 + \sigma_{u_1}^2 & \sigma_d^2 \\ \sigma_d^2 & \sigma_d^2 + \sigma_{u_2}^2 \end{pmatrix} \right)$$
$$\begin{pmatrix} e_{u_1} \\ e_{u_2} \end{pmatrix} | H_{1_{CS}} \sim NMV\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \sigma_d^2 + \sigma_{u_1}^2 & 0 \\ 0 & \sigma_d^2 + \sigma_{u_2}^2 \end{pmatrix} \right)$$

- Definição 12:
  O segundo e último cenário chama-se *Specific Source Scenario* e tipicamente envolve a comparação entre uma impressão recuperada e_u e uma impressão controle, de um indivíduo conhecido, e_s. As hipóteses, assim, são as seguintes:
  - H_{0SS} : e_u e e_{us} foram deixados pela mesma pessoa, conhecida;
  - H_{1SS} : e_u foi deixado por outra pessoa, desconhecida, que não seja a dona de e_s
  Percebe-se que, nesse cenário, o interesse está mais em testar um suspeito do que testar a evidência em si. Assim

como no *Common Source Scenario*, assume-se modelos generativos da seguinte forma:

$$e_u = \mu + d + u, \text{ onde } d \sim N(0, \sigma_d^2) \text{ e } u \sim N(0, \sigma_u^2)$$
$$e_s = \mu_d + s, \text{ onde e } s \sim N(0, \sigma_s^2)$$

em que μ é a média da distribuição de características das impressões de toda população, d é a distância entre as características em diferentes indivíduos e a média da população , u e s são efeitos aleatórios que afetam a aparência final (depois do desenvolvimento, transferência, fotografia etc.) das impressões da cena e controle, respectivamente, e μ_d a média de características de um indíviduo específico (nesse caso o suspeito).
Com isso em mente e a natureza do cenário, temos que:

$$\begin{pmatrix} e_u \\ e_s \end{pmatrix} | H_{0_{SS}} \sim NMV\left(\begin{pmatrix} \mu_d \\ \mu_d \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & 0 \\ 0 & \sigma_s^2 \end{pmatrix} \right)$$

$$\begin{pmatrix} e_u \\ e_s \end{pmatrix} | H_{1_{SS}} \sim NMV\left(\begin{pmatrix} \mu \\ \mu_d \end{pmatrix}, \begin{pmatrix} \sigma_d^2 + \sigma_u^2 & 0 \\ 0 & \sigma_{u_s}^2 \end{pmatrix} \right)$$

Definição 13:

- As razões de verossimilhança, de um paradigma Bayesiano, para as situações de *Common Source Scenario* e *Specific Source Scenario* são definidas da seguinte forma, respectivamente:

$$LR_{CS} = \frac{f(e_u, e_s | H_{0_{CS}})}{f(e_u, e_s | H_{1_{CS}})} = \frac{f(e_u, e_s | H_{0_{CS}})}{f(e_u | H_{1_{CS}}) f(e_s | H_{1_{CS}})}$$

$$LR_{SS} = \frac{f(e_u, e_s | H_{0_{SS}})}{f(e_u, e_s | H_{1_{SS}})} = \frac{f(e_u | H_{0_{SS}})}{f(e_u | H_{1_{SS}})}$$

  Vale comentar que, na maioria dos casos, os examinadores de impressões digitais trabalham no *Specific Source Scenario*.

3.4 Quimiometria

Outro problema forense importantíssimo é a adulteração de substâncias por exemplo, e para tal estudo fora realizada a revisão breve de conceitos de absorbância e estimação de curvas (dados funcionais). De modo geral, toda substância reflete a luz de alguma forma de acordo com diferentes níveis de luz (infra-vermelho por exemplo), a partir da estimação da curva de reflexão (absorbância) de uma substância, pode-se, em um contexto forense, facilmente identificar possíveis substâncias comportando-se de maneira ineseperada (adulteração presente por exemplo).

Definição 14:

- A absorbância de uma substância é definida por:

$$x = -\log_{10}\left(\frac{I}{I_0}\right)$$

em que I é a intensidade da luz transmitida, após testada na substância ("refletida") e I_0 é a intensidade da luz incidente, antes da interação com a substância em questão.

Definição 15:

- A lei de Beer-Lambert é definida como:

$$x_j = \sum_{l=1}^{m} y_l a_{lj} + \varepsilon_j$$

em que x_j é a absorbância da amostra no j-ésimo comprimento de onda ($j = 1, \ldots, k$), a_{lj} é a absorbância no l-ésimo constituinte puro no j-ésimo comprimento de onda, y_l é a concentração do l-ésimo constituinte e um erro aleatório ε_j para $l = 1, \ldots, m$ e $j = 1, \ldots, k$.
Em outras palavras, essa lei diz que cada substância pode ser particionada em constituintes e, assim, sua absorbância total também. Assim, a estimação torna-se mais simples a medida que as absorbâncias dos constituintes pode ser conhecida (como água etc.) e basta, agora, utilizar das devidas concentrações.

4 Análise descritiva do conjunto de dados de comparação de resíduos

Para obter um panorama geral dos dados que serão futuramente alvo dos métodos forenses, a leitura do texto-livro e de artigos foi realizada e além disso, uma análise descritiva de um banco de dados[5] que continha diversas observações de medidas de fragmentos de vidro de duas empresas diferentes foi realizada. O objetivo dessa análise foi encontrar quais compostos químicos têm concentrações características que diferem sua origem pois, sob ponto de vista forense, tal análise é necessária quando o objetivo é comparar fragmentos encontrados em duas fontes distintas (uma cena de crime e um suspeito por exemplo). A aplicação dos métodos de estatística para inferência está em andamento em conjunto com a leitura previamente citada.

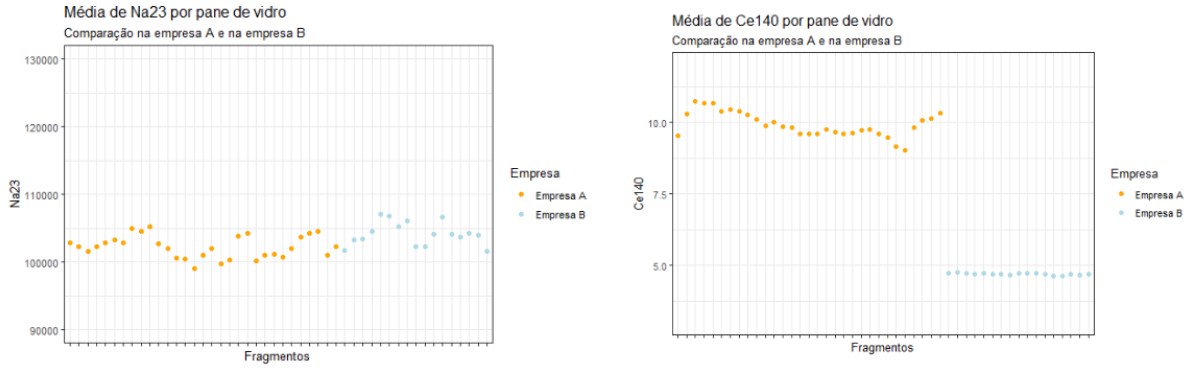

Figure 1: Comparação na distribuição de alguns compostos químicos presentes na composição do vidro entre as empresas

Embora haja 18 compostos primordiais que compõem o vidro, a partir da análise feita, consegue-se observar alguns que caracterizam mais um vidro diferente de outro, podendo assim, serem aproveitados no futuro na aplicação de métodos como visto na *Figure 1.*

Composto	Proporção no vidro A	Proporção no vidro B
Na23	0.525890272459	0.513312239580
Ca42	0.328634980465	0.347569012832
Mg25	0.119408672166	0.122883376080
Al27	0.014533100102	0.008525898344
K39	0.006478834621	0.003390014348
Fe57	0.003465719839	0.002626889221
Mn55	0.000909907548	0.000633387087
Ti49	0.000242604417	0.000605019017
Zr90	0.000163732480	0.000159129827
Sr88	0.000093914566	0.000155693771
Ba137	0.000056221698	0.000065518507
Ce140	0.000048685355	0.000023410727
La139	0.000021559232	0.000013443667
Nd146	0.000018291170	0.000011711831
Rb85	0.000011007671	0.000010689919
Li7	0.000010422971	0.000005179384
Hf178	0.000006342123	0.000005166298
Pb208	0.000005731116	0.000004219562

Table 1: Comparação entre as proporções de compostos químicos de dois painéis de vidro aleatórios da empresa A e da empresa B

5 Simulação dos panoramas em análise de impressões digitais

Partindo para outro cenário, fora realizada simulações de cenários distintos (*Common Source* ou *Specific Source*),a partir da referência *Handbook of Forensic Statistics*[12], a fim de avaliar as razões de verossimilhança, isto é, a quantificação da evidência e como ela converge ou não (tem mesma significância forense) de acordo com o cenário escolhido.

Para isso, consideraremos pares de e_u e e_s gerados pelo modelo dito anteriromente, sob H_{0SS} ou H_{1SS} (para o teste de convergência mesmo) e será computado as razões de verossimilhança.

Para calcular a razão de verossimilhança do caso *Common Source Scenario* será definido $e_{u1} = e_u$, $e_{u2} = e_s$,$\sigma^2_{u_1} = \sigma^2_u$ e $\sigma^2_{u_2} = \sigma^2_s$.

Por último, será fixado $\mu = 10$, $\sigma^2_d = 10$ e $\sigma^2_u = 2$ e ocorrerão 1000 repetições da simulação, primeiramente com $\mu_d = 0$ (isto é, evidência relativamente rara a média populacional, ou seja, mais propensa a ser distinguível) e $\sigma^2_s = 1$ (variância 'instrumental' existente e considerável) e no segundo caso, mantendo $mu_d = 0$ e com $\sigma^2_s = 10^{-5}$ (variância da recuperação da impressão digital quase inexistente, negligenciável).

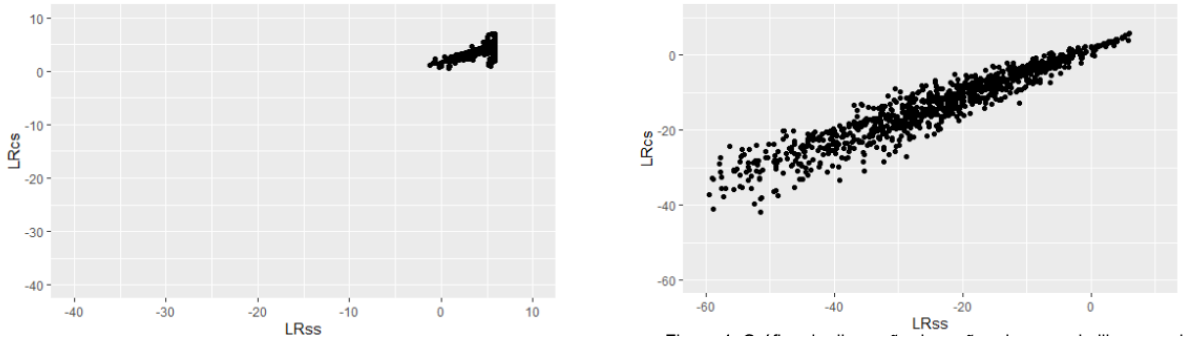

Figure 2: Gráfico de dispersão de razões de verossimilhança sob a hipótese nula e alternativa de dois cenários forenses diferentes com $\mu_d = 0$ e $\sigma_s^2 = 1$

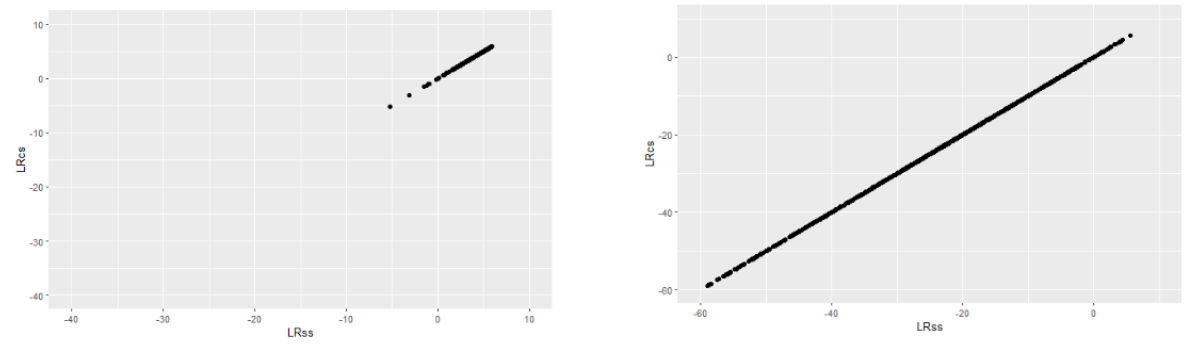

Figure 3: Gráfico de dispersão de razões de verossimilhança sob a hipótese nula e alternativa de dois cenários forenses diferentes com $\mu_d = 0$ e $\sigma_s^2 = 10^{-5}$

Percebe-se pelas Figuras 2 e 3, que não ocorre a convergência das situações, exceto quando a variância instrumental é negligenciável (um cenário praticamente utópico). Daí nota-se a importância na identificação da situação forense-estatística e o póstumo modelo adotado em conjunto com as hipóteses assumidas.

6 Considerações Finais

Como visto, o projeto teve como foco inicial o aprendizado e leitura de conceitos estatísticos assim como problemas do universo forense, as próximas etapas, intuitivamente, focam na aplicação de tais métodos e conceitos em diferentes situações encontradas nos bancos de dados, por exemplo, em problemas de fragmentos de vidros ou rastros de sangue. Entretanto, por ser uma área relativamente recente, há muito a considerar-se quanto a abordagem estatística, daí surge a necessidade da revisão literária até então. Em diversos contextos, entretanto, vale o alto cuidado na escolha do teste de hipóteses como visto nas simulações, para que evite diferentes viés como de seleção, ou de confirmação, espalhando assim um erro e consequentemente uma conclusão incorreta.

Em suma, a importância dos métodos estatísticos em ciência forense fora demonstrada através dos estudos, visto que, ambas as naturezas das áreas envolve, principalmente, incerteza e variações.

Referências

[1] Aitken C.G.G, Taroni F (1995) Statistics and the Evaluation of Evidence for Forensic Scientists. John Wiley & Sons, Ltd.
[2] Nobles R, Schiff D (2005) Misleading statistics within criminal trials: The Sally Clark case . Significance magazine
[3] Lindley DV (1977) A Problem in Forensic Science. Biometrika, pp. 207-213
[4] Royall R (2000) On the Probability of Observing Misleading Statistical Evidence. Journal of the American Statistical Association Vol. 95, No.451 , pp 760-768
[5] Open Source Forensic Data: CSAFE Forensic Science Dataset Portal : https://forensicstats.org/data/
[6] Walsh KAJ, Buckleton JS (1991) Calculating the frequency of occurrence of a blood type for a 'random man' . Journal of the Forensic Science Society.
[7] RStudio Desktop - Posit : https://posit.co/download/rstudio-desktop/
[8] Balding DJ, Donnely P (1996) Evaluating DNA Profile Evidence When the Suspect Is Identified Through a Database Search. Journal of Forensic Sciences
[9] Neumann C et al. (2012) Quantifying the weight of evidence from a forensic fingerprint comparison : a new paradigm. Journal of the Royal Statistical Society
[10] Saraiva MA (2009) Análise não-paramétrica de dados funcionais : uma aplicação a quimiometria. Universidade Estadual de Campinas
[11] Evett IW, Buckleton J (1990) The interpretation of glass evidence. A practical approach. Journal of the Forensic Science Society
[12] Banks DL (2001) Handbook of Forensic Statistics. Chapman & Hall/CRC
[13] Kern J (2016) Fingerprinting: A Study in Cognitive Bias and its Effects on Latent Fingerprint Analysis. Undergraduate Honors College Theses of Long Island University
[14] National Research Council DNA Technology in Forensic Science (1992)
[15] Morton NE (1993) DNA in court. European Journal of Human Genetics, pp. 172-178

Anexo

O presente anexo apresenta o código das análises e simulações vistas anteriormente.

Carregamento dos pacotes utilizados

```
library(MASS)
library(ggplot2)
library(tidyverse)
library(mvtnorm)
library(readr)
```

Código para análise do banco de dados

Leitura e limpeza inicial dos arquivos .csv contendo os fragmentos de vidro

```
tabelaA - list.files(path = "~
CompanyA", pattern = "*ċsv") % > %
```

```
lapply(read_csv, show_col_types = FALSE) % > %
bind_rows % > %
as_data_frame() % > %
select(-c(Rep))

    tabelaB ¡- list.files(path = ”
CompanyB”, pattern = ”*.csv”) % > %
lapply(read_csv, show_col_types = FALSE) % > %
bind_rows % > %
as_data_frame() % > %
select(-c(Rep))
```

Criação de variáveis juntando arquivos lidos para manipulação

```
    juncao = bind_rows(tabelaA, tabelaB)
juncaoLonger = juncao % > % select(-c(fragment)) % > %
pivot_longer(!c(pane,Company), names_to = ”composto”, values_to = ”valores”)
```

Criação de tabelas contendo proporção de compostos químicos na composição dos fragmentos

```
    aggregate(valores composto, data = filter(juncaoLonger, pane == ”AA”), FUN = sum) % > %
mutate(prop = format(valores / sum(valores), scientific = FALSE)) % > % arrange(desc(prop))

    aggregate(valores composto, data = filter(juncaoLonger, pane == ”BC”), FUN = sum) % > %
mutate(prop = format(valores / sum(valores), scientific = FALSE)) % > % arrange(desc(prop))
```

Gráficos de comparação de concentrações médias das empresas por Na23 e Ce140

```
    x = juncao % > % select(Ce140,Company,pane) % > % group_by(pane, Company) % > %
summarise(mean(Ce140))

    x % > % ggplot(mapping = aes(x = pane, y = ‘mean(Ce140)‘, col = Company))+
geom_point(ylim = c(90000,130000))+
labs(title = ”Média de Ce140 por pane de vidro”,
subtitle = ”Comparação na empresa A e na empresa B”,
x = ”Fragmentos”, y = ”Ce140”)+
theme_bw()+
theme(axis.text.x = element_blank())+
ylim(3,12)+
scale_color_manual(name = ”Empresa”, values = c(”orange”,”lightblue”),labels = c(”Empresa
A”,”Empresa B”))

    x = juncao % > % select(Na23,Company,pane) % > % group_by(pane, Company) % > %
summarise(mean(Na23))

    x % > % ggplot(mapping = aes(x = pane, y = ‘mean(Na23)‘, col = Company))+
geom_point(ylim = c(90000,130000))+
labs(title = ”Média de Na23 por pane de vidro”,
subtitle = ”Comparação na empresa A e na empresa B”,
x = ”Fragmentos”, y = ”Na23”)+
```

```
theme_bw()+
theme(axis.text.x = element_blank())+
scale_color_manual(name = "Empresa", values = c("orange","lightblue"),labels = c("Empresa
A","Empresa B"))
```

Código para Simulação

Definição de variáveis fixas para simulação

```
mu = 10
sigma_d = 10
sigma_u = 2
mu_d = 9
sigma_s = 1
```

Geração de amostras e computação das razões de verossimilhança para o primeiro caso

```
sigma = matrix(c(sigma_u,0,0,sigma_s),2,2)
sample_h0ss = mvrnorm(1000, mu = c(mu_d, mu_d), Sigma = sigma) % > % as.data.frame()
colnames(sample_h0ss) = c("eu","es")

sigma = matrix(c(sigma_u+sigma_d,0,0,sigma_s),2,2)
sample_h1ss = mvrnorm(1000, mu = c(mu, mu_d), Sigma = sigma) % > % as.data.frame()
colnames(sample_h1ss) = c("eu","es")

sigma = matrix(c(sigma_u + sigma_d,sigma_d,sigma_d,sigma_s + sigma_d),2,2)

a = dmvnorm(sample_h0ss, mean = c(mu, mu), sigma = sigma)
b = dnorm(sample_h0ss$eu, mean = mu, sd = sqrt(sigma_d + sigma_u))
c = dnorm(sample_h0ss$es, mean = mu, sd = sqrt(sigma_d + sigma_s)) LRcs = log(a/(b*c))

a = dnorm(sample_h0ss$eu, mean = mu_d, sd = sqrt(sigma_u))
b = dnorm(sample_h0ss$eu, mean = mu, sd = sqrt(sigma_d + sigma_u)) LRss = log(a/b)
```

Gráfico para verificar convergência para o primeiro caso

```
ggplot(data.frame(LRcs, LRss), mapping = aes(x = LRss, y = LRcs))+
geom_point()+
xlim(-40,1)+
ylim(-40,1)
```

Geração de amostras e computação das razões de verossimilhança para o segundo caso

```
mu_d = 9
sigma_s = 10^(-5)

sigma = matrix(c(sigma_u,0,0,sigma_s),2,2)
sample_h0ss = mvrnorm(1000, mu = c(mu_d, mu_d), Sigma = sigma) % > % as.data.frame()
colnames(sample_h0ss) = c("eu","es")
```

```
    sigma = matrix(c(sigma_u+sigma_d,0,0,sigma_s),2,2)
sample_h1ss = mvrnorm(1000, mu = c(mu, mu_d), Sigma = sigma) % > % as.data.frame()
colnames(sample_h1ss) = c("eu","es")

    sigma = matrix(c(sigma_u + sigma_d,sigma_d,sigma_d,sigma_s + sigma_d),2,2)
a = dmvnorm(sample_h0ss, mean = c(mu, mu), sigma = sigma)
b = dnorm(sample_h0ss$eu, mean = mu, sd = sqrt(sigma_d + sigma_u))
c = dnorm(sample_h0ss$es, mean = mu, sd = sqrt(sigma_d + sigma_s))
LRcs = log(a/(b*c))

    a = dnorm(sample_h0ss$eu, mean = mu_d, sd = sqrt(sigma_u))
b = dnorm(sample_h0ss$eu, mean = mu, sd = sqrt(sigma_d + sigma_u))
LRss = log(a/b)

    ggplot(data.frame(LRcs, LRss),mapping = aes(x = LRss, y = LRcs))+
geom_point()+
xlim(-40,10)+
ylim(-40,10)

    sigma = matrix(c(sigma_u + sigma_d,sigma_d,sigma_d,sigma_s + sigma_d),2,2)

    a = dmvnorm(sample_h1ss, mean = c(mu, mu), sigma = sigma)
b = dnorm(sample_h1ss$eu, mean = mu, sd = sqrt(sigma_d + sigma_u))
c = dnorm(sample_h1ss$es, mean = mu, sd = sqrt(sigma_d + sigma_s))
LRcs = log(a/(b*c))

    a = dnorm(sample_h1ss$eu, mean = mu_d, sd = sqrt(sigma_u))
b = dnorm(sample_h1ss$eu, mean = mu, sd = sqrt(sigma_d + sigma_u))
LRss = log(a/b)
```

Gráfico para verificar convergência para o segundo caso

```
    ggplot(data.frame(LRcs, LRss),mapping = aes(x = LRss, y = LRcs))+
geom_point()+
xlim(-60,10)+
ylim(-60,10)
```